MASARYK UNIVERSITY

FACULTY OF SCIENCE

Department of Chemistry

Dynamical Features of Biomolecular Complexes

HABILITATION THESIS

JOZEF HRITZ

Brno 2021

Bibliographic record

Author:	JOZEF HRITZ Faculty of Science Masaryk University
Title of Thesis:	Dynamical Features of Biomolecular Complexes
Habilitation Field:	Physical Chemistry
Year:	2021
Number of Pages:	72+176
Keywords:	biophysical chemistry, computational simulations, molecular dynamics, biomolecules, statistical physics, free energy, binding affinity, enhanced sampling, nuclear magnetic resonance spectroscopy, intrinsically disordered proteins

Bibliografický záznam

Autor:	JOZEF HRITZ Faculty of Science Masaryk University
Název práce:	Dynamické Vlastnosti Biomolekulárnich Komplexů
Obor habilitačního řízení:	Fyzikální Chemie
Rok:	2021
Počet stran:	72+176
Klíčová slova:	biofyzikální chemie, výpočetní simulace, molekulová dynamika, biomolekuly, statistická fyzika, volní energie, vazební afinita, zesílené vzorkování, nukleární magnetická rezonanční spektroskopie, přirozene neuspořádané proteíny

Commentary to habilitation thesis

This habilitation thesis, submitted to the Dean of Faculty of Science at the Masaryk University as a part of the application for the academic title "Docent" (Associated Professor), documents the most important research activities of the applicant after his Ph.D. study. The main research activities described here cover a variety of novel methods developed in the field of biomolecular simulations and the corresponding experimental studies that could be used to experimentally validate parts of the computational predictions of macromolecules in solution. Biomolecular simulations are very useful tools to rationalize experimental findings and to direct future experiments. Computer simulations offer insight at an atomic resolution and a femtosecond timescale, which are usually beyond experimental means. On the other hand, computational simulations use a range of approximations and simplifications and are restricted to simulating only very short time periods. Therefore, it is crucial to validate the results obtained from computational simulations by comparison to experimentally obtained data as far as possible. To achieve this the applicant deployed mostly solution NMR and biophysical techniques such as measurements of thermostability and binding affinities to obtain experimental data.

From the 29 research articles published after the applicant's Ph.D. thesis defense, 15 were selected to be included as the attached reprintsP1-P15 and are briefly commented in four scientific chapters (3-6). Out of these 15 scientific peer-review publications, the applicant was the first author in 6 of them and the corresponding author in 5 of them. The first scientific chapter here, **"3. Efficient incorporation of plasticity and water molecules into the molecular docking approach"**, describes our developed method combining ensemble ligand docking and molecular dynamics (MD) simulations, incorporating the plasticity and flexibility of cytochrome P450 2D6 structures in a highly efficient way and leading to a significant increase in the reliability of predicted binding poses.[P1] The approach was extended to include the unbiased addition of the most important explicit water molecules in the docking procedure.[P2] The developed techniques were successfully applied also for the ligand docking into another cytochrome P450s[P3] as well as for RNA-protein docking.[P4] The described docking approaches are restricted by the limited sampling potential of standard MD. This can be significantly increased by applying some enhanced sampling approach such as temperature or Hamiltonian replica-exchange MD (T/H-REMD). This is the content of the second scientific chapter **"4. Methodological developments of enhanced sampling methodology H-REMD"** where we describe H-REMD using soft-core interactions.[P5] In contrast to T-REMD, H-REMD needs a priori knowledge of the origin of energy barriers in the system, which implies that different systems require different H-REMD schemes. For this particular problem, we introduced the "fast mimicking" procedure for providing the sought parameters.[P6] The applicability of described methods to produce the canonical ensemble efficiently was shown using GTP analogs. This chapter also describes an approach combining H-REMD with the distance-field distance restraints leading to a powerful tool for the unbiased simulation of binding pathways and conformational changes induced by phosphopeptides binding or unbinding from the 14-3-3zeta protein.[P7] The reliable sampling of binding/unbinding events also allowed the calculation of the corresponding binding affinity of ligands to the protein. Unfortunately, this approach is computationally very expensive. In many applications rather than calculating the absolute binding affinity, it is sufficient to calculate the relative free energy differences by alchemical free energy calculations between quite similar compounds. However, the efficiency may drop dramatically for compounds that have multiple stable conformations separated by high energy barriers. This problem is addressed in the third scientific chapter **"5. Alchemical (relative) free energy calculations"**.[P8-P10] Another class of biomolecules that represent a challenge for conformational sampling are intrinsically disordered proteins (IDPs), not because of high energy barriers, but because they possess a huge number (>thousands) of conformational states. They are the topic of the fourth scientific chapter. **"6. Structural and interaction properties of IDPs determined by experimental NMR and computational studies"**.[P11-P15] Here we describe the combination of experimental NMR and MD techniques in a closely co-ordinated fashion allowing the characterization of this important class of proteins.

My contribution to the selected 15 articles to this habilitation thesis is summarized in the following tables with special attention to the experimental work, supervision of students, manuscript preparation, and research direction.

[P1][1] Hritz, J.; de Ruiter, A.; Oostenbrink, C. Impact of plasticity and flexibility on docking results for Cytochrome P450 2D6: a combined approach of molecular dynamics and ligand docking. *J. Med. Chem.* **2008**, 51, 7469-7477 (IF=4.9)

Experimental work (%)	Supervision (%)	Manuscript (%)	Research direction (%)
80%	70%	60%	50%

[P2] Santos, R.; Hritz, J.; Oostenbrink, C. The role of water in molecular docking simulations of Cytochrome P450 2D6. *J. Chem. Inf. Model.* **2010**, 50, 146-154 (IF=3.8)

Experimental work (%)	Supervision (%)	Manuscript (%)	Research direction (%)
20%	70%	40%	50%

[P3] Oostenbrink C.; de Ruiter A.; Hritz J.; Vermeulen N.P.E Malleability and versatility of Cytochrome P450 active sites studied by molecular simulations.
Curr. Drug Metab. **2012**, 13, 190-196 (IF=4.2)

Experimental work (%)	Supervision (%)	Manuscript (%)	Research direction (%)
-	20%	20%	20%

[P4] Byeon I-J. , Ahn J., Mitra M., Byeon C-H., Hercík K., Hritz J., Charlton L., Levin J., Gronenborn A.M. NMR structure of human restriction factor APOBEC3A reveals substrate binding and enzyme specificity. *Nat. Commun.* **2013**, 4, 1890 (IF=10.7)

Experimental work (%)	Supervision (%)	Manuscript (%)	Research direction (%)
10%	-	10%	-

[P5] Hritz J., Oostenbrink C. Hamiltonian replica exchange molecular dynamics using soft-core interactions. *J. Chem. Phys.* **2008,** 128, 144121 (IF=3.1)

Experimental work (%)	Supervision (%)	Manuscript (%)	Research direction (%)
80%	-	70%	50%

[1] Bibliographic record of a published scientific result, which is part of the habilitation thesis.

[P6] Hritz J, Oostenbrink C. Optimization of Replica Exchange Molecular Dynamics by Fast Mimicking. *J. Chem. Phys.* **2007,** 127, 204104 (IF=3.1)

Experimental work (%)	Supervision (%)	Manuscript (%)	Research direction (%)
80%	-	70%	50%

[P7] Nagy,G.; Oostenbrink, C.; Hritz, J.*: Exploring the Binding Pathways of the 14-3-3ζ Protein: Structural and Free-Energy Profiles Revealed by Hamiltonian Replica Exchange Molecular Dynamics with Distance Field Distance Restraints. *PLoS ONE* **2017**,12(7), e0180633 (IF=2.8)

Experimental work (%)	Supervision (%)	Manuscript (%)	Research direction (%)
10%	80%	40%	50%

[P8] Hritz, J.; Lappchen T.; Oostenbrink, C. Calculations of binding affinity between C8-substituted GTP analogs and the bacterial cell-division protein FtsZ.
Eur.Biophys. J. **2010**, 39, 1573-1580 (IF=2.4)

Experimental work (%)	Supervision (%)	Manuscript (%)	Research direction (%)
70%	-	60%	50%

[P9] Hritz, J.; Oostenbrink, C. Efficient free energy calculations for compounds with multiple stable conformations separated by high energy barriers.
J. Phys. Chem. B **2009**, 113, 12711-12720 (IF=3.5)

Experimental work (%)	Supervision (%)	Manuscript (%)	Research direction (%)
80	-	70	50

[P10] Jandova, Z.; Trosanova, Z.; Weisova, V.; Oostenbrink, C.*, Hritz, J.*:Free energy calculations on the stability of the 14-3-3ζ protein. *BBA - Proteins and Proteomics*, **2018**, 1866, 442-450 (IF=2.5)

Experimental work (%)	Supervision (%)	Manuscript (%)	Research direction (%)
20%	50%	30%	40%

[P11] Hritz J.; Byeon I-J.; Krzysiak T.; Martinez A.; Sklenář V.; Gronenborn A.M. Dissection of binding between a phosphorylated tyrosine hydroxylase peptide and 14-3-3ζ: a complex story elucidated by NMR. *Biophys. J.* **2014**, 107, 2185-2194 (IF=4.0)

Experimental work (%)	Supervision (%)	Manuscript (%)	Research direction (%)
70%	-	70%	50%

[P12] Louša, P.; Nedozrálová, H.; Župa, E.; Nováček, J.; Hritz, J.*: Phosphorylation of the regulatory domain of human tyrosine hydroxylase 1 monitored using non-uniformly sampled NMR. *Biophysical Chemistry* **2017**, 223, 25-29 (IF=1.9)

Experimental work (%)	Supervision (%)	Manuscript (%)	Research direction (%)
10	100	40	60

[P13] Zapletal, V.; Mládek, A.; Melková, K.; Louša, P.; Nomilner, E.; Jaseňáková, Z.; Kubáň, V.; Makovická, M.; Laníková, A.; Žídek L.; Hritz, J.* Choice of force field for proteins containing structured and intrinsically disordered regions. *Biophys. J.* **2020**, 118, 1621–1633 (IF=3.9)

Experimental work (%)	Supervision (%)	Manuscript (%)	Research direction (%)
10	50	30	40

[P14] Pavlíková Přecechtělová, J.; Mládek, A.; Zapletal, V.; Hritz, J. Quantum Chemical Calculations of NMR Chemical Shifts in Phosphorylated Intrinsically Disordered Proteins, JCTC **2019**, 15, 5642-5658 (IF=5.0)

Experimental work (%)	Supervision (%)	Manuscript (%)	Research direction (%)
-	30	20	50

[P15] Jansen, S.;Melková, K.; Trošanová, Z.; Hanáková, K.; Zachrdla, M.; Nováček, J.; Župa, E.; Zdráhal, Z.; Hritz, J.*; Žídek, L.*: Quantitative Mapping of MAP2c Phosphorylation and 14-3-3ζ Binding Sites Reveals Key Differences Between MAP2c and Tau. *J. Biol. Chem.* **2017**, 292, 6715-6727 (IF=4.0)

Experimental work (%)	Supervision (%)	Manuscript (%)	Research direction (%)
10	30	30	20

Komentář

Tato habilitační práce, předložená děkanovi Přírodovědecké fakulty Masarykovy univerzity v rámci žádosti o akademický titul "Docent" (Přidružený profesor), dokumentuje nejdůležitější výzkumnou činnost uchazeče po jeho doktorském studiu. Zde popsané hlavní výzkumné činnosti zahrnují řadu nových metod vyvinutých v oblasti biomolekulárních simulací a odpovídajících experimentálních studií, které by mohly být použity k experimentální validaci částí výpočetních předpovědí makromolekul v roztoku. Biomolekulární simulace jsou velmi užitečnými nástroji pro racionalizaci experimentálních nálezů a pro detailní návrh budoucích experimentů. Počítačové simulace nabízejí přehled v atomovém rozlišení a současně v femtosekundovém časovém měřítku, které jsou obvykle mimo experimentální prostředky.

Na druhou stranu výpočetní simulace používají řadu aproximací a zjednodušení a jsou omezeny na simulaci pouze velmi krátkých časových období. V tomto případě je nezbytné co nejvíce ověřit výsledky získané z výpočetních simulací ve srovnání s experimentálně získanými daty. K dosažení tohoto cíle žadatel použil převážně roztokovou NMR spektroskopii a biofyzikální techniky, jako jsou měření termostability a vazebné afinity. Z 29 výzkumných originálních článků byly vybrány 15(přiložené reprinty[P1-P15]) a jsou stručně komentovány ve čtyřech vědeckých kapitolách(3-6). **Z těchto 15 vědeckých recenzovaných publikací byl žadatel prvním autorem v 6 z nich a korespondujícím autorem v 5 z nich.**

První vědecká kapitola zde, **"3. Efektivní začlenění plasticity a molekul vody do molekulárních dokovacích metod"**, popisuje nami vyvinutou metodu kombinující dokování ligandů do souboru proteinových struktur a simulace molekulární dynamiky (MD), zahrnující plasticitu a flexibilitu cytochromových struktur P450 2D6 vysoce účinným způsobem a vedoucí k významnému zvýšení spolehlivosti předpokládaných vazebných pozic.[P1] Tato metodológie byla dál rozšířena tak, aby zahrnovala dynamické přidání nejdůležitějších explicitních molekul vody do dokovací procedury.[P2] Vyvinuté výpočetní techniky byly úspěšně použity rovněž pro dokování ligandů do jiných cytochromů P450[P3] a rovněž pro dokování proteinů flexibilní RNA molekulou.[P4] Popsané dokovací přístupy jsou omezeny omezeným vzorkovacím potenciálem standardní MD. To lze významně zvýšit napr. použitím MD se zámenou replik(REMD). To je obsahem druhé vědecké kapitoly **"4. Metodický vývoj Hamiltonovskej REMD (H-REMD)"**, kde popisujeme H-REMD využívajíci tzv. změkčené interakce.[P5] Na rozdíl od teplotní REMD (T-REMD), nastavení H-REMD vyžaduje předchozí znalost původu energetických bariér v systému, což znamená, že různé systémy vyžadují různá schémata H-REMD. Pro tento konkrétní problém jsme popsali efektivní postup "rychlého mimikování" pro poskytování hledaných interakčních parametrů.[P6] Použitelnost popsaných metod pro efektivní generování kanonického strukturního souboru byla prokázána na sade GTP analogů. Nasledovná podkapitola popisuje přístup kombinující H-REMD s varirováním vzdálenosti distančního pole vedoucím k výkonnému nástroji pro nezaujatou simulaci vazebných drah a konformačních změn vyvolaných vázáním fosfopeptidů do 14-3-3ζ proteinu.[P7] Dostateční vzorkování vazebných/nespojitých příhod rovněž umožnil výpočet odpovídající vazebné afinity ligandů k proteinu. Bohužel tento přístup je výpočetně velmi nákladný. V mnoha aplikacích spíše než výpočet absolutní vazebné afinity stačí vypočítat relativní volné energetické rozdíly alchymickými výpočty volné energie mezi poměrně podobnými sloučeninami. Účinnost však může dramaticky klesnout u sloučenin, které mají více stabilních konformací oddělených vysokými energetickými bariérami. Tento problém je řešen ve třetí vědecké kapitole **"5. Alchemické výpočty (relativní) volné energie"**.[P8-P10] Další třídou biomolekul, které představují výzvu pro konformační odběr vzorků, jsou přirozene neuspožádané proteiny (IDP), nikoli kvůli vysokým energetickým bariérám, ale proto, že mají obrovské množství (>tisíce) konformačních stavů. Jsou tématem čtvrté vědecké kapitoly: **"6. Určení strukturních a interakčných IDP proteinů určovaných kombinaci experimentální NMR a výpočetních studii"**.[P11-P15] V této kapitole je detailne popsána úzka kombinace experimentální roztokové NMR a MD simulací, která umožňuje charakterizaci této důležité třídy proteinů.

Declaration

I hereby confirm that I have written the habilitation thesis independently, that I have not used other sources than the ones mentioned and that I have not submitted the habilitation thesis elsewhere.

In Brno, June 10, 2021

....

JOZEF HRITZ

ACKNOWLEDGMENTS

First of all, I would like to thank my family. My parents and brother have been practical people and did not have much appreciation for scientific research. Therefore they were not pleased when I decided to study physics and later biophysics at the University of P.J.Safarik in Kosice. However, when they understood how much science means to me they fully supported me in my decision. My scientific career has not been easy either for my wife Renata to whom belongs my warmest thanks because it meant to live in several different countries where she followed me and had to raise our kids there: our son Alex who was born in the Netherlands, our daughter Daniela in the USA and now they attend their schools in the Czech Republic. Thank you for all your support and understanding of my scientific mission.

I feel sorry there is not enough space for thanking explicitly all my colleagues, students, and supervisors that supported me at various stages of my career. Still, I wish to rise the warmest thanks to at least five supervisors that are great researchers but at the same time great people who helped me greatly many times: to prof. Chris Oostenbrink was my supervisor during my post-doctoral stay at Vrije University in Amsterdam for 4.5 years. He always had time for a little chat if needed. Thank you for all your excellent ideas, your trust, and great collaboration! To prof. Angela Gronenborn was my supervisor during my post-doctoral stay at the University of Pittsburgh for 2.5 years. Thank you for that great opportunity to work in your team! I thank prof. Vladimir Sklenar allowed me to join CEITEC-MU in Brno. And to prof. Lukas Zidek who is my current boss and colleague at CEITEC-MU. I feel very thankful and privileged that these great scientists still do collaborate with me and my research team and that I always feel long-term support from their side. Regarding my pedagogical activities, I wish to thank prof. Libuse Trnkova who supported my interest to lecture the courses from Biophysical Chemistry from the beginning and that we could shape this interdisciplinary study program together.

Last but not least, I thank all students under my supervision. Not only because we can perform very exciting research together but they contributed to a very nice and friendly atmosphere. They also let me realize how much supervision and lecturing activities fulfill me.

As mentioned before there are more than a hundred scientists and friends that I feel very thankful to and despite I cannot list them all here explicitly let me do it at least in this implicit way THANK YOU ALL VERY MUCH!.

Table of Contents

List of Figures

Abbreviations

AAAH	–	aromatic amino acid hydroxylases
ACT	–	aspartate kinase, chorismate mutase and TyrA
A-EDS	–	accelerated enveloping distribution sampling
BioEN	–	Bayesian ensemble reweighing
CD	–	circular dichroism / catalytic domain
cryoEM	–	cryo electron microscopy
CS	–	chemical shift
EDS	–	enveloping distribution sampling
FRET	–	Förster Resonance Energy Transfer
H-REMD	–	Hamiltonian replica exchange molecular dynamics
IDP	–	intrinsically disordered protein
IDR	–	intrinsically disordered region
Map2c	–	microtubule associated protein 2c
MD	–	molecular dynamics
MM	–	molecular mechanics
MTSL	–	methanethiosulfonate
NMR	–	nuclear magnetic resonance
PKA	–	protein kinase A
PRE	–	paramagnetic relaxation enhancement
RD	–	regulatory domain
RDC	–	residual dipolar coupling
REMD	–	replica exchange molecular dynamics
SAXS	–	small-angle X-ray scattering
TD	–	tetramerization domain
TyrH	–	tyrosine hydroxylase

1 Introduction

I decided on the title of my habilitation thesis “Dynamical features of biomolecular complexes” due to three factors: i) it is the common denominator of the majority of the research papers and projects I was involved in after my Ph.D. ii) it is included in multiple lectures that I give during the Biophysical Chemistry course when describing theoretical or experimental methods of this interdisciplinary field iii) interestingly, the topic of the dynamics of biomolecular complexes seems to be viewed quite differently in computational biophysical chemistry from the view in structural biology.

The final statement above reminds me to place a disclaimer here, namely – opinions presented here are subjective and based on my experience when working in both research fields. Researchers may disagree with some of the statements presented. While I was involved in computational physical chemistry during my first post-doctoral stay at the research group of prof. Chris Oostenbrink, Ph.D. at Vrije University in Amsterdam; I worked in the field of structural biology during my second post-doc in the research group of prof. Angela Gronenborn, Ph.D. at University of Pittsburgh (USA). At Masaryk University, I try to combine and explore experiences from both of those fields in my running research at CEITEC-MU in the group of prof. Lukas Zidek, Ph.D. as well as in my lectures from the division of Biophysical Chemistry led by prof. Libuse Trnkova, Ph.D. I am involved in the following courses: Fundamentals of Biophysical Chemistry; Experimental methods of Biophysical Chemistry I; Advanced theoretical methods of Biophysical Chemistry; Experimental methods of Biophysical Chemistry II; and contribute to the Protein Expression and Purification courses.

For my undergraduate and graduate studies, I decided on Biophysics because I always loved and admired physics but at the same time wished to apply its powerful arsenal of methods and rigorous view on complex entities such as biomolecular complexes, that I was taught can explain the biological function. I must confess that I never liked chemistry during my high school and undergraduate studies because of the need to memorize tons of names, only later did I acknowledge the beauty and usefulness of chemistry and its subfields such as biochemistry. Another confession is the fact that I am not fond of strict definitions between the different fields in natural sciences. In the current interdisciplinary world of science, I tend to say to students – focus on your scientific problem and use the benefits of the most suitable methods and technologies you know about. This spirit lies in the name of the study program Biophysical Chemistry which covers three traditional study fields (biology, physics and chemistry) but in my personal view also mathematics, computational sciences, bioinformatics and partially medicine.

There should be a combination of not only the different fields but also of theoretical computational and experimental approaches. I think that almost every theoretician or computational scientist experiences frustration when obtaining interesting results from computational approaches but facing disbelief from experimentalist colleagues.

> ***Scientific joke*:** There is a running joke among scientists that when a theoretician comes with a new theory there is only one person that believes in it – him/herself. On contrary – experimental results are typically trusted by the whole community except the experimentalist who performed the measurements and gave their interpretation. He/she is aware of all the technical problems faced and the simplifications used in the interpretation of measured data.

I still remember confusion at my undergraduate lectures when first hearing that proteins are flexible and dynamic but at the very same time all functional features we were learning were explained by one static figure of the corresponding 3D structure. We learned that NMR in contrast to the "artificial" X-ray crystallography allows the determination of the real structural ensemble of the studied protein in solution. When I learned it, I immediately downloaded some of such ensembles from the pdb database but have to admit my disappointment. It seemed to me like some small fluctuations around a central structure. Even, today I have the feeling that many structural biologists have the feeling that one static structure is sufficient for correlation within the structure-function paradigm.

A much more dynamic view of biomolecules tends to prevail within computational biophysics and biophysical chemistry. "Bio" is here quite important because computational chemists often study small quite rigid molecules by quantum mechanical (QM) methods. And often consider a molecule by single structure sometimes extended with small fluctuations in the harmonic approximation. More complex systems such as biomacromolecules or in general biomolecular complexes in an explicit water environment are instead treated by classical molecular mechanics (MM) methods following classical Newtonian laws. There are some exceptions in which one needs to apply QM methods e.g. when dealing with enzymatic activity or in general when electronic properties should be considered explicitly. However, when this is not the case – classical MM is typically used despite the fact that QM would be more precise than classical MM. The underlying reason is that conformational sampling is more important than "precision" in most cases.

Pedagogical note: In my lectures for undergraduate students I compare this choice to the following problem – you should determine the average height of buildings in Brno (or the lowest one) within 24 hours where you can choose between three of the following measuring tools. The first measuring tool has micrometer precision but one measurement takes 12 hours. The second measuring tool has decimeter precision but one measurement takes only 1 minute and allows systematic scanning of building from north and row by row from west to east. The third measuring tool is similar to the second one, i.e. has a precision of one decimeter but it does not allow systematic scanning through the city – it rather randomly selects buildings in Brno but thanks to this factor one measurement takes just one second. (note: In my example, an American city would be used rather than Brno due to their planned design consisting of perpendicular streets and avenues …). The fourth approach is even faster – one measurement taking the only ms and in addition, it can quite quickly recognize different areas with a larger variety of buildings. However, its (un)precision is only 10m. Which of these four tools is most suitable for the given task? Probably the first tool may be needed when installing satellite or GSM antenna for the mobile network on the particular building. The second tool may be rather useful when comparing how heights of buildings are changing along with particular districts or along the streets or avenues. The third approach may be best when one cares about finding reliable values of the building heights in the whole city thanks to its high speed. Sure, the fourth method is even much faster and can provide a useful overview picture or a larger variety of buildings in a variety of districts and hope to find the lowest ones. But in the case of typical buildings, its inability to distinguish family homes from four-story apartment buildings is in many applications a discriminating factor. In this pedagogical example, I tried to present the following analogy: first tool – QM; second tool – MD; third tool – enhanced sampling MD; fourth tool – ligand-protein and protein-protein docking methods using scoring functions.

Because the dynamics of biomacromolecules and their complexes are understood differently in different research fields here is my view of the meaning of this term. There are multiple layers to the dynamics of biomolecular complexes:

i) with few exceptions, proteins should fold and unfold reversibly

ii) association and dissociation of biomolecular complexes (protein/ligand)

iii) the flexibility of protein side-chains around a stable folded conformation

iv) transitions between multiple conformational states

v) plasticity of the binding site allowing it to adapt to a variety of ligands (e.g. drug-like molecules)

vi) induced fit, conformational selection

vii) regardless of the higher or lower rigidity of biomolecules, there are always highly mobile water molecules surrounding studied molecules at room or physiological temperatures

viii) Conformational ensembles. Most of the above factors have a direct consequence on conformational ensembles and thus affect the thermodynamic/statistical physics properties such as entropy, enthalpy, and free energy.

The habilitation thesis is divided into four scientific chapters summing up 15 selected peer-review papers of the applicant after his Ph.D. defense till this date. Given that each of these chapters' uses the molecular dynamics method in some way – it is described with the related issues in the general theoretical background section preceding the scientific ones.

2 General Theoretical Background – Molecular Dynamics

P0: Jozef Hritz* & Arnost Mladek: Plant Structural Biology: Hormonal Regulations **2018,** 295-322, (ch: Computational Molecular Modeling Techniques of Biomacromolecular Systems), Eds. Hejátko, J., Hakoshima, T.

MD simulations allow the calculation of the trajectory of all atoms in the simulated system during a period of time, where the time evolution of a set of interacting atoms is followed by integrating their equations of motion. By following the dynamics of a molecular system in space and time, we can obtain information about molecular geometries and energies, mean atomic fluctuations, local fluctuations (like formation/breakage of hydrogen bonds), protein/ligand binding, free energies, and the nature of various types of concerted motions. Classical mechanics is sufficiently precise for most practical cases, except for those cases where is not possible to neglect QM effects. At the heart of molecular dynamics simulations lay the empirical force-fields that allow calculating the potential energy of the simulated system as a function of the nuclear coordinates. An overview of the commonly used biomolecular force fields, and how they were developed over the last five decades, was reviewed in e.g.[1] The applicant also recently reviewed in the book chapter[P0] - several aspects of MD in the context of convergence of produced conformational ensembles and enhanced sampling methods. In the following part of this general theoretical chapter, only the most relevant sections from[P0] are listed in the view of scientific chapters. For more general aspects such as empirical force-fields, periodic boundary conditions, geometry optimizations, MD algorithms and protocols, the reader is referred either to the book chapter itself [P0] or to the textbooks describing a variety of molecular modeling techniques from QM level through the classical MM to the fluid dynamics, e.g.[2–5].

2.1 Ensemble averages – link to the experimental data

The MD simulation is in many aspects very similar to the real experiment. It is the reason, why it is often called a computational experiment or experiment in the white and why the calculation of averages of properties of our interest is very similar. In the case of an experimentally measured value of the property $X(r^N, p^N)$, it is the average value for the duration of the experiment. In principle, the real average value would be gained in the measurement in the limit of infinite time duration:

$$X_{average} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} X\left(\overrightarrow{r^N}(t), \overrightarrow{p^N}(t)\right) dt \tag{2.1}$$

We know from experience, that the determination of the average value with sufficient precision measurement within a finite time duration is sufficient (this time has to be much longer than the relaxation time of the measured quantities). Similarly, in the MD simulation with the high enough number of steps M we can "measure" many properties of a system as the average value of discrete values in the individual steps:

$$X_{average} \cong \frac{1}{M} \sum_{i=1}^{M} X\left(\overrightarrow{r^N}_i, \overrightarrow{p^N}_i\right) \tag{2.2}$$

The dilemma appears to be that one can calculate time averages using MD simulation, but the experimental observables are assumed to be ensemble averages as outlined above. Resolving this leads us to one of the most fundamental axioms of statistical mechanics, the **ergodic hypothesis,** which states that the time average equals the ensemble average in case of infinite simulations:

$$\langle X \rangle_{ens} = \langle X \rangle_{time} \tag{2.3}$$

The basic idea is that if one allows the system to evolve in time **indefinitely**, that system will eventually pass through all possible states. One goal, therefore, of a molecular dynamics simulation is to generate enough representative conformations such that this equality is satisfied. If this is the case, experimentally relevant information concerning structural, dynamic and thermodynamic properties may then be calculated using a feasible amount of computer resources. Because the simulations are of fixed duration, one must be certain to sample a sufficient amount of phase space.

In general, the **ensemble average** may be written as:

$$\langle X \rangle_{ens} = \sum_{r^N, p^N} X(r^N, p^N) \rho(r^N, p^N) \tag{2.4}$$

where $X(r^N, p^N)$ stands for the observable of interest expressed as a function of positions r^N and momenta $p^N$of N particles, in the case of classical MD simulations of all the atoms, a system is comprised of. Note that the summation runs over all possible values of the variables r^N and p^N, or, in other words, overall possible system states. The second term in equation (2.4) is the **ensemble probability density** reflecting the intuitive fact that each microstate of the system is adopted with a different probability. According to the Boltzmann law, the **canonical ensemble** probability density is given by:

$$\rho(r^N, p^N) \sim e^{\frac{-E(r^N, p^N)}{k_B T}} \tag{2.5}$$

where $E(r^N, p^N)$ is the total energy of the system including potential energy ($E_p(r^N)$) and kinetic ($E_k(p^N)$) as a function of positions and momenta, T is the absolute temperature of the system, k_B is the Boltzmann constant. Thanks to the fact that Boltzmann factor is factorizable, i.e.:

$$e^{\frac{-E(r^N, p^N)}{k_B T}} = e^{\frac{-E_p(r^N)}{k_B T}} e^{\frac{-E_k(p^N)}{k_B T}} \tag{2.6}$$

we can write:

$$\rho(r^N) \sim e^{\frac{-E_p(r^N)}{k_B T}} \tag{2.7}$$

Considering that probability density has to be normalized we get:

$$\rho(r^N) = \frac{1}{Z} e^{\frac{-E_p(r^N)}{k_B T}} \tag{2.8}$$

where Z is the **partition function** and serve here as a normalization factor:

$$Z = \sum_{r^N} e^{\frac{-E_p(r^N)}{k_B T}} \tag{2.9}$$

<u>*Pedagogical note:*</u> An exercise from the advanced theoretical methods of Biophysical Chemistry, students are expected to calculate <U$_{\text{harm}}$> for the quadratic form of U$_{\text{harm}}$

$$U_{harm}(x) \equiv U_0 + \frac{\kappa}{2}(x - x_0)^2 \tag{2.10}$$

by using equation:[6]

$$\langle U_{harm} \rangle = \frac{\int_{-\infty}^{\infty} U_{harm}(x) e^{-\frac{U_{harm}(x)}{k_B T}} dx}{\int_{-\infty}^{\infty} e^{-\frac{U_{harm}(x)}{k_B T}} dx} \tag{2.11}$$

after several steps of integration, they finally derive $\langle U_{harm} \rangle = U_0 + \frac{k_B T}{2}$ Students are impressed that this final result does not depend on spring constant κ, i.e. the width of potential energy. This is content of **equipartition theorem** – energy tends to distribute equally among available modes (roughly degrees of freedom), with. $\frac{k_B T}{2}$ allocated to each. Also, kinetic energy that depends harmonically, averages as $\frac{k_B T}{2}$ per degree of freedom.

2.2 Replica-exchange molecular dynamics (REMD)

In addition to MD, there are also other methods that allow the generation of the canonical ensemble. Another commonly used method is the Monte-Carlo (MC) approach. In the first step of MC random step/move (e.g. of an atom) is generated and subsequently

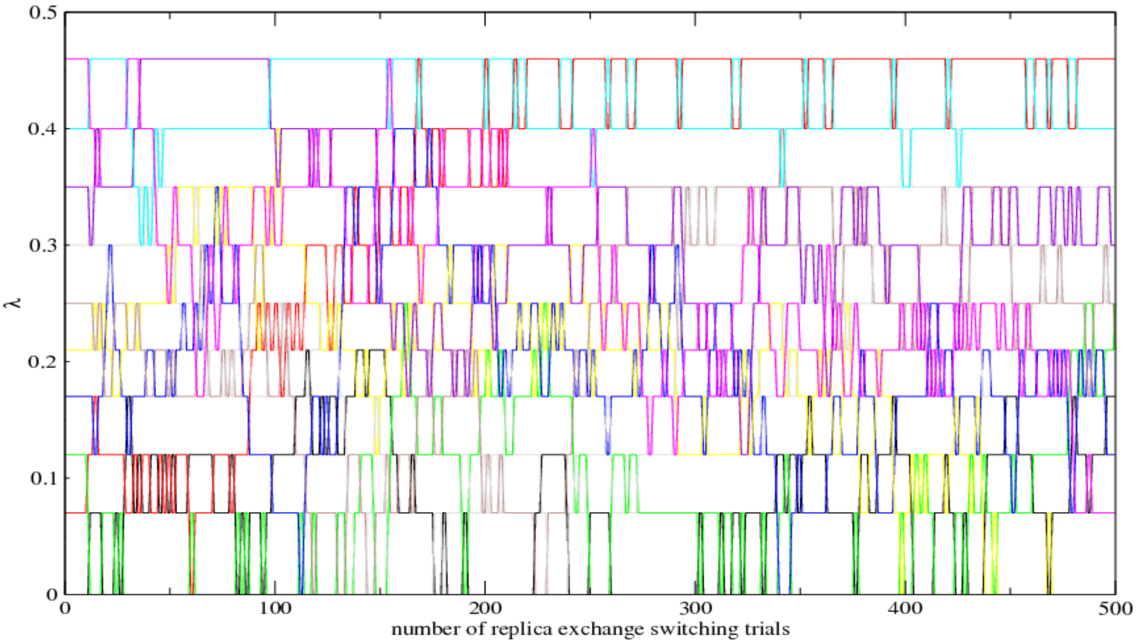


Figure 1: Schematic example of Hamiltonian replica exchange molecular dynamics (H-REMD) with ten replicas.

Different Hamiltonians of individual replicas are controlled by the parameter λ.

this state 2 (having potential energy $E_p(2)$) is accepted or rejected based on Metropolis criterion:

$$p(1 \rightarrow 2) = \begin{cases} 1 & for\ E_p(2) \leq E_p(1) \\ e^{-\frac{E_p(2)-E_p(1)}{k_bT}} & for\ E_p(2) > E_p(1) \end{cases} \tag{2.12}$$

The MC method can be very efficient at simulating liquids or systems in a vacuum. However, simulations of complex biomolecular systems in explicit water environments suffer from the fact that most of the trial moves are rejected because they lead to a high increase of potential energy. This is the very problem that the replica-exchange molecular dynamics method (REMD) tries to solve.[7]

The principle of REMD combines the ideas and advantages of MD and MC methods.[8,9] In REMD simulations, multiple non-interacting replicas of the system are simulated in parallel by MD, each at a different condition. Conservation of the Boltzmann canonical structural ensemble for individual conditions is guaranteed by the application of the Metropolis criterion for exchange trials of complete configurations between neighboring replicas. Schematically it is presented in Fig. 1. The REMD methods differ in the chosen parameter that changes among the simulated replicas. There are two types of REMD distinguished by varying condition: temperature REMD (T-REMD)[10] and Hamiltonian REMD (H-REMD).[11–13]

T-REMD can be applied straightforwardly to any system, although the maximal temperature difference between neighboring replicas depends inversely on the square root of the number of degrees of freedom, essentially prohibiting the application of T-REMD for systems in explicit solvent. H-REMD has a much better potential for systems in explicit solvent as the Hamiltonians of individual replicas may differ only in selected interactions, resulting in a smaller number of replicas to be used. Once well-chosen modifications of the interactions contributing most to high energy barriers are made, H-REMD can provide a sampling efficiency that is several orders of magnitude higher as compared to T-REMD.[7] However, efficient H-REMD needs a priori knowledge of the origin of energy barriers in the system, which implies that different systems require different H-REMD schemes.[14]

Temperature replica-exchange molecular dynamics (T-REMD)[9,10,15]

A T-REMD scheme starts with choosing the temperatures. Values of temperature must cover a wide range including the biologically relevant ones (e.g. 37°C) as well as temperatures at which sampling is significantly enhanced, i.e. elevated temperatures. Intuitively, systems at higher temperatures can overcome energy barriers easily, compared to a situation at low temperatures. However, the temperature does not only change free energy barriers, but it also changes the free energy values of individual minima. This is secured by the replica-exchange part. At certain time intervals, for example, every picosecond, energies of neighboring replicas are compared. A replica simulated at a lower temperature tends to have lower potential energy than the one simulated at a higher temperature. However, as temperature differences between replicas are small, it may happen that the potential energy of the high-temperature replica is lower. In this case, coordinates of replicas are exchanged and the simulation of low-temperature continues at higher temperature and vice versa (probability of exchange is 1). If not, replica-exchange probability is calculated from the potential energy difference as:

$$P = \exp\left[\left(E_i - E_j\right)\left(\frac{1}{k_B T_i} - \frac{1}{k_B T_j}\right)\right] \tag{2.13}$$

where E_i and E_j are energies of the i-th and j-th replica, respectively. Similarly, T_i and T_j are respective absolute temperatures of the i-th and j-th replica.[16] The computer then generates a random number in the range from 0 to 1. If this number is lower than the calculated probability, replicas are exchanged and the simulation continues. At the end of the simulation, it is possible to collect all coordinate snapshots at a certain temperature. The exchange criterion is designed in the way that such collected snapshots sample the studied system canonically at a given temperature. This means that it is possible to calculate probabilities of different conformational families and the free energy surface, not only for the biological temperature.

Hamiltonian replica-exchange molecular dynamics (H-REMD)[11,17]

As the temperature-based REMD method suffers from the need of high numbers of replicas to be efficient, the Hamiltonian replica-exchange (H-REMD) is often the method of choice.[12,18] The H-REMD method is grounded on the consideration that, since the different parallel simulations do not interact (i.e. independent simulations), there is no need to use the same Hamiltonian for all of the replicas. In H-REMD the different replicas are usually simulated at a constant temperature, while the Hamiltonian of the system is used as the replica coordinate. While the standard part of the Hamiltonian (the kinetic and the potential energy) is usually common for all simulated replicas, the additional bias term is replica-dependent. The switching probability α_{ij} is determined based on the energy difference following the Metropolis criterion:

$$\alpha_{ij} = \min\left\{1, e^{-\frac{H_i(q_j)+H_i(q_i)}{k_B T_i}+\frac{-H_j(q_i)+H_j(q_j)}{k_B T_j}}\right\} \tag{2.14}$$

where Hamiltonian H is defined as the sum of the force field and the bias potential. There are various ways to modify the Hamiltonian. For protein-protein interactions, it is often advantageous to complement the standard Hamiltonian with a distance restrain potential.[19] To overcome energy barriers, it is possible to use variable soft-core potentials for the interactions that contribute most to the energy barriers.[14] It should be stressed, however, that the setup of the replica-coordinate parameters may be non-trivial and it is often necessary to go through a trial and error tuning process. Another possibility is to utilize some kind of optimization algorithm to get the most appropriate parameter values.[20]

Summing 15 selected peer-review papers in 4 scientific chapters:

3 Efficient incorporation of plasticity and water molecules into the molecular docking approach

Paper 1: Hritz, J.; de Ruiter, A.; Oostenbrink, C. Impact of plasticity and flexibility on docking results for Cytochrome P450 2D6: a combined approach of molecular dynamics and ligand docking. *J. Med. Chem.* **2008**, 51, 7469-7477

Paper 2: Santos, R.; Hritz, J.; Oostenbrink, C. The role of water in molecular docking simulations of Cytochrome P450 2D6. *J. Chem. Inf. Model.* **2010**, 50, 146-154

Paper 3: Oostenbrink C.; de Ruiter A.; Hritz J.; Vermeulen N.P.E Malleability and versatility of Cytochrome P450 active sites studied by molecular simulations. *Curr. Drug Metab.* **2012**, 13, 190-196

Paper 4: Byeon I-J. , Ahn J., Mitra M., Byeon C-H., Hercík K., Hritz J., Charlton L., Levin J., Gronenborn A.M. NMR structure of human restriction factor APOBEC3A reveals substrate binding and enzyme specificity. *Nat. Commun.* **2013**, 4, 1890

Ligand-protein and protein-protein docking are very popular techniques because of their high speed and because they provide a reasonable variety of bound conformations. The majority of ligand-protein docking approaches treat the proteins as mostly rigid and ligands as flexible. Typically, their limitations lie in lacking the possibility to incorporate the induced fit effect or explicit water molecules in the expected binding cavity. This limitation is quite obvious in the case of the crystal structure of the apo form of Cytochrome P450 2D6 (pdb code: 2F9Q),[21] an important target of pharmaceutical companies trying to avoid drugs for which human liver metabolism is solely dependent on only this enzyme. The reason is that about 10% of the western population lacks this enzyme and thus may easily be overdosed if the drug is degraded only by Cytochrome P450 2D6. For this reason, companies like to exclude those cases in the early stages of drug development preferably in initial computational screening utilizing ligand docking amongst other approaches. Here the problem is that the traditional ligand-protein docking approaches lead to only a 20% reliability of prediction of the site of metabolism (SOM). In [P1] we introduced an approach combining ensemble ligand docking and molecular dynamics (MD) simulations, incorporating the plasticity and flexibility of cytochrome P450 2D6 structures in a highly efficient way and leading to a significant increase in the reliability of the predicted binding poses.

The presented approach selects the three most suitable structures out of several thousand generated by MD. **We have also developed a binary decision tree to decide which protein structure to dock the substrate into, such that each ligand needs to be docked only once, leading to a successful site of metabolism prediction in 80% of the substrates.** This approach has wide applicability and became quite popular in the community as is documented by over a hundred citations of this paper. Another encouraging fact was that based on an existing crystal structure of CYP2D6 in the apo form, our computational results revealed different conformational states of Phe483 for different sets of substrates. Four years later our data were confirmed by crystallography studies of CYP2D6 in complex with various compounds.[22]

While being quite successful in the incorporation of induced fit and flexibility, our effort to incorporate the presence of critical water molecules into the docking approach for CYP2D6.[23,P2] was less successful. The paper of my master student describes a possible way to consider possibly critical water molecules but then let a docking algorithm "decide" whether to keep it or not based on the scoring function.[P2] Although this approach works, the reliability of outcomes was increased by only 1-2 % which considering the complexity of the proposed approach is probably not worthwhile for the application field. Rather the very direct and much faster approach that was already described above[P1] would be preferred. The comparison of a variety of computational approaches applied in the field of Cytochrome P450 was reviewed by our group in[P3].

Here it is important to emphasize that the described incorporation of induced fit and flexibility in molecular docking has very general applicability outside of the Cytochrome P450 field. We for example used a version of this approach for the prediction of the complex structure between highly flexible oligonucleotides in complex with human restriction factor APOBEC3A protein. In this approach, we first generated a relatively large structural ensemble of the most flexible parts of the studied system, performed molecular docking, and then refined them through the use of NMR data. This approach has been developed during my post-doctoral stay in the research group of Prof. Angela Gronenborn.[P4]

While the modifications listed in this chapter provide improved reliability in structural information, they were still unsatisfactory in terms of the prediction of binding affinities from docking simulations due to the limiting potential of scoring functions. Methods allowing for more reliable predictions of absolute and relative binding affinities and related to this, sufficient conformational sampling, are addressed in the following two chapters.

Paper 1

Hritz, J.; de Ruiter, A.; Oostenbrink, C. Impact of plasticity and flexibility on docking results for Cytochrome P450 2D6: a combined approach of molecular dynamics and ligand docking.
J. Med. Chem. **2008**, 51, 7469-7477

Journal of
Medicinal Chemistry

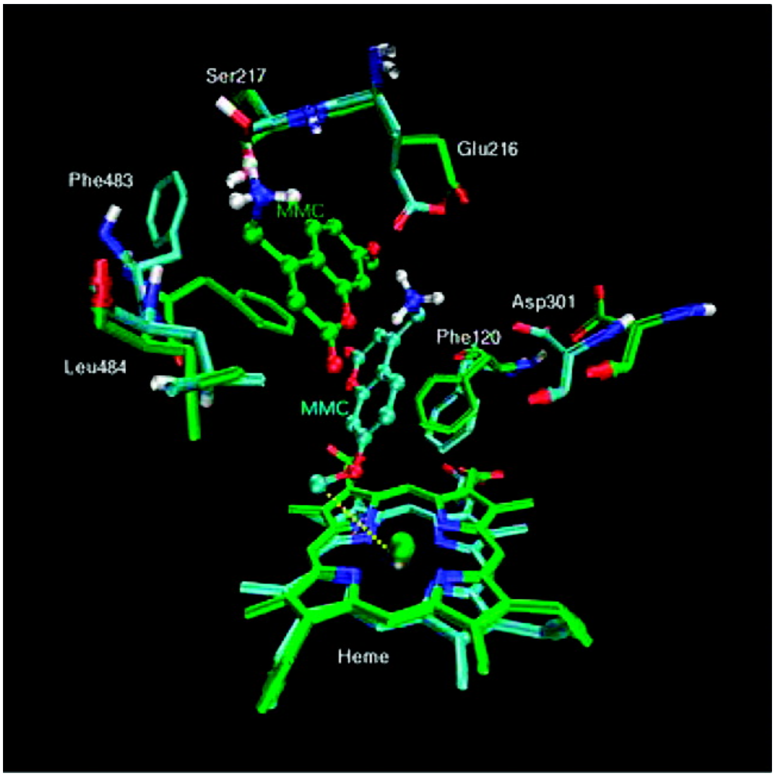
Article

Impact of Plasticity and Flexibility on Docking Results for Cytochrome P450 2D6: A Combined Approach of Molecular Dynamics and Ligand Docking

Jozef Hritz, Anita de Ruiter, and Chris Oostenbrink

J. Med. Chem., **2008**, 51 (23), 7469-7477 • Publication Date (Web): 11 November 2008

Downloaded from http://pubs.acs.org on December 15, 2008



Impact of Plasticity and Flexibility on Docking Results for Cytochrome P450 2D6: A Combined Approach of Molecular Dynamics and Ligand Docking

Jozef Hritz, Anita de Ruiter, and Chris Oostenbrink*

Leiden-Amsterdam Center for Drug Research, Section of Molecular Toxicology, Department of Chemistry and Pharmacochemistry, Vrije Universiteit, De Boelelaan 1083, 1081 HV Amsterdam, The Netherlands

Received August 8, 2008

Cytochrome P450s (CYPs) exhibit a large plasticity and flexibility in the active site allowing for the binding of a large variety of substrates. The impact of plasticity and flexibility on ligand binding is investigated by docking 65 known CYP2D6 substrates to an ensemble of 2500 protein structures. The ensemble was generated by molecular dynamics simulations of CYP2D6 in complex with five representative substrates. The effect of induced fit, the conformation of Phe483, and thermal motion on the accuracy of site of metabolism (SOM) predictions is analyzed. For future predictions, the three most essential CYP2D6 structures were selected which are suitable for different kinds of ligands. We have developed a binary decision tree to decide which protein structure to dock the ligand into, such that each ligand needs to be docked only once, leading to successful SOM prediction in 80% of the substrates.

Introduction

Cytochrome P450s (CYPs)[a] form a superfamily of heme−thiolate-containing proteins which play a crucial role in the metabolism of drugs and other xenobiotics.[1] CYPs generally detoxify potentially hazardous compounds, but in a number of cases nontoxic parent compounds are bioactivated into toxic metabolites or procarcinogens into the ultimate carcinogens. The human isoform cytochrome P450 2D6 (CYP2D6) constitutes only ~2% of the hepatic CYPs but is responsible for the metabolism of 15−20% of currently marketed drugs.[2,3] CYP2D6 substrates and inhibitors are commonly characterized by a protonated basic nitrogen and an aromatic system, which are also common features in drugs addressing the central nervous or cardiovascular system.[4] Inhibition of CYP2D6 may easily lead to adverse drug−drug interactions. Large interindividual differences exist in CYP2D6 activity, due to gene multiplicity and genetic polymorphisms, emphasizing the need to include metabolism prediction early in the drug discovery process.[5,6] Early considerations of AMDET properties (absorption, metabolism, distribution, excretion, and toxicology) include attempts to model the CYP activity *in silico*.[7−9] For many years, structure-based approaches had to rely on homology models[10] until the first (apo) structure of the enzyme was published in 2006.[11]

Many CYPs seem to show a larger extent of active site plasticity and flexibility when compared to many other proteins. This may be explained from the ability of CYPs to bind and metabolize a large variety of substrates.[12,13] Even if standard docking approaches are quite reliable for less flexible proteins, special care has to be taken for proteins with very malleable active sites. This paper describes the impact of different kinds of flexibility of CYP2D6 on the prediction of the site of metabolism (SOM). We will describe strategies to incorporate the most essential structural variety into efficient ligand docking methods. We use a variant of ligand docking into an ensemble of protein structures generated by molecular dynamics (MD) simulations. This approach was introduced by Pang and Kozikovski in their docking study of huperzine A to 69 snapshots of a MD trajectory of acetylcholinesterase.[14] The rationale behind this approach is supported by recent advances in the description of ligand−protein binding. The traditional lock-and-key picture seems to be gradually replaced by a model in which an ensemble of protein conformations is described and a ligand selectively binds to the most appropriate shape of the binding site.[15,16]

In silico structure-based predictions of drug metabolism are usually considered to be made up of (1) the binding orientation of the substrate in the active site, placing the reactive group in close vicinity of the heme iron atom, and (2) the intrinsic reactivity of this group.[17,18] The relative importance of these two issues will differ between different CYPs. The substrate binding pose will play a crucial role for the tight binding site of CYP1A2 but seems to be less relevant for the large binding site of CYP3A4. Here we address the prediction of the site of metabolism (SOM) based on the most likely poses of substrates within the active site of CYP2D6. In particular, we will address the effects of protein plasticity and flexibility on such predictions. For this, many thousands of docking experiments are performed, based on which a small number of protein structures will be selected which are most optimal for SOM predictions.

The only available crystal structure of CYP2D6 is an apo structure, for which the crystallographers already observed that the active site is too small to fit known CYP2D6 substrates.[11] Also, other groups reported better results when docking into CYP2D6 structures refined from MD simulations with ligand bound.[19,20] The active site needs to be opened up, to accommodate a variety of substrates. This raises the questions how much the active site needs to be enlarged and, more importantly, if different substrates should be fitted into the same shape of the active site. Because of their versatile function in the metabolism of diverse compounds, CYPs are known to show large plasticity in their active sites, being able to accommodate many different substrates.[12] In this work, we distinguish three different kinds of protein flexibility.

* To whom correspondence should be addressed. Phone: +31 20 5987606. Fax: +31 20 5987610. E-mail: c.oostenbrink@few.vu.nl.

[a] Abbreviations: CYP, cytochrome P450; CYP2D6, cytochrome P450 isoform 2D6; EDR, (*R*)-3,4-methylenedioxy-*N*-ethylamphetamine; LPBS, ligand probing binding site method; MD, molecular dynamics; MMC, 7-methoxy-4-(aminomethyl)coumarin; PPD, (*R*)-propranolol; CHZ, chlorpromazine; TMF, tamoxifen; SOM, site of metabolism.

10.1021/jm801005m CCC: $40.75
Published on Web 11/11/2008

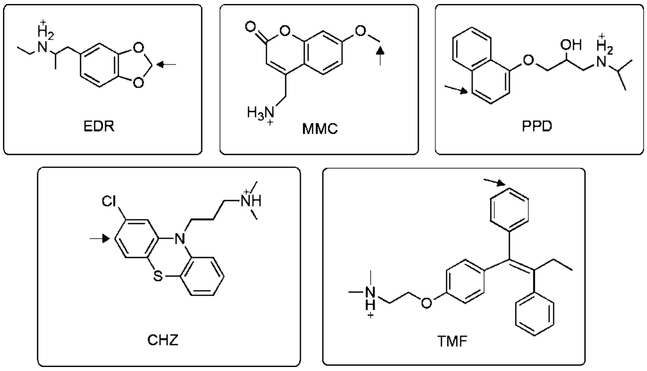


Figure 1. Chemical structures of the five representative substrates: (*R*)-3,4-methylenedioxy-*N*-ethylamphetamine (EDR), 7-methoxy-4-(aminomethyl)coumarin (MMC), (*R*)-propranolol (PPD), chlorpromazine (CHZ), and tamoxifen (TMF).

(1) Structurally different substrates will induce different conformations of the protein active site or, alternatively, specifically bind to different conformations from the overall structural ensemble of the protein.[16]

(2) Within the conformations suitable to accommodate a specific substrate, one may still distinguish different subclasses of conformations, which are separated by relatively high energetic barriers. In the case of CYP2D6 two such conformations are observed for the side chain of Phe483, as will be explained below.

(3) All atoms constituting the active site are continuously in thermal motion, leading to smaller fluctuations in the shape of the active site. We will show that also thermal fluctuations may have a strong effect on SOM predictions.

To incorporate these three types of plasticity and flexibility, we generate a library of 2500 protein structures, which covers the various conformations of the CYP2D6 active site. Subsequent docking experiments on all members of this library allows for an analysis of the effects of the different phenomena on the accuracy of SOM predictions. Finally, we will select the essential ensembles of protein structures and propose an optimal set of protein structures to be used in docking experiments. A simple decision tree model is also presented to select for any given putative substrate, which protein structure is most likely to yield an accurate SOM prediction.

Results

The library of protein conformations was built up in the following way. To account for induced fit effects, five representative substrates (Figure 1) were taken from a database of 65 known CYP2D6 substrates. Forty-five substrates could be grouped into five clusters, represented by the substrates in Figure 1. The remaining 20 substrates were placed in the sixth cluster. See Figure S1 in the Supporting Information for the complete set of substrates, how they are clustered, and an indication of the major SOM. We remark that for some substrates multiple SOMs were considered. One of the hypotheses to be investigated here is that the active site adopts a conformation around the representative substrates which is similar to the conformation it adopts around the other members of the cluster.

Opening of CYP2D6 Catalytic Site. Docking the smallest representative substrate (EDR) into the apo structure of CYP2D6 resulted in the SOM being placed at 4.5 Å from the heme iron. Subsequently, the system was solvated and electroneutralized by adding counterions and relaxed by molecular dynamics (MD) simulation. All waters, counterions, and EDR atoms were removed from the structure obtained after 100 ps of unrestrained MD, which was subsequently used for the docking of the larger representative substrate MMC. This procedure was repeated for the other representative substrates of increasing size (PPD, CHZ, TMF), thereby inducing CYP2D6 conformational changes to accommodate a wider range of substrates.

Phe483 Conformations. The largest conformational changes of the protein (mainly side chains) around the representative substrates occurred during the 100 ps of MD described in the previous paragraph. Subsequent MD simulations were performed to include thermal motions in equilibrium. If different conformations present at room temperature are separated by a low-energy barrier, then they should be adequately sampled within 0.5 ns of MD, but in the case of larger barriers the sampling within this time may be inadequate.

MD simulations of CYP2D6 in complex with each of the representative substrates revealed that two conformations of the Phe483 side chain are possible but that transitions take place on a time scale (~1 ns) that is too long to accurately determine the relative populations of these conformations. Therefore, we have performed 1 ns MD simulations for each of the representative substrates in which the χ_1 angle of Phe483 (χ_1^{483}) was restrained to either 70° or 170°. This leads to 10 structural ensembles of active site conformations, from each of which we stored 250 structures, sampled from the last 500 ps of the simulations. We refer to the ensembles by the name of the representative substrate used to generate them, followed by the approximate value of χ_1^{483}: EDR_170 represents the ensemble of 250 CYP2D6 structures which were obtained from MD of CYP2D6 in complex with EDR in which the dihedral angle χ_1 of the Phe483 side chain was restrained to 170°. The complete structural library now contains 2500 protein structures.

Docking into CYP2D6 Structural Ensembles. Sixty-five substrates of CYP2D6 with known sites of metabolism were docked into 250 protein structures for each of the 10 ensembles. An example of SOM prediction for the representative substrates for one particular docking simulation to one protein structure is shown under "single docking" in Table 1. This table presents the distance of the SOM from the heme iron atom (third column) for the first ranked pose (score in the fourth column).

As described in the Experimental Section, five docking simulations were performed for every protein structure, leading to the final prediction for one particular protein structure shown under "five dockings" in Table 1. It does not only show the SOM prediction (fifth column) but also the corresponding consensus between the five docking simulations (sixth column). These data were averaged over 250 CYP2D6 structures for every ensemble. The columns labeled as "250 frame ensemble" in Table 1 present for each representative substrate the fraction of frames for which the binding mode was predicted with the SOM within 6 Å from the heme iron atom. The same percentages for all 65 substrates docked into the PPD_70 ensemble giving the most reliable SOM prediction (as shown below) are presented in Figure 2. This figure is divided into six parts; the first five parts of the table contain the substrates belonging to clusters characterized by the listed representative ligands. Ligands which do not belong to any of these five clusters are listed in the sixth part of the figure. Tables with the same statistical analysis extended with consensus results for the highest ranked pose, over the five highest ranked poses, and over all 10 docking poses for all 10 ensembles are available in the Supporting Information (Tables S1−S10). For all 10 ensembles we note that the average reliability of binding mode prediction selecting the binding pose with the highest score from five independent dockings (third column, labeled "first rank") is slightly higher than the average

Table 1. Best Scored Results from One Particular Docking Simulation, Final Prediction Based on Five Independent Docking Simulations into the Last Frame, and Averages over the Complete PPD_70 Ensemble of CYP2D6 Structures, Showing Only the Five Representative Substrates

		single docking[c]		five dockings[f]		250 frame ensemble[i]	
code[a]	name[b]	dist[d]	score[e]	dist[g]	perc[h] (%)	first rank[j] (%)	consens[k] (%)
EDR	*R*-MDEA	3.5	24.4196	3.5	80	52.0	53.5
MMC	MAMC	3.5	26.7666	3.9	100	98.0	98.1
PPD	(*R*)-propranolol	4.6	36.1397	3.2	100	99.2	98.2
CHZ	chlorpromazine	3.5	41.9164	3.5	100	31.2	39.8
TMF	tamoxifen	3.3	42.7726	3.7	100	86.4	86.6
total[l]		46/65		44/65	70.2	62.0	61.8

[a] Three letter code for the substrate. [b] Name of the substrate. [c] First ranked docking pose over 10 poses from a single docking simulation. [d] Distance from the SOM to the heme iron in the first ranked docking pose. [e] Value of the Chemscore docking score for the first ranked pose. [f] Results based on five independent dockings into the same frame. [g] Distance from the SOM to the heme iron in the best ranked docking pose over all 50 poses. [h] Fraction of the first ranked poses in individual dockings (10 poses) giving the same SOM prediction as the overall best ranked pose. [i] Results based on five independent dockings into the ensemble of 250 frames. [j] Percentage of frames (over 250 frames) where the highest scored binding pose (over five docking simulations; 50 poses) places the SOM within 6 Å distance from the heme iron atom. [k] Same percentage as in footnote *j*, averaged over the first ranked poses of all docking simulations. [l] Statistics for all 65 substrates; number of substrates for which the SOM was predicted within 6 Å from the heme iron for the best scoring pose in one docking simulation, over five docking simulations; the last three columns are averages over corresponding columns.

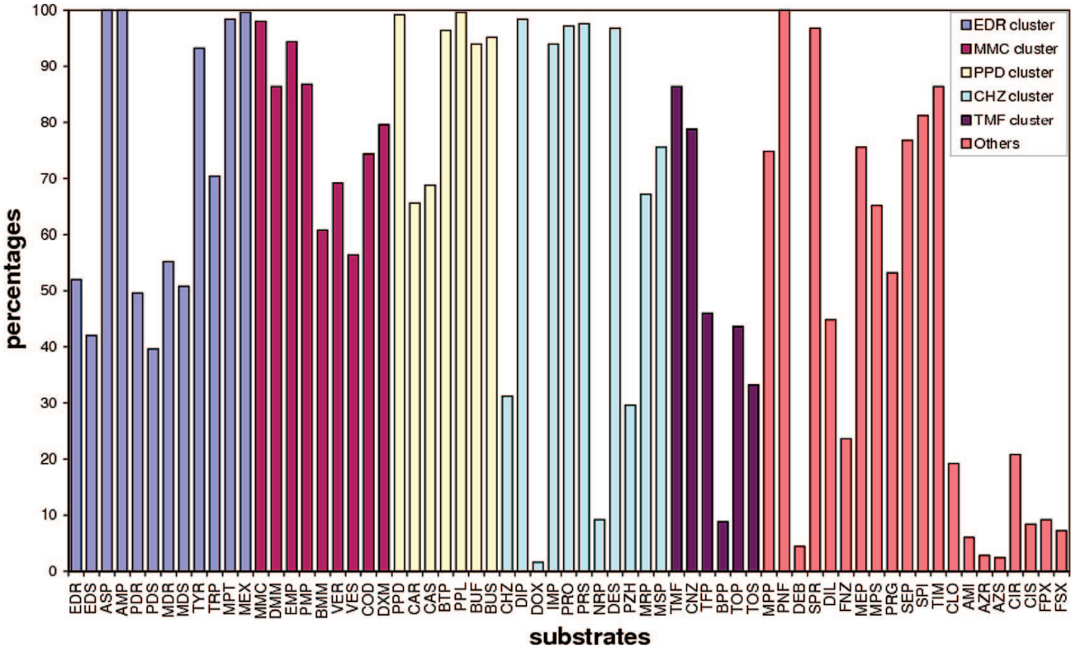


Figure 2. Percentage of frames within the PPD_70 ensemble in which the highest scored binding pose (over five docking simulations) places the SOM within 6 Å distance from the heme iron atom. Chemical structures and names of all 65 substrates corresponding to the three letter codes are listed in Figure S1 (a−f) and Tables S1−S10, respectively, in the Supporting Information. Substrates are divided into six groups according to five clusters named by representative substrates, indicated with different colors in this figure.

over all first ranked poses of the individual dockings (fourth column, labeled "consens").

Discussion

The structural ensemble of CYP2D6 covers a wide range of active site plasticity and flexibility. The effect of thermal motion, induced fit, and the conformation of Phe483 on the reliability of the SOM prediction is discussed in the following sections.

Impact of Thermal Fluctuations on CYP2D6 Docking Predictions. Figure 3 shows the variation in the number of substrates for which the first ranked pose places the SOM within 6 Å distance from the heme iron atom for the PPD_70 ensemble. A sampling time of 2 ps between subsequent structures allows for only small conformational changes. Still, it can be seen that even such small changes in the protein structure have an impact on the SOM prediction. For most ensembles the deviations from the average values (summarized in Table 2) are within 10%, but more extreme fluctuation can be observed. For instance, in the EDR_70 ensemble, the percentage of substrates that is correctly placed in the active site changes from 0% in structure 239 (indicated as EDR_70_fr_239) to 41.5% in the next structure (EDR_70_fr_240). We do not see any obvious structural difference between these protein structures (Figure 4), which indicates that very subtle differences in protein structures can have tremendous effects on the SOM prediction. In Figure 4, we show the binding mode prediction for substrate MMC into both frames, which demonstrates the observation that many substrates are bound to a subpocket between Gln244 and Ala300 when docked to the EDR_70_fr_239 structure. It seems that such erratic behavior in the pose predictions can be traced to the fact that the scoring functions are not continuous, so that small structural changes can have big effects on the prediction of the scores. We also emphasize that a correct inclusion of entropic effects would prevent rare events from becoming the first ranked predictions. This example demonstrates the risk of blindly selecting a single protein structure and strengthens us in the belief that one should consider the values that are averaged over the structural ensemble (last two

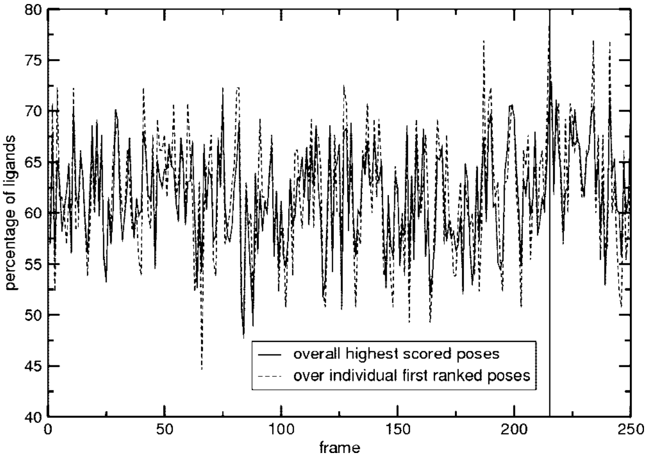


Figure 3. Docking into individual frames of the PPD_70 ensemble of CYP2D6 structures. Solid line: The fraction of substrates for which the binding pose with the overall highest score (from five docking simulations) positions the SOM within 6 Å from the heme iron atom. Dashed line: The average value when considering the first ranked poses for individual docking simulations separately. Frame 216 (indicated by the vertical line) was finally taken as the single CYP2D6 structure giving the highest docking prediction reliability.

Table 2. Average Percentage over All Structures and 65 Substrates for Each of the 10 Generated Ensembles

	CYP2D6/ EDR (%)	CYP2D6/ MMC (%)	CYP2D6/ PPD (%)	CYP2D6/ CHZ (%)	CYP2D6/ TMF (%)
$\chi_1^{483} \approx 70°$	45.4	42.2	62.0	57.4	47.4
$\chi_1^{483} \approx 170°$	48.6	48.2	62.0	59.6	51.4

columns of Table 1 and Tables S1−S10) or make a careful selection of single frames as shown below.

Impact of Induced Fit Effects on CYP2D6 Docking Predictions. Table 3 presents the ligand−probe binding site (LPBS) distance (see Experimental Section) between the 10 structural ensembles, based on the percentages of correct predictions in Tables S1−S10. The largest difference of 36% is observed between the MMC_70 and PPD_70 ensembles. Different structural ensembles obtained from simulations of different representative substrates are not equally well able to accommodate the various substrates. This can be seen from the percentage of frames in which individual substrates are positioned with their SOM within 6 Å from the heme iron atom (Tables S1−S10). We can determine if a substrate is indeed most often correctly placed in the ensemble generated for the representative substrate of the cluster to which it belongs. Table 4 shows how the percentages of protein structures in which a correct pose was predicted for the representative substrates docked into all ensembles. It can be seen that for MMC, PPD, and CHZ we indeed obtain the highest reliability when docking into protein structures refined from MD of CYP2D6 in complex with themselves. For TMF a higher reliability (~90%) is obtained when docking into the CHZ_170 ensemble. Still, the TMF_70 ensemble itself also leads to a quite high reliability (~70%). On the other hand, the reliability of docking EDR into the EDR_70 and EDR_170 ensembles is significantly lower (<40%) than when docking to the PPD_170 ensemble (~63%).

Table 4 also shows that one of the most promiscuous substrates is MMC because a high reliability is obtained when docking to most of the structural ensembles. On the other hand, E-doxepine seems to be very sensitive to the shape of the protein binding site. Successful binding poses were only obtained to a significant level (70%) when it was docked into the MMC_70 ensemble.

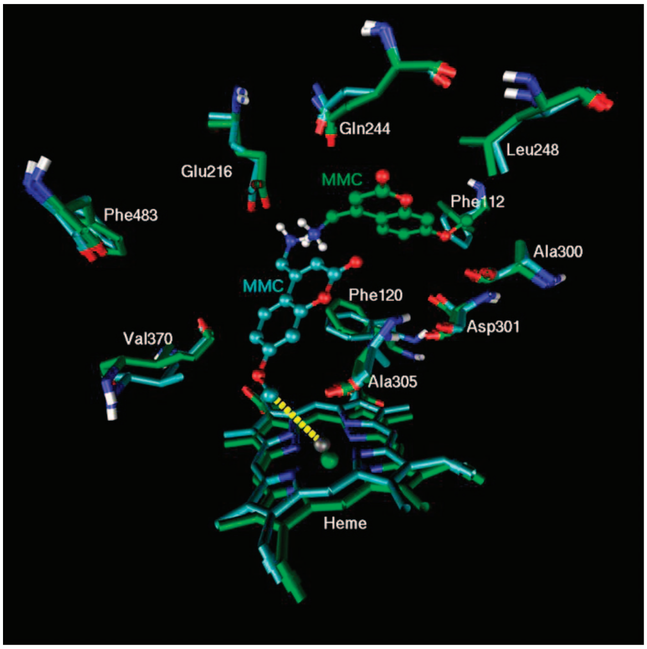


Figure 4. Binding mode prediction for MMC (ball and stick) when docked into the EDR_70_fr_239 structure (shown in green) and in the very similar EDR_70_fr_240 structure (shown in cyan). No successful correct poses were observed for any of 65 substrates in the green structure, but many occupy a subpocket between Gln244 and Ala300. In the cyan structure 27 substrates docked with the SOM within 6 Å from heme iron (indicated by the yellow dashed line).

Impact of the Conformation of Phe483 on CYP2D6 Docking Predictions. From the ensemble tables (Tables S1−S10) and from Table 4, it can be seen that substrates show different sensitivity toward the side-chain conformation of Phe483. The most significant difference is observed for substrates belonging to the MMC cluster, which are significantly more often placed correctly in the MMC_170 ensemble than in the MMC_70 ensemble (Table 5). A closer inspection revealed that Phe483 with $\chi_1^{483} \approx 170°$ does not stabilize the correct binding poses but rather destabilizes the incorrect binding pose (Figure 5). It is interesting to note that an experimental study on the effect of the F483A mutation on the metabolism of four substrates shows the largest effect for MMC. The F483A mutant does not show any metabolism of this substrate anymore.[21] The molecular explanation of this observation may very well be that the mutant binds MMC in binding poses similar to the not catalytically active pose (green) in Figure 5, as this is no longer destabilized by Phe483.

Essential Conformational Ensembles. The 65 CYP2D6 substrates can be reclustered according to the ensemble of structures that accommodates them most often in catalytically active poses (Table S11 in the Supporting Information). This reclustering shows that, for almost all substrates, at least one ensemble exists that correctly predicts the SOM with high reliability, leading to an overall reliability of 75.4%. We also investigated which are the essential conformational ensembles; i.e., how many ensembles should be included to approach the value of 75.4%. Table 2 presents the average reliability over all substrates when considering single ensembles. The ensemble leading to the largest number of substrates with their SOM within 6 Å from the heme iron atom are the PPD_70 and the PPD_170 ensemble (~62%), followed by the CHZ_170 and the CHZ_70 ensemble (~58%). Because these are averages over 250 protein structures and over 65 substrates, the conformation of Phe483 has only a small effect on these percentages. Next,

Table 3. Ligand-Probed Binding Site (LPBS) Distance between the 10 Protein Ensembles[a]

	EDR_70 (%)	EDR_170 (%)	MMC_70 (%)	MMC_170 (%)	PPD_70 (%)	PPD_170 (%)	CHZ_70 (%)	CHZ_170 (%)	TMF_70 (%)	TMF_170 (%)
EDR_70	0	9.3	17.1	17.3	30.3	28.4	26.7	26.8	21.7	20.6
EDR_170		0	18.4	17.1	28.1	24.7	24.6	23.5	20.7	18.0
MMC_70			0	23.0	36.0	33.5	29.6	30.8	21.6	23.3
MMC_170				0	27.2	29.4	30.3	30.4	25.2	23.1
PPD_70					0	16.0	26.5	26.8	30.8	27.6
PPD_170						0	18.4	18.0	27.9	21.7
CHZ_70							0	10.9	26.8	18.0
CHZ_170								0	26.5	16.3
TMF_70									0	21.0
TMF_170										0

[a] For a description of the LPBS distance, see the Experimental Section.

Table 4. Fraction of Structures for Which the Substrates Dock Correctly into the Different Structural Ensembles[a]

CYP2D6 ensemble	EDR (%)	MMC (%)	PPD (%)	CHZ (%)	TMF (%)	E-doxepine[b] (%)
EDR_70	36.8	58.8	40.4	56.0	41.2	10.0
EDR_170	39.6	78.0	39.2	36.8	37.6	18.4
MMC_70	16.0	59.6	62.8	62.0	9.2	**70.0**
MMC_170	29.2	**98.4**	50.0	39.6	63.2	5.2
PPD_70	52.0	98.0	**99.2**	31.2	86.4	1.6
PPD_170	**62.8**	97.2	97.6	78.8	60.4	0.4
CHZ_70	62.4	74.0	91.2	98.0	52.4	0.4
CHZ_170	60.8	95.6	86.0	**100.0**	**90.4**	0.0
TMF_70	6.8	71.6	96.0	4.0	65.2	4.8
TMF_170	41.2	85.2	84.0	86.0	70.0	1.6

[a] A docking pose is considered to be correct if the SOM of the substrate is placed within 6 Å of the heme iron atom. [b] E-doxepine is the substrate which is the most sensitive to the structural ensemble to which it is docked.

Table 5. Effect of the Phe483 Side-Chain Conformation on the Fraction of Structures for Which the Substrates Dock Correctly into Ensembles of Structures[a]

code[b]	name[c]	MMC_70 (%)	MMC_170 (%)
MMC	MAMC	59.6	98.4
DMM	diMAMC	30.4	93.2
EMP	EMAMC	45.2	90.4
PMP	PMAMC	39.2	67.6
BMM	BMAMC	27.2	55.6
VER	(*R*)-venlafaxine	8.8	24.0
VES	(*S*)-venlafaxine	7.2	18.0
COD	codeine	3.2	13.6
DXM	dextromethorphan	18.4	76.4
average%		26.6	59.7

[a] A docking pose is considered to be correct if the SOM of the substrate is placed within 6 Å of the heme iron atom; data shown for the substrates belonging to the MMC cluster where the strongest effect was observed only. [b] Three letter code of the substrate. [c] Name of the substrate.

we have analyzed all combinations of two and three ensembles and select for every substrate the ensemble for which the highest value was obtained. This gives an upper limit to the averaged reliability over all substrates. The combinations with the best overall prediction are given in Table 6, with the optimal reliability given in the third row. The assignment of the substrates to the ensembles for all combinations is available in Table S12 of the Supporting Information. The upper bound to the SOM predictions converge to the overall upper bound based on ensembles of 75.4%, with a value of 72.3% when including only three ensembles. Note that even though all combinations of ensembles were considered, the optimal combinations always contain the same ensembles as the combination formed with one ensemble less (Table 6).

For almost all substrates there is at least one structural ensemble for which the SOM prediction is quite good. The only exceptions are (*R*)- and (*S*)-fluoxetine with a maximum reliability of only ~8%. These substrates show a trifluoro group and have their SOM at the positively charged nitrogen atom. For most CYP2D6 substrates the positively charged nitrogen atom is expected to interact with the negatively charged side chain Glu216 (see, e.g., Figures 4 and 5), and the major SOM is elsewhere in the molecule. This is also the case for the five representative substrates used in this study. However, there are also CYP2D6 substrates following different binding modes (e.g., substrates metabolized close to the positively charged nitrogen atom (N^+)) which would not be found if one would constrain the Glu216−N^+ distance during docking simulation. For all substrates that are metabolized close to the positively charged nitrogen atom, other than (*R*)- or (*S*)-fluoxetine, there is at least one ensemble that still shows a reasonable SOM prediction. In most cases, this is the TMF_70 ensemble, in which the active site was opened up most, allowing the substrate to dock in a larger variety of poses. The fact that successful ensembles are found for most substrates indicates that the CYP2D6 structural library is extensive enough to dock substrates that are quite distinct from the representative substrates.

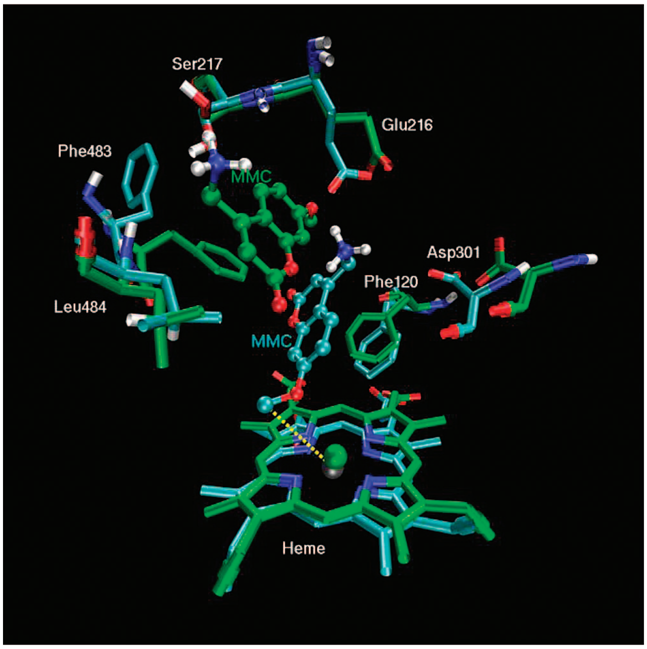


Figure 5. The number of frames into which MMC is incorrectly docked is significantly higher for the MMC_70 ensemble (40.4%) (in green) than for the MMC_170 ensemble (1.6%) (in cyan). The side chain of Phe483 with $\chi_1^{483} \approx 170°$ seems to destabilize the incorrect binding mode.

Correlation of Molecular Properties with Essential Conformational Ensembles. The observation that the reliability of SOM predictions can be as high as 72.3%, using only three structural ensembles, is promising, but it is far from trivial to

Table 6. Percentage of Correctly Docked Substrates for Essential Groups of Ensembles and Single Structures[a]

essential groups	1	2	3	all
ensembles[b]	PPD_70	PPD_70 CHZ_170	PPD_70 CHZ_170 TMF_70	all 10 ensembles
optimal reliability[c] (%)	62.0	70.3	72.3	75.4
reliability based on decision tree[d] (%)	62.0	63.6	64.2	
single structures[e]	PPD_70_fr_216	PPD_70_fr_216 CHZ_170_fr_79	PPD_70_fr_216 CHZ_170_fr_79 TMF_70_fr_3	65 structures
optimal reliability[f] (%)	71.3	86.2	89.8	100
reliability based on decision tree[g] (%)	71.3	79.4	80.3	

[a] A docking pose is considered to be correct if the SOM of the substrate is placed within 6 Å of the heme iron atom. [b] Sets of the essential ensembles. [c] Maximum reliability over 65 substrates where for each substrate the optimal value toward the ensembles in the given set is taken. [d] Averaged reliability over 65 substrates where for each substrate the ensembles are selected based on the decision tree in Figure 6. [e] Sets of the essential structures. [f] Maximum reliability over 65 substrates where for each substrate the optimal value toward the structures in the given set is taken. [g] Averaged reliability over 65 substrates where for each substrate the structures are selected based on the decision tree in Figure 6.

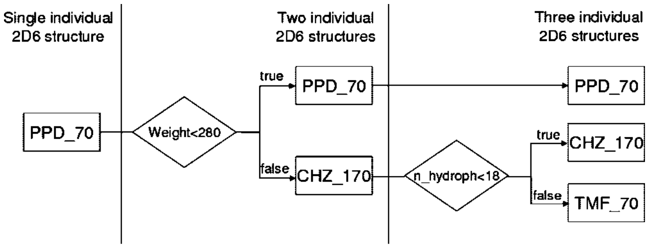

Single individual 2D6 structure
Two individual 2D6 structures
Three individual 2D6 structures
PPD_70
Weight<280
true
PPD_70
false
CHZ_170
n_hydroph<18
PPD_70
true
CHZ_170
false
TMF_70

Figure 6. Binary decision tree using the molecular weight and the number of hydrophobic atoms to separate substrates into two or three groups corresponding to different individual structures or ensembles. This way, the optimal ensemble or single protein structure can be selected for a substrate *a priori*.

determine *a priori* to which ensemble a ligand should be docked. PPD-like compounds are best docked to the PPD_70 ensemble, and CHZ-like compounds are best docked to the CHZ_170 ensemble, but there are also some stereoisomers that give better results when docked to the PPD_70 ensemble. As noted before, the TMF_70 ensemble seems most suited for substrates which are (atypically) metabolized close to the protonated nitrogen.

For new, possibly more diverse, sets of compounds it would very advantageous if general molecular descriptors could be used to predict the optimal ensemble of protein structures. Binary decision trees were developed to divide the substrates over two (PPD_70 and CHZ_170) or three (PPD_70, CHZ_170, TMF_70) ensembles. The molecular weight was found to be the most important descriptor to determine if substrates should be docked into the PPD_70 or the CHZ_170 ensemble. A further improvement of the overall reliability can be obtained, including the TMF_70 ensemble and using the number of hydrophobic atoms to divide the substrates between the CHZ_170 and TMF_70 ensembles. This decision tree is schematically drawn in Figure 6, and the average number of successful SOM predictions following this decision tree is given in the fourth row of Table 6. Note that the increase in these values is considerably lower than the upper bounds to the predictions discussed in the previous section, which shows that the separation according to the decision tree is still quite far from the optimal one.

Single Frames with the Highest Reliability of Prediction. The previous sections describe the selection of essential CYP2D6 ensembles to dock substrates with a high reliability. For many applications, such as virtual screening of large compound libraries, it is not possible to perform docking into complete structural ensembles. Therefore, it is of enormous practical interest to select only very few protein structures, which lead to accurate SOM predictions. To find the single CYP2D6 structure that offers the best SOM predictions, we first selected from the library of 2500 structures the 15 protein structures with the highest fraction of correct SOM predictions. Two additional docking simulations (each of five docking runs) were performed for these structures, to obtain a higher consensus between the different runs. Two single frames were found in which >71% of the substrates docked correctly. These originated from the PPD_70 ensemble (structure 216: PPD_70_fr_216) and from the PPD_170 ensemble (structure 173: PPD_170_fr_173).

The fact that a single structure can be selected is extremely important if one wishes to perform docking experiments for many putative inhibitors or substrates and no specific knowledge is available for the ligand. These two structures offer the most reliable docking poses out of all 2500 available protein structures even though we note that these structures are only slightly better than the 15 initially selected protein structures. We stress that it is important to select protein structures carefully. As was mentioned above, docking into the EDR_70_fr_239 structure did not position the SOM of any substrate within 6 Å of the heme iron atom within the first ranked pose. The choice of the optimal structures will depend on the docking protocol and scoring function that is used. It is unlikely that when using, e.g., a different docking program this protein structure will fail completely, but different structures may perform slightly better.

We compare the single protein structures that were selected here to a homology model that was developed several years ago in our group.[10,22] The 65 substrates were docked to this model using the same settings as before, and for 60% of the substrates the SOM was placed within 6 Å of the heme iron atom for the first ranked pose. This is significantly lower than the 71% obtained here for the newly refined structures. When docking directly to the original apo crystal structure of CYP2D6, this percentage is only 20%. Using CYP2D6 structures from a MD simulation starting from this apo structure without adding a substrate decreases the reliability of SOM predictions to 0−9%, depending on the chosen frame (data not shown).

Essential Single Protein Structures. Similar to what was described for the ensembles, we can search for a small set of essential single protein structures. The single structures selected in the previous section represent an essential set of size 1. Sets of two and three single protein structures were selected from the 2500 structures as well and are given in the fifth row of Table 6. For the combination of three structures, an exhaustive search involves searching through 2500^3 combinations. Rather, we have limited our search to sets of three structures that contain the set of two optimal protein structures. From an exhaustive search, this set already contains the PPD_70_fr_216 structure, which is the best single structure. The sixth row of Table 6 gives an upper bound to the docking reliability when using these sets of structures, by selecting for every substrate an optimal structure. Table S13 in the Supporting Information specifies for each of these sets which substrate is best docked in which protein structure. Note that the essential single structures stem from the essential ensembles determined earlier, although this was not imposed in the search algorithm. Therefore, it is not so surprising that the binary classification tree that was developed to divide the substrates over the single protein structures is the same as the one described for the essential ensembles (Figure 6). The seventh row of Table 6 presents the reliability in the docking predictions, which increases from 71.3% for a single structure to 79.4% and 80.3% when taking two or three structures into account, respectively. This increase is significantly higher than observed for the essential ensembles (fourth row of Table 6). We stress that these values can be obtained without any additional increase of computational demands as compared to docking into a single protein structure, as every substrate is docked only once to a protein structure selected based on the decision tree (Figure 6).

On the other hand, if substrates are docked to all three essential protein structures, the consensus between predictions can be used. The reliability of the binding mode prediction is 90% if the best frame is selected for every substrate. This means that if substrates are docked to all three structures, and we obtain the same binding mode, one can be almost sure that it will be correct. From the 65 substrates, 29 showed complete consensus when docking to the three essential structures. For 23 of these, the SOM is within 6 Å from heme iron atom. Two of the remaining six substrates are (*R*)- and (*S*)-fluoxetine, which were already described above as failing to dock in the majority of the 2500 protein structures. Among the 29 substrates, there are 11 substrates for which the consensus between the five individual docking runs (second part of Table 1) was 100%. All of these have the SOM within 6 Å from the heme iron atom. In cases where no consensus is reached between the different protein structures, we recommend to consider the two or three docking poses for further analysis (one of them is most likely correct). As mentioned earlier, if the substrate is docked to the TMF_170_fr_3 structure and places a charged nitrogen close to the heme iron atom, this docking pose deserves attention, even if such poses are not observed for the other two protein structures.

Conclusions

Structural refinements were performed on the crystal structure of cytochrome P450 2D6 (CYP2D6). The active site in this apo structure is too small to be used directly for predictions of the site of metabolism (SOM) in substrates (only 20% reliability), so we have adopted the active site to five representative substrates. Ten structural ensembles of 250 protein structures each were generated using molecular dynamics (MD) simulations of CYP2D6 in complex with each of the five representative substrates, in which two different conformations of the side chain of Phe483 were considered (χ_1^{483} ~70° and χ_1^{483}~170°). Docking experiments were performed for 65 substrates into all 2500 protein structures. Statistical analysis of the obtained docking poses revealed that even though thermal motion generally involves only small conformational changes, these may have a dramatic effect on the resulting docking poses. This strongly warns against performing docking experiments on any single (MD) protein structure and rather suggests to consider averages over structural ensembles or carefully selected protein structures only.

The effect of the side-chain conformation of Phe483 was seen to be strongest for the cluster of substrates represented by MMC. We suggest that Phe483 destabilizes substrate orientations far away from the heme iron, in accordance with site-directed mutagenesis data. The ensembles obtained from MD of CYP2D6 in complex with PPD and CHZ display the highest number of substrates with their SOM placed within 6 Å from the heme iron atom in the first ranked pose. The ensembles PPD_70, CHZ_170, and TMF_70 were determined to be the most essential ones.

These ensembles may be used for SOM prediction, but it is probably computationally too demanding to screen large databases for putative inhibitors or ligands by docking into complete ensembles. Therefore, we have selected a single CYP2D6 structure (PPD_70_fr_216) which offers a SOM prediction reliability of 71%. This reliability can be theoretically further increased up to 90% if part of the substrates are docked into the CHZ_170_fr_79 and TMF_70_fr_3 structures. A simple and robust decision tree was developed increasing the SOM prediction reliability to 80% without any additional computational costs. This offers a very powerful tool to perform SOM predictions efficiently, where every compound needs to be docked to one protein structure only.

The effects that we described here are not only valid for CYPs but are highly relevant for a much wider range of proteins. The large sensitivity of docking reliability not only due to larger conformational changes but also due to thermal motion should be kept in mind in any docking experiment, regardless of the protein target. Also, we have suggested techniques to open the binding site of apo structures and to select the most appropriate structures for high-throughput docking experiments. These findings are relevant for docking studies on any protein target.

Experimental Section

Clustering and Selection of Representative Substrates. A set of 65 known substrates of cytochrome P450 2D6 that was previously used in docking experiments was taken from the literature.[22] These substrates were clustered based on their expected binding mode. We assume that a conformation of the CYP2D6 active site that accommodates one representative substrate will also be able to accommodate the other substrates from the same cluster. Clustering was based on the observation that most substrates contain a positively charged nitrogen in a flexible tail, connected to a rigid body, usually consisting of several aromatic rings. The width of the substrates perpendicular to the flexible tail was used to cluster the compounds into five groups, according to Table S1 in the Supporting Information. The representative substrates for these five clusters are given in Figure 1. Twenty substrates could not be assigned to any of these clusters.

CYP2D6 Structure Preparation. The crystal structure of CYP2D6[11] was downloaded from the Protein Data Bank (www.pdb.org; code 2F9Q). We selected chain A of this apo structure, in which some atoms of a few side chains were missing. These were modeled using the program MOE.[23] Three mutations required for the crystallization were reverted to the wild-type amino acids (Asp230Leu, Arg231Leu, and Met374Val). Coordinates for the missing loop between amino acids 42 and 51 were taken from an earlier homology model.[10] Atoms for which new coordinates were assigned were minimized and equilibrated for 10 ps of molecular dynamics (MD) simulation at 300 K, while position restraints were assigned to the rest of the protein structure.

Molecular Dynamics Simulations. Molecular dynamics simulations of CYP2D6 in complex with the five representative substrates (EDR, MMC, PPD, CHZ, TMF) were performed using the GROMOS05 simulation package[24] in combination with the GROMOS 45A4 force field.[25,26] For every substrate two simulations of 1 ns were performed in which the χ_1 angle of the Phe483 was restrained to either 70° or 170° using a force constant of 30.0 kJ mol^{-1} deg^{-2}. All bonds were constrained, using the SHAKE algorithm[27] with a relative geometric accuracy of 10^{-4}, allowing for a time step of 2 fs used in the leapfrog integration scheme.[28] The system was solvated in 20292 explicit SPC water molecules[29] and 7 Na^+ ions in a periodic rectangular simulation box. After a steepest descent minimization to remove clashes between molecules, initial velocities were randomly assigned from a Maxwell–Boltzmann distribution at 300 K, according to the atomic masses. The temperature was kept constant using weak coupling to a bath of 300 K with a relaxation time of 0.1 ps.[30] The CYP2D6–substrate complex and the solvent (including the ions) were independently coupled to the heat bath. The pressure was controlled using isotropic weak coupling to atmospheric pressure with a relaxation time of 0.5 ps.[30] Van der Waals and electrostatic interactions were calculated using a triple range cutoff scheme. Interactions within a short-range cutoff of 0.8 nm were calculated every time step from a pair list that was generated every five steps. At these time points, interactions between 0.8 and 1.4 nm were also calculated and kept constant between updates. A reaction field contribution was added to the electrostatic interactions and forces to account for a homogeneous medium outside the long-range cutoff, using the relative permittivity 61 of SPC water.[31]

Docking. From the last 500 ps of each of the 10 MD simulations described above, 250 structures were stored to disk. The GROMOS 45A4 force field is a united atom force field, meaning that aliphatic hydrogen atoms are treated implicitly. Coordinates for the implicit hydrogen atoms were added, all waters, counterions, and the representative substrates were removed, and the resulting files were converted into mol2 file format using Sybyl (version 6.8) and the standard Tripos atom and bond types for the amino acids and the heme group.

Docking was performed by GOLD (Genetic Optimization for Ligand Docking)[32] version 3.3.1 in combination with the Chemscore scoring function[33] parametrized for heme-containing proteins.[34] This scoring function outperformed the Goldscore scoring function in our preliminary work (data not shown). Ten docking runs with maximally 1000 genetic algorithm operations were performed using a population of 100 genes. The center point for the docking was placed in the middle of the cavity in between residues Phe120 and Phe483. The radius from this point was set to 18 Å to include the solvent channel in the accessible volume for the docking. We stress that a shorter radius will automatically result in predicted binding modes that are closer to the heme and thus positively bias the SOM predictions. No water molecules, or a sixth ligand on the heme iron atom, were taken into account.

Docking Reproducibility. Because the genetic algorithm used for docking utilizes a considerable amount of randomness, results from individual docking simulations (consisting of 10 runs) are not reproducible. For about 20% of substrates, the evaluation of the first ranked docking pose (see below) changed when two independent docking experiments were performed. Therefore, we have performed five independent docking experiments (10 docking runs each) for all 65 substrates in all 2500 protein structures. For every experiment 10 docking poses were obtained, and the SOM prediction was based on the first ranked pose. The final SOM prediction is subsequently based on the binding mode with the highest score over all 50 binding poses. Furthermore, we assigned a weight (sixth column in Table 1) for the prediction, based on the consensus with the prediction according to the other four first-ranked poses. Note that this approach is not the same as performing a single experiment with 50 docking runs, because the docking program explicitly searches for new solutions that differ from previous runs.

Analysis. Any docking pose of a substrate in a protein structure is considered to be correct if the known SOM is within 6 Å of the heme iron atom.[22] This relatively wide criterion accounts for thermal fluctuations and dioxygen binding. In order to use this criterion in predictions for new substrates, it is important that one only needs to consider the first ranked pose, which should correspond to the binding pose with the highest affinity. Unfortunately, scoring functions are often not accurate enough to correctly predict the first ranked pose. Also, the highest affinity binding pose does not necessarily have to be the catalytically active pose. For these reasons, we have performed the distance analysis not only for the first ranked pose, but we also present the shortest distance between the heme iron atom and the substrate's SOM when considering the five highest ranked poses or all 10 generated poses in a docking experiment (see Table 1).

Ligand Probing of the Binding Sites (LPBS). We want to emphasize that the effects of conformational changes of the protein will depend on both the exact region in the protein that is modified and the substrate under consideration. Therefore, a characterization of the conformational changes in terms of an atom-positional root-mean-square deviation (rmsd) after superposition will not reflect the chances in the docking results. Rather, we calculate the root-mean-square distance between the structural features of a set of substrates when docked to two protein structures. To describe the structural differences between two protein structures, we assign to each of the 65 substrates a value of 1 if the SOM lies within 6 Å of the heme iron atom and a value of 0 otherwise. The ligands probed binding site (LPBS) distance between the two protein structures is now expressed as the rmsd over the 65 assignments in both proteins. Similarly, the LPBS distance between two ensembles is calculated using the percentage of structures in which the substrates are successfully docked. The LPBS distance truly reflects the conformational changes of the protein that make a difference for this set of substrates. Of course, any structural feature may be used to calculate such a LPBS distance.

Decision-Tree Classification. The quality of the predictions of the SOM of substrates varies strongly between the different ensembles of protein structures and between the individual protein structures involved. In order to be able to predict the SOM for new putative substrates, it is highly advantageous if the ensemble or protein structure that most likely leads to the correct prediction could be determined *a priori*. For the 65 substrates under consideration here, the major SOM is known, and we can easily determine which ensemble or individual protein structure is most successful in predicting its SOM. We have used this information to train a decision tree using QSAR classification with 2-fold cross-validation as implemented in MOE. Our database of 65 substrates was randomly divided into two mutually exclusive subsets of size 33 (learning set) and 32 substrates (test set). As initial set of descriptors, we have considered the number of atoms, hydrophobic atoms, hydrogen atoms, halogens, rings, hydrogen bond acceptors, and hydrogen bond donors, as well as SlogP, molecular volume, total hydrophobic surface area, water-accessible surface area, weight, water-accessible surface area divided by the weight, and the first, second, and third kappa shape indexes (Kier 1, Kier 2, Kier 3).[35] Several decision trees were generated at different reliability levels and complexity. To avoid overfitting, only classification trees of depth 1 or 2 were considered. In the end the most simple, robust decision tree was selected (Figure 6), which was valid for both the selection of the best performing ensemble of protein structures (from a selection of the three essential ensembles) and for the selection of the best performing protein structures (from a selection of the three essential structures).

Acknowledgment. We gratefully acknowledge financial support from The Netherlands Organization for Scientific Research, Horizon Breakthrough grant 935.18.018 (JH) and VENI Grant 700.55.401 (CO).

Supporting Information Available: Figures of 65 substrates, with indications of their SOM, clustered into six groups; tables containing the docking results averaged over the 10 protein structure ensembles; substrates reclustered according to SOM predictions into ensembles and into single protein structures; and force field parameters for the representative substrates. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Ortiz de Montellano, P. R. *Cytochrome P450: Structure, Mechanism, and Biochemistry*, 3rd ed.; Kluwer Academic/Plenum Publishers: New York, 2005.

(2) Williams, J. A.; Hyland, R.; Jones, B. C.; Smith, D. A.; Hurst, S.; Goosen, T. C.; Peterkin, V.; Koup, J. R.; Ball, S. E. Drug-drug interactions for UDP-glucuronosyltransferase substrates: A pharmacokinetic explanation for typically observed low exposure (AUC(i)/AUC) ratios. *Drug Metab. Dispos.* **2004**, *32*, 1201–1208.

(3) Bazeley, P. S.; Prithivi, S.; Struble, C. A.; Povinelli, R. J.; Sem, D. S. Synergistic use of compound properties and docking scores in neural network modeling of CYP2D6 binding: Predicting affinity and conformational sampling. *J. Chem. Inf. Model.* **2006**, *46*, 2698–2708.

(4) Rendic, S. Summary of information on human CYP enzymes: Human P450 metabolism data. *Drug Metab. Rev.* **2002**, *34*, 83–448.

(5) Pirmohamed, M.; Park, B. K. Cytochrome P450 enzyme polymorphisms and adverse drug reactions. *Toxicology* **2003**, *192*, 23–32.

(6) Ingelman-Sundberg, M. Pharmacogenetics of cytochrome P450 and its applications in drug therapy: The past, present and future. *Trends Pharmacol. Sci.* **2004**, *25*, 193–200.

(7) de Groot, M. J.; Kirton, S. B.; Sutcliffe, M. J. In silico methods for predicting ligand binding determinants of cytochromes P450. *Curr. Top. Med. Chem.* **2004**, *4*, 1803–1824.

(8) De Graaf, C.; Vermeulen, N. P. E.; Feenstra, K. A. Cytochrome P450 in silico: An integrative modeling approach. *J. Med. Chem.* **2005**, *48*, 2725–2755.

(9) Stjernschantz, E.; Vermeulen, N. P. E.; Oostenbrink, C. Computational prediction of drug binding and rationalisation of selectivity towards cytochromes P450. *Expert Opin. Drug Metabol. Toxicol.* **2008**, *4*, 513–527.

(10) de Graaf, C.; Oostenbrink, C.; Keizers, P. H. J.; van Vugt-Lussenburg, B. M. A.; van Waterschoot, R. A. B.; Tschirret-Guth, R. A.; Commandeur, J. N. M.; Vermeulen, N. P. E. Molecular modeling-guided site-directed mutagenesis of cytochrome P450 2D6. *Curr. Drug Metab.* **2007**, *8*, 59–77.

(11) Rowland, P.; Blaney, F. E.; Smyth, M. G.; Jones, J. J.; Leydon, V. R.; Oxbrow, A. K.; Lewis, C. J.; Tennant, M. G.; Modi, S.; Eggleston, D. S.; Chenery, R. J.; Bridges, A. M. Crystal structure of human cytochrome P450 2D6. *J. Biol. Chem.* **2006**, *281*, 7614–7622.

(12) Guengerich, F. P. A malleable catalyst dominates the metabolism of drugs. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 13565–13566.

(13) Ekroos, M.; Sjogren, T. Structural basis for ligand promiscuity in cytochrome P450 3A4. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 13682–13687.

(14) Pang, Y. P.; Kozikowski, A. P. Prediction of the binding-sites of huperzine-A in acetylcholinesterase by docking studies. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 669–681.

(15) Ma, B. Y.; Shatsky, M.; Wolfson, H. J.; Nussinov, R. Multiple diverse ligands binding at a single protein site: A matter of pre-existing populations. *Protein Sci.* **2002**, *11*, 184–197.

(16) Carlson, H. A. Protein flexibility and drug discovery: How to hit a moving target. *Curr. Opin. Chem. Biol.* **2002**, *6*, 447–452.

(17) Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. MetaSite: Understanding metabolism in human cytochromes from the perspective of the chemist. *J. Med. Chem.* **2005**, *48*, 6970–6979.

(18) Afzelius, L.; Arnby, C. H.; Broo, A.; Carlsson, L.; Isaksson, C.; Jurva, U.; Kjellander, B.; Kolmodin, K.; Nilsson, K.; Raubacher, F.; Weidolf, L. State-of-the-art tools for computational site of metabolism predictions: Comparative analysis, mechanistical insights, and future applications. *Drug Metab. Rev.* **2007**, *39*, 61–86.

(19) Ito, Y.; Kondo, H.; Goldfarb, P. S.; Lewis, D. F. V. Analysis of CYP2D6 substrate interactions by computational methods. *J. Mol. Graphics Model.* **2008**, *26*, 947–956.

(20) Saraceno, M.; Coi, A.; Bianucci, A. M. Molecular modelling of human CYP2D6 and molecular docking of a series of ajmalicine- and quinidine-like inhibitors. *Int. J. Biol. Macromol.* **2008**, *42*, 362–371.

(21) Lussenburg, B. M. A.; Keizers, P. H. J.; De Graaf, C.; Hidestrand, M.; Ingelman-Sundberg, M.; Vermeulen, N. P. E.; Commandeur, J. N. M. The role of phenylalanine 483 in cytochrome P450 2D6 is strongly substrate dependent. *Biochem. Pharmacol.* **2005**, *70*, 1253–1261.

(22) De Graaf, C.; Oostenbrink, C.; Keizers, P. H. J.; Van der Wijst, T.; Jongejan, A.; Vermeulen, N. P. E. Catalytic site prediction and virtual screening accuracy of cytochrome P450 2D6 substrates by consideration of water and rescoring in automated docking. *J. Med. Chem.* **2006**, *49*, 2417–2430.

(23) Molecular Operating Environment, v., Chemical Computing Group, Montreal, Canada.

(24) Christen, M.; Hunenberger, P. H.; Bakowies, D.; Baron, R.; Burgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Krautler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; Van Gunsteren, W. F. The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem.* **2005**, *26*, 1719–1751.

(25) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular simulation: The GROMOS96 manual and user guide*, Vdf Hochschulverlag AG an der ETH Zürich: Zürich, 1996.

(26) Schuler, L. D.; Daura, X.; van Gunsteren, W. F. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.* **2001**, *22*, 1205–1218.

(27) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

(28) Hockney, R. W. The potential calculations and some applications. *Methods Comput. Phys.* **1970**, *9*, 136–211.

(29) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction models for water in relation to protein hydration. In *Intermolecular Forces*, Pullman, B., Ed.; Reidel: Dordrecht, The Netherlands, 1981; pp 331−342.

(30) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(31) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. A generalized reaction field method for molecular-dynamics simulations. *J. Chem. Phys.* **1995**, *102*, 5451–5459.

(32) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748.

(33) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions 0.1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.

(34) Kirton, S. B.; Murray, C. W.; Verdonk, M. L.; Taylor, R. D. Prediction of binding modes for ligands in the cytochromes P450 and other heme-containing proteins. *Proteins: Struct., Funct., Bioinf.* **2005**, *58*, 836–844.

(35) Hall, L. H.; Kier, L. B. The molecular connectivity chi indices and kappa shape indices in structure-property modeling. In *Reviews in Computational Chemistry*, 2nd ed.; Lipkowitz, K. B.; Boyd, D. B., Eds.; VCH Publishers: New York, 1991; p 367.

JM801005M

Paper 2

Santos, R.; Hritz, J.; Oostenbrink, C. The role of water in molecular docking simulations of Cytochrome P450 2D6. *J. Chem. Inf. Model.* **2010**, 50, 146-154

Role of Water in Molecular Docking Simulations of Cytochrome P450 2D6

Rita Santos, Jozef Hritz, and Chris Oostenbrink*

Leiden-Amsterdam Center for Drug Research, Section of Molecular Toxicology, Department of Chemistry and Pharmaceutical Sciences, Vrije Universiteit, De Boelelaan 1083, 1081 HV Amsterdam, The Netherlands

Received August 7, 2009

Active-site water molecules form an important component in biological systems, facilitating promiscuous binding or an increase in specificity and affinity. Taking water molecules into account in computational approaches to drug design or site-of-metabolism predictions is currently far from straightforward. In this study, the effects of including water molecules in molecular docking simulations of the important metabolic enzyme cytochrome P450 2D6 are investigated. The structure and dynamics of water molecules that are present in the active site simultaneously with a selected substrate are described, and based on this description, water molecules are selected to be included in docking experiments into multiple protein conformations. Apart from the parent substrate, 11 similar and 53 dissimilar substrates are included to investigate the transferability of active-site hydration sites between substrates. The role of water molecules appears to be highly dependent on the protein conformation and the substrate.

INTRODUCTION

Water is a highly versatile component at the interface of biomolecular complexes because of its unique physical chemical properties: it can act both as a hydrogen-bond donor and as a hydrogen-bond acceptor; it imposes few steric constraints on bond formation; and it is able to form hydrogen-bond networks, occupying less space than the polar side chains of a protein. Therefore, water can offer a high level of adaptability to a surface, allowing promiscuous binding, or at the same time, it can provide increased specificity and affinity to an interaction.[1] The role of water molecules can be as a solvent, involved in the stabilization of a biomolecular complex; as a bridge between two molecules; as a reagent, being a substrate or product of many enzymatically catalyzed reactions; as a lubricant, through the formation of a network linking distant residues or by promoting the flexibility required for the events of the catalytic cycle;[2] and as building block of macromolecules, being part of the receptor binding site and, thereby, altering the topological surface and recognition determinants.[3]

Because of the importance of water molecules in biological systems, discussion has arisen in the literature as to whether these molecules should or should not be included in computational approaches for structure-based drug design, mainly in molecular docking simulations.[4−10] However, no consensus has been reached about the role of water molecules in such simulations. For instance, Huang et al. reported that the docking accuracy of 12 targets from the DUD database was improved by including water molecules,[11] whereas Birch et al. described the opposite in a study of neuraminidase structures.[12]

Cytochromes P450 are a family of heme-containing oxygenases and are considered to be highly important in phase I biotransformation.[13,14] There are several human subfamilies of cytochromes P450, mainly involved in the synthesis of critical signaling molecules for homeostasis of the organism and in the metabolism of endogenous (fatty acids, steroids, prostaglandins) and exogenous (natural products contained in plants, drugs, environmental pollutants) compounds. By rendering the compounds more water-soluble, their excretion is facilitated.[13]

Human cytochrome P450 2D6 (CYP2D6) is considered to be an important drug-metabolizing enzyme, despite the fact that it corresponds to only approximately 2% of the cytochrome P450 liver content. CYP2D6 is responsible for the metabolism of approximately 15−20% of the current drugs on the market, such as beta blockers, neuroleptics, antidepressants, and others.[15] Genetic polymorphisms of CYP2D6 are known. For instance, 7% of the Caucasian population does not possess this functional enzyme, resulting in the defective metabolism of many important drug molecules.[16] Therefore, the assessment of the metabolism of drugs under development is desirable as early as possible to prevent adverse drug effects. Metabolism prediction in silico can speed up the identification of compounds that might be relevant for further investigation. We distinguish methods that are based on the molecular structure and reactivity of substrates solely from methods that take the protein structure explicitly into account.[17] The commercial program MetaSite combines both approaches.[18] As an alternative, molecular docking simulations are most commonly used to predict the orientation of substrates in the enzyme active site.

The apo crystal structure of human cytochrome P450 2D6 was recently resolved by Rowland et al.,[19] and so far, this is the only available crystal structure of CYP2D6. Eleven water molecules can be found in the crystal structure of CYP2D6; however, they are positioned more than 1.0 nm away from the heme iron. A previous study using a homology model of CYP2D6 indicated that strategically placed static water molecules in the active site can significantly influence the binding pose of 65 substrates.[20] On the

* Corresponding author phone: +31 20 5987606; fax: +31 20 5987610; e-mail c.oostenbrink@few.vu.nl.

10.1021/ci900293e CCC: $40.75 © 2010 American Chemical Society
Published on Web 11/09/2009

other hand, we showed recently that active-site flexibility and plasticity could account for similar effects.[21] It is unclear what the role of active-site water molecules is in the orientation of the substrates in the active site. Should water molecules be included explicitly in molecular docking simulations, or is their effect negative? Is this role substrate-dependent, and can the selection of water molecules in the active site be automated?

In this study, we tried to address some of these questions by analyzing the behavior of water molecules in the active site of cytochrome P450 2D6 in a complex with *R*-3,4-methylenedioxy-N-ethylamphetamine (MDEA) using molecular dynamics (MD) simulations. The dynamic properties of these water molecules were determined for a few selected protein conformations, and the water molecules that seemed to influence the position of the ligand most were identified. It was our aim to identify the role of these waters and their influence on the reliability of predictions of the site of metabolism (SOM) by molecular docking simulations for MDEA, along with sets of similar and dissimilar compounds. MDEA was chosen as a representative because, in our previous work, we found that MD-generated structures with this substrate can accommodate many substrates while still leaving room for improvement of the SOM predictions.[21]

MATERIALS AND METHODS

All docking experiments were performed using protein structures that were extracted from a molecular dynamics (MD) simulations of cytochrome P450 2D6 in complex with the substrate (*R*)-3,4-methylenedioxy-N-ethylamphetamine (MDEA). Initial coordinates of the protein were obtained from the protein databank (www.pdb.org) entry code 2F9Q.[19] For details concerning the preparation of the complex structure and simulation settings, we refer to our earlier publication.[21] In short, the protein was solvated in 20292 SPC water molecules[22] and 7 Na^+ ions. Using the GROMOS simulation package,[23] a 10-ns simulation was performed at a constant temperature of 300 K and a pressure of 1 atm. The temperature and pressure were kept constant by weak coupling, using relaxation times of 0.1 and 0.5 ps, respectively.[24] The isothermal compressibility was set to 4.575×10^{-4} (kJ mol^{-1} $nm^{-3})^{-1}$. All bond lengths were constraint to their optimal values using the SHAKE algorithm[25] with a relative geometric accuracy of 10^{-4}. All interactions were calculated according to the GROMOS force field, parameter set 45A4.[26,27] Nonbonded interactions were calculated using a triple-range cutoff scheme. At every time step (2 fs), nonbonded interactions at distances shorter than 0.8 nm were calculated using a pair list that was constructed every five steps. Upon pair-list construction, interactions at distances up to 1.4 nm were also calculated and kept constant between pair-list updates. A reaction-field contribution[28] was added to the electrostatic interactions and forces to account for a homogeneous medium outside the cutoff sphere, using a relative dielectric permittivity of 61.[29] No additional constraints were added to the simulation.

Coordinates were stored every 10 ps. Starting at 2 ns, eight structures (A−H) were selected every 1 ns to comprise the test set for docking experiments. For each of these structures, the water behavior was analyzed by considering the 500 ps before and the 500 ps directly following the snapshot. All protein structures were superposed based on the backbone atoms prior to analysis. The distribution of water molecules in the active site was obtained by placing the protein on a regular grid with a spacing of 0.05 nm and monitoring the number of water molecules closest to the grid points. For every static protein conformation, the water molecules occupying clusters of grid points with a probability at least 30 times larger than the probability of finding a water molecule at a similar grid point in bulk water (0.4% for SPC at a density of 970 g/L) were selected initially. Even though the affinity of the water molecules was not calculated explicitly,[30,31] the high probability of occurrence at specific sites can be considered a thermodynamic measure. Hydrogen bonds were analyzed for these water molecules using a geometric criterion. A hydrogen bond was defined as having a minimum donor−hydrogen−acceptor angle of 135° and a maximum hydrogen−acceptor distance of 0.25 nm. The diffusion of water molecules relative to the protein structure, *D*, was calculated from the mean-square displacement using the Einstein equation

$$D = \lim_{t \to \infty} \frac{|\mathbf{r}_0 - \mathbf{r}(t)|^2}{2N_d t} \tag{1}$$

where $\mathbf{r}_0$ and $\mathbf{r}(t)$ are the water positions in a reference configuration and at time *t*, respectively. N_d is the number of dimensions that are being considered, which is 3 here.

From the protein structures A−H, the substrate, ions, and water molecules that were not selected were removed. Coordinates for hydrogen atoms that were implicitly treated in the simulations were added, and the resulting files were converted into mol2 file format using the standard Tripos atom and bond types for the amino acids and the heme group. Docking was performed using GOLD (Genetic Optimization for Ligand Docking),[32] version 3.3.1, in combination with the Chemscore scoring function[33] parametrized for heme-containing proteins.[34] The center point for docking was placed in the middle of the cavity between residues Phe120 and Phe483. The radius from this point was set to 1.8 nm to include the solvent channel in the accessible volume for the docking. At most, 100000 operations in the genetic algorithm were performed using a population of 100 genes. At least five independent docking simulations were performed to reach statistical significance. At least 50 poses were stored from every docking simulation, except if the best three docking poses had root-mean-square deviations smaller than 0.15 nm. To prevent an early convergence, a relative pressure of 1.1 was specified, and to account for diversity, the number of niches was set to 2. Ligand flexibility was specified as follows: The flipping of the free corners of ligand rings, the flipping of amide bonds, and the flipping of planar nitrogens were allowed. Intramolecular hydrogen bonds were also allowed. GOLD offers the possibility of automatically determining whether a specific water molecule should be bound or displaced by turning its interactions ON or OFF during the docking simulation.[35] In addition, the water molecules can be allowed to spin around their three principle axes. Because this increases the number of degrees of freedom, GOLD developers advise that it be done for at most three water molecules at a time (private communication). For structures in which more than three waters were selected based on local densities, preliminary docking experiments

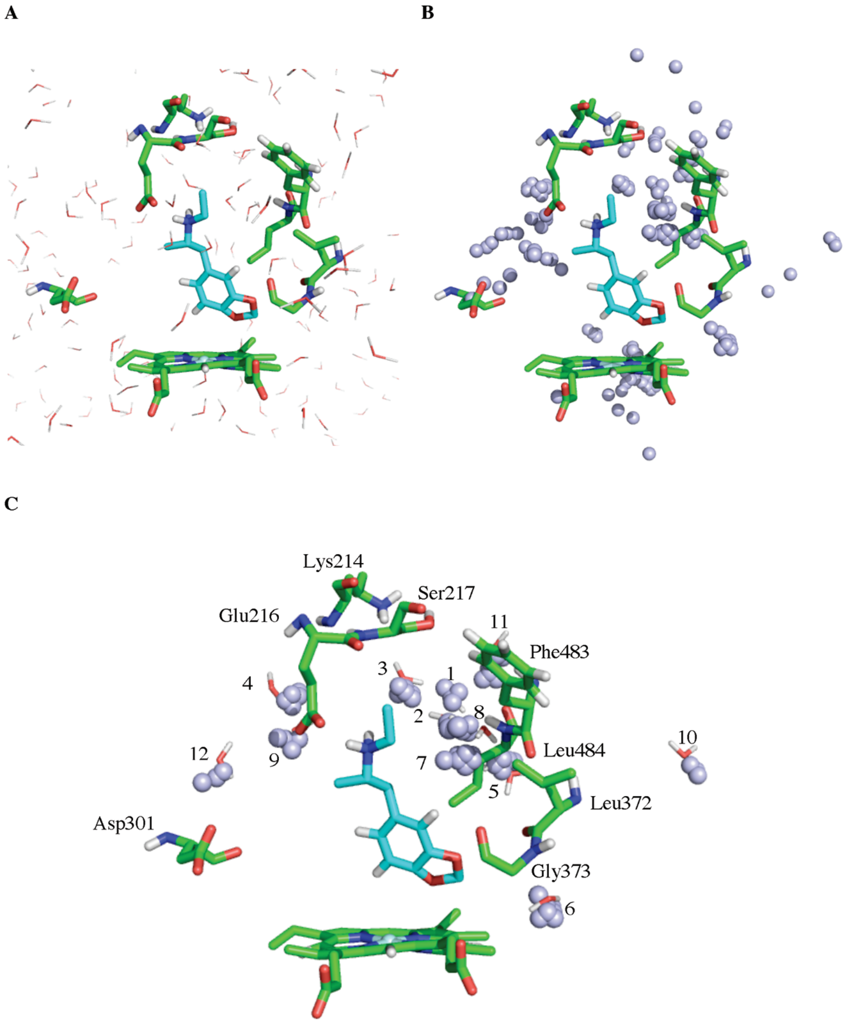


Figure 1. (A) Snapshot (denoted conformation A) of the molecular dynamics simulations of the complex of cytochrome P450 2D6 with MDEA. (B) Identification of hydration sites in the cavity of cytochrome P450 2D6 by determining the probability of occurrence over a grid. (C) Selection of the most static hydration sites in protein conformation A. Selected residues in the binding site are highlighted, and the hydration sites are numbered in accordance with Table 1.

were performed using MDEA, to determine the three water molecules for which the possibility of toggling between ON and OFF was most crucial for finding the correct binding pose. A water molecule was, therefore, considered to be critical if it increased the reliability of the docking prediction when determined to be ON and decreased the reliability prediction when determined to be OFF. In subsequent docking simulations, the effects of these water molecules on the binding poses for 11 substrates that are similar to MDEA and for 53 substrates that are dissimilar to MDEA were explored.[21] For each of these substrates, the site of metabolism (SOM) was determined experimentally; for complete references, see ref 20. As before, we considered a docking pose to be incorrect if the substrate site of metabolism was farther from the heme iron atom than 0.6 nm.[20,21] The final SOM prediction was subsequently based on the binding mode with the highest score among the five independent docking simulations.

RESULTS

Identification of Hydration Sites. Analysis of the trajectory of CYP2D6 with MDEA (Figure 1a) revealed hydration sites in the active site of CYP2D6 close to the substrate (Figure 1b). To discriminate between genuine hydration sites and false-positive hydration sites, the regions identified for protein conformation A were analyzed by visual inspection with VMD.[36] A genuine hydration site was defined as a position where a water molecule would remain for at least 80% of the time, which corresponds to an occurrence at least 30 times larger than that for a similar site in bulk water. Water molecules that were within 0.15 nm of a grid point belonging to a hydration site were initially selected for each protein conformation. In this way, dynamic changes in the shape and size of the hydration site were taken into consideration. This increase in the hydration site region also accounts for the dynamics of the water molecules, as even static water molecules still move considerably around a certain position (0.05−0.1 nm).

Twelve hydration sites were identified in protein conformation A (Figure 1c and Table 1). Between 10 and 17 hydration sites with dynamics similar to those described for protein conformation A could be identified in the other protein conformations.

Selection of Water Molecules for Docking. Despite our attempt to select only the most static water molecules from the molecular dynamics simulations of CYP2D6, the number of water molecules selected was still too high to reliably include in the docking experiments. To reduce the number

Table 1. Parameters That Reflect the Dynamics of the Water Molecules Identified in Protein Conformation A

hydration site	N_{HS}[a] (%)	V_{HS}[b] (10^{-4} nm^3)	r_{HS}[c] (nm)	HB_{pro}[d](%)	HB_{lig}[e](%)	HB_{HOH}[f] (%)	f_p[g]	HOH[h]	D[i] (10^{-3} nm^2 ps^{-1})
1	14	3.75	0.054	100	0	0	11.7	466	0.094
2	18	5.00	0.060	5.6	5.6	100	11.3	467	0.148
3	26	7.50	0.068	0	96.2	3.8	10.8	468	0.238
4	27	7.50	0.068	81.5	0	50	11.3	469	0.098
5	22	6.25	0.064	0	0	100	11.0	470[j]	0.224
6	39	10.00	0.075	100	0	0	12.2	471	0.064
7	32	8.75	0.072	0	0	90.6	11.4	472[j]	0.154
8	18	3.75	0.054	0	16.7	100	15.0	473	0.048
9	16	5.00	0.060	6.3	0	100	10.0	474	0.082
10	16	3.75	0.054	81.3	0	56.3	13.3	475	1.37
11	35	8.75	0.072	91.4	0	88.6	12.5	476[j]	0.156
12	15	3.75	0.054	0	0	100	12.5	477	0.338
bulk water	0.4	1.25	–	–	–	–	1.00	–	2.3[k]

[a] Hydration site occupancy (N_{HS}), defined as the sum of all occupancies of the grid points belonging to the hydration site. [b] Hydration site volume (V_{HS}), defined as the volume of one grid point (1.25×10^{-4} nm^3) multiplied by the number of grid points contributing to the hydration site. [c] Hydration site radius (r_{HS}), defined as the radius of a sphere with the same volume as the hydration site. [d] Occurrence of hydrogen bonds between the water molecules in the cluster and the protein. [e] Occurrence of hydrogen bonds between the water molecules in the cluster and the ligand. [f] Occurrence of water−water hydrogen bonds. [g] Spot density ratio, $f_p = \rho_{HS}/\rho_B$, where ρ_{HS} is the spot density for the hydration site, $\rho_{HS} = N_{HS}/V_{HS}$, and ρ_B is the bulk water density in molecules per nm^3 (32 nm^{-3}, calculated from the density of SPC water, 970 g/L). [h] Sequence numbers of the water molecule occupying the site in protein conformation A. [i] Relative diffusion constant (D) determined using eq 1. [j] Selected water molecules to be used for molecular docking simulations. [k] Value taken from ref 41.

of waters to be use for docking, preliminary docking studies were performed to identify the 3 water molecules that most influenced in a positive way the docking poses of MDEA. Therefore, water molecules that have a tendency to be ON when the SOM is further than 0.6 nm from the iron and at the same time have a tendency to be OFF when the SOM is within 0.6 nm from the iron, were removed from the complex, and in the end, only the water molecules responsible for improved docking poses remained. For some protein conformations, it was not clear which water molecules to remove from the initial selection, in these cases a visual inspection of the preliminary docking poses was carried out to identify the water molecule(s) responsible for a pose where the SOM was far from the iron. The improvement observed in the preliminary docking results for protein conformation A, could not be achieved by using the water molecules from the 3 hydration sites with the highest occupancy (data not shown).

The selected waters for protein conformation A were HOH470, HOH472, HOH476 belonging to cluster 5, 7, and 11 respectively (Figure 1c and Table 1).

Dynamic Properties of Selected Water Molecules. On the 1-ns time scale, the hydration sites identified are quite static positions, as can be seen for protein conformation A in Table 1. Between different protein conformations, the identified hydration sites could differ though. The spot density ratios are at least 10 times larger than for bulk water. The same trend can be observed for the diffusion constants of the water molecules within the hydration sites. A more thorough analysis of bulk water simulations revealed that no hydration site could be identified with a density greater than 640 molecules per cubic nanometer, and the hydration sites in the active site of CYP2D6 were found to be even more pronounced than that (minimum of 960 molecules/nm^3). However, only a few hydration sites revealed a high probability to form hydrogen bonds between the water molecules contributing to the hydration sites and the protein, and even fewer for the ligand. This indicates the possibility of a hydrogen-bond network between water molecules, or the existence of trapped water molecules in the cavity of CYP2D6 with a lubricant role. Because of the low resolution of the sampling times in MD, the average residence time can not be calculated with precision, but it was observed that the average residence time was above 10 ps only in some cases, being below 10 ps in the majority of the cases. For 70% of the hydration sites, it was found that more than one water molecule contributed to it over 1 ns. On average, three water molecules contributed to the same hydration site, with a maximum of five water molecules per site. Even though the density of water molecules at the hydration sites was quite high, the water molecules themselves appeared to be quite mobile.

Table 2. Comparison of Molecular Docking Results for MDEA, 11 MDEA-like Compounds, and 53 Non-MDEA-like Compounds in Protein Conformation A, When the Water Molecules Were Specified as OFF, Toggle, or ON

compound	water scenario	correctly docked[a] (%)
MDEA	HOH OFF	0
	HOH toggle	100
	HOH ON	100
MDEA-like	HOH OFF	90.9
	HOH toggle	90.9
	HOH ON	81.8
non-MDEA-like	HOH OFF	52.8
	HOH toggle	56.6
	HOH ON	54.7

[a] Percentage of correctly docked substrates based on the highest-ranked pose over the five independent simulations.

ON vs Toggle. Preliminary docking studies with MDEA and protein conformation A were done in order to determine the best treatment for water molecules during the simulations. The results are displayed in Table 2 and indicate that allowing the constant presence of water molecules during the docking runs (HOH ON) does not alter the docking results for MDEA, but it decreases the performance for MDEA-like and non-MDEA-like compounds when compared to the possibility of displacing these water molecules during the run (HOH toggle). Therefore, in this project, we allowed GOLD to determine during the docking run whether the water molecules would be displaced by the ligand.

Table 3. Molecular Docking Simulation Results for MDEA, MDEA-like and Non-MDEA-like Compounds Per Protein Conformation[a]

	MDEA		MDEA-like		non-MDEA-like	
protein conformation	HOH OFF	HOH toggle	HOH OFF	HOH toggle	HOH OFF	HOH toggle
A	0	100	90.9	90.9	52.8	56.6
B	100	100	81.8	90.9	62.3	54.7
C	0	0	36.4	72.7	45.3	50.9
D	0	0	45.5	63.6	41.5	49.1
E	0	100	90.9	63.6	62.3	58.5
F	0	0	45.5	90.9	45.3	41.5
G	100	0[b]	45.5	54.6	58.5	54.7
H	100	100	100	100	56.6	54.7

[a] Percentage of correctly docked substrates based on the highest-ranked pose over the five independent simulations in the (possible) presence and absence of water molecules. [b] No water was decided to be ON. The inconsistency arises because of a threshold problem; see text and Figure 2 for details.

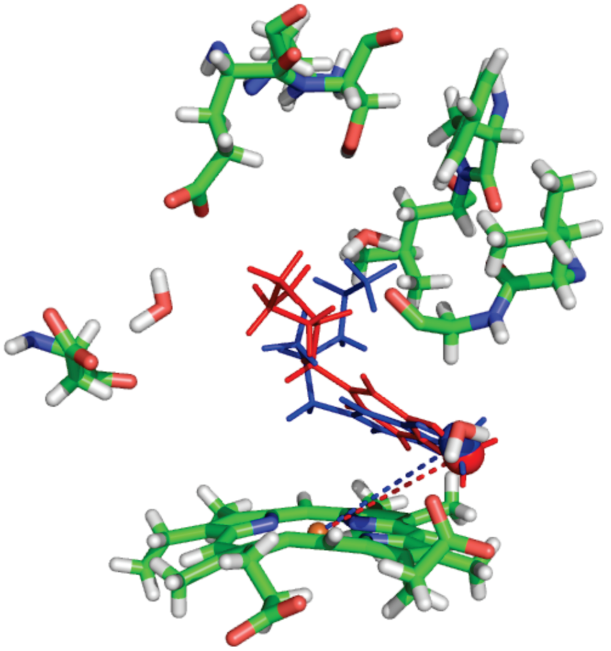


Figure 2. MDEA in frame G is a good example that the 0.6-nm rule might not be ideal for monitoring the correctness of the docking poses. By monitoring only the distance and not the poses, it seems that the presence of water worsens the SOM prediction (going from 2/5 correct to 0/5), whereas, in reality, the poses are very similar but the Fe−SOM distance increases just slightly over the threshold (0.61 nm with water vs 0.60 nm without water). Red and blue poses are achieved without water and with water, respectively. The sphere represents the SOM.

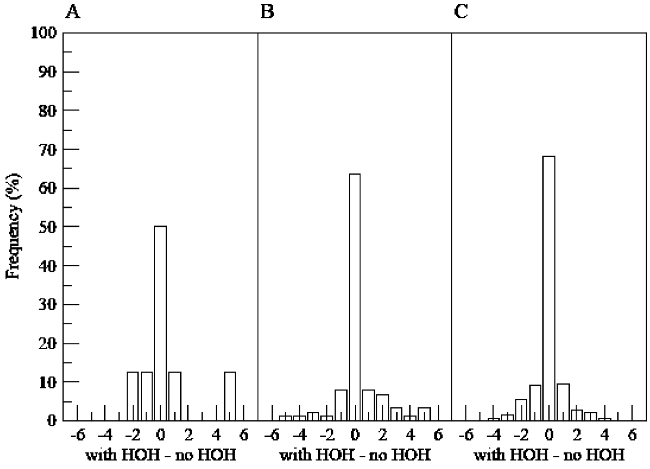


Figure 3. Distribution of reliability difference over all protein conformations between docking simulations both including possible water molecules and excluding water for (A) MDEA and (B) MDEA-like and (C) non-MDEA-like compounds. The scale of the *x* axis is as follows: A value of +5 indicates that, in five independent simulations, the substrate always docked in the first ranked pose with the SOM within 0.6 nm of the iron in the presence of water but always farther from the iron in the absence of water; −5 indicates that, in five independent simulations, the substrate always docked in the first ranked pose with the SOM far from the iron in the presence of water but within 0.6 nm of the iron in the absence of water; and 0 indicates no difference in the reliability of the docking prediction, regardless of whether it was docking with the SOM far or near the iron in the absence of water. On the *y* axis, the frequency is represented.

Docking Results for MDEA. Docking results for MDEA back into the active site of CYP2D6, in the (possible) presence and absence of water molecules (Table 3), showed that, overall, the presence of water molecules improved the reliability of the docking prediction for some protein conformations and did not decrease the reliability for others, except in protein conformation G. In this conformation, the results apparently worsen. However, this is due to a pitfall of the 0.6 nm rule, which only monitors the distance between the SOM and the iron and not the geometrical similarity between docking poses. In this protein conformation, no water molecule was switched ON by GOLD, and when analyzing the docking poses, we observed that this pose was just above the limit (0.6 vs 0.61 nm), but that the poses were quite similar (Figure 2).

A distribution of the effect of water molecules on the docking results of MDEA over all protein conformations with different degrees of confidence can be found in Figure 3A. A positive value on the *x* axis indicates an improvement of the reliability of SOM predictions when water is included. Larger values indicate a better reproducibility between individual docking simulations and, thus, a larger statistical significance. A negative *x* value indicates a worsening of the results when water is included, again with more statiscal confidence for larger values. A slightly higher frequency toward the positive side of the graph can be seen, meaning that the presence of water molecules might increase the accuracy of the docking results for MDEA or at least not alter them, given that the highest frequency is at 0.

Docking Results for MDEA-like and Non-MDEA-like Compounds. Analysis of the docking results of similar and dissimilar compounds in the active site of CYP2D6 (Table 3) shows that there is a greater effect for MDEA-like compounds and almost no effect for non-MDEA-like compounds. Further analysis reveals that there is a slight tendency for an improvement of the docking results for MDEA-like compounds when water molecules are included, because the distribution is slightly shifted to the positive side of the histogram (Figure 3B, C).

Effect of Water Molecules on the Reliability of SOM Predictions. As can be seen from the results above, the effect of water molecules in the docking results is ambiguous, being highly dependent on the protein conformation and substrate. In Figure 4, we identify four effects of water molecules: improvement, worsening, no effect, and no improvement. In Figure 4A, there is an improvement of the docking results for MDEA, as water molecules occupy a position that

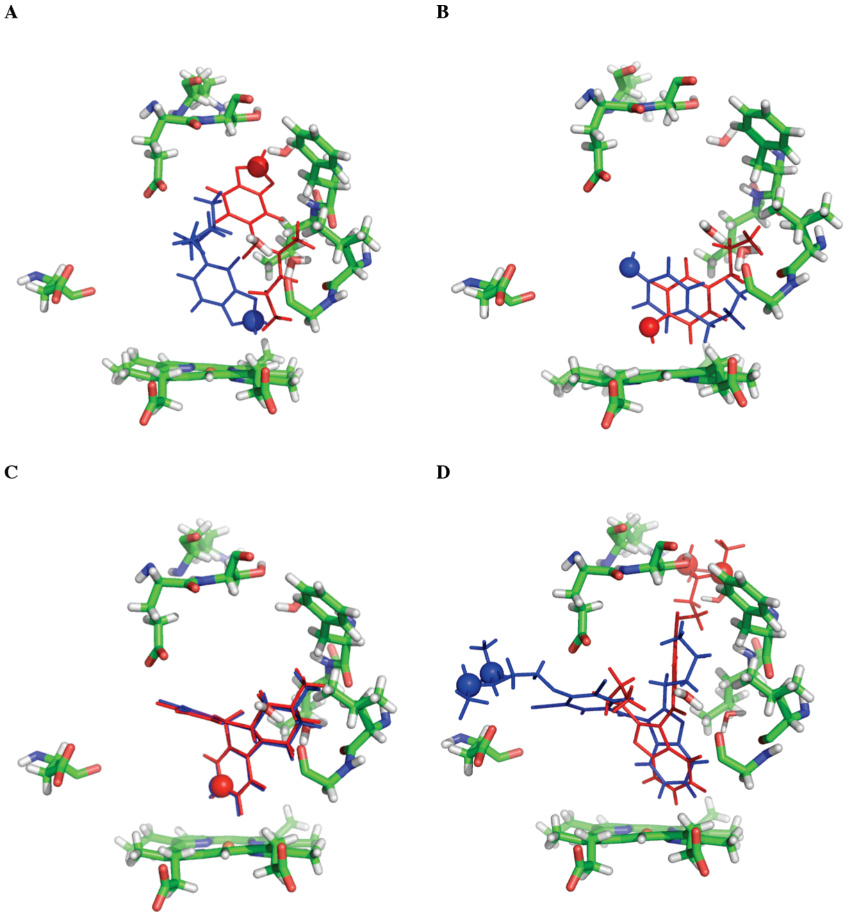


Figure 4. Effect of water molecules on the docking poses. Red poses and blue were achieved without water and with water, respectively. The sphere represents the experimentally determined SOM. (A) Improvement, MDEA in frame A: water molecules are preventing the substrate from binding in a region far above the heme group, leading to an improvement of 100%. (B) Worsening, TRP in frame A: water molecules are forming hydrogen bonds with the substrate, which leads to incorrect poses. (C) No effect, MRP in frame A: the presence of water molecules does not influence the poses. (D) No improvement, AMI in frame A: the presence of water molecules leads to different poses, but not to any with the SOM within 0.6 nm of the iron.

prevents the ligand to dock in a subpockect between Ser217 and Phe483. HOH476 is the main water molecule responsible for this improvement and is involved in a hydrogen bond with the protein 91% of the time. In Figure 4B, the substrate TRP no longer finds a catalytically active pose because water molecules are occupying a position that prevents the ligand from forming a hydrogen bond essential for GOLD to find a correctly docked pose. In Figure 4C, the docking of MRP is not affected by the (possible) presence of water molecules, as the poses are perfectly superimposed. Finally, in Figure 4D, the best-ranked pose of AMI is altered in the presence of water molecules, but the new poses still do not bring the SOM closer to the iron. Similar cases were observed for other protein conformations as well.

DISCUSSION

Water molecules play an important role in biological systems. Studies of the water molecules in the ligand-binding cavities of cytochromes P450 indicate that their high mobility facilitates the movement of the substrates and products into and out of the active site.[37] In this study, we identified hydration sites in the cavity of human CYP2D6 with an occupation probability at least 30 times larger than that in bulk water. However, water molecules contributing to it have an average residence time below 10 ps, being quite mobile, and having a low probability of forming hydrogen bonds with the protein or ligand. Rather, they are involved in hydrogen-bonded networks with other water molecules in the vicinity, occupying the empty space in the cavity. The dynamic nature of water molecules might add to the promiscuity of CYP metabolism and to the reported malleability of the active site.[38] The position of the hydration sites changes quite a lot from one protein conformation to the other, supporting the idea that the water molecules are quite mobile and considerably change the hydration shell over the span of 1 ns. However, regions with high probability of finding water molecules can be identified, and inclusion of these water molecules in a docking protocol leads to an improvement of the reliability of the SOM prediction for the compound that was used to optimize the selection of water molecules (MDEA) in several protein conformations and does not alter the results for the others. For compounds that are similar to MDEA, an effect on the pose prediction is also observed when water molecules are included, leading to a slight improvement in the results. There is virtually no effect when water molecules are included for compounds that are unlike MDEA. The water molecules seem to offer yet another way to allow CYPs to catalyze the metabolism of a wide range of structurally diverse compounds.[21,38]

It is also clear from the results that the effect of including water molecules in the docking protocol is rather dependent on the protein conformation and substrate. For instance, in protein conformations A and E, there is a clear improvement of the docking results for MDEA, but for the other protein conformations, there are no alterations in the results, except for protein conformation G. In protein conformation G, no water molecule was decided to be ON by GOLD, and therefore, the inconsistency of the results arises as a result of boundary effects of the 0.6-nm rule used to discriminate between correct and incorrect docking poses. This approach monitors only the SOM−Fe distance and does not account for the geometrical dissimilarity between poses. Because the metabolic site of the substrates is the only available experimental indication of the binding orientation, a root-mean-square geometrical measure cannot be used to distinguish "correct" and "incorrect' poses".

There are, however, a few pitfalls in this approach, that might influence the accuracy of the results and, consequently, our conclusions: (i) the GOLD parametrization to determine whether a water molecule should be displaced or bound, (ii) the fairness of comparing simulations in the presence and absence of water molecules using the same maximum number of genetic algorithm operations when the number of degrees of freedom differs, and (iii) comparison of docking results with different numbers of water molecules.

According to the literature, a water molecule is determined to be ON by GOLD if the intrinsic binding affinity of the water molecule outweighs the free energy associated with the loss of rigid-body entropy upon binding to the target.[35] The free energy associated with the loss of entropy upon binding to the target varies for different water binding sites, as tightly binding water molecules will loose more rigid-body entropy than loosely binding water molecules. However, this term in GOLD is treated as a constant for simplification. This constant is optimized for a set of water molecules in 58 crystal structures such as HIV-1, FXa, TK, and OppA, which include both highly structured (e.g., HIV-1) and rather promiscuous hydration sites (e.g., OppA). It might be too high to be used accurately with relatively mobile water molecules in the promiscuous CYP2D6, selected from molecular dynamics simulations. A more accurate determination of the water binding thermodynamics[31] is too computationally expensive to perform within a docking approach.

When docking in the possible presence of water molecules, the number of degrees of freedom increases (water ON/OFF, rotation around the principal axes), so to ensure sufficient sampling, the maximum number of genetic algorithm operations (maxops) needs to be increased. This does not hold when docking without water, because the number of degrees of freedom is smaller. However, no reliable estimate could be made of the value of maxops that would lead to a fair comparison because the degrees of freedom are very ligand-dependent and the exact relation between the number of configurations and the number of degrees of freedom is unknown.

Finally, it has been mentioned in the literature that an accurate comparison of the energy scores seems to rely on an equally homogeneous environment.[8] Therefore, the unequal numbers of water molecules among the docking poses might introduce some noise in the comparison, because it does not mimic the full hydration shell that is expected in aqueous solution. However, currently, no solution has been proposed for this problem.[8,39]

Taking these pitfalls into account, it still seems clear that water molecules do influence the docking orientations of the substrates in the CYP2D6 active site. The relevant water molecules and their presence, however, seem to be strongly dependent on the protein conformation and the substrate under consideration.

In contrast, de Graaf et al.[20] came to a different conclusion. The reliability of the SOM predictions for CYP2D6 based on docking simulations were improved by including explicit water molecules (specified as ON) for all 65 compounds (MDEA, MDEA-like, and non-MDEA-like). The positions of the water molecules were based on grid-based energy calculations, in which a region surrounding the heme iron atom was disallowed for water molecules. Because of the constant presence of water molecules during the docking simulations, the search space was restricted, and the substrate was pushed toward the heme, thereby increasing the chance of a successful SOM prediction. The improvement that was observed might not be due only to the presence of water molecules and their favorable interactions with the substrate or the protein, but might also be caused by space restrictions. This hypothesis is supported by our preliminary docking results for MDEA in protein conformation A of CYP2D6, where we see a decrease in the performance if water molecules cannot be displaced by the MDEA-like and non-MDEA-like sets of substrates. Overall, the role of water molecules in a promiscuous protein such as CYP2D6 does not appear to be straightforward. From the results for the MDEA and MDEA-like compounds, it is clear that water molecules should not be excluded from the calculations completely. On the other hand, for new substrates, different water molecules might be relevant. Therefore, we do not recommend to use a single set of water molecules for all substrates.

Analysis of the docking poses allowed us to classify the role of water molecules in four groups: improvement, worsening, no effect, and no improvement. For instance, water molecules can lead to improvement by preventing the substrate from being docked in subpockets or by forming hydrogen bonds that are involved in the stabilization of the correct pose. However, water molecules can also lead to worsening by forming hydrogen bonds that destabilize the correct pose or by forming hydrogen bonds with the protein that subsequently prevent a favorable hydrogen bond with the substrate that is essential for the correct pose to occur. However, water molecules can also not have any impact on the results, or they can simply not lead to an improvement of the docking results, meaning new poses can occur, but they do not bring the SOM closer to the heme. This effect of water molecules has been observed earlier. Some docking studies in proteases and kinases have also pointed out that, indeed, the inclusion of water molecules can lead to improvement, worsening, or no effect depending on the protein under study and its conformation.[8] We observed similar effects previously for a crystallographic water molecule in cytochrome P450 1A2.[40]

CONCLUSIONS

In this study, we demonstrated that hydration sites can be found in the cavity of CYP2D6, by monitoring the positions of water molecules in molecular dynamics simulations of CYP2D6 with MDEA. Even though the probability of finding a water molecule at these sites was at least 30 times larger than expected for bulk water, the water molecules themselves were rather mobile and dynamic, possibly adding to the well-known promiscuity of CYPs. Inclusion of selected water molecules in molecular docking simulations of CYP2D6 had an effect on the reliability of the site of metabolism prediction. However, their role is not unique, sometimes leading to a slight improvement or to no overall alterations depending on the protein conformation and the substrate. The larger effect was observed for the substrate that was used to select the water molecules. The same waters seemed to be transferable to similar substrates, still leading to a small overall improvement. However, for dissimilar compounds, no net effect could be observed. Our study sheds light on the relevant yet highly versatile role of water molecules in the CYP2D6 active site.

Abbreviations. AMI, amiodarine [(2-{4-[(2-butyl-1-benzofuran-3-yl)carbonyl]-2,6-diiodophenoxy}-ethyl)diethylamine]; CYP, cytochrome P450; CYP2D6, cytochrome P450 isoform 2D6; MDEA, *R*-3,4-methylenedioxy-N-ethylamphetamine; MD, molecular dynamics; MRP, *R*-mianserine (2-methyl-1,2,3,4,10,14b-hexahydrodibenzo[*c,f*]pyrazino[1,2-*a*]azepine); SOM, site of metabolism; TRP, *p*-tyramine (4-hydroxyphenethylamine).

REFERENCES AND NOTES

(1) Ladbury, J. E. Just add water! The effect of water on the specificity of protein−ligand binding sites And its potential application to drug design. *Chem. Biol.* **1996**, *3*, 973–980.

(2) Helms, V. Protein Dynamics Tightly Connected to the Dynamics of Surrounding and Internal Water Molecules. *ChemPhysChem* **2007**, *8*, 23–33.

(3) Cozzini, P.; Fornabaio, M.; Mozzarelli, A.; Spyrakis, F.; Kellogg, G. E.; Abraham, D. J. Water: How to Evaluate Its Contribution in Protein−Ligand Interactions. *Internat. Quant. Chem.* **2006**, *106* (3), 647–651.

(4) Poornima, C. S.; Dean, P. M. Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein−ligand interactions. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 500–512.

(5) Poornima, C. S.; Dean, P. M. Hydration in drug design. 2. Influence of local site surface shape on water binding. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 513–520.

(6) McConkey, B. J.; Sobolev, V.; Edelman, M. The performance of current methods in ligand−protein docking. *Curr. Sci.* **2002**, *83* (7), 845–856.

(7) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein−ligand docking: Current status and future challenges. *Proteins* **2006**, *65* (1), 15–26.

(8) Roberts, B. C.; Mancera, R. L. Ligand−Protein Docking with Water Molecules. *J. Chem. Inf. Model.* **2008**, *48* (2), 397–408.

(9) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *Br. J. Pharmacol.* **2008**, *153* (S1), S7–S26.

(10) Corbeil, R. C.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 3. Impact of input ligand conformation, protein flexibility, and water molecules on the accuracy of docking programs. *J. Chem. Inf. Model.* **2009**, *49*, 997–1009.

(11) Huang, N.; Shoichet, B. K. Exploiting Ordered Waters in Molecular Docking. *J. Med. Chem.* **2008**, *51* (16), 4862–4865.

(12) Birch, L.; Murray, C. W.; Hartshorn, M. J.; Tickle, I. J.; Verdonk, M. L. Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase. *J. Comput.-Aided Mol. Des.* **2002**, *16* (12), 855–869.

(13) Denisov, I. G.; Makris, T. M.; Sligar, S. G.; Schlichting, I. Structure and Chemistry of Cytochrome P450. *Chem. Rev.* **2005**, *105*, 2253–2277.

(14) Sono, M.; Roach, M. P.; Coulter, E. D.; Dawson, J. H. Heme-Containing Oxygenases. *Chem. Rev.* **1996**, *96* (7), 2841–2888.

(15) Williams, J. A.; Hyland, R.; Jones, B. C.; Smith, D. A.; Hurst, S.; Goosen, T. C.; Peterkin, V.; Koup, J. R.; Ball, S. E. Drug−drug interactions for UDP-glucuronosyltransferase substrates: A pharmacokinetic explanation for typically observed low exposure (AUCi/AUC) ratios. *Drug Metab. Dispos.* **2004**, *32*, 1201–1208.

(16) Ghoneim, M. M.; Korttila, K.; Chiang, C. K.; Jacobs, L.; Schoenwald, R. D.; Mewaldt, S. P.; Kayaba, K. O. Diazepam effects and kinetics in Caucasians and Orientals. *Clin. Pharmacol. Ther.* **1981**, *29*, 749–756.

(17) Stjernschantz, E.; Vermeulen, N. P. E.; Oostenbrink, C. Computational prediction of drug binding and rationalisation of selectivity towards cytochromes P450. *Exp. Opin. Drug Metab. Toxicol.* **2008**, *4* (5), 513–527.

(18) Cruciani, G.; Carosati, E.; DeBoeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. MetaSite: Understanding Metabolism in Human Cytochromes from the Perspective of the Chemist. *J. Med. Chem.* **2005**, *48* (22), 6970–6979.

(19) Rowland, P.; Blaney, F. E.; Smyth, M. G.; Jones, J. J.; Leydon, V. R.; Oxbrow, A. K.; Lewis, C. J.; Tennant, M. G.; Modi, S.; Eggleston, D. S.; Chenery, R. J.; Bridges, A. M. Crystal Structure of Human Cytochrome P450 2D6. *J. Biol. Chem.* **2006**, *281* (11), 7614–7622.

(20) de Graaf, C.; Oostenbrink, C.; Keizers, P. H. J.; van der Wijst, T.; Jongejan, A.; Vermeulen, N. P. E. Catalytic Site Prediction and Virtual Screening of Cytochrome P450 2D6 Substrates by Consideration of Water and Rescoring in Automated Docking. *J. Med. Chem.* **2006**, *49* (8), 2417–2430.

(21) Hritz, J.; de Ruiter, A.; Oostenbrink, C. Impact of Plasticity and Flexibility on Docking Results for Cytochrome P450 2D6: A Combined Approach of Molecular Dynamics and Ligand Docking. *J. Med. Chem.* **2008**, *51* (23), 7469–7477.

(22) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction models for water in relation to protein hydration. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, The Netherlands, 1981; pp 331−342.

(23) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem.* **2005**, *26*, 1719–1751.

(24) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690.

(25) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of *n*-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341.

(26) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; vdf Hochschulverlag AG an der ETH Zürich: Zürich, 1996.

(27) Schuler, L. D.; Daura, X.; van Gunsteren, W. F. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.* **2001**, *22* (11), 1205–1218.

(28) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. A generalized reaction field method for molecular dynamics simulations. *J. Chem. Phys.* **1995**, *102* (13), 5451–5459.

(29) Heinz, T. N.; van Gunsteren, W. F.; Hünenberger, P. H. Comparison of four methods to compute the dielectric permittivity of liquids from molecular dynamics simulations. *J. Chem. Phys.* **2001**, *115* (3), 1125–1136.

(30) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J. Am. Chem. Soc.* **2008**, *130*, 2817–2831.

(31) Michel, J.; Tirado-Rives, J.; Jorgenson, W. J. Prediction of the water content in protein binding sites. *J. Phys. Chem. B* **2009**, *113*, 13337–13346.

(32) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727–748.

(33) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11* (5), 425–445.

(34) Kirton, S. B.; Murray, C. W.; Verdonk, M. L.; Taylor, R. Prediction of binding modes for ligands in the cytochromes P450 and other heme-containing proteins. *Proteins* **2005**, *58*, 836–844.

(35) Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Modeling Water Molecules in Protein−Ligand Docking Using GOLD. *J. Med. Chem.* **2005**, *48* (20), 6504–6515.

(36) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14* (1), 33–38.

(37) Rydberg, P.; Rod, T. H.; Olsen, L.; Ryde, U. Dynamics of Water Molecules in the Active-Site Cavity of Human Cytochromes P450. *J. Phys. Chem. B* **2007**, *111* (19), 5445–5457.

(38) Guengerich, F. P. A malleable catalyst dominates the metabolism of drugs. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (37), 13565–13566.

(39) de Graaf, C.; Pospisil, P.; Pos, W.; Folkers, G.; Vermeulen, N. P. E. Binding Mode Prediction of Cytochrome P450 and Thymidine Kinase Protein−Ligand Complexes by Consideration of Water and Rescoring in Automated Docking. *J. Med. Chem.* **2005**, *48* (7), 2308–2318.

(40) Vasanthanathan, P.; Hritz, J.; Taboureau, O.; Olsen, L.; Steen Jorgensen, F.; Vermeulen, N. P. E.; Oostenbrink, C. Virtual Screening and Prediction of Site of Metabolism for Cytochrome P450 1A2 Ligands. *J. Chem. Inf. Model.* **2009**, *49* (1), 43–52.

(41) Harris, K.; Woolf, L. Pressure and temperature dependence of the self diffusion coefficient of water and oxygen-18 water. *J. Chem. Soc., Faraday Trans. 1* **1980**, *76*, 377–385.

CI900293E

Paper 3

Oostenbrink C.; de Ruiter A.; Hritz J.; Vermeulen N.P.E Malleability and versatility of Cytochrome P450 active sites studied by molecular simulations.
Curr. Drug Metab. **2012**, 13, 190-196

Malleability and Versatility of Cytochrome P450 Active Sites Studied by Molecular Simulations

Chris Oostenbrink[1,2,*], Anita de Ruiter[2], Jozef Hritz[3] and Nico Vermeulen[1]

[1]Leiden-Amsterdam Centre for Drug Research, Division of Molecular and Computational Toxicology, De Boelelaan 1083, 1081 HV Amsterdam, The Netherlands; [2]Institute of Molecular Modeling and Simulation, University of Natural Resources and Life Sciences, Muthgasse 18, 1190 Vienna, Austria; [3]Department of Structural Biology, School of Medicine, University of Pittsburgh, 3501 Fifth Avenue, Pittsburgh PA 15260, USA

Abstract: As the most important phase I drug metabolizing enzymes, the human Cytochromes P450 display an enormous versatility in the molecular structures of possible substrates. Individual isoforms may preferentially bind specific classes of molecules, but also within these classes, some isoforms show remarkable levels of promiscuity. In this work, we try to link this promiscuity to the versatility and malleability of the active site at the hand of examples from our own work. Mainly focusing on the flexibility of protein structures and the presence or absence of water molecules, we establish molecular reasons for observed promiscuity, determine the relevant factors to take into account when predicting binding poses and rationalize the role of individual interactions in the process of ligand binding. A high level of active site flexibility does not only allow for the binding of a large variety of substrates and inhibitors, but also appears to be important to facilitate ligand binding and unbinding.

Keywords: Site of metabolism prediction, protein flexibility, molecular docking, molecular dynamics simulations, replica exchange.

INTRODUCTION

The superfamily of Cytochrome P450 (CYP) enzymes is known to be highly diverse and versatile in its substrate specificity. This allows the enzymes to metabolize a wide range of xenobiotics. This holds true for the large diversity of the various isoforms for which general descriptions of typical substrates may be described, e.g. in the form of pharmacophore descriptions [1]. Individual isoforms, however, still show varying degrees of promiscuity towards potential substrates. The level of promiscuity has been attributed to properties of the active site, such as size, flexibility and malleability [2]. With the increased number of crystal structures of mammalian CYPs, a structural rationalization becomes feasible based on computer visualizations or molecular dynamics simulations to monitor the structure and dynamics of their catalytic sites [3-5].

CYP3A4 is probably the most promiscuous of the drug metabolizing enzymes and is known to have the largest active site, which in addition appears to be highly flexible [6]. CYP2C9 and CYP2D6 have smaller active site pockets and have a more tightly defined substrate specificity, including mostly negatively and positively charged substrates, respectively. However, the enzymes are not restricted to substrates bearing a full charge, and even within this restriction, a large variety of substrates have been reported. The 2A subfamily of CYPs is among the more specific drug metabolizing enzymes, mostly acting on flat aromatic compounds. Even within this class considerable variability exists [1].

Here, we report and summarize some of our own rationalizations for the versatility observed in various CYP isoforms, using computer simulations of the molecular interactions between the proteins and small molecule inhibitors or substrates. Molecular simulations offer insight at a resolution that is often not reachable by experimental means [7]. The molecular detail available in molecular dynamics simulations of proteins and protein-ligand complexes allows us to understand the subtle differences that enable the protein to discriminate between ligands. At a first level, the binding of ligands to the enzymes is explained in terms of the binding complementarity, according to the classical lock-and-key model. However, in enzymes as flexible as Cytochrome P450, this model needs to be extended to include induced fit effects. Here, we use the term *flexibility* to describe the overall degrees of freedom that are accessible to the protein. We use the term *malleability* to specifically refer to the possibility of accommodating different substrates or inhibitors as a result of the protein flexibility.

It is useful to describe the binding in the context of conformational selection. This latter paradigm considers a protein to consist not of a single static structure, but rather to be present in an ensemble of different conformations [8]. By considering multiple protein conformations to understand the binding of small molecule inhibitors or substrates, the complexity of the models increases. However, if predictions are to be performed at a quantitative rather than qualitative level, the protein-ligand interactions are to be described in terms of binding free energies, including both enthalpic and entropic contributions [9]. All contributions from direct protein-ligand interactions, loss of conformational freedom and desolvation of the ligand and the protein should be taken into account in calculations of the binding free energies.

Low-energy conformations of the protein are more often observed in the ensemble of protein structures, while higher-energy conformations will occur only rarely. A ligand that binds to the active site may only be able to do so if the protein is in a specific conformation. If this happens to be a low-energy conformation, then the protein-ligand interactions and desolvation effects may be sufficient to explain most of the binding free energy. However, for tightly binding ligands, the favorable ligand-protein enthalpy may be sufficient to stabilize the protein when it is in a high-energy conformational state. From the perspective of the protein, this is an unfavorable enthalpic contribution. As the ligand now locks the protein in this conformational state, its flexibility is also reduced, leading to an additional unfavorable entropic contribution.

In the following, we shortly describe the methods used to place substrates into the active site and to monitor the flexibility of the enzymes. In particular, we highlight Hamiltonian replica exchange molecular dynamics (H-REMD) simulations using distance restraining potentials. This approach allows for sufficient conformational sampling along the ligand exit pathway to calculate a potential of mean force (PMF). Following is a description of the effect of

[*]Address correspondence to this author at the Institute of Molecular Modeling and Simulation, University of Natural Resources and Life Sciences, Muthgasse 18, 1190 Vienna, Austria; E-mail: chris.oostenbrink@boku.ac.at

1875-5453/12 $58.00+.00

active site flexibility and the presence of water molecules on the accurate prediction of binding poses and an analysis of the versatility of protein-ligand interactions. Finally, the role of Phe483 in substrate orientation and ligand unbinding in CYP2D6 is analysed in more detail as an example of active site flexibility.

METHODS

Unless stated differently, all simulations were based on the following X-ray structures available in the protein databank[10]: CYP1A2 (PDB-entry 2HI4 [11]); CYP2C9 (PDB-entry 1R9O [12]); CYP2D6 (PDB-entry 2F9Q [13]) for details on the modeling of missing loops or side-chains and specific back-mutations, we refer to our original publications [14-16].

Automated docking experiments described in this work were performed with the program GOLD (Genetic Optimization for Ligand Docking) [17] version 3.2 or 3.3.1 in combination with the Chemscore scoring function [18]. For the docking experiments in the X-ray structures of CYP2D6 and CYP2C9, the optimized parameters for heme-containing proteins was used [19]. The radius for docking was set to 1.8 – 2.0 nm around a point in the center of the active site. Using a population of 100 genes, docking runs were performed with maximally 1000 genetic algorithm operations. Water molecules were either absent, included in the active site, or switched on and off by the program [20]. In the docking experiments the Heme group was modeled without a sixth ligand on the iron atom.

All molecular dynamics simulations described in this work were performed using a modified version of the GROMOS simulation package [21, 22]. Typically, the protein structure, with or without docked substrate was solvated in rectangular periodic boxes containing SPC water molecules [23]. The minimum solute-to-wall distance was at least 0.8 nm. After careful thermalization and equilibration the simulations were performed under a constant temperature of 298 K using the weak-coupling method with a relaxation time of 0.1 ps [24]. The pressure was kept constant at 1 atm using weak-coupling with a relaxation time of 0.5 ps and an estimated isothermal compressibility of $4.575{}_{\cdot}10^{-4}$ (kJ mol^{-1} nm^{-3})$^{-1}$ [24]. Bond lengths were restrained using the SHAKE algorithm, allowing for a time step of 2 fs [25]. Nonbonded interactions up to a distance of 0.8 nm were calculated every time step using a group-based cutoff that was constructed every fifth time step. At these times, the interactions up to 1.4 nm were also calculated and kept constant between pairlist updates. A reaction-field contribution [26] was added to the energies and forces to account for a homogeneous medium with relative dielectric constant of 61 outside the cutoff sphere [27].

In order to monitor the egress of substrate 7-methoxy-4-(aminomethyl)-coumarin (MAMC) out of the active site of CYP2D6 [28, 29], we used the following approach. Hamiltonian replica exchange (H-REMD) simulations [30] were performed using 26 parallel MD simulations in which distance restraints were applied between the Heme Fe-ion in Cytochrome P450 2D6 and the centre of geometry of 4 atoms in the ligand MAMC using reference distances of 0.720, 0.820, 0.910, 1.020, 1.101, 1.170, 1.220, 1.269, 1.340, 1.390, 1.430, 1.470, 1.539, 1.580, 1.620, 1.701, 1.800, 1.881, 1.950, 2.020, 2.110, 2.180, 2.270, 2.350, 2.440, and 2.530 nm, respectively. The force constant for the distance restraints was 1000 kJ mol^{-1} nm^{-2}. Initial structures of these replicas were obtained from a simulation in which the distance restraint was gradually increased from 0.4 to 2.4 nm. In this simulation, the MAMC molecule was observed to leave the active site through the solvent channel (according to the nomenclature of Wade [31]). After an initial equilibration of 25 ps, switching attempts were performed every 5 ps between neighbouring replica's using the metropolis criterion to accept or reject the switch. All replicas were simulated for 0.4 ns. This approach allows the individual MAMC molecules to move in and out of the active site, potentially also exiting through alternative egress channels. Simultaneously, the replica exchange approach ensures that at all conformations sampled at the individual restraining lengths correspond to a correct ensemble in a statistical mechanical sense, i.e. have the proper likelihood of occurrence given the Hamiltonian of the (restrained) system. This allows us to calculate a potential of mean force using the weighted histogram analysis method (WHAM) [32, 33].

PREDICTION OF BINDING POSES

In the context of Cytochromes P450, the binding orientation of substrates may directly influence their regiospecific metabolism. The reaction that takes place should not only be energetically feasible, the proper site of the substrate should also be in the vicinity of the reactive centre [9, 34-36]. A very simple geometrical test to see if binding orientations are in agreement or disagreement with catalytic activity is to measure the distance between the site of metabolism (SOM) in the substrate to the Heme Fe ion. As an empirical criterion, a pose is considered to agree with the observed metabolism if this distance is shorter than 6 Å, while larger distances indicate that the substrate orientation is incompatible with its experimentally determined metabolism [37, 38].

This 6 Å rule, however crude it is, allows us to check for known substrates how well automated docking simulations perform and to check the effect of slight modifications in the protein structure or the presence of water molecules in the active site.

CYP1A2

The active site of CYP1A2 is relatively narrow and contains an aromatic cluster of three Phenylalanine residues, allowing mostly flat aromatic compounds to bind to it. In the X-ray structure, the protein forms a complex with the inhibitor α-naphthoflavone, with a structured water molecule bridging a hydrogen bond to the active site [11]. 20 known substrates were docked to the active site using the GOLD docking program [17]. For three substrates (15%) no docking solutions were found in which the SOM was placed within 6 Å of the Heme iron using the chemscore scoring function and including the water molecule. For 13 substrates the top-ranked pose did correspond to a catalytically feasible pose, while for the remaining 4 substrates the second ranked pose needed to be included to understand the metabolism [39]. Interesting results were obtained when the presence of the water molecule was determined based on the scoring function, in the course of the docking experiment [20]. Overall, the number of 'correct' predictions slightly decreased to 10 first ranked poses, 6 substrates for which the second or third ranked pose needed to be considered and 4 substrates for which no catalytically active pose could be identified. As an example, the binding poses of imipramine were studied in more detail. When the water molecule was not taken into account, the first ranked pose corresponded to an orientation in agreement with the major metabolite of this molecule. In the presence of the water molecule, alternative orientations were found, in agreement with the formation of the minor metabolite. When the presence or absence of the water molecule was determined by the docking program, a mixture of both binding orientations was found, but the water was always determined to be off, also for the poses that were in agreement with the minor metabolite. This demonstrates on the one hand the effect of water molecules on the prediction of docking poses, while on the other side it shows that the presence or absence of such water molecules is not necessarily predicted in a consistent manner [39].

CYP2D6

Fig. (**1**) compares different experiments for CYP2D6 based on the percentage of poses that agree with observed metabolite formation for a set of 65 CYP2D6 substrates [38], using different protein models.

Before the availability of this X-ray structure, a homology model was described based on the rabbit CYP2C5 structure [40]. In

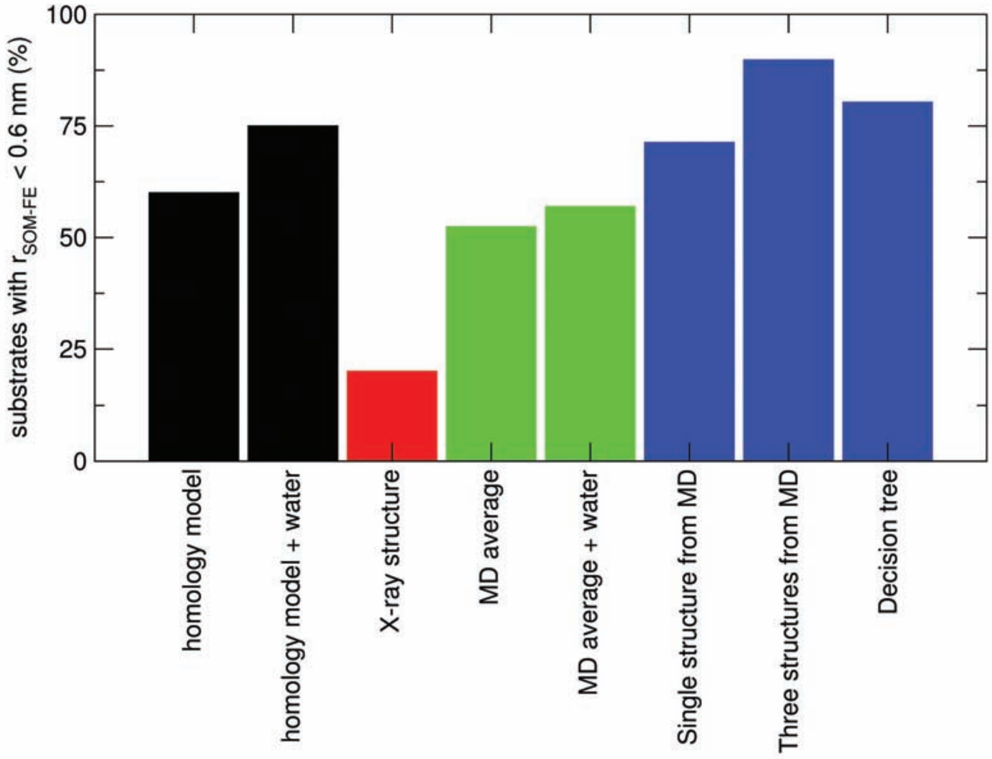


Fig. (1). Percentage of compounds for which the first-ranked pose from a docking experiment agrees with the experimentally determined site of metabolism. A set of 65 CYP2D6 substrates were docked into various protein structures obtained from homology modeling or X-ray crystallography possibly followed by molecular dynamics simulations, and/or the addition of water molecules. The sixth and seventh bars involve one or three protein structures, where in the latter case the optimal protein structure is selected (out of three) for every substrate. For the last bar a simple decision tree was used to select the appropriate protein structure from the same set of three.

this structure the active site was modeled in the presence of the substrate codeine. About 60% of the substrates could be docked to this homology model in poses that agree with the experimentally observed metabolite formation. This value could be further increased by adding static water molecules to the active site [38], probably by restricting the available space in the active site. The X-ray structure of an *apo* form of CYP2D6 was subsequently solved [13]. Using this structure, only 20% of the substrates could be docked in catalytically active poses. In a series of molecular dynamics simulations in which substrates of increasing volume were subsequently added, the active site was expanded to accommodate larger substrates as well. 2500 protein structures were extracted from 10 simulations containing five different substrates and two different initial protein conformations of Phe483 in the active site [14]. The average quality of the SOM predictions over all these protein conformations was 52 %.

In a subsequent experiment, water molecules were carefully selected based on their presence during MD simulations for 8 diverse protein structures. The docking program was used to guide the presence or absence of these water molecules. Averaged over the 8 protein structures, the overall percentage of binding poses in agreement with experiment marginally increased to 56 % [41]. It was observed that this percentage fluctuates strongly over the different protein snapshots, indicating that small thermal fluctuations as observed in the protein active site may have a profound effect on the predicted binding orientations of the substrates.

In fact, it was possible to identify a single, water-free, protein conformation for which 71% of the substrates were found to dock in a catalytically active pose. By reducing the complete ensemble of 2500 protein conformations to three structures, an overall accuracy of maximally 90% could theoretically be obtained, if one could determine which protein structure is most appropriate for a given substrate. Docking a substrate into three protein structures is usually feasible, but the selection of the proper protein structure cannot be made on the basis of the scoring function (unpublished results). A simple decision tree was established, which suggests which protein structure to use based on the molecular mass and the number of hydrophobic atoms in the substrate. Using the decision tree, the agreement of the predicted poses with experiment remains at an impressive 80% [14].

Using a very similar approach for CYP3A4, similar results were obtained [42]. When docking 16 substrates into 125 protein structures as obtained from MD simulations, on average 26% of the docking experiments yielded a pose in agreement with experiments. A single protein structure could be identified in which 62% of the compounds were correctly placed in the protein.

In summary, docking experiments on CYPs show that it is highly sensitive to the exact protein structure that is used in the presence or absence of water molecules. For the selection of an appropriate protein structure for CYP2D6 from a small set of three structures a simple decision tree was established [14], while the proper selection of which water molecules to take into account for any given substrate remains elusive for this versatile enzyme [41].

CYP2C9

In CYP2C9 a different approach was used to predict the binding modes for a series of thiourea-containing inhibitors. This series of compounds contains both an imidazole moiety and a thiourea group, both of which are known to interact favourably with the heme iron. As these compounds are inhibitors rather than substrates, there is no direct indication of a proper binding mode. All 12 compounds were docked into the active site of CYP2C9 and up to four different poses, coordinating the heme with the imidazole or the thiourea moiety, were identified. In an effort to calculate the binding affinities for these compounds, MD simulations were performed for all docking poses and an iterative scheme using the linear interaction energy methodology [43] was used. In short, this scheme gives weights to the individual simulations, based on the free energy of the ligand in the active site. By averaging the interaction energies according to the weights, an overall free energy of binding may be established. The weights may subsequently be interpreted as the probability that a given pose occurs [15]. From this we observed that 9 of the 12 compounds in this series prefer to bind in very similar orientations, while 3 bind in alternative poses. Moreover, for 2 of the compounds, multiple binding orientations contribute to the total free energy of binding, suggesting that they are equally likely to occur. This demonstrates that the active site of CYP2C9 is quite versatile in the binding modes that it accepts and that a dynamic equilibrium between multiple binding modes needs to be taken into account to explain the experimental data [15]. Similar observations were previously made based on MD simulations of CYP2D6 [44].

ANALYSIS OF INTERACTIONS

Once reasonable binding modes of substrates or inhibitors have been determined, it is interesting to investigate which amino acids are most important for the interactions between the ligands and the protein. From docking experiments this can be done by establishing protein-ligand interaction fingerprints. One striking feature from such analyses is the diversity of the interactions. In general, only one or two amino acids really interact with the majority of substrates, while many others only form interactions with less than 50% of the substrates. This has been observed for CYP1A2[39], CYP2D6 [40] and CYP3A4 [42] and may indicate another source of observed promiscuity. The free energy of binding to CYP1A2 was seen to correlate mostly with the non-polar protein-ligand interaction energy, while the electrostatic contribution to the free energy became negligible in linear interaction energy (LIE) models [16]. Even in a hydrophobic cavity as observed in CYP1A2, the protein is able to present a variety of suitable hydrogen bonding partners when this is required by the ligand, therefore the shape complementarity becomes more important to describe differences in binding than the actual electrostatic interactions.

Phe483 in CYP2D6

As was outlined in the introduction, the strength of computational approaches is that it allows us to investigate the flexibility and dynamics of protein-ligand complexes at a submolecular level. In the following, we will shift our focus to the behaviour of Phe483 in 2D6. Experimentally this residue was seen to play a role in the recognition of *R*- and *S*-propranolol. While the wildtype protein binds these substrates with very comparable affinity, the affinity of the F483A mutant was stereospecifically reduced for *R*-propranolol by a factor 20, or 7 .7 kJ mol^{-1} [45]. In molecular dynamics simulations of CYP2D6 with various substrates bound in the active site, the χ_1 torsional angle of the Phe483 sidechain was observed to change considerably (see Fig. (**2**)). Over 4 ns, this dihedral angle preferentially samples a value of 170° for *R*-3,4-methylenedioxy-N-ethylamphetamine (EDR), 7-methoxy-4-(aminomethyl)coumarin (MAMC) and chlorpromazine (CHZ), while it prefers a value of 70° for *R*-propranolol (PPD) and tamoxifen (TMF). Note that these are not static pictures. During several simulations, the dihedral angle occasionally takes on different values, indicating an intrinsic variability of this residue within the active site for which the equilibrium may be shifted depending on the substrate bound.

The χ_1 torsional angle of Phe483 was restrained to 70° or 170° in the 10 conformational ensembles that were described earlier to perform the docking experiments in [14]. From these docking experiments conformational preferences could be observed. Some substrate, like e.g. a group of substrates similar to MAMC, could be found to dock catalytically active poses significantly more often in protein structures for which χ_1 of Phe483 fluctuates around 170° [14], which agrees to the observations in Fig. (**2**). Fig. (**3**) shows the structure of MAMC in the CYP2D6 active site. It forms a salt bridge with Glu216, also indicated are the Heme group and Phe483.

In subsequent simulations, the unbinding of MAMC from the active site was simulated. In a preliminary simulation, a distance restraint was applied between the Heme Fe ion and the MAMC molecule. This distance was gradually increased, thereby forcing the MAMC to leave from the active site. As observed earlier, the

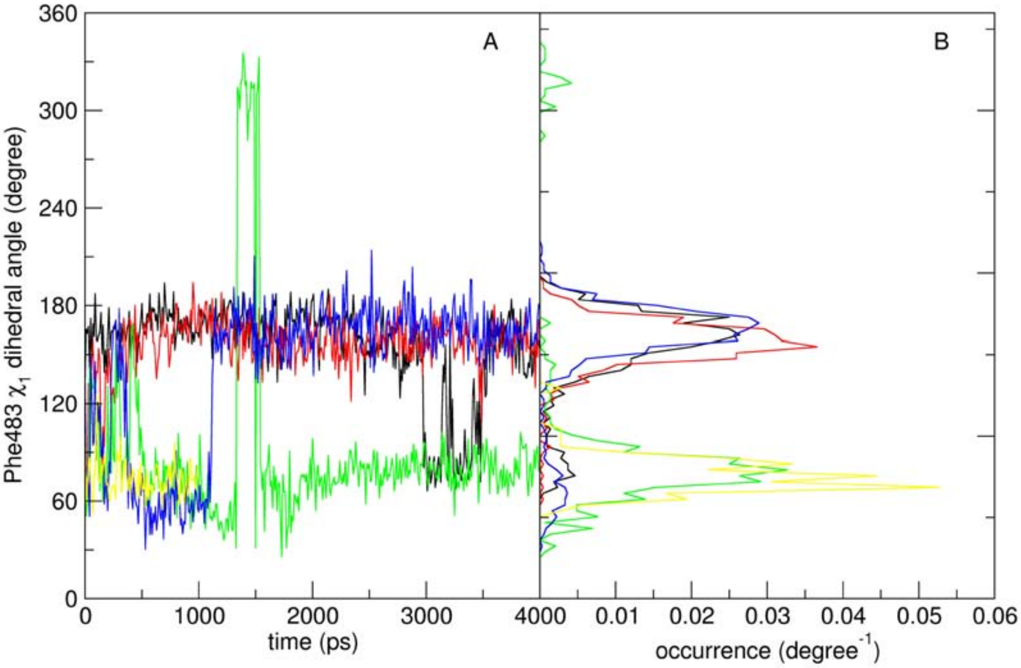


Fig. (2). Time series (**A**) and normalized distributions (**B**) of the x_1 torsional dihedral angle of Phe483 in CYP2D6, with different substrates bound to the active site: *R*-3,4-methylenedioxy-N-ethylamphetamine (EDR) in black; 7-methoxy-4-(aminomethyl)coumarin (MAMC) in red; *R*-propranolol (PPD) in green; chlorpromazine (CHZ) in blue and tamoxifen (TMF) in yellow.

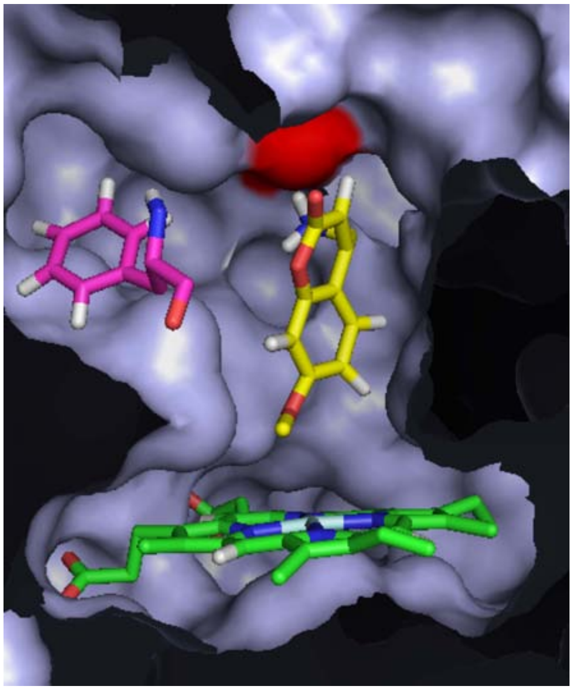


Fig. (3). Cytochrome P450 2D6 active site with 7-methoxy-4-(aminomethyl)coumarin (MAMC) bound. The surface of the protein is drawn in light grey, the heme group in green sticks, MAMC in yellow and Phe483 in magenta. The surface of the negatively charged Glu216 is colored red.

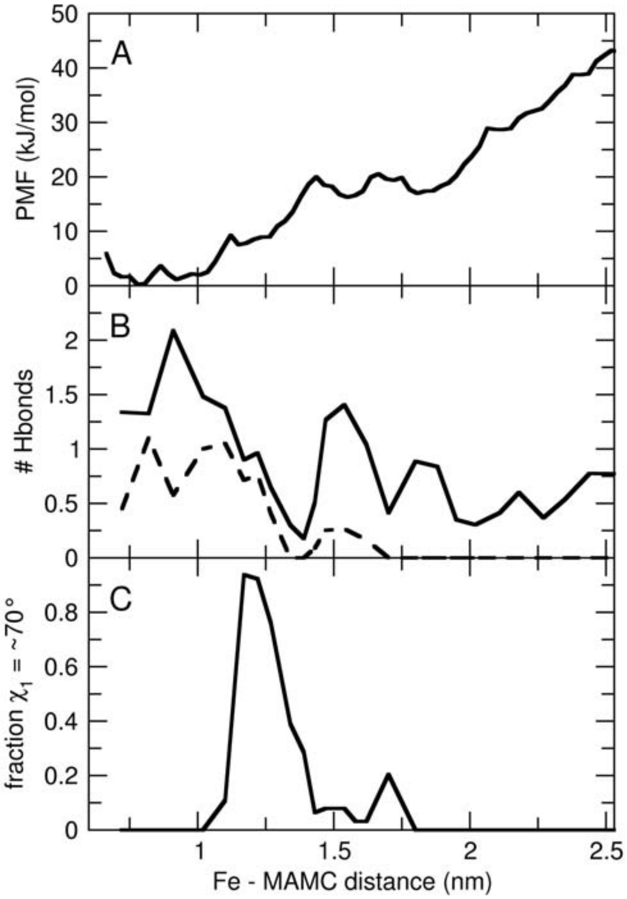


Fig. (4). A) Potential of mean force along the unbinding pathway of MAMC from CYP2D6 as a function of the Fe - MAMC distance; **B**) number of hydrogen bonds between MAMC and CYP2D6 (solid curve) and between MAMC and Glu216 (dashed curve); **C**) fraction of Phe483 conformations in which the χ_1 torsional angle takes a value of ~70° (between 0 and 120°).

molecule exited the protein through the so called solvent channel [31]. Using initial coordinates taken from this pathway, Hamiltonian Replica Exchange MD (H-REMD) simulations were performed with 26 replicas differing in the reference value of the distance restraint (see methods section). Every 5 ps a switch between the replicas is attempted, thereby allowing configurations at shorter distances to move to longer distances and *vice versa*. As no restrictions with respect to the path are given, the molecule theoretically can also sample different exit pathways in this process. However, this was not observed in the current simulations. From the H-REMD simulations, correct ensembles at every restraining distance were obtained which include rather diverse protein conformations. The potential of mean force (PMF) that is calculated from this set of 26 ensembles is shown in panel A of Fig. (**4**).

Even at a distance of 2.5 nm, the MAMC molecule is still interacting with the surface of the protein, forming on average 0.77 hydrogen bond, explaining that the PMF does not yet level off at this distance. Up to a distance of 1.1 nm, MAMC interacts through an average of 1.5 hydrogen bonds with the protein, of which 0.8 are formed with Glu216 (see panel B in Fig. (**4**)). At larger distances, the average number of hydrogen bonds rapidly decreases to 0.2 at 1.4 nm, corresponding to the steep increase in the PMF. Interestingly, the ensemble at 1.5 nm contains snapshots with two distinct orientations of the MAMC molecule. One in which the molecule has rotated completely with respect to the orientation in Fig. (**3**), such that a hydrogen bond to Glu216 may still be formed and another orientation in which the positively charged ammonium group is pointing towards bulk solvent. At even larger distances, alternative interactions may be formed, leading to a plateau in the PMF, but the average number of hydrogen bonds steadily decreases from 1 to 0.5 over the distance range 1.5 to 2.3 nm. Only at the very end it increases again to 0.8 when MAMC diffuses slightly over the surface of the protein. The current PMF cannot be used as such to estimate the free energy of binding. It does not level off to a plateau corresponding to the unbound substrate. Moreover, corrections would need to be estimated corresponding to the release of the distance restraint and to the transfer of the substrate from the accessible volume in these simulations to the standard state [33]. Still, it is reassuring to note that the free energy change corresponding to moving the ligand from the binding site to the distance at which all native interactions are broken, amounts to roughly 20 kJ/mol, which is in the same order of magnitude as an estimated experimentally determined affinity of 27 kJ/mol [46].

Panel C in Fig. (**4**) represents the fraction of the time that the Phe483 χ_1 dihedral angle was observed to be around 70° (using limits of 0 and 120°). It can be seen that as long as MAMC is bound to the active site, with a MAMC to Fe distance up to 1 nm, Phe483 prefers a conformation with χ_1 = ~170°, similar to the one in Fig. (**3**), and as observed in the simulation containing MAMC (Fig. (**2**)). Once the MAMC molecule starts to move out, the sidechain of Phe483 rotates into the active site (χ_1 = ~70°), thereby forming favourable hydrophobic interactions with the coumarin backbone of MAMC. Once the MAMC has left the active site at a distance of 1.5 nm, this interaction is no longer possible and Phe483 returns to its initial position (χ_1 = ~170°).

CONCLUSION

Above we have summarized and described some of our own molecular modeling work on Cytochrome P450 active sites and the interactions with substrates or inhibitors. As the most important phase I drug metabolizing enzymes, the human Cytochromes P450 display an enormous versatility in the molecular structures of substrates. Individual isoforms may preferentially bind specific classes of molecules, but also within these classes, some isoforms show remarkable levels of promiscuity. Even the relatively rigid and well-structured active site of CYP1A2 allows for the creation of a versatile network of hydrogen bonds, allowing the protein to accommodate quite diverse substrates. This observation is also made for more malleable proteins in which smaller or larger fluctuations in the protein structure were additionally seen to facilitate a large substrate promiscuity. In particular for CYP2D6 and CYP3A4 different substrates prefer binding to different conformations of the active site, in agreement with the conformational selection model. For CYP2C9 it was shown that it is well possible that multiple orientations of inhibitors contribute similarly to the overall binding affinity, which may complicate predictions of binding orientations even more.

A more detailed analysis of different conformational states of a single amino acid in CYP2D6 reveals that the active site malleability is not only crucial to understand how substrates may bind in catalytically active poses, but may also play a role during the binding and unbinding processes of inhibitors, substrates and products.

ACKNOWLEDGEMENTS

C.O. and A.d.R. gratefully acknowledge financial support of the Vienna Science and Technology Fund (WWTF) and the European Research Council (ERC) of the European Union (grant number 260408). J.H. acknowledges financial support by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme (Contract No. PIOF-GA-2009-235902).

REFERENCES

[1] Ortiz de Montellano, P. R., *Cytochrome P450: structure, mechanism and biochemistry*. Kluwer Academic/Plenum Publishers: New York, 2005.

[2] Guengerich, F., P,, A malleable catalyst dominates the metabolism of drugs. *Proc. Nat. Acad. Sci. USA,* **2006**, *103,* 13565 - 13566.

[3] Mestres, J., Structure conservation in Cytochromes P450. *Proteins,* **2005**, *58*, 596 - 609.

[4] Skopalik, J.; Anzenbacher, P.; Otyepka, M., Flexibility of human Cytochromes P450: Molecular dynamics reveals differences between CYPs 3A4, 2C9 and 2A6, which correlate with their substrate preferences. *J. Phys. Chem. B.*, **2008**, *112*, 8165 - 8173.

[5] Hendrychová, T.; Berka, K.; Navrátilova, V.; Anzenbacher, P.; Otyepka, M., Dynamics and hydration of the active sites of mammalian cytochromes P450 probed by molecular dynamics simulations. *Curr. Drug Metab.*, **2012**, 13(2): 177-189.

[6] Ekroos, M.; Sjögren, T., Structural basis for ligand promiscuity in cytochrome P450 3A4. *Proc. Nat. Acad. Sci. USA,* **2006**, *103,* 13782 - 13687.

[7] van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D.; Glättli, A.; Hünenberger, P. H.; Kastenholz, M. A.; Oostenbrink, B. C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F. A.; Yu, H. B., Biomolecular modeling: goals, problems, perspectives. *Angew. Chem. Intnl. Ed.* **2006**, *45*, 4064 - 4092.

[8] Carlson, H. A., Protein flexibility and drug design: how to hit a moving target. *Curr. Opin. Chem. Biol.*, **2002**, *6*, (447 - 452).

[9] Stjernschantz, E.; Vermeulen, N. P. E.; Oostenbrink, C., Computational prediction of drug binding and rationalisation of selectivity towards cytochromes P450. *Exp. Opin. Drug Metab. Tox.*, **2008**, *4*, 513 - 527.

[10] Rose, P. W.; Beran, B.; Bi, C. X.; Bluhm, W. F.; Dimitropoulus, D.; Goodsell, D. S.; Prlic, A.; Quesada, M.; Quinn, G. B.; Westbrook, J. D.; Young, J.; Yukick, B.; Zardecki, C.; Berman, H. M.; Bourne, P. E., The RCSB protein data bank: redesighned web site and web services. *Nucl. Acids. Res.,* **2011**, *39,* D392 - D401.

[11] Sansen, S.; Yano, J. K.; Reynald, R. L.; Schoch, G. A.; Griffin, K. J.; Stout, C. D.; Johnson, E. F., Adaptations for the oxidation of polycyclic aromatic hydrocarbons exhibited by the structure of human P450 1A2. *J. Biol. Chem.*, **2007**, *282*, 14348 - 14355.

[12] Wester, M. R.; Yano, J. K.; Schoch, G. A.; Yang, C.; Griffin, K. J.; Stout, C. D.; Johnson, E. F., The structure of human cytochrome P450 2C9 complexed with flurbiprofen at 2.0-Å resolution. *J. Biol. Chem.*, **2004**, *279,* 35630 - 35637.

[13] Rowland, P.; Blaney, F. E.; Smyth, M. G.; Jones, J. J.; Leydon, V. R.; Oxbrow, A. K.; Lewis, C. J.; Tennant, M. M.; Modi, S.; Eggleston, D. S.; Chenery, R. J.; Bridges, A. M., Crystal structure of human cytochrome P450 2D6. *J. Biol. Chem.*, **2005**, *281*, 7614 - 7622.

[14] Hritz, J.; de Ruiter, A.; Oostenbrink, C., Impact of plasticity and flexibility on docking results for Cytochrome P450 2D6: a combined approach of molecular dynamics and ligand docking. *J. Med. Chem.*, **2008**, *51*, 7469 - 7477.

[15] Stjernschantz, E.; Oostenbrink, C., Improved ligand-protein binding affinity predictions using multiple binding modes. *Biophys. J.*, **2010**, *98*, 2682 - 2691.

[16] Vasanthanathan, P.; Olsen, L.; Jorgensen, F. S.; Vermeulen, N. P. E.; Oostenbrink, C., Computational prediction of binding affinity for CYP1A2-ligand complexes using empirical free energy calculations. *Drug Metab. Disp.,* **2010**, *38*, 1347 - 1354.

[17] Jones, G.; Willet, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, **1997**, *267*, 727 - 748.

[18] Eldritch, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P., Empricial scoring functions 0.1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comp.-Aid. Mol. Des.*, **1997**, *11*, 425 - 445.

[19] Kirton, S. B.; Murray, C. W.; Verdonk, M. L.; Taylor, R., Prediction of binding modes for ligands in the Cytochromes P450 and other heme-containing proteins. *Proteins,* **2005**, *58*, 836 - 844.

[20] Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R., Modeling water molecules in protein-ligand docking using GOLD. *J. Med. Chem.*, **2005**, *48*, (20), 6504 - 6415.

[21] van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G., *Biomolecular simulation: The GROMOS96 manual and user guide*. Vdf Hochschulverlag AG an der ETH Zürich: Zürich, 1996.

[22] Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F., The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem.*, **2005**, *26*, 1719 - 1751.

[23] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J., Interaction models for water in relation to protein hydration. In *Intermolecular Forces*, Pullman, B., Ed. Reidel: Dordrecht, The Netherlands, 1981; pp 331-342.

[24] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R., Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.*, **1984**, *81*, (8), 3684-3690.

[25] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C., Numerical integration of cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.*, **1977**, *23*, (3), 327-341.

[26] Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F., A generalized reaction field method for molecular-dynamics simulations. *J. Chem. Phys.*, **1995**, *102*, (13), 5451-5459.

[27] Heinz, T. N.; van Gunsteren, W. F.; Hünenberger, P. H., Comparison of four methods to compute the dielectric permittivity of liquids from molecular dynamics simulations. *J. Chem. Phys.*, **2001**, *115*, (3), 1125-1136.

[28] Ludemann, S. K.; Lounnas, V.; Wade, R. C., How do substrates enter and products leave the buried active site of cytochrome P450cam? 1. Random expulsion molecular dynamics investigation of ligand access channels and mechanisms. *J. Mol. Biol.*, **2000**, *303*, 797 - 811.

[29] Ludemann, S. K.; Lounnas, V.; Wade, R. C., How do substrates enter and products exit the buried active site of cytochrome P450cam? 2. Steered molecular dynamics and adiabatic mapping of substrate pathways. *J. Mol. Biol.*, **2000**, *303*, 813 - 830.

[30] Sugita, Y.; Kitao, A.; Okamoto, Y., Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.*, **2000**, *113*, (15), 6042 - 6051.

[31] Cojocaru, V.; Winn, P. J.; Wade, R. C., The ins and outs of cytochrome P450s. *Biochim. Biophys. Acta*, **2007**, *1170*, 390 - 401.

[32] Souaille, M.; Roux, B., Extension to the weighted histogram analysis method: combining umbrella sampling ith free energy calculations. *Comput. Phys. Comm*., **2001**, *135*, 40 - 57.

[33] Doudou, S.; N.A., B.; Henchman, R. H., Standard free energy of binding from a one-dimensional potential of mean force. *J. Chem. Theor. Comp.,* **2009**, *5*, 909 - 918.

[34] Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R., MetaSite: Understanding metabolism in human cytochromes from the perspective of the chemist. *J. Med. Chem.,* **2005**, *48*, 6970 - 6979.

[35] De Graaf, C.; Vermeulen, N. P. E.; Feenstra, K. A., Cytochrome P450 in silico: An integrative modeling approach. *J. Med. Chem.*, **2005**, *48*, (8), 2725 - 2755.

[36] Afzelius, L.; Hasselgren Arnby, C.; Broo, A.; Carlsson, L.; Isaksson, C.; Jurva, U.; Kjellander, B.; Kolmodin, K.; Nilsson, K.; Reaubacher, F.; Weidolf, L., State-of-the-art tools for compuational site of metabolism predictions: comparative analysis, mechanistical insights, and future applications. *Drug Metab. Rev.,* **2007**, *39*, 61 - 86.

[37] De Graaf, C.; Pospisil, P.; Pos, W.; Folkers, G.; Vermeulen, N. P. E., Binding mode prediction of cytochrome P450 and thymidine kinase protein-ligand complexes by consideration of water and rescoring in automated docking. *J. Med. Chem.*, **2005**, *48*, 2308 - 2318.

[38] De Graaf, C.; Oostenbrink, C.; Keizers, P. H. J.; Van der Wijst, T.; Jongejan, A.; Vermeulen, N. P. E., Catalytic site prediction and virtual screening accuracy of cytochrome P450 2D6 substrates by consideration of water and rescoring in automated docking. *J. Med. Chem.,* **2006**, *49*, 2417 - 2430.

[39] Vasanthanathan, P.; Hritz, J.; Taboureau, O.; Olsen, L.; Jorgensen, F. S.; Vermeulen, N. P. E.; Oostenbrink, C., Virtual screening and prediction of site of metabolism for cytochrome P450 1A2 ligands. *J. Chem. Inf. Model.,* **2009**, *49*, 43 - 52.

[40] De Graaf, C.; Oostenbrink, C.; Keizers, P. H. J.; van Vugt-Lussenburg, B. M. A.; van Waterschoot, R.; Tschirret-Guth, R.; Commandeur, J. N. M.; Vermeulen, N. P. E., Molecular modeling-guided site-directed mutagenesis of cytochrome P450 2D6. *Curr. Drug Metab.,* **2007**, *8*, 59 - 77.

[41] Santos, R.; Hritz, J.; Oostenbrink, C., The role of water in molecular docking simulations of Cytochrome P450 2D6. *J. Chem. Inf. Model.* **2010**, *10*, 55 - 66.

[42] Teixeira, V. H.; Ribeiro, V.; Martel, P. J., Analysis of binding modes of ligands to multiple conformations of CYP3A4. *Biochim. Biophys. Acta,* **2010**, *1804*, 2036 - 2045.

[43] Åqvist, J.; Medina, C.; Samuelsson, J. E., New method for predicting binding affinity in computer-aided drug design. *Protein Eng.*, **1994**, *7*, (3), 385-391.

[44] Keizers, P. H. J.; De Graaf, C.; de Kanter, F. J. J.; Oostenbrink, C.; Feenstra, K. A.; Commandeur, J. N. M.; Vermeulen, N. P. E., Metabolic regio- and stereoselectivity of cytochrome P450 2D6 towards 3,4-methylenedioxy-N-alkylamphetamines: in silico predictions and experimental validation. *J. Med. Chem.,* **2005**, *48*, 6117 - 6127.

[45] De Graaf, C.; Oostenbrink, C.; Keizers, P. H. J.; van Vugt-Lussenburg, B. M. A.; Commandeur, J. N. M.; Vermeulen, N. P. E., Free energies of binding of *R*- and *S*-propranolol to wildtype and F483A mutant Cytochrome P450 2D6 from molecular dynamics simulations. *Eur. Biophys. J.,* **2007**, *36*, 589 - 599.

[46] Venhorst, J.; Onderwater, R. C. A.; Meerman, J. H. N.; Commandeur, J. N. M.; Vermeulen, N. P. E., Influence of N-substitution of 7-methoxy-4-(aminomethyl)-coumarin on Cytochrome P450 metabolism and selectivity. *Drug Metab. Disp.,* **2000**, *28*, 1524 - 1532.

Received: October 15, 2010 Revised: March 28, 2011 Accepted: April 29, 2011

Paper 4

Byeon I-J. , Ahn J., Mitra M., Byeon C-H., Hercík K., Hritz J., Charlton L., Levin J., Gronenborn A.M. NMR structure of human restriction factor APOBEC3A reveals substrate binding and enzyme specificity. *Nat. Commun.* **2013**, 4, 1890

nature COMMUNICATIONS

ARTICLE

Received 29 Jan 2013 | Accepted 12 Apr 2013 | Published 21 May 2013

DOI: 10.1038/ncomms2883

NMR structure of human restriction factor APOBEC3A reveals substrate binding and enzyme specificity

In-Ja L. Byeon[1,2], Jinwoo Ahn[1,2], Mithun Mitra[3], Chang-Hyeock Byeon[1,2], Kamil Hercík[3,†], Jozef Hritz[1,†], Lisa M. Charlton[1,2], Judith G. Levin[3,*] & Angela M. Gronenborn[1,2,*]

Human APOBEC3A is a single-stranded DNA cytidine deaminase that restricts viral pathogens and endogenous retrotransposons, and has a role in the innate immune response. Furthermore, its potential to act as a genomic DNA mutator has implications for a role in carcinogenesis. A deeper understanding of APOBEC3A's deaminase and nucleic acid-binding properties, which is central to its biological activities, has been limited by the lack of structural information. Here we report the nuclear magnetic resonance solution structure of APOBEC3A and show that the critical interface for interaction with single-stranded DNA substrates includes residues extending beyond the catalytic centre. Importantly, by monitoring deaminase activity in real time, we find that A3A displays similar catalytic activity on APOBEC3A-specific TT<u>C</u>A- or A3G-specific CC<u>C</u>A-containing substrates, involving key determinants immediately 5′ of the reactive C. Our results afford novel mechanistic insights into APOBEC3A-mediated deamination and provide the structural basis for further molecular studies.

[1] Department of Structural Biology, University of Pittsburgh School of Medicine, 3501 Fifth Avenue, Pittsburgh, PA 15261, USA. [2] Pittsburgh Center for HIV Protein Interactions, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA. [3] Section on Viral Gene Regulation, Program on Genomics of Differentiation, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892-2780, USA. † Present addresses: Fraunhofer IME – Division Molecular Biology, Biological Operating Systems: Infectious Diseases, Forckenbeckstraβe 6 , 52074 Aachen, Germany (K.H.); CEITEC, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic (J.H.). * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to A.M.G. (email: amg100@pitt.edu).

The human APOBEC3 (A3) proteins are a family of deoxycytidine deaminases that convert dC residues in single-stranded DNA (ssDNA) to dU, and act as DNA mutators. These proteins, which have an important role in the innate immune response, function as host restriction factors and display a broad range of activities against endogenous and exogenous retroelements[1–3]. There are seven members in the A3 family, each having one (A3A, A3C, A3H) or two (A3B, A3D, A3F, A3G) zinc (Zn)-binding domains with $\mathbf{H}X_1\mathbf{E}X_{23\text{-}24}\mathbf{C}X_{2\text{-}4}\mathbf{C}$ motifs, where X is any amino acid[4]. The histidine and cysteine residues coordinate Zn^{2+}, while glutamic acid is thought to function as a proton shuttle during the deaminase reaction[5].

The single-domain A3A protein, the subject of this study, has multiple activities. A3A degrades foreign DNA introduced into human cells[6,7] and blocks replication of exogenous viruses such as human papilloma virus[8], Rous sarcoma virus[9], parvoviruses[10,11] and human T-lymphotropic virus type 1 (ref. 12). In addition, it strongly inhibits retrotransposition of LINE-1, *Alu* and LTR retroelements[10,13–16], which cause insertional mutations. Unlike A3G (ref. 17), A3A is capable of deaminating 5-methylcytosine[18,19], an epigenetic marker in genomic DNA, as well as inducing cell cycle arrest[20] and somatic hypermutation of nuclear and mitochondrial DNAs in a dynamic interplay between A3 editing and DNA catabolism[20,21]. How all these activities are regulated, however, is not fully understood, although Tribbles 3, a human protein, was recently reported to protect nuclear DNA from A3A-mediated deamination[22].

A3A is highly expressed in cells of the myeloid lineage, such as monocytes and macrophages, and its expression is upregulated by treatment with interferon-alpha[10,23–26]. Interestingly, silencing of A3A in monocytes is associated with increased susceptibility to human immunodeficiency virus (HIV)-1 infection, suggesting that the presence of A3A may be protective against HIV-1 (ref. 23). Recent studies indicate that endogenous A3A in macrophages restricts HIV-1 replication by reducing synthesis of viral DNA during reverse transcription[27]. This result is consistent with an independent observation that HIV-1 transcripts in interferon-alpha-treated infected macrophages seem to be edited predominantly by A3A (ref. 28).

Although abundant information on the biological activities of the A3 proteins has been reported, only limited structural data are available. For example, the structure of the C-terminal domain (CTD) of A3G, which contains the catalytic centre for its deaminase activity[17,29,30], was solved by nuclear magnetic resonance (NMR)[31–33] and X-ray crystallography[34,35], but the structure of full-length A3G has been more elusive. While this manuscript was in preparation, the X-ray structure of another single-domain A3 protein, A3C, was reported[36].

Given human A3A's function as an inhibitor of retroviruses and retroelements with significant effects on cellular activities, the availability of an atomic, three-dimensional structure clearly is of significant importance. Here we report the NMR solution structure of human A3A, define the interface that is critical for its interaction with single-stranded oligonucleotide substrates and characterize its catalytic activity. Detailed analysis of A3A binding to nucleic acids and real-time monitoring of the deamination reaction by NMR allow us to propose a mechanism for substrate selection and specificity. These studies provide the structural basis for a deeper understanding of A3A's biological activities and broaden our knowledge of the molecular properties of the A3 proteins.

Results

Biochemical characterization of purified A3A. To ensure that the purified, recombinant wild-type A3A (199 aa), (containing a C-terminal His$_6$-tag (LEHHHHHH)), was enzymatically active, a uracil DNA glycosylase (UDG)-dependent gel-based deaminase assay was performed (Fig. 1a,b). The catalytic activity of this tagged protein was evaluated using a fluorescently labelled 40-nt

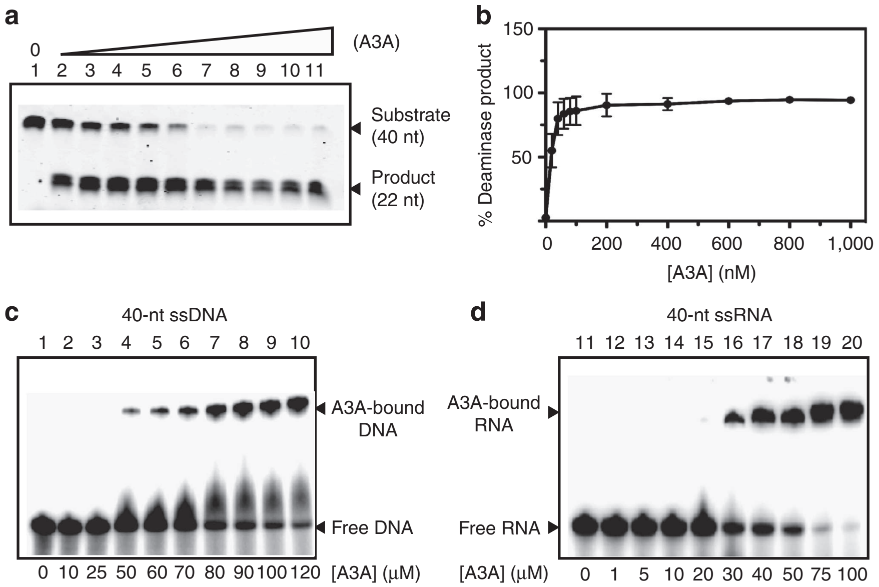


Figure 1 | Deaminase activity and ssDNA and ssRNA binding of A3A. (**a**) Deamination of a 40-nt ssDNA as a function of A3A concentration measured in a UDG-dependent assay. Reactions were performed as described in Methods with increasing concentrations of A3A. Lanes: 1, no A3A; 2, 20 nM; 3, 40 nM; 4, 60 nM; 5, 80 nM; 6, 100 nM; 7, 200 nM; 8, 400 nM; 9, 600 nM; 10, 800 nM; 11, 1000 nM, and relative amounts of substrate and product were assessed by gel electrophoresis. A representative gel (of three independent assays) was chosen for the figure. (**b**) Quantification of the relative amounts of deaminase product versus A3A concentration from gel analysis as shown in (**a**). The error bars represent the s.d. for three independent measurements. (**c**,**d**) Binding of A3A to 40-nt ssDNA (**c**) or 40-nt ssRNA (**d**) evaluated by electrophoretic mobility shift assay (EMSA; details in Methods). The positions of the A3A-bound as well as free DNA or RNA are indicated. A3A concentrations are listed under each lane. In each case, a representative gel (of five independent assays) was chosen for the figure.

ssDNA substrate, containing the TTCA deaminase recognition site, in the presence of increasing concentrations of protein, ranging from 20 to 1000 nM. Over 50% of the substrate was converted to the deaminated product with as little as 20 nM A3A, and complete conversion was seen with 200 nM A3A. These results demonstrate that the recombinant A3A protein is highly active as a cytidine deaminase.

Binding of A3A to ssDNA and ssRNA was evaluated in electrophoretic mobility shift assay experiments with ^{32}P-labelled 40-nt oligonucleotides and varying concentrations of A3A (Fig. 1c,d). Complexes are seen at A3A concentrations of $\geq 50\,\mu$M with ssDNA (Fig. 1c, lane 4) and $\geq 20\,\mu$M with ssRNA (Fig. 1d, lane 15). To achieve ~50% complexation, twofold more A3A (80 μM) was required for ssDNA (Fig. 1c, lane 7), compared with ssRNA (40 μM) (Fig. 1d, lane 17), suggesting that A3A has a somewhat higher binding affinity for ssRNA than that for ssDNA. Estimated K_d values of ~80 μM for the A3A–ssDNA complex agree well with values derived from NMR titration data (see below).

NMR sample preparation and behaviour in solution. Full-length A3A was monomeric and soluble up to ~0.2 mM at pH 6.5, 200 mM NaCl, and the ^{1}H–^{15}N heteronuclear single quantum coherence (HSQC) spectrum exhibited well dispersed, narrow resonances (Fig. 2). Nearly complete backbone and side-chain assignments were obtained using unlabelled, uniformly ^{15}N-, ^{13}C/^{15}N- and ^{2}H/^{13}C/^{15}N-labelled A3A with specific-protonations (see Methods) for three different conditions: (1) ~0.2 mM A3A at pH 6.5, 200 mM NaCl (Fig. 2); (2) ~0.2 mM A3A at pH 6.9; and (3) ~0.3 mM A3A at pH 8.1, all in 25 mM sodium phosphate buffer. As essentially identical nuclear Överhauser effect (NOE) patterns and chemical shifts were observed for the different conditions (only the amide resonances exhibited small, pH-dependent chemical shift differences), the structure was assumed to be unaffected by these different conditions. A mutant A3A, L63N/C64S/C171Q, was also used to help resolve ambiguous assignments (see Methods).

A3A structure. The structure of full-length A3A was calculated on the basis of 3279 NMR-derived experimental constraints. The final model satisfies all experimental constraints, displays excellent covalent geometry and the 30-conformer ensemble exhibits atomic r.m.s. deviations of 0.60 ± 0.05 and 1.18 ± 0.05 Å with respect to the mean coordinate positions for the backbone (N, C$_\alpha$, C′) and all heavy atoms, respectively (Table 1). A stereoview of the ensemble of conformers as well as a ribbon representation of the lowest energy structure from the ensemble are depicted in Fig. 3a,b, respectively.

Overall, the A3A structure consists of six helices surrounding a central β sheet of five strands (Fig. 3a,b), common to all APOBEC proteins whose structures are known; that is, A3G-CTD, A3C and A2 (Supplementary Table S1). The structured region of A3A is limited to residues 10–194 (Fig. 3a,b), as very intense, sharp amide resonances, exhibiting random-coil chemical shifts (Fig. 2), are present for the N-terminal nine (1–9) and final five (195–199) residues in the A3A sequence as well as the His$_6$-tag residues, indicating a substantial degree of flexibility. In addition to the termini, the loop connecting β2′ and α2 (loop 3, residues 57–70; Fig. 3a) is also highly plastic and undergoes motions on an

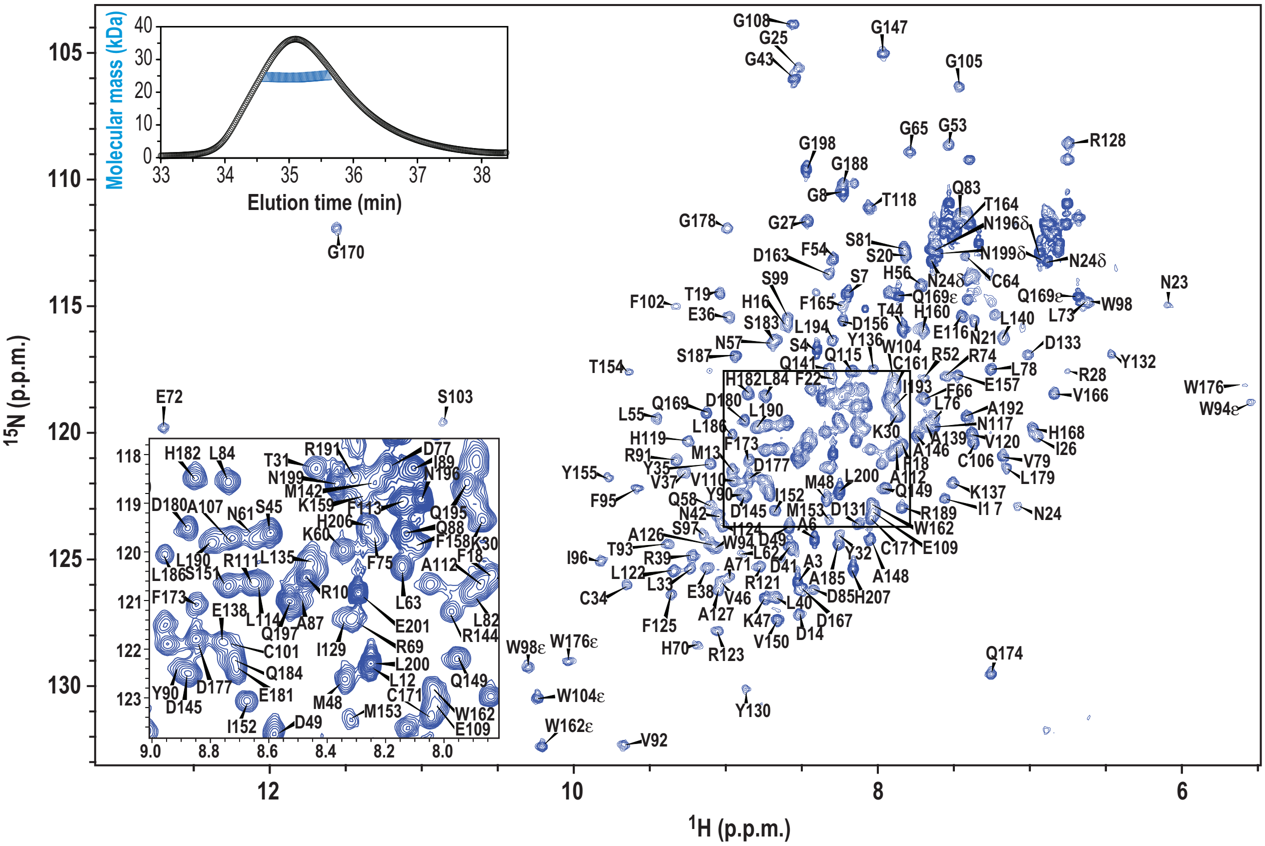


Figure 2 | A3A NMR assignments. 600 MHz ^{1}H-^{15}N HSQC NMR spectrum of 0.17 mM ^{13}C/^{15}N-labelled A3A in 25 mM sodium phosphate, 200 mM NaCl, pH 6.5, 25 °C. Assignments are indicated by residue name and number. An expansion of the boxed region is provided in the lower left corner. Size-exclusion chromatography/multi-angle light scattering data are shown in the inset in the upper left corner, with the elution profile shown with black circles and the estimated molecular masses across the peak with blue triangles.

Table 1 | Statistics for the final 30 conformer ensemble of A3A.

Number of NOE distance constraints	
Intra-residue (i–j = 0)	1114
Sequential (|i–j| = 1)	655
Medium range (2 ≤ |i–j| ≤ 4)	309
Long range (|i–j| ≥ 5)	749
Total	2827
Number of hydrogen bond constraints	152
Number of dihedral angle constraints	
ϕ	151
ψ	149
Total	300
Structural quality	
Violations*	
Distances constraints (Å)	0.016 ± 0.002
Dihedral angles constraints (°)	0.299 ± 0.037
Deviation from idealized covalent geometry	
Bond lengths (Å)	0.001 ± 0.000
Bond angles (°)	0.398 ± 0.006
Improper torsions (°)	0.222 ± 0.005
Average r.m.s.d. of atomic coordinates (Å)†	
Backbone heavy atoms	0.60 ± 0.05
All heavy atoms	1.18 ± 0.05
Ramachandran plot analysis (%)‡	
Most favourable region	73.7 ± 2.3
Additional allowed regions	22.5 ± 2.6
Generously allowed regions	2.9 ± 1.1
Disallowed regions	0.9 ± 0.6

*No individual member of the ensemble exhibited distance violations > 0.5 Å or dihedral angle violations > 5°.
†Average r.m.s.d. of atomic coordinates for residues (11–57 and 70–194) with respect to the mean structure. A3A regions (1–9, 58–69 and 195–199) were excluded from the statistics because they exhibit a flexible, random-coil conformation.
‡Statistics were calculated using PROCHECK for full-length A3A (residues 1–199).

intermediate (μ-ms) timescale, as many amide resonances exhibit severe line broadening, and are of low intensity (Q58, N61, L62) or entirely missing (A59, Y67 and G68) in the ^{1}H–^{15}N HSQC spectrum (Fig. 2). This is suggestive of multiple conformations undergoing chemical exchange.

Comparison with other APOBEC protein structures. We carried out a detailed comparison of the present A3A solution structure with other available structures of APOBEC proteins (Supplementary Table S1). Among these structures, the present A3A NMR structure is most similar to the X-ray structure of the A3G-CTD quintuple mutant (residues 191–384, PDB: 3IR2)[35] (Fig. 3c,d), exhibiting an average pairwise backbone atomic r.m.s. difference of 1.86 Å. The next closest structures are the X-ray structures of wild-type A3C, another single-domain human A3 protein (1–190, PDB: 3VOW)[36], and wild-type A3G-CTD (residues 197–380, PDB: 3E1U/3IQS)[34], both exhibiting average backbone r.m.s. differences of ~2.3–2.4 Å. The NMR solution structures of monomeric A2 proteins (murine A2, PDB: 2RPZ and a CS-HM Rosetta model for human A2 (ref. 37)) are also similar to A3A (atomic r.m.s. differences of ~3.4 Å). In the previously solved A3G-CTD structures, some local differences were reported[31–34]. Our current A3A structure clearly possesses an interrupted β2 strand (β2-bulge-β2′; Fig. 3b and Supplementary Fig. S1), as well as an N-terminal α-helix (α1, residues 15–21).

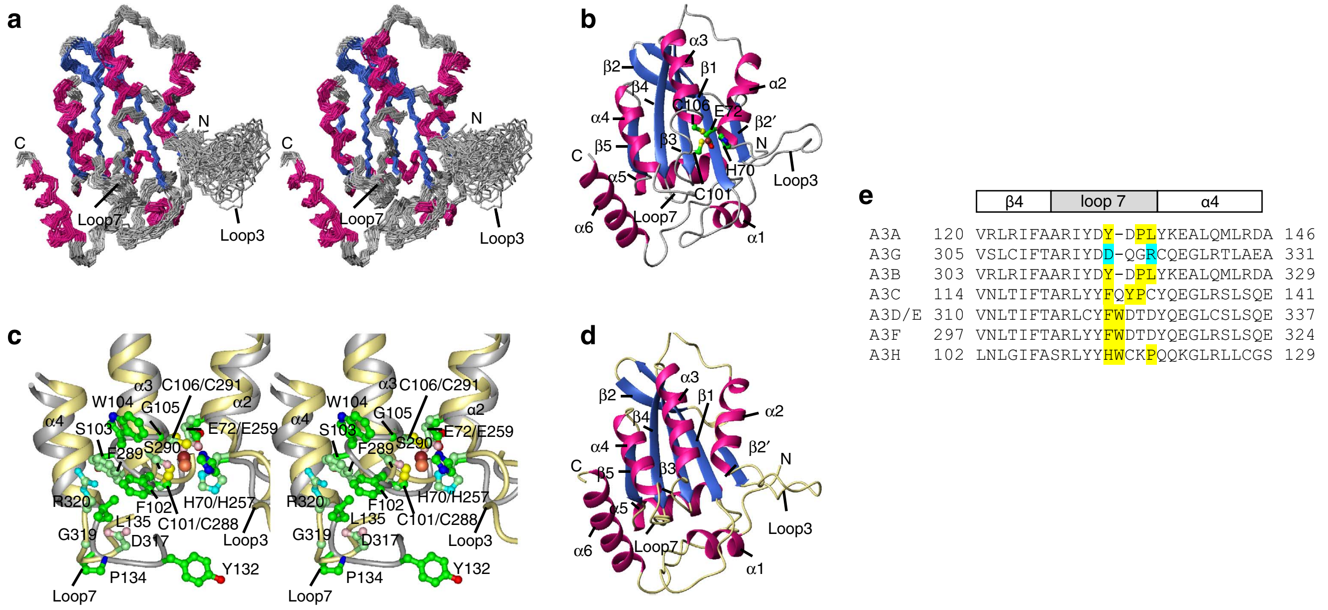
A more detailed comparison was carried out between the current A3A and the most similar A3G-CTD structures (A3G191-384-2K3A; Fig. 3c,d and Supplementary Table S1). We focused on the regions around the active site where the A3A and A3G sequences differ. A two amino-acid (W104 and G105) insertion, located between two Zn-coordinating cysteines (C101 and C106) in the active site of A3A, is the most prominent sequence change. Despite this insertion, the positions of the catalytic site residues H70 (H257 in A3G), E72 (E259), C101 (C288) and C106 (C291) are very similar in both structures (Fig. 3c). The insertion, however, distorts the N-terminal end of helix α3, with W104 bulging out, causing positioning of the backbone and/or side chain atoms of S103 and G105 in A3A equivalent to those of F289 and S290 in A3G. This structural adjustment places the hydrophobic F102 and W104 side chains at the protein surface, next to the active site. In contrast, A3G possesses only one hydrophobic residue (F289) in this region. Loop 7 is located near the active site (Fig. 3), and its C-terminal half in A3G is known to have an important role in substrate selection (TC or CC)[38]. In A3A, loop 7 (127–135) adds additional hydrophobic residues (Y132, P134 and L135), whereas in A3G two of the equivalent residues (D317 and R320) are polar, and the proline is replaced by glycine, resulting in conformational differences between the A3A and A3G loop 7 (Fig. 3c). Note that among the human A3 proteins, the polar characteristic of this region is unique for A3G (Fig. 3e).

A3A interacts with dCTP and dUTP. Although A3A is a well-known cytidine deaminase, the determinants of substrate specificity (that is, preferential recognition of TC (refs 7,10,39)) have not been characterized in detail. As an initial approach, we investigated binding of A3A to dCTP, several ssDNA oligonucleotide substrates and/or their dU-containing products by NMR.

dUTP (Fig. 4a) and dCTP (Fig. 4e) binding was monitored by ^{1}H–^{15}N HSQC spectroscopy, and affected regions were mapped onto the A3A structure (insets). Perturbations of amide resonances of the active site amino acids H70, E72, C101 and C106 as well as other surrounding ones (K30, L55, H56, N57, Q58, K60, N61, L63, C64, H70, A71, W98, S99 and Y132) were very similar, indicating that both the substrate and the product bind to the active site. Several of the amino acids whose resonances were affected reside in loop 3, with N61, L63 and C64 located distal to the active site, indicating that a remote conformational change was induced by dCTP/dUTP binding or that a particular preferred conformation was selected from the flexible ensemble in the substrate-free A3A state. Titration curves of selected perturbed resonances upon dCTP (Fig. 4e) and dUTP (Fig. 4a) binding were used to extract K_d values of 536 ± 72 μM and 578 ± 115 μM, respectively (Table 2). Note that as up to 30% of the dCTP was converted to dUTP during the titration, the value for dCTP should be regarded as an approximation.

A3A binds ssDNA substrates via an extended surface. The interaction of 5′-ATTT$\underline{\text{C}}$ATTT-3′ (~0.1–3 mM) with A3A (~0.15 mM) under our NMR conditions resulted in conversion to 5′-ATTT$\underline{\text{U}}$ATTT-3′ within < 1 min, rendering it impossible to measure a true K_d value for 5′-ATTT$\underline{\text{C}}$ATTT-3′. However, given the similarity between dUTP and dCTP binding to A3A (Fig. 4a versus Fig. 4e), it seemed reasonable to assume that 5′-ATT T$\underline{\text{U}}$ATTT-3′ and 5′-ATTT$\underline{\text{C}}$ATTT-3′ would also have similar binding properties. The K_d value for the ATTT$\underline{\text{U}}$ATTT product was 58 ± 8 μM (Fig. 4b, Table 2), and its interaction surface on the A3A structure (Fig. 4b, inset) covers a larger surface, extending beyond the catalytic site.

This extended interface involves an area to the left side of the active site, opposite of loop 3 (where residues are coloured in hot pink ($\Delta\delta$ > 0.050 p.p.m.) or cyan ($\Delta\delta$ between 0.028 and 0.050 p.p.m.) in the ribbon diagram). Interestingly, the two-



```
A3A   120 VRLRIFAARIYDY-DPLYKEALQMLRDA 146
A3G   305 VSLCIFTARIYDD-QGRCQEGLRTLAEA 331
A3B   303 VRLRIFAARIYDY-DPLYKEALQMLRDA 329
A3C   114 VNLTIFTARLYYFQYPCYQEGLRSLSQE 141
A3D/E 310 VNLTIFTARLCYFWDTDYQEGLCSLSQE 337
A3F   297 VNLTIFTARLYYFWDTDYQEGLRSLSQE 324
A3H   102 LNLGIFASRLYYHWCKPQQKGLRLLCGS 129
```

Figure 3 | A3A NMR solution structure. (**a**) Stereoview of the final 30 conformer ensemble (N, Cα, C′). Regions of helical and beta sheet structures are coloured hot pink and royal blue, respectively, and the remainder of the structure in grey. (**b**) Ribbon representation of the lowest energy structure of the ensemble, using the same colour scheme as in (**a**). Secondary structure elements are labelled and the active site residues (H70, E72, C101 and C106) and the Zn^{2+} ion are shown in ball-and-stick representation with carbon, nitrogen, oxygen, sulphur and Zn atoms in green, blue, red, yellow and brown, respectively. (**c**) Stereoview of the superimposition of the active site regions of the current A3A NMR and the A3G-CTD (PDB: 3IR2) X-ray structures. The backbone traces of A3A and A3G-CTD are coloured grey and khaki, respectively. Side chains are shown in ball-and-stick representation with carbon, nitrogen, oxygen, sulphur and Zn atoms of A3A and A3G in green, blue, red, yellow and brown, and pale green, cyan, pink, yellow and orange, respectively. A3A residues are labelled in bold. (**d**) Ribbon representation of the A3G-CTD (A3G191-384-2K3A, PDB: 3IR2) X-ray structure[35]. (**e**) Amino-acid sequences of the loop 7 region in different A3 proteins. Large, hydrophobic residues are highlighted in yellow and the polar residues D317 and R320 in A3G are highlighted in cyan.

residue insertion (W104/G105), unique to A3A, is part of the interface: the W104 backbone amide ($\Delta\delta = 0.122$ p.p.m.) and the W104 ε1 side chain resonance ($\Delta\delta = 0.079$ p.p.m.), as well as the G105 ($\Delta\delta = 0.066$ p.p.m.) and S103 ($\Delta\delta = 0.072$ p.p.m.) amide resonances exhibit the largest perturbations. In addition, loop 7 and helix α4 are also part of the extended interface, as effects on the amide resonances of residues D133 ($\Delta\delta = 0.112$ p.p.m.), L135 ($\Delta\delta = 0.102$ p.p.m.) and E138 ($\Delta\delta = 0.079$ p.p.m.) were observed. Additional binding studies of A3A using a 15-nt oligonucleotide, 5′-ATTATTTUATTTATT-3′ yielded essentially the same binding site (Fig. 4c, inset) and affinity ($K_d = 57 \pm 11$ μM; Fig. 4c, Table 2), as found with the 9-mer.

Although smaller perturbations were observed for some additional resonances, these changes were non-saturable and most likely reflect non-specific binding (for example, the I17 resonance marked by a dashed line in Fig. 4c). The non-specific binding site of the 15-mer oligonucleotide mapped onto the A3A structure involves residues 12–19 (α1), 36–37, 40–42 (β1), 44–49 (β2), 53–56 (β2′) and 179–186 (α6) (Supplementary Fig. S2). To validate the binding data, the interaction of an A3A catalytic site mutant, E72Q, with a 15-nt substrate 5-′(ATTATTTCAT TTATT)-3′ was evaluated. The mutant exhibited the same affinity ($K_d = 55 \pm 11$ μM, data not shown) as wild-type A3A with the 5′-ATTATTTUATTTATT-3′ product, confirming that binding of substrate and product occurs with essentially identical affinities.

Further titration experiments using smaller oligonucleotides (5′-ATTT(C/U)A-3′, Fig. 4d; 5′-TCATTT-3′, Fig. 4f) delineated a binding site similar to the one observed with the nona-nucleotide. This implies that three nucleotides, (TCA), constitute the essential moiety for binding to A3A. The affinities of the hexanucleotides, however, were approximately threefold weaker ($K_d \sim 190$ μM) than the one measured for 5′-ATTT(C/U)ATTT-3′ ($K_d = 58 \pm 8$ μM; Table 2), suggesting a stabilizing effect of the flanking nucleotides in the complex.

Interaction of A3A with a 9-nt substrate, 5′-AAACCCAAA-3′, containing the A3G deaminase-specific recognition site[2], maps to the same surface (Fig. 4g, inset) as 5′-ATTT(C/U)ATTT-3′ (Fig. 4b, inset). This suggests that both dT and dC at the −1 and −2 positions, that is, pyrimidine bases, can interact with the extended interface next to the catalytic site of A3A. However, slightly weaker binding ($K_d = 91 \pm 15$ μM, Fig. 4g, Table 2) was observed for 5′-AAACCCAAA-3′ compared with 5′-ATTTCATTT-3′. Replacement of the central CCC in 5′-AAACCCAAA-3′ by CCA (5′-AAACCAAAA-3′, Fig. 4h) or CAA (5′-AAACAAAAA-3′, Table 2) resulted in identical ($K_d = 94 \pm 11$ μM) or slightly weaker ($K_d = 161 \pm 19$ μM) affinities, respectively.

Structural model for A3A–ssDNA complexes. Model structures of the A3A-oligonucleotide complexes were created by flexible docking[40] and selecting binding poses compatible with chemical shift perturbations (see Methods for details). Initial docking results for 5′-ATTTCATTT-3′ and 5′-AAACCCAAA-3′ indicated that only the central region of the oligonucleotides interacts specifically with A3A, leaving the two ends free. Therefore, we only generated final structural models of A3A with the pentanucleotides 5′-TTCAT-3′ (Fig. 5a) and 5′-CCCAA-3′ (Fig. 5b). In the A3A/TTCAT complex model, the central reactive C occupies the deep pocket delineated by the active site and surrounding residues T31, N57, Q58, H70, E72, W98, C101, C106, D131 and the Zn^{2+} ion. The thymidine (T_{-1}) immediately

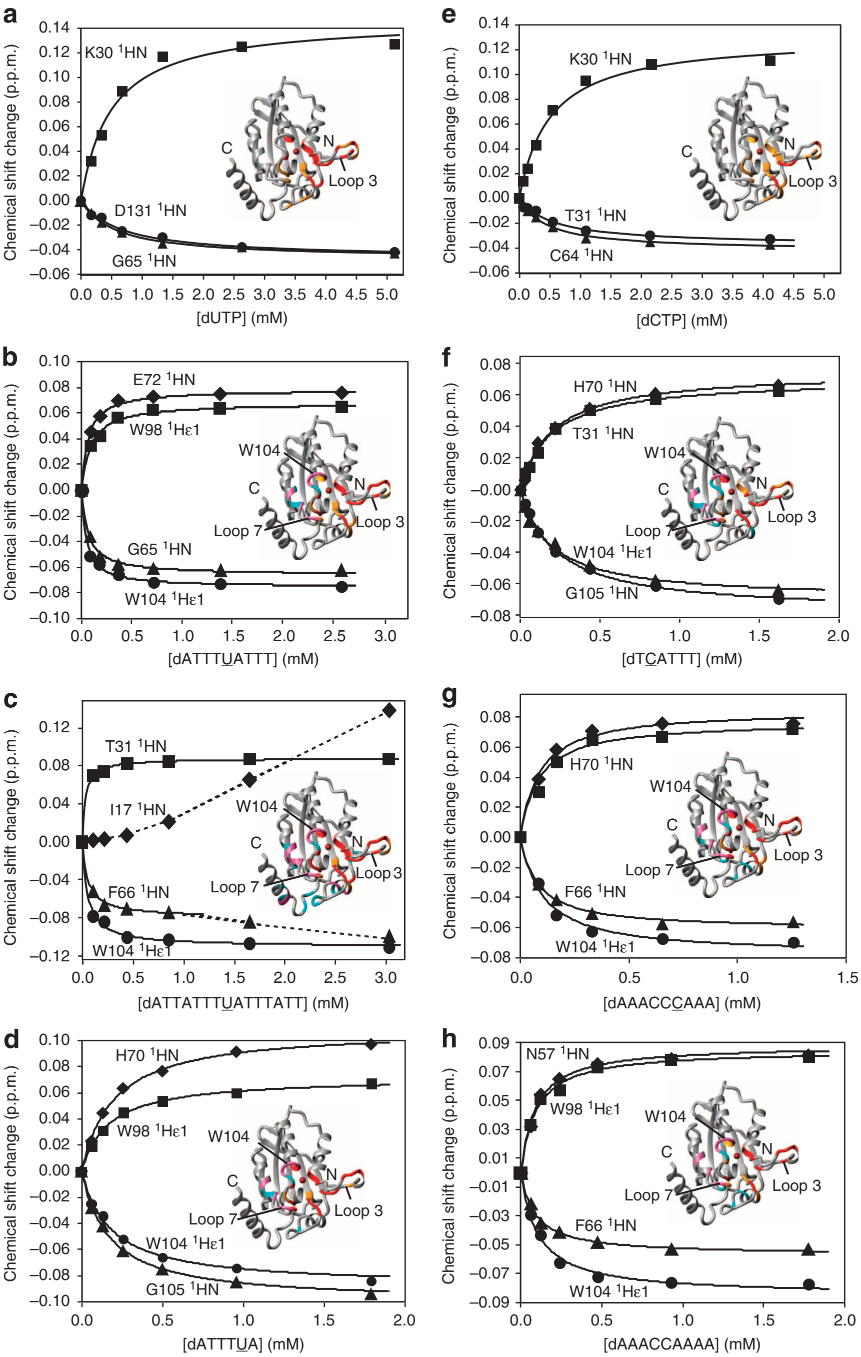


Figure 4 | Binding of A3A to mononucleotides and ssDNAs. Titration curves for representative HN resonances and binding site mapping (inset) for binding of dUTP (**a**), ATTTUATTT (**b**), ATTATTTUATTTATT (**c**), ATTTUA (**d**), dCTP (**e**), TCATTT (**f**), AAACCCAAA (**g**) and AAACCAAAA (**h**). A3A residues whose resonances exhibit large ^{1}H,^{15}N-combined chemical shift changes upon nucleotide addition are coloured red (>0.050 p.p.m.) and orange (0.028–0.050 p.p.m.). Those only affected by ssDNA binding, but not by dCTP/dUTP, are shown in dark pink (>0.050 p.p.m.) and cyan (0.028–0.050 p.p.m.). All ^{1}H-^{15}N HSQC spectra were recorded at 25°C using 25 mM sodium phosphate buffer, pH 6.9. ^{1}H,^{15}N-combined chemical shift changes were calculated using $\sqrt{\Delta\delta_{HN}^2 + (\Delta\delta_N/6)^2}$, with $\Delta\delta_{HN}$ and $\Delta\delta_N$, the ^{1}HN and ^{15}N chemical shift differences observed for A3A before and after adding ligands (~90% saturation).

preceding the C interacts with a surface formed by residues D133, P134, L135 (loop 7) and F102 (loop 5), whereas T_{-2} contacts residues F102, S103, W104, and G105 (loop 5), L135 (loop 7) and E138 (α4). The adenosine (A_{+1}) at the 3′ side of the C interacts with residues D131, Y132 and D133 (loop 7). The 5-methyl groups of T_{-1} and T_{-2} are in close contact with several hydrophobic residues, for example, F102, W104 and L135, suggesting that this hydrophobic interaction may contribute to tighter binding of TTCA- compared with CCCA-containing ssDNA. The model for the A3A/CCCAA complex (Fig. 5b) is very similar to that of A3A/TTCAT (Fig. 5a), including the localization of the individual nucleotides.

Table 2 | Dissociation and catalytic constants for A3A interaction with ss deoxymono- and deoxyoligonucleotides.*

Nucleotide sequences	K_d (μM)	K_M (μM)	k_{cat} (min^{-1})	k_{cat}/K_M (M^{-1}s^{-1})
dCTP	536 ± 72	600†	0.033 ± 0.003†	0.91
dUTP	578 ± 115			
5′-dATTT**C**ATTT-3′	58 ± 8‡	66 ± 7	71 ± 6	1.8 × 10^4
5′-dATTATTT**C**ATTTATT-3′	57 ± 11‡	62 ± 5	66 ± 5	1.8 × 10^4
5′-dATTT**C**A-3′	184 ± 20‡	210 ± 17	65 ± 6	5.2 × 10^3
5′-dT**C**ATTT-3′	194 ± 18‡	219 ± 29	55 ± 5	4.2 × 10^3
5′-dAAA**CCC**AAA-3′	91 ± 15‡	100†	20 ± 3§	3.3 × 10^3
5′-dAAA**CC**AAAA-3′	94 ± 11‡	100†	13 ± 4§	2.2 × 10^3
5′-dAAA**C**AAAAA-3′	161 ± 19‡	192 ± 32	5.5 ± 1.1	4.7 × 10^2

*Dissociation constants (K_d) and catalytic constants (Michaelis–Menten) derived from NMR data at 25 °C (pH 6.9), as described in Methods.

†As accurate measurements of K_M are precluded when more than one dC is present, the K_M values of these substrates are assumed to be similar to the K_d values and are set to 100 (K_d values rounded up to the nearest hundred). This assumption is based on data obtained with substrates containing a single dC.

‡The K_d values for these dC-containing oligonucleotides reflect values for their dU-containing products, as these substrates are rapidly deaminated under the NMR experimental conditions (A3A ~ 0.1 mM).

§k_{cat} values are derived from the initial rate in a sample with 1000 μM substrate and 0.197 μM A3A.

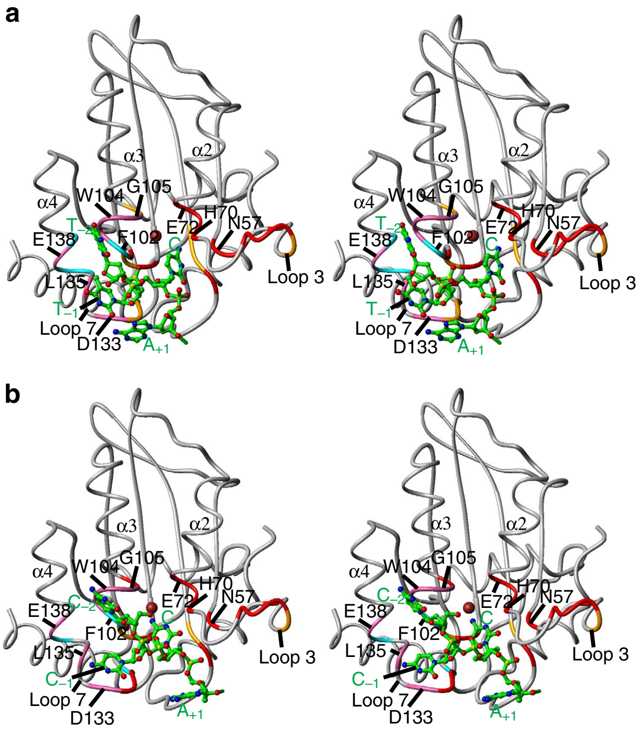


Figure 5 | Model of A3A complexed with TTCAT or CCCAA. Stereoviews of the A3A backbone structure and the 5-nt ssDNAs (TTCAT (**a**) and CCCAA (**b**)), as well as interacting residues (from Fig. 4b or Fig. 4g) in ball-and-stick representation, with carbon, nitrogen, oxygen, phosphorus and Zn atoms in green, blue, red, gold and brown, respectively. The 3′-end thymidine (**a**) and adenosine (**b**), which appear to be random, are omitted for clarity.

Deamination by real-time NMR. The A3A-catalyzed deamination reaction with several substrates was followed by two-dimensional (2D) ^{1}H–^{13}C HSQC (Fig. 6a) or 1D ^{1}H (Fig. 6b–d) NMR. Using dCTP as the substrate, we determined that slow but measurable deaminase activity is present (Fig. 6a): at 25 °C, 5.2 mM dCTP was completely converted to dUTP by 170 μM A3A in ~50 h. Longer substrates, such as the 9-nt 5′-ATTTCATTT-3′, were deaminated much more rapidly (Fig. 6b), and in such cases 1D ^{1}H NMR spectroscopy was used (Fig. 6b inset). The 9-nt substrate (0.98 mM) was completely deaminated by 0.196 μM A3A in ~1.7 h, with an initial rate of ~0.8 mM h^{-1}. Other substrates, a 15-mer, 5′-ATTATTTCATTTATT-3′, and two hexanucleotides, 5′-ATTTCA-3′ and 5′-TCATTT-3′, (Supplementary Fig. S3a–c, respectively) exhibited essentially identical rates to that observed with the nona-nucleotide.

A3A and A3G purportedly possess a difference in substrate specificity[2,6,7,10,39,41]. To quantify A3A's substrate specificity, we also assayed the deamination of an A3G-specific substrate, AAA$\mathrm{C}_{-2}\mathrm{C}_{-1}$CAAA, by A3A. Interestingly, all three cytidines were deaminated in a sequential 3′→5′ manner (Fig. 6c), with the third dC (C) converted with an initial rate of ~0.2 mM h^{-1}, only fourfold slower than the dC in the A3A-specific 5′-ATTTCATTT-3′ substrate. When comparing substrates with varying numbers of cytosines, such as 5′-AAACCCAAA-3′ (Fig. 6c), 5′-AAACCAAAA-3′ (Fig. 6d) and 5′-AAACAAAAA-3′ (Supplementary Fig. S3d), deamination of the 3′ dC (C) was about two- to four-fold faster in 5′-AAACCCAAA-3′ than in the other two substrates, indicating a preference for two pyrimidine bases at the 5′ side of the reactive C.

We also noted that the kinetics of the C→U conversions of C_{-1} and C_{-2} are much slower, quite distinct from that of C in the reaction with the AAA$\mathrm{C}_{-2}\mathrm{C}_{-1}$CAAA substrate. Similar behaviour was observed with AAA$\mathrm{C}_{-2}\mathrm{C}_{-1}$UAAA (Supplementary Fig. S3e), a substrate containing a uridine instead of the most reactive C. In contrast, when dC is followed by dA, and not by dU, for example, the C in 5′-AAAC_{-1}CAAA-3′ (Fig. 6d) and 5′-AAACAAAAA-3′ (Supplementary Fig. S3d), the reaction proceeds readily, albeit somewhat slower than with AAA$\mathrm{C}_{-2}\mathrm{C}_{-1}$CAAA (Fig. 6c). Given our observation that dC and dU essentially possess the same A3A-binding affinities and interfaces, it seems reasonable to assume that once dC is converted to dU, competitive binding and inhibition occurs, slowing the deamination of C_{-1} and C_{-2} in AAA$\mathrm{C}_{-2}\mathrm{C}_{-1}$UAAA. Therefore, optimal substrates for A3A contain a cytidine preceded by two pyrimidine bases and followed by an adenine; that is, (T/C)(T/C)CA.

Further studies were performed to determine the catalytic constants for A3A deaminase activity on single cytidine-containing substrates; that is, 5′-ATTTCATTT-3′, 5′-ATTATTTCATTTATT-3′, 5′-ATTTCA-3′, 5′-TCATTT-3′ and 5′-AAACAAAAA-3′. The kinetics of deamination revealed that all TCA-containing oligonucleotide substrates, regardless of length, exhibited identical turnover rates (k_{cat} ~ 60–70 min^{-1}; Table 2, Fig. 6e–g). The hexanucleotides, however, displayed an approximately threefold lower apparent second-order rate constant (k_{cat}/K_M ~ 5 × 10^3 M^{-1}s^{-1}) than the one measured for the 9-nt and 15-nt substrates (k_{cat}/K_M ~ 1.8 × 10^4 M^{-1}s^{-1}), caused mainly by an increase in K_M. For all A3A substrates, K_M and K_d values were very similar. A large reduction (~40-fold) of k_{cat}/K_M was seen only with a substrate that is lacking pyrimidine bases at the 5′ side of the reactive C; that is, 5′-AAACAAAAA-3′, caused by a large (13-fold) reduction in k_{cat} and a small (approximately threefold) increase in K_M (less-favourable binding), highlighting the importance of a pyrimidine base at the 5′ side.

Collectively, the data in Fig. 6 demonstrate that the nucleotide context surrounding the reactive C is a major determinant of enzymatic activity.

Discussion

Here we present the NMR solution structure of human A3A as well as a detailed analysis of its nucleic acid interaction surface, providing new insights into substrate selection and binding. The A3A NMR structure very closely resembles the A3G-CTD and A3C X-ray structures[31–36] (Supplementary Table S1) and not surprisingly, there are also similarities between the activities of

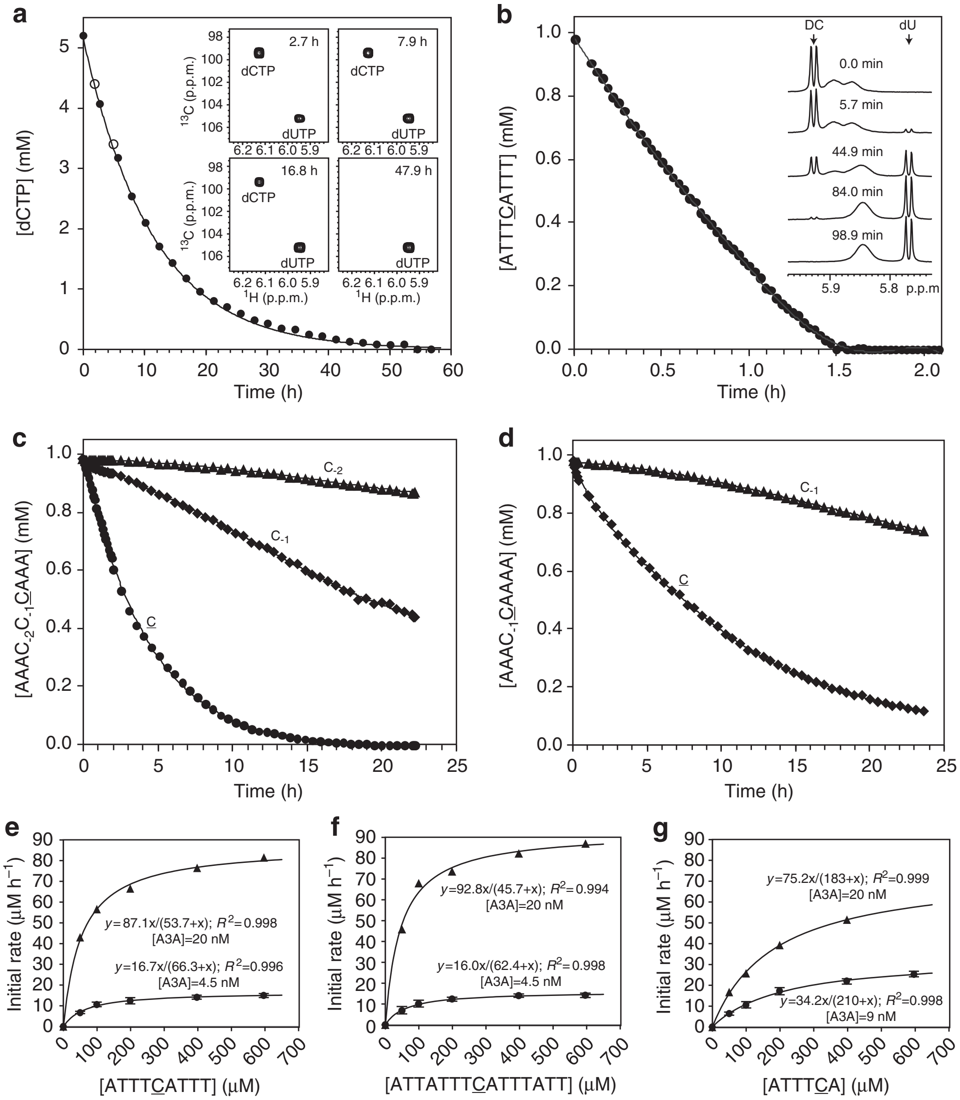


Figure 6 | A3A-catalyzed deamination of dCTP and several ssDNA substrates monitored by real-time NMR. Concentrations of dCTP (**a**), 9-nt ssDNA ATTT$\underline{C}$ATTT (**b**), AAA$\underline{C}_{-2}\underline{C}_{-1}\underline{C}$AAA (**c**) and AAA$\underline{C}_{-1}\underline{C}$AAAA (**d**) versus incubation time are provided. All concentrations of unreacted substrates (cytidine) and end products (uridine) were determined by measuring the intensities of the ^{13}C-5-^{1}H resonances of cytosine and uracil in the 2D ^{1}H-^{13}C HSQC spectra (**a**; 600 MHz) or the ^{1}H-5 resonances in 1D ^{1}H spectra (**b**–**d**; 900 MHz), as a function of time. The A3A concentration was 0.17 mM (**a**) or 0.197 μM (**b**–**d**). Best fit curves are shown by a solid line. Representative 2D ^{1}H-^{13}C HSQC (**a**) or 1D ^{1}H NMR (**b**) spectra acquired at the indicated times are shown in the inset. (**e**–**g**) Kinetics of A3A-catalyzed deamination reactions for ATTT$\underline{C}$ATTT (**e**), ATTATTT$\underline{C}$ATTTATT (**f**) and ATTT$\underline{C}$A (**g**); initial reaction rates ($<$5%) are plotted versus substrate concentrations. Reactions were monitored by 1D ^{1}H real-time NMR (900 MHz) at 25 °C in 25 mM sodium phosphate buffer, pH 6.9. Two different A3A concentrations were used: (**e** and **f**), 4.5 nM (circles) and 20 nM (triangles); (**g**), 9.0 nM (circles) and 20 nM (triangles). Two independent experiments were performed with 4.5 and 9 nM A3A and average values with s.d. are shown. Reactions with the higher A3A concentration (20 nM) resulted in K_M and k_{cat} values very similar to those obtained with low A3A concentrations (4.5 or 9 nM).

A3A and the A3G-CTD. Using NMR, we show that A3A binds TTCA- or CCCA-containing single-stranded oligonucleotides ($\geq$9 nt) with K_d values ranging from 50–100 μM (Table 2), in excellent agreement with binding affinity data ($\sim$80 μM) estimated from electrophoretic mobility shift assay (Fig. 1c) and with fluorescence depolarization results obtained by Love *et al.*[39]

Intriguingly, A3A binds ssDNA much more weakly ($\sim$1000-fold) than A3G, for which K_d values ranging from 50 to 240 nM have been reported[17,39,42–44]. However, like A3A, K_d values (200–450 μM) were obtained for the single-domain A3G-CTD protein[31,32], implying that the tighter binding of full-length A3G is associated with its double domain structure. Indeed, the A3G-N-terminal domain contains numerous positively charged amino acids that contribute to high efficiency binding to ssDNA[17,29], whereas both A3A and the A3G-CTD are slightly acidic and are unable to interact in this manner.

Importantly, titration of A3A with diverse ssDNA substrates by NMR made it possible to distinguish specific binding from non-specific binding (Fig. 4c). Our results showed that A3A can bind both TT$\underline{C}$A- and CC$\underline{C}$A-containing oligonucleotides, using the same five A3A contacts, namely the active site, loop 3, loop 5 including the exposed dipeptide W104-G105, loop 7 and helix α4. Surprisingly, although the W104NεH side chain resonance experiences the largest chemical shift changes upon substrate binding (Fig. 4), W104 mutations do not seem to influence or abrogate A3A's catalytic activity[16,18].

Note that A3A oligonucleotide-binding regions are highly localized and clustered around the active site, in stark contrast to results in the two studies that mapped ssDNA binding to A3G-CTD (refs 31,32). Interestingly, our A3A–ssDNA complex model (Fig. 5) suggests that the DNA bends to insert the reactive cytidine into the active site, permitting only the immediate neighbouring (− 1, − 2 and + 1) nucleotides to interact locally near the catalytic site. Interestingly, there are reports of RNA bending in the crystal structure of a transfer RNA adenosine deaminase/RNA complex[45] and DNA contraction during A3G scanning of the ssDNA substrate[46].

NMR real-time kinetic data for the deamination reaction of A3A using 5′-ATTTCATTT-3′ and 5′-ATTATTTCATTT-3′ as substrates yielded values of $k_{\mathrm{cat}} \sim 70\,\mathrm{min}^{-1}$ and $K_{\mathrm{M}} \sim 60\,\mu\mathrm{M}$, which differ from the values ($k_{\mathrm{cat}} \sim 15\,\mathrm{min}^{-1}$ and $K_{\mathrm{M}} \sim 230\,\mathrm{nM}$) obtained using a 43-nt TTCT-containing ssDNA[18]. These differences may be related to variations in the ssDNA sequences and experimental conditions. Interestingly, Carpenter *et al.* reported that A3A is a stronger (~ 200-fold) deaminase than full-length A3G (ref. 18). Indeed, although similar amounts of deaminase product were observed in the NMR deamination assays described here for A3A (Fig. 6b) and for A3G-CTD in Furukawa *et al.*[32], the amount of A3A used was 1000-fold less than the amount of A3G-CTD.

The fact that A3A also deaminates the most 3′ dC (C) in 5′-AAACCCAAA-3′ (thought to be an A3G-specific substrate[2]) with only an approximate fivefold reduction in $k_{\mathrm{cat}}/K_{\mathrm{M}}$, compared with TTCA-containing 9-nt and 15-nt oligonucleotides (Table 2) appears puzzling at first. However, these results are in agreement with the deaminase specificity observed for A3A in cell-based assays. For example, in an investigation of foreign DNA restriction in human primary cells, A3A preferentially deaminated TC and CC sequences in the green fluorescent protein gene of the transfected plasmid DNA (ref. 6). In addition, a strong bias for deamination of TC (~ 50–70%) and CC (~ 15–25%) dinucleotide sequences was detected in nascent HIV-1 complementary DNAs isolated from infected macrophages[28] and in an *in vitro* HIV-1 model replication assay, performed using purified A3A protein[39]. Furthermore, a preference for TC and CC substrate recognition sites was also observed for A3A editing of human T-lymphotropic virus type 1 (ref. 12). Thus, A3A's active site possesses the flexibility to accommodate and deaminate dC residues (C) in both TC and CC dinucleotides.

Interestingly, in a study by Shinohara *et al.*[7], it was shown that A3A mediates genomic DNA editing in human cells, but no editing site preference was detected. In contrast, Suspène *et al.*[21] found that A3A preferentially deaminates cytidines in TC and CC dinucleotides in genomic DNA, when cells are exposed to UDG inhibitor. APOBEC-mediated genomic DNA mutations have been implicated in carcinogenesis[47] and, for example, A3B was shown to be a source of DNA mutations in breast cancer[48]. These observations suggest that the strong mutagenic potential of A3A might be detrimental to the stability of the human genome.

Thus, the dual function of A3A as a host restriction factor and as a DNA mutator that can potentially act on genomic DNA, an activity that may be associated with malignancies, suggests that A3A can act as a 'double-edged sword'. The high resolution NMR structure of A3A presented here is a first step in aiding future structure-function studies for addressing these seemingly diverse A3A functions. Furthermore, the addition of the A3A structure to the still limited list of currently known APOBEC structures contributes to efforts towards elucidating the molecular mechanisms of the innate immune response.

Methods

Protein expression and purification. Wild-type (Accession number NM_145699), E72Q and L63N/C64S/C171Q mutant synthetic A3A genes with a C-terminal His$_6$-tag (LEHHHHHH) were inserted into the *Nde*I–*Xho*I site of the pET21 plasmid (Novagen) for expression in *Escherichia coli* Rosetta 2 (DE3). Uniform ^{15}N- and ^{13}C-labelling of the proteins was carried out by growth in modified minimal medium at 18 °C, using ^{15}NH$_4$Cl and ^{13}C$_6$-glucose as the sole nitrogen and carbon sources, respectively. Uniform ^{2}H-, ^{15}N- and ^{13}C-labelling of the proteins was achieved using ^{2}H$_2$O, ^{15}NH$_4$Cl and ^{13}C$_6$/^{2}H$_7$-glucose as deuterium, nitrogen and carbon sources, respectively, with two different selective protonation of the side chains of (1) Tyr/Phe/Ile residues and (2) Tyr/Phe/Trp/Ile/Val/Leu residues, by adding 0.10–0.15 mg of ^{13}C/^{15}N-tyrosine, -phenylalanine and -isoleucine (for sample 1), and ^{13}C/^{15}N-tyrosine and -phenylalanine, unlabelled tryptophan, 2-keto-butyrate (1,2,3,4-^{13}C, 98%; 3,3'-^{2}H, 98%, CIL, Andover, MA, USA) and 2-keto-3-methyl-butyrate (1,2,3,4-^{13}C, 99%; 3,4,4',4''-^{2}H, 98%, CIL; for sample 2), respectively. These chemicals were added to the culture 1 h before induction with 0.4 mM isopropyl-1-thio-β-D-galactopyranoside (total induction time = 16 h). Proteins were purified over a 5-ml Hi-Trap His column (GE Healthcare) and Hi-Load Superdex 200 (1.6 cm × 60 cm) column, equilibrated in buffer containing 25 mM Tris–HCl (pH 7.5), 50 mM NaCl, 5% glycerol, 2 mM dithiothreitol (DTT) and 0.02% NaN$_3$. Fractions containing A3A were further purified over an 8-ml MONO-Q column (GE Healthcare) in 25 mM Tris-HCl buffer (pH 8.5), 5% glycerol, 2 mM DTT and 0.02% sodium azide, employing a linear gradient of 0–1 M NaCl. The final A3A preparations were > 99% pure, as estimated by SDS–polyacrylamide gel electrophoresis. The molecular mass of the A3A proteins were confirmed by LC-ESI-TOF mass spectrometry (Bruker Daltonics, Billerica, MA, USA).

Multi-angle light scattering. Size-exclusion chromatography/multi-angle light scattering data were obtained at room temperature using an analytical Superdex 200 (S200) column with in-line multi-angle light-scattering refractive index (Wyatt Technology, Inc., Santa Barbara, CA, USA) and ultraviolet (Agilent Technologies, Santa Clara, CA) detectors. One hundred microlitres of 78.4 μM A3A were applied to the S200 column pre-equilibrated and eluted with 25 mM sodium phosphate buffer (pH 6.5), 200 mM NaCl, 0.02% sodium azide and 1 mM DTT at a flow rate of 0.5 ml min^{-1}.

Deaminase assay using fluorescent-tagged ssDNA substrates. Deaminase assay conditions were adapted from Iwatani *et al.*[17] Forty microlitres reactions, containing 180 nM of a 40-nt ssDNA (JL913, 5′-ATT ATT ATT ATT ATT ATT ATT TCA TTT ATT TAT TTA TTT A-3′), labelled at its 5′ end with Alexa Fluor 488, (Integrated DNA Technologies (IDT, Coralville, IA, USA) and varying amounts of A3A in 10 mM Tris–HCl buffer, pH 8.0, 50 mM NaCl, 1 mM DTT, 1 mM EDTA, pH 8.0, and 10 units of *E. coli* UDG (New England BioLabs) were incubated at 37 °C for 1 h. The reaction was stopped by incubation with Proteinase K (40 μg, Ambion) at 65 °C for 20 min, followed by sequential addition of 10 μl of 1 N NaOH for 15 min at 37 °C and 10 μl of 1 N HCl. Ten microlitres aliquots of the final mixture were subjected to electrophoresis in a 10% denaturing polyacrylamide gel. Gels were scanned in fluorescence mode on a Typhoon 9400 Imager and the data were quantified using ImageQuant software (GE Healthcare).

Electrophoretic mobility shift assay. Ten microlitres reactions, containing varying amounts of A3A, 20 nM of a 5′ ^{32}P-labelled 40-nt ssDNA (JL895, identical sequence to JL913 but without the Alexa Fluor 488 label; Lofstrand Labs Ltd, Gaithersburg, MD, USA) or 40-nt ssRNA (JL931, identical sequence to JL895, except that U was substituted for T; IDT), were incubated with 4 U SUPERase IN (Ambion) at 37 °C for 10 min in 50 mM Tris–HCl buffer, pH 7.0, 100 mM NaCl, 1 mM DTT, 1% Ficoll-400 and 2.5 mM EDTA. Aliquots from each reaction were loaded onto an 8% native polyacrylamide gel in 40 mM Tris-acetate buffer, pH 8.4, 1 mM EDTA and 5% glycerol. One microlitre DNA Loading Gel Solution (Quality Biological, Inc.), containing bromphenol blue and xylene cyanol, was added to a control sample without A3A. Gels were run at 4 °C (5 mA) until the bromphenol blue dye had migrated ~ 2/3 through the gel. Radioactive products were detected with a Typhoon 9400 Imager and were quantified using ImageQuant software.

NMR spectroscopy. All NMR spectra for the structure determination of A3A were recorded at 25 °C on Bruker AVANCE900, AVANCE800, AVANCE700 and AVANCE600 spectrometers, equipped with 5 mm triple resonance, Z-axis gradient cryoprobes. The NMR samples contained unlabelled, ^{13}C/^{15}N- or ^{2}H/^{13}C/^{15}N-labelled A3A with two types of selective protonations (see above). At pH 6.9 and 8.1, A3A showed higher solubility (~ 0.5 and ~ 1 mM, respectively) than that at pH 6.5 (~ 0.2 mM). However, at these higher protein concentrations, soluble aggregation occurred, as evidenced by severe line broadening. We therefore performed all of our NMR experiments at concentrations of ~ 0.2–0.3 mM. The sample temperature in the spectrometer was calibrated with 100% methanol. Backbone and side chain resonance assignments were carried out using 2D ^{1}H–^{15}N HSQC, ^{1}H–^{13}C HSQC and nuclear Överhauser enhancement spectroscopy (NOESY) and three-dimensional (3D) HNCACB, HN(CO)CACB, HNCA, HN(CO)CA and HCCH-total correlation spectroscopy experiments[49]. Distance constraints were derived from 3D simultaneous ^{13}C- and ^{15}N-edited NOESY (ref. 50) and 2D NOESY experiments. All NOESY spectra were acquired at 800 or 900 MHz, using a mixing time of 100 ms (non-perdeuterated samples) or 150 ms (perdeuterated samples). Spectra were processed with TOPSPIN 2.1 (Bruker) and

NMRPipe[51], and analyzed using SPARKY3 (version 3.113; T.D. Goddard and D.G. Kneller, University of California, San Francisco) and NMRView J (version 8.0.3)[52].

NMR structure calculation. All NOE cross peaks were assigned from the 3D and 2D NOESY spectra using the SPARKY3 assignment tool. Structure calculations were performed for A3A residues 1–199, using the anneal.py protocol in XPLOR-NIH (ref. 53). An iterative approach with extensive, manual cross-checking of all distance constraints against the NOESY data and the generated structures was employed. The final number of the NMR-derived constraints were 3279, with 2827 NOE distances, 152 H-bond distances identified from NOE patterns for helices and β-sheets, and 300 ϕ and ψ backbone torsion angles from TALOS calculations[54]. In addition, a Zn^{2+} ion was added at a late stage in the calculations to coordinate with H70, C101 and C106 (refs 31,32) using constraints based on the X-ray structures of A3G-CTD (refs 34,35). Five hundred and twelve structures were generated and the 30 lowest energy structures were selected and analyzed using PROCHECK-NMR (ref. 55) (Table 1). All structure figures were generated with MOLMOL (ref. 56).

A3A nucleotide-binding site mapping and titration by NMR. To monitor nucleotide binding to A3A, aliquots of 14–100 mM mono- or oligodeoxynucleotide stock solutions were added to 0.10–0.18 mM $^{13}C/^{15}N$-labelled A3A. 100 mM dCTP and dUTP stock solutions were purchased from Promega (Madison, WI, USA) and HPLC-purified DNA oligonucleotides were obtained from IDT or Midland Co. (Midland, TX, USA). A series of 2D $^{1}H–^{15}N$ HSQC titration spectra were acquired and binding isotherms were obtained by plotting ^{1}HN proton chemical shift change versus nucleotide concentrations for 5–8 unambiguously traceable amide resonances. Dissociation constants were calculated by non-linear best fitting of the isotherms using KaleidaGraph (Synergy Software, Reading, PA, USA), and values for 5–8 resonances were averaged.

Real-time studies of A3A-catalyzed deamination. A series of 2D $^{1}H–^{13}C$ HSQC and/or 1D ^{1}H NMR spectra were acquired as a function of time, after addition of A3A to solutions of mono- or oligodeoxynucleotide. Concentrations of A3A and deoxynucleotides for the different samples are provided in the figure captions. The intensities of well-resolved ^{13}C-5-^{1}H resonances (volumes in 2D $^{1}H–^{13}C$ HSQC spectra) and/or ^{1}H-5 resonances (integrals in 1D ^{1}H spectra) of cytosine and uracil were used for quantification. Real-time monitoring of the A3A-catalyzed deamination reaction by NMR permitted the extraction of initial ($<5\%$ dC → dU conversion) rates for a series of substrate concentrations (50, 100, 200, 400 and 600 μM). k_{cat} (V_{max}/[A3A]) and K_M values were obtained using the Michaelis–Menten module in Kaleidagraph.

Molecular docking. Structures of 5′-TTCAT-3′, 5′-CCCAA-3′, 5′-ATTTCATTT-3′ and 5′-AAACCCAAA-3′ ssDNAs were generated by MacroMoleculeBuilder (version 2.8)[57]. The oligonucleotide and A3A NMR structures were converted to a mol2 file format using the Hermes programme (version 1.4) and standard Tripos atom and bond types. Docking was performed with the flexible docking programme GOLD version 5.1 (ref. 40), in combination with the Chemscore scoring function[58]. The Zn^{2+} ion was chosen as the centre point in the docking and the radius was set to 20 Å. For each nucleotide, five independent docking runs into the five lowest energy NMR structures of A3A, each producing fifty binding poses, were performed using a population of 100 ligands, with a maximal number of genetic algorithm operations dependent on ligand size: 50,000 for 5′-TTCAT-3′ and 5′-CCCAA-3′, and 150,000 for 5′-ATTTCATTT-3′ and 5′-AAACCCAAA-3′. Docking poses (1250) for the pentanucleotides were subjected to a two-step selection process utilizing the NMR data. In the first step, complexes were discarded if the oligonucleotide engaged in contacts with residues that did not exhibit experimental chemical shift changes. In the second step, structures were selected that exhibited good agreement with the NMR binding-site mapping data, taking the size of the chemical shift changes and the scoring value of the binding pose into account. The final structural models for A3A, complexed with 5′-TTCAT-3′ (Fig. 5a) and 5′-CCCAA-3′ (Fig. 5b), represent ~80% of all the structures that result from the selection process.

References

1. Chiu, Y. L. & Greene, W. C. The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements. *Annu. Rev. Immunol.* **26,** 317–353 (2008).
2. Malim, M. H. APOBEC proteins and intrinsic resistance to HIV-1 infection. *Philos. Trans. R Soc. Lond. B Biol. Sci.* **364,** 675–687 (2009).
3. Duggal, N. K. & Emerman, M. Evolutionary conflicts between viruses and restriction factors shape immunity. *Nat. Rev. Immunol.* **12,** 687–695 (2012).
4. Jarmuz, A. *et al.* An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. *Genomics* **79,** 285–296 (2002).
5. Betts, L., Xiang, S., Short, S. A., Wolfenden, R. & Carter, Jr C. W. Cytidine deaminase. The 2.3 Å crystal structure of an enzyme: transition-state analog complex. *J. Mol. Biol.* **235,** 635–656 (1994).
6. Stenglein, M. D., Burns, M. B., Li, M., Lengyel, J. & Harris, R. S. APOBEC3 proteins mediate the clearance of foreign DNA from human cells. *Nat. Struct. Mol. Biol.* **17,** 222–229 (2010).
7. Shinohara, M. *et al.* APOBEC3B can impair genomic stability by inducing base substitutions in genomic DNA in human cells. *Sci. Rep.* **2,** 806 (2012).
8. Vartanian, J. P., Guetard, D., Henry, M. & Wain-Hobson, S. Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions. *Science* **320,** 230–233 (2008).
9. Wiegand, H. L. & Cullen, B. R. Inhibition of alpharetrovirus replication by a range of human APOBEC3 proteins. *J. Virol.* **81,** 13694–13699 (2007).
10. Chen, H. *et al.* APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons. *Curr. Biol.* **16,** 480–485 (2006).
11. Narvaiza, I. *et al.* Deaminase-independent inhibition of parvoviruses by the APOBEC3A cytidine deaminase. *PLoS Pathog.* **5,** e1000439 (2009).
12. Ooms, M., Krikoni, A., Kress, A. K., Simon, V. & Münk, C. APOBEC3A, APOBEC3B, and APOBEC3H haplotype 2 restrict human T-lymphotropic virus type 1. *J. Virol.* **86,** 6097–6108 (2012).
13. Bogerd, H. P. *et al.* Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. *Proc. Natl Acad. Sci. USA* **103,** 8780–8785 (2006).
14. Muckenfuss, H. *et al.* APOBEC3 proteins inhibit human LINE-1 retrotransposition. *J. Biol. Chem.* **281,** 22161–22172 (2006).
15. Kinomoto, M. *et al.* All APOBEC3 family proteins differentially inhibit LINE-1 retrotransposition. *Nucleic Acids Res.* **35,** 2955–2964 (2007).
16. Bulliard, Y. *et al.* Structure-function analyses point to a polynucleotide-accommodating groove essential for APOBEC3A restriction activities. *J. Virol.* **85,** 1765–1776 (2011).
17. Iwatani, Y., Takeuchi, H., Strebel, K. & Levin, J. G. Biochemical activities of highly purified, catalytically active human APOBEC3G: correlation with antiviral effect. *J. Virol.* **80,** 5992–6002 (2006).
18. Carpenter, M. A. *et al.* Methylcytosine and normal cytosine deamination by the foreign DNA restriction enzyme APOBEC3A. *J. Biol. Chem.* **287,** 34801–34808 (2012).
19. Wijesinghe, P. & Bhagwat, A. S. Efficient deamination of 5-methylcytosines in DNA by human APOBEC3A, but not by AID or APOBEC3G. *Nucleic Acids Res.* **40,** 9206–9217 (2012).
20. Landry, S., Narvaiza, I., Linfesty, D. C. & Weitzman, M. D. APOBEC3A can activate the DNA damage response and cause cell-cycle arrest. *EMBO Rep.* **12,** 444–450 (2011).
21. Suspène, R. *et al.* Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism. *Proc. Natl Acad. Sci. USA* **108,** 4858–4863 (2011).
22. Aynaud, M. M. *et al.* Human Tribbles 3 protects nuclear DNA from cytidine deamination by APOBEC3A. *J. Biol. Chem.* **287,** 39182–39192 (2012).
23. Peng, G. *et al.* Myeloid differentiation and susceptibility to HIV-1 are linked to APOBEC3 expression. *Blood* **110,** 393–400 (2007).
24. Koning, F. A. *et al.* Defining APOBEC3 expression patterns in human tissues and hematopoietic cell subsets. *J. Virol.* **83,** 9474–9485 (2009).
25. Refsland, E. W. *et al.* Quantitative profiling of the full *APOBEC3* mRNA repertoire in lymphocytes and tissues: implications for HIV-1 restriction. *Nucleic Acids Res.* **38,** 4274–4284 (2010).
26. Thielen, B. K. *et al.* Innate immune signaling induces high levels of TC-specific deaminase activity in primary monocyte-derived cells through expression of APOBEC3A isoforms. *J. Biol. Chem.* **285,** 27753–27766 (2010).
27. Berger, G. *et al.* APOBEC3A is a specific inhibitor of the early phases of HIV-1 infection in myeloid cells. *PLoS Pathog.* **7,** e1002221 (2011).
28. Koning, F. A., Goujon, C., Bauby, H. & Malim, M. H. Target cell-mediated editing of HIV-1 cDNA by APOBEC3 proteins in human macrophages. *J. Virol.* **85,** 13448–13452 (2011).
29. Navarro, F. *et al.* Complementary function of the two catalytic domains of APOBEC3G. *Virology* **333,** 374–386 (2005).
30. Newman, E. N. *et al.* Antiviral function of APOBEC3G can be dissociated from cytidine deaminase activity. *Curr. Biol.* **15,** 166–170 (2005).
31. Chen, K.-M. *et al.* Structure of the DNA deaminase domain of the HIV-1 restriction factor APOBEC3G. *Nature* **452,** 116–119 (2008).
32. Furukawa, A. *et al.* Structure, interaction and real-time monitoring of the enzymatic reaction of wild-type APOBEC3G. *EMBO J.* **28,** 440–451 (2009).
33. Harjes, E. *et al.* An extended structure of the APOBEC3G catalytic domain suggests a unique holoenzyme model. *J. Mol. Biol.* **389,** 819–832 (2009).
34. Holden, L. G. *et al.* Crystal structure of the anti-viral APOBEC3G catalytic domain and functional implications. *Nature* **456,** 121–124 (2008).
35. Shandilya, S. M. D. *et al.* Crystal structure of the APOBEC3G catalytic domain reveals potential oligomerization interfaces. *Structure* **18,** 28–38 (2010).
36. Kitamura, S. *et al.* The APOBEC3C crystal structure and the interface for HIV-1 Vif binding. *Nat. Struct. Mol. Biol.* **19,** 1005–1010 (2012).
37. Krzysiak, T. C., Jung, J., Thompson, J., Baker, D. & Gronenborn, A. M. APOBEC2 is a monomer in solution: implications for APOBEC3G models. *Biochemistry* **51,** 2008–2017 (2012).

38. Bransteitter, R., Prochnow, C. & Chen, X. S. The current structural and functional understanding of APOBEC deaminases. *Cell. Mol. Life Sci.* **66,** 3137–3147 (2009).
39. Love, R. P., Xu, H. & Chelico, L. Biochemical analysis of hypermutation by the deoxycytidine deaminase APOBEC3A. *J. Biol. Chem.* **287,** 30812–30822 (2012).
40. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins* **52,** 609–623 (2003).
41. Aguiar, R. S., Lovsin, N., Tanuri, A. & Peterlin, B. M. Vpr.A3A chimera inhibits HIV replication. *J. Biol. Chem.* **283,** 2518–2525 (2008).
42. Chelico, L., Pham, P., Calabrese, P. & Goodman, M. F. APOBEC3G DNA deaminase acts processively $3' \rightarrow 5'$ on single-stranded DNA. *Nat. Struct. Mol. Biol.* **13,** 392–399 (2006).
43. Nowarski, R., Britan-Rosich, E., Shiloach, T. & Kotler, M. Hypermutation by intersegmental transfer of APOBEC3G cytidine deaminase. *Nat. Struct. Mol. Biol.* **15,** 1059–1066 (2008).
44. Iwatani, Y. *et al.* Deaminase-independent inhibition of HIV-1 reverse transcription by APOBEC3G. *Nucleic Acids Res.* **35,** 7096–7108 (2007).
45. Losey, H. C., Ruthenburg, A. J. & Verdine, G. L. Crystal structure of Staphylococcus aureus tRNA adenosine deaminase TadA in complex with RNA. *Nat. Struct. Mol. Biol.* **13,** 153–159 (2006).
46. Senavirathne, G. *et al.* Single-stranded DNA scanning and deamination by APOBEC3G cytidine deaminase at single molecule resolution. *J. Biol. Chem.* **287,** 15826–15835 (2012).
47. Roberts, S. A. *et al.* Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46,** 424–435 (2012).
48. Burns, M. B. *et al.* APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494,** 366–370 (2013).
49. Clore, G. M. & Gronenborn, A. M. Determining the structures of large proteins and protein complexes by NMR. *Trends Biotechnol.* **16,** 22–34 (1998).
50. Sattler, M., Maurer, M., Schleucher, J. & Griesinger, C. A Simultaneous ^{15}N, ^{1}H-HSQC and ^{13}C, ^{1}H-HSQC with sensitivity enhancement and a heteronuclear gradient-echo. *J. Biomol. NMR* **5,** 97–102 (1995).
51. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6,** 277–293 (1995).
52. Johnson, B. A. & Blevins, R. A. NMR View - a computer-program for the visualization and analysis of NMR data. *J. Biomol. NMR* **4,** 603–614 (1994).
53. Brunger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta. Crystallogr. D Biol. Crystallogr.* **54**(Pt 5): 905–921 (1998).
54. Cornilescu, G., Delaglio, F. & Bax, A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13,** 289–302 (1999).
55. Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R. & Thornton, J. M. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8,** 477–486 (1996).
56. Koradi, R., Billeter, M. & Wüthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14,** 29–32 (1996).
57. Flores, S. C., Sherman, M. A., Bruns, C. M., Eastman, P. & Altman, R. B. Fast flexible modeling of RNA structure using internal coordinates. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8,** 1247–1257 (2011).
58. Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V. & Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.* **11,** 425–445 (1997).

Acknowledgements

We thank Maria DeLucia for help with A3A purification, Doug Bevan and Phil Greer for computer technical support and Michael J Delk for NMR instrumental support. J.H. acknowledges financial support by an International Outgoing Fellowship of the European Community program 'Support for training and career development of researchers (Marie Curie)', under Contract No. PIOF-GA-2009-235902. This work was supported in part by the National Institutes of Health NIGMS Grant P50GM082251 (A.M.G.) and the Intramural Research Program at the National Institutes of Health, Eunice Kennedy Shriver National Institute of Child Health and Human Development (M.M., K.H. and J.G.L.).

Author contributions

The study was conceived by A.M.G. and J.G.L. The NMR experiments were designed by A.M.G. and I.-J.L.B.; I.-J.L.B. and C.-H.B. performed experiments and data analysis; I.-J.L.B. and A.M.G. interpreted the NMR data; J.A., L.M.C. and C.-H.B. prepared purified A3A; J.H. built structural models of A3A, complexed with short oligonucleotides; the biochemical assays were designed by J.G.L., M.M. and K.H.; M.M. and K.H. performed the experiments and M.M., K.H. and J.G.L. analyzed the results; the manuscript was written by I.-J.L.B., M.M., J.G.L. and A.M.G.

Additional information

Accession codes: The A3A atomic coordinates and NMR constraints have been deposited in the RCSB Protein Data Bank under accession code 2m65, and the NMR chemical shift data have been deposited in the Biological Magnetic Resonance Bank under accession code 19108.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors claim no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Byeon, I.-J. L. *et al.* NMR structure of human restriction factor APOBEC3A reveals substrate binding and enzyme specificity. *Nat. Commun.* 4:1890 doi: 10.1038/ncomms2883 (2013).

4 Methodological developments of enhanced sampling methodology H-REMD

Paper 5: Hritz J., Oostenbrink C. Hamiltonian replica exchange molecular dynamics using soft-core interactions. *J. Chem. Phys.* **2008,** 128, 144121

Paper 6: Hritz J., Oostenbrink C. Optimization of Replica Exchange Molecular Dynamics by Fast Mimicking. *J. Chem. Phys.* **2007,** 127, 204104

Paper 7: Nagy G., Oostenbrink C., Hritz J.*: Exploring the Binding Pathways of the 14-3-3ζ Protein: Structural and Free-Energy Profiles Revealed by Hamiltonian Replica Exchange Molecular Dynamics with Distance Field Distance Restraints. *PLoS ONE* **2017**,12(7), e0180633

Two types of Hamiltonian changes within H-REMD were addressed by the applicant: increasing softness of the selected non-bonded interactions and distance restraints applied to the ligand with respect to its binding site.

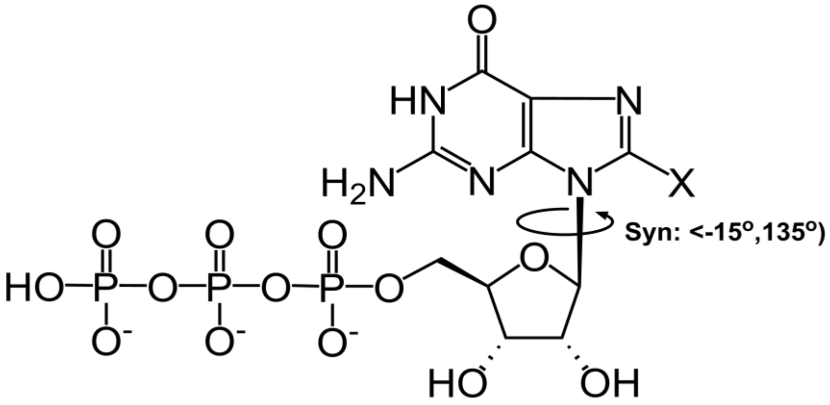


Figure 2: Structure of GTP and 8-Br-GTP in syn conformation.

Conformational transitions between the syn and anti states occur by rotation around the glycosidic bond (indicated). The glycosidic dihedral angle φ is defined over atoms: C4-N9-C1-O4.

]

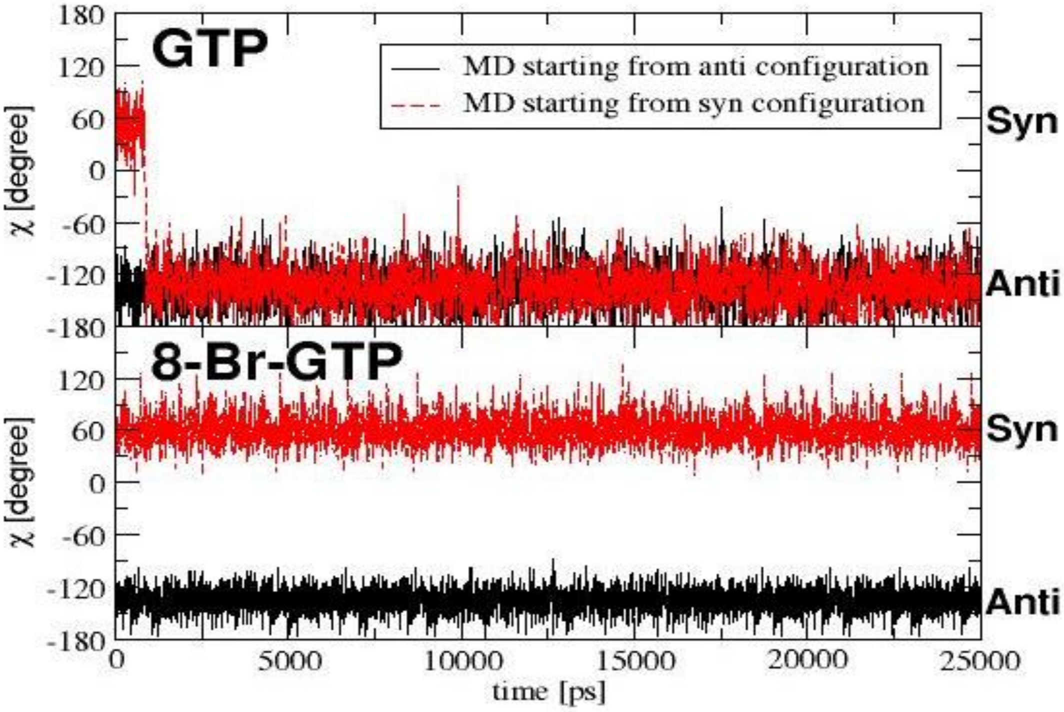


Figure 3: Time dependence of dihedral angle around the glycosidic bond during two MD simulations for GTP and 8-Br-GTP

Values for MD simulations starting from anti conformation are shown in black and from syn conformation in red color.

4.1 H-REMD using soft-core interactions

High intramolecular energy barriers of complex biomolecules restrict their efficient conformational sampling. A millisecond MD of even such simple biomolecules as GTP does not provide sufficient conformational sampling due to the high energy barrier for rotation around the glycosidic bond between the base and sugar part of GTP (Fig. 2). This energy barrier is even higher for C8- analogs of GTP (Fig. 3). Therefore the applicant has developed the novel scheme of H-REMD, where the main idea is to use soft-core interactions[24] between atoms of the base and sugar part of GTP analog in the following forms:[P5]

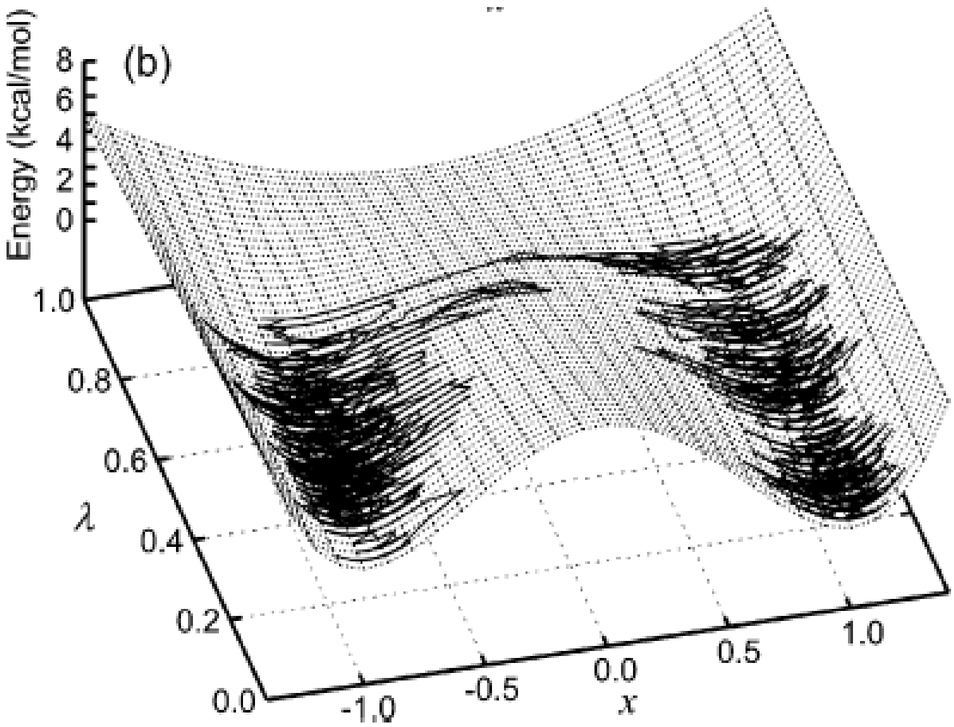


Figure 4: Schematic representation of the energy profile as a function of the softness of nonbonded interactions.

Replicas at higher levels of softness convert more easily from one stable conformation to the other, which are separated by a high energy barrier if no softness is applied.

$$E_{ij}^{vdw}(r_{ij},\lambda) = \left(\frac{C_{ij}^{(12)}}{A_{ij}(\lambda)+r_{ij}^{6}} - C_{ij}^{(6)}\right)\frac{1}{A_{ij}(\lambda)+r_{ij}^{6}}; A_{ij}(\lambda) = \frac{C_{ij}^{(12)}}{C_{ij}^{(6)}}\lambda^2 \quad (4.1)$$

$$E_{ij}^{el}(r_{ij},\lambda) = \frac{q_i q_j}{4\pi\varepsilon}\frac{1}{\sqrt{B_{ij}(\lambda)+r_{ij}^{2}}}; B_{ij}(\lambda) = \lambda^2 \quad (4.2)$$

The softness of Van der Walls and electrostatic interactions are controlled here by the λ parameter.[24] Because soft-core potentials are almost the same as regular ones at longer distances, most of the interactions between atoms of the perturbed parts are changed only slightly. Rather, the strong repulsion between atoms that are close in space, which results in high energy barriers in most cases, is weakened within higher replicas of our scheme. Their desired effect on the overall energy landscape between two stable conformations (e.g. anti and syn conformation of C8-analogs of GTP) can be seen in Fig. 4.

The advantage of this approach over other REMD schemes is the possibility to use a relatively small number of replicas (6 for GTP) with larger local differences between the individual Hamiltonians. A huge computational gain is also harnessed since it is possible to apply the soft-core interaction for only a selected region and therefore to allow for enhanced conformational sampling only in this region, with the rest of the system still moving on the timescale of the MD. The replica with unmodified Hamiltonian produces unbiased canonical ensembles of structures in a highly efficient way.[P5] The problem with H-REMD using soft-core interactions is its setup (how many replicas are needed, at what softness, etc.) Therefore, the optimization of REMD setup parameters was developed.[P6] This approach has general applicability for any REMD, not only H-REMD using soft-core interactions.

4.2 Umbrella sampling

The actual binding process of ligands with the accompanying free energy differences can be studied by methods such as umbrella sampling.[25,26] Such methods are computationally more demanding than molecular docking by several orders of magnitude as described in the previous scientific chapter. However, they are in principle capable of determining reliable binding affinities. Umbrella sampling attempts to overcome the sampling problem by modifying the original Hamiltonian (unbiased) - where $H_u(\vec{r}^N)$ by a bias term ($V_b(\vec{r}^N)$) resulting in the overall (biased) Hamiltonian $H_b(\vec{r}^N)$. A bias, an additional potential energy term, is applied to the system to ensure efficient sampling along the **reaction coordinate,** a path connecting two end-point states separated by an energy barrier, e.g. bound and unbound state of a ligand of our interest. MD simulation of the biased system provides the biased (unrealistic) distribution along the reaction coordinate. It can still be used for the calculation of the unbiased average of a property A by the following equation:

$$\langle A\rangle_u = \frac{\langle A(\vec{r}^N)e^{+\frac{V_b(\vec{r}^N)}{k_B T}}\rangle_b}{\langle e^{+\frac{V_b(\vec{r}^N)}{k_B T}}\rangle_b} \tag{4.3}$$

where angle brackets indicate the ensemble averaging. The effect of the bias potential to connect energetically separated regions in phase space gave rise to the name **umbrella sampling**.[26,27]

4.3 H-REMD using distance (field)restraints

A typical problem when applying the umbrella sampling method to biomolecular systems is that the acting bias very often “damages” the biomacromolecule. For instance, when we applied the distance restraints of substrates within the CYP 2D6 catalytic site to simulate the unbinding process we often observed the ligand hitting against the closed binding channel. This motivated us to apply the distance restraint within the H-REMD scheme. Idea was that the switches in the H-REMD protocol allow for the reversible binding and unbinding of the ligand, ensuring that all relevant pathways are sampled. However, this may lead to unfavorable situations as an unbound ligand may be directed towards the active site along a path that would lead through the protein, rather than along an entrance channel. To circumvent this problem, a distance-field was defined, using Dijkstra’s algorithm to determine the optimal path to the active site without going through the protein as indicated in Fig. 5.[19] The applicant successfully applied this approach to the sampling of binding pathways of phosphopeptides to the 14-3-3 protein, calculating the corresponding binding affinities that showed reasonable agreement with the available experimental data.[P7]

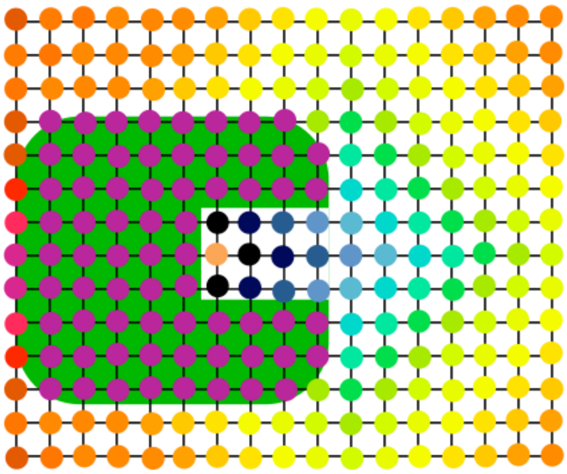


Figure 5: Schematic representation of distance-field restraints.

A protein is represented by a green shape, with the active site opening showing the strongest distance-field restraints. The figure is taken from[19]

Paper 5

Hritz J., Oostenbrink C. Hamiltonian replica exchange molecular dynamics using soft-core interactions. *J. Chem. Phys.* **2008,** 128, 144121

Hamiltonian replica exchange molecular dynamics using soft-core interactions

Jozef Hritz and Chris Oostenbrink[a)]
Leiden Amsterdam Center for Drug Research (LACDR), Division of Molecular Toxicology, Vrije Universiteit, Amsterdam NL-1081 HV, The Netherlands

(Received 19 December 2007; accepted 5 February 2008; published online 14 April 2008)

To overcome the problem of insufficient conformational sampling within biomolecular simulations, we have developed a novel Hamiltonian replica exchange molecular dynamics (H-REMD) scheme that uses soft-core interactions between those parts of the system that contribute most to high energy barriers. The advantage of this approach over other H-REMD schemes is the possibility to use a relatively small number of replicas with locally larger differences between the individual Hamiltonians. Because soft-core potentials are almost the same as regular ones at longer distances, most of the interactions between atoms of perturbed parts will only be slightly changed. Rather, the strong repulsion between atoms that are close in space, which in many cases results in high energy barriers, is weakened within higher replicas of our proposed scheme. In addition to the soft-core interactions, we proposed to include multiple replicas using the same Hamiltonian/level of softness. We have tested the new protocol on the GTP and 8-Br-GTP molecules, which are known to have high energy barriers between the *anti* and *syn* conformation of the base with respect to the sugar moiety. During two 25 ns MD simulations of both systems the transition from the more stable to the less stable (but still experimentally observed) conformation is not seen at all. Also temperature REMD over 50 replicas for 1 ns did not show any transition at room temperature. On the other hand, more than 20 of such transitions are observed in H-REMD using six replicas (at three different Hamiltonians) during 6.8 ns per replica for GTP and 12 replicas (at six different Hamiltonians) during 8.7 ns per replica for 8-Br-GTP. The large increase in sampling efficiency was obtained from an optimized H-REMD scheme involving soft-core potentials, with multiple simulations using the same level of softness. The optimization of the scheme was performed by fast mimicking [J. Hritz and C. Oostenbrink, J. Chem. Phys. **127**, 204104 (2007)]. © *2008 American Institute of Physics*. [DOI: 10.1063/1.2888998]

I. INTRODUCTION

The energy landscape for biomolecules in explicit solvent exhibits many local free energy minima.[1] While many minima are readily sampled in a molecular dynamics (MD) simulation, some are separated by high free energy barriers.[2] A molecule can be trapped in local energy minimum conformations for times comparable to or longer than typical simulation times that are reachable by conventional MD. In these cases, regular MD simulations will not lead to a complete conformational sampling of the studied system.[3]

One possible solution is the application of temperature replica exchange MD (T-REMD), in which several parallel simulations are performed at different temperatures. At regular intervals switches of the temperature between simulations are attempted, allowing high temperature simulations to cool down and low temperature simulations to heat up. A gain in efficiency should be obtained if the free energy barriers can be easily crossed in high temperature simulations.[4] T-REMD, however, suffers from the fact that the number of replicas needed to cover the necessary temperature range is proportional to the square root of the number of degrees of freedom of the system. This means that T-REMD simulations of biomolecules in explicit solvent can be computationally very demanding. The requirement of a large number of replicas in T-REMD can be overcome by applying Hamiltonian REMD (H-REMD),[5,6] where not the temperature but the Hamiltonian is varied over the replicas, through a perturbation of the original Hamiltonian. Note that T-REMD is a special case of H-REMD in which all the terms in the Hamiltonian are multiplied by the same scaling factor.

Hamiltonians can be perturbed in different ways. Useful approaches are those, in which the free energy landscape of the higher replicas allows for faster conformational conversions compared to the unperturbed Hamiltonian at the lowest replica.

We choose to perturb the Hamiltonian by describing selected interactions with a soft-core potential[7] and vary the level of softness over the replicas through a softness parameter λ. This approach allows for a diminishing of free energy barriers at higher replicas with higher levels of softness. The advantage of our approach is that we can specifically use soft-core interactions for those biomolecular parts that are of interest, e.g., parts contributing the most to the free energy barriers between different conformations. In this study we have also applied the novel concept of a degenerate highest level of softness $\lambda_{\max}$. This involves multiple (n) replicas at

[a)] Electronic mail: c.oostenbrink@few.vu.nl.

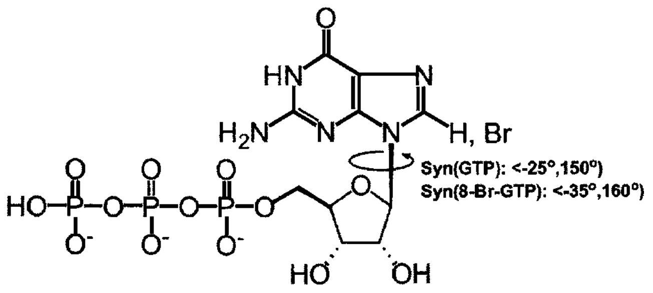


FIG. 1. Structure of GTP and 8-Br-GTP in *syn* conformation. Conformational transitions between the *syn* and *anti* states occur by rotation around the glycosidic bond (indicated). The glycosidic dihedral angle (ϕ) is defined over atoms: C4-N9-C1′-O4′.

$\lambda_{\max}$, allowing the system to spend more time at this λ-value so that alternative conformations can be reached in shorter overall simulation time. This is reminiscent of the J-walking method[8] or finite reservoir simulations,[9,10] where a reservoir of structures is pregenerated at the highest temperature or Hamiltonian ($T_{\max}/\lambda_{\max}$) and then "coupled" to REMD to get correct ensembles at lower values of T or λ. However, the concept of a degenerate $T_{\max}/\lambda_{\max}$ is an integral part of REMD requiring no precalculation. It allows for the most efficient balance between conformational transitions at $T_{\max}/\lambda_{\max}$ and replica exchanges toward T_0/λ_0. Refinement of the optimal n and the optimal selection of λ-values by fast REMD mimicking is described in our previous paper.[11]

We have tested the H-REMD scheme using soft-core interactions on two biologically relevant systems: GTP and 8-Br-GTP (Fig. 1). These systems can adopt two stable conformations by rotation around the glycosidic bond: *Anti* and *syn*. The boundaries of these conformations are not exactly the same for GTP and 8-Br-GTP. Based on the dihedral angle distributions shown in the Results section, we consider GTP to be in a *syn* conformation if the dihedral angle around the glycosidic bond is within the interval $\langle -25°, 150°\rangle$. For 8-Br-GTP we use the interval $\langle -35°, 160°\rangle$ to define the *syn* state. NMR studies indicate that the dominant conformation for GTP is *anti*, while it is *syn* for 8-Br-GTP.[12] Both GTP and 8-Br-GTP have high free energy barriers between the *syn* and *anti* conformations. However, H-REMD simulations using a few different λ-values (3 for GTP and 6 for 8-Br-GTP) suffice to enhance the conformational sampling enormously. The preference of *syn* and *anti* should follow quantitatively from the ensemble generated at the lowest λ-value corresponding to the unperturbed Hamiltonian of the REMD scheme. We compare the relative population of the two states with direct free energy calculations between them using hidden restraints along the glycosidic dihedral angle.

II. METHODS

The Methods section is divided into four parts. Firstly we discuss the implementation of soft-core interactions in REMD simulations. Secondly, we discuss the REMD simulations in more detail, and thirdly the use of hidden restraints to calculate the free energy difference between *syn* and *anti*. The section is concluded with a description of the exact simulation settings.

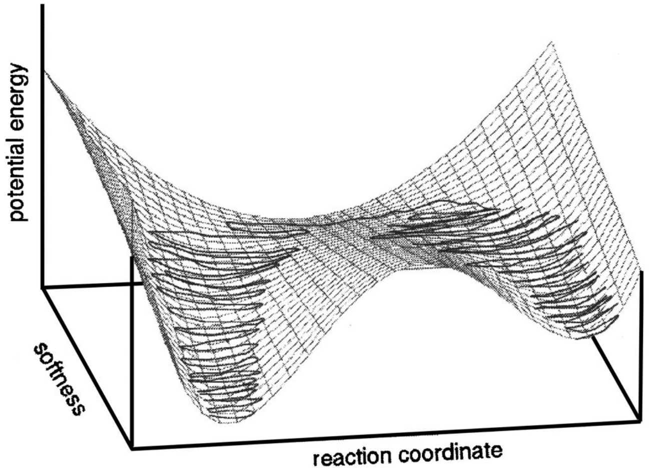


FIG. 2. Schematic representation of the energy profile as function of the softness of nonbonded interactions. Replicas at higher levels of softness convert more easily from one stable conformation to the other, which are separated by a high energy barrier if no softness is applied.

A. Implementation of soft-core interactions

We have used the following functional form for van der Waals and electrostatic soft-core interactions[7] between atoms i and j as a function of the interatomic distance r_{ij}

$$E_{ij}^{\mathrm{vdW}}(r_{ij},\lambda) = \left(\frac{C12_{ij}}{A_{ij}(\lambda)+r_{ij}^6} - C6_{ij}\right)\frac{1}{A_{ij}(\lambda)+r_{ij}^6}, \tag{1}$$

$$E_{ij}^{\mathrm{el}}(r_{ij},\lambda) = \frac{q_i q_j}{4\pi\varepsilon}\frac{1}{\sqrt{B_{ij}(\lambda)+r_{ij}^2}}, \tag{2}$$

with $A_{ij}(\lambda)=\alpha_{\mathrm{vdW}}(C12_{ij}/C6_{ij})\lambda^2$ and $B_{ij}(\lambda)=\alpha_{\mathrm{el}}\lambda^2$. $C12_{ij}$, $C6_{ij}$ are the Lennard–Jones parameters for atom pair i and j, q_i and q_j the partial charges of particles i and j, and α_{vdW} and α_{el} are the softness parameters. In the current study we used in all simulations $\alpha_{\mathrm{vdW}}=\alpha_{\mathrm{el}}=1$, and the softness of the interactions was controlled through the parameter λ. It can be seen that at longer distances $[r_{ij} \gg A(\lambda)$ and $r_{ij} \gg B(\lambda)]$ the soft-core interaction approximates the interaction for normal atoms and that they differ mostly at short distances between the atoms $[r_{ij} \leqslant A(\lambda)$ or $r_{ij} \leqslant B(\lambda)]$. Potential energy barriers are mostly the result of a short-ranged repulsion between atoms, which can strongly be reduced by increasing the levels of softness. An idealized potential energy landscape for two states separated by a high barrier when no softness is applied, but part of the same energy valley at higher levels of softness is sketched in Fig. 2.

In GROMOS96 (Ref. 13) as well as GROMOS05 (Ref. 14) one can use soft-core interactions in the context of a free energy perturbation in which the interaction energy E_{ij} is written as a linear combination of the potential energy for two different states A and B,

$$E_{ij}(r_{ij},\lambda) = [1-\lambda]^n E^A(r_{ij},\lambda) + \lambda^n E^B(r_{ij},1-\lambda). \tag{3}$$

In cases where the Lennard–Jones parameters and partial charges are identical in states A and B and we set $\lambda=0.5$ and $n=1$, this reduces to the functional forms of Eqs. (1) and (2). For other λ-values one gets a mixture of potentials at different levels of softness. In order to be able to control the level of softness by λ, which is convenient in a REMD setting, we extended the implementation of free energy perturbations in GROMOS05 to include the possibility to set individual λ-values for Lennard–Jones interaction (λ_{vdW}), the Lennard–Jones softness level $\lambda^{\text{vdW}}_{\text{soft}}$, Coulombic interactions (λ_{el}), and the Coulombic softness level ($\lambda^{\text{el}}_{\text{soft}}$) as a polynomial function (up to fourth order) of an overall or "global" λ-value. For the Lennard–Jones interaction we write the interaction energy as

$$E^{\text{vdW}}_{ij}(r_{ij},\lambda) = [1-\lambda_{\text{vdW}}(\lambda)]^n E^{\text{vdW},A}_{ij}(r_{ij},\lambda^{\text{vdW}}_{\text{soft}}(\lambda)) + [\lambda_{\text{vdW}}(\lambda)]^n E^{\text{vdW},B}_{ij}(r_{ij},1-\lambda^{\text{vdW}}_{\text{soft}}(\lambda)). \quad (4)$$

An analogous form can be written for the Coulombic interaction E^{el}_{ij}. For the REMD simulations of GTP and 8-Br-GTP reported here, we used individual λ-dependencies as follows:

$$\lambda_{\text{vdW}}(\lambda)=0, \quad \lambda^{\text{vdW}}_{\text{soft}}(\lambda)=\lambda, \quad (5)$$

$$\lambda_{\text{el}}(\lambda)=0, \quad \lambda^{\text{el}}_{\text{soft}}(\lambda)=\lambda. \quad (6)$$

With these settings the softness at individual replicas is easily controlled by the global lambda λ, while the actual interactions are only defined by the parameters defined for state A.

B. REMD

A REMD simulation involves M noninteracting copies (replicas) of MD of one system running in parallel at various conditions.[15] Let us mark the state of a REMD ensemble of simulations as $X=\{\ldots,x^i_m,\ldots,x^j_n,\ldots\}$, where the indices indicate that the ith replica is simulated at the mth condition and the jth replica is simulated at the nth condition. x represents the collected positions q and momenta p of all particles ($x\equiv(q,p)$). The weight factor for this state of the REMD ensemble is given by the product of Boltzmann factors for individual noninteracting replicas,

$$W_{\text{REMD}}(X)=\prod_{k=1}^{M} e^{-\beta_{m(k)}H_{m(k)}(q^k,p^k)}, \quad (7)$$

where $\beta_{m(k)}=1/k_B T_{m(k)}$ and $H_{m(k)}(q^k,p^k)$ is the Hamiltonian of the system given by sum of the kinetic energy $K(p)$ and the potential energy $E(q)$. k_B is the Boltzmann constant. The subscript $m(k)$ indicates the condition m at which replica k is currently simulated. In T-REMD, conditions m differ only by the temperature T, while in H-REMD, the condition represents different Hamiltonians, in this case described by the softness parameter λ. Conditions (λ or T) do not have to be different for all replicas. In a degenerate $\lambda_{\max}/T_{\max}$ scheme,[11] multiple replicas are simulated at the same $\lambda_{\max}/T_{\max}$ simultaneously, i.e., $m(i)=m(j)$.

At regular time intervals we attempt to exchange (switch) the conditions at which replicas i,j are simulated,

$$X=\{\ldots,x^i_m,\ldots,x^j_n,\ldots\}\rightarrow X'=\{\ldots,x^j_m,\ldots,x^i_n,\ldots\}. \quad (8)$$

In order to maintain the proper weight of the REMD ensemble as described in Eq. (7) a detailed balance condition must be imposed on the exchange probability,

$$w(X\rightarrow X')=\min[1,\exp(-\Delta)], \quad (9)$$

where

$$\Delta=\beta_m[H_m(q^j,p^j)-H_m(q^i,p^i)] - \beta_n[H_n(q^j,p^j)-H_n(q^i,p^i)]. \quad (10)$$

Sugita and Okamoto showed that the kinetic energy parts of the Hamiltonians cancel,[16] leading to

$$\Delta=\beta_m[E_{\lambda_m}(q^j)-E_{\lambda_m}(q^i)]-\beta_n[E_{\lambda_n}(q^j)-E_{\lambda_n}(q^i)]. \quad (11)$$

In the case of H-REMD, where all replicas are run at the same temperature, Δ reduces further to

$$\Delta=\beta[E_{\lambda_m}(q^j)-E_{\lambda_m}(q^i)-E_{\lambda_n}(q^j)+E_{\lambda_n}(q^i)]. \quad (12)$$

In our study the Hamiltonian applied for individual replicas is varied by introducing the soft-core interactions (1-2) controlled by parameter λ.

C. Thermodynamic integration using hidden dihedral angle restraints

The use of hidden dihedral angle restraints[17] offers an alternative approach to calculate the free energy difference between the *anti* and *syn* conformations of GTP/8-Br-GTP. A hidden dihedral angle restraint around the glycosidic bond was used to propagate the GTP/8-Br-GTP systems from one stable conformation to the other, thereby overcoming the energy barrier and sampling conformations on the transition path in an efficient way. The restraining energies for the dihedral angle ϕ(C4-N9-C1′-O4′) were calculated as a function of the coupling parameter λ according to Ref. 16,

$$V^{\text{dihres}}(\phi,\lambda)=4\lambda(1-\lambda)V^{\text{dihres},AB}_{\text{restr}}(\phi,\lambda), \quad (13)$$

where

$$V^{\text{dihres},AB}_{\text{restr}}(\phi,\lambda)=\begin{cases} V^{\text{dihres},AB}_{\text{harm}}(\phi,\lambda) & \text{if } |\Delta\phi_\lambda|\leq\phi^0_{\text{lin}} \\ V^{\text{dihres},AB}_{\text{lin}}(\phi,\lambda) & \text{if } |\Delta\phi_\lambda|>\phi^0_{\text{lin}}, \end{cases} \quad (14)$$

with

$$\Delta\phi_\lambda=\phi-(1-\lambda)\phi_{0,A}-\lambda\phi_{0,B}+2n\pi, \quad (15)$$

and

$$V^{\text{dihres},AB}_{\text{harm}}(\phi,\lambda)=1/2[(1-\lambda)K^A+\lambda K^B](\Delta\phi_\lambda)^2, \quad (16)$$

$$V^{\text{dihres},AB}_{\text{lin}}(\phi,\lambda)=[(1-\lambda)K^A+\lambda K^B](\zeta\Delta\phi_\lambda-1/2\phi^0_{\text{lin}})\phi^0_{\text{lin}}, \quad (17)$$

with $\zeta=-1$ if $\Delta\phi_\lambda<-\phi^0_{\text{lin}}$ and $\zeta=1$ if $\Delta\phi_\lambda>-\phi^0_{\text{lin}}$ for the linearized part of the restraint. We have used $K^A=K^B=100$ kJ rad^{-2} and $\phi^0_{\text{lin}}=30°$. For both GTP and 8-Br-GTP, four thermodynamic integration calculations were done: From −120° (*anti*) to 60° (*syn*) and back ($\phi^{0,A}=-120°$, $\phi^{0,B}=60°$ or $\phi^{0,A}=60°$, $\phi^{0,B}=-120°$), from 60° (*syn*) to 240° (*anti*) and back ($\phi^{0,A}_k=60°$, $\phi^{0,B}_k=240°$ or $\phi^{0,A}_k=240°$, $\phi^{0,B}_k=60°$).

As can be seen from Eq. (13) the restraint energy reduces to zero at the beginning ($\lambda=\lambda_A=0$) and end ($\lambda=\lambda_B=1$) states. Therefore an unperturbed relative free energy

TABLE I. Optimal REMD setting for GTP and 8-Br-GTP.

Molecule	$\lambda_{\max}$	n	$t^{\text{total}}_{\lambda_{\max}}$ (ps)	Optimal λ-set
GTP	0.45	4	100	[0.0,0.25,0.45,0.45,0.45,0.45]
8-Br-GTP	0.7	7	100	[0.0,0.2,0.4,0.55,0.65,0.7,0.7,0.7,0.7,0.7,0.7,0.7]

ΔG_{BA} of state B with respect to state A can be calculated using the thermodynamic integration method,[18]

$$\Delta G_{BA} = \int_{\lambda_A}^{\lambda_B} \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{\lambda} d\lambda, \qquad (18)$$

where H is the Hamiltonian of the system including the restraining potential energy term, $V^{\text{dihres}}(\phi,\lambda)$, and the angular brackets indicate an ensemble average obtained at λ. The integration was performed by calculating the ensemble average $\langle \partial H / \partial \lambda \rangle_{\lambda}$ at discrete λ points, corresponding to angular changes of 20° ($\Delta\lambda = 0.111$). In regions where the curvature of $\langle \partial H / \partial \lambda \rangle$ was high, the step size was halved. The initial sampling time for each λ point was 1 ns (after 100 ps relaxation which started from the equilibrated conformation of the simulation at the previous λ value). For λ points in which the ensemble averages did not converge to the same value in forward and reverse thermodynamic integration calculations, the sampling time was extended to 6 ns.

D. MD and REMD settings

All MD and REMD simulations were conducted using the GROMOS05 simulation package running on a linux cluster.[14] All bonds were constrained, using the SHAKE algorithm[19] with a relative geometric accuracy of 10^{-4}, allowing for a time step of 2 fs used in the leapfrog integration scheme.[20] Periodic boundary conditions, with a truncated octahedral box (average volume of 59.6 nm^3), were applied. After a steepest descent minimization to remove bad contacts between molecules, initial velocities were randomly assigned from a Maxwell–Boltzmann distribution at 298 K, according to the atomic masses. The temperature was kept constant using weak coupling to a bath of 298 K with a relaxation time of 0.1 ps.[21] The solute molecules (GTP or 8-Br-GTP) and solvent (i.e., 1926 explicit SPC water molecules[22] and three Na^+ counterions) were independently coupled to the heat bath. The pressure was controlled using isotropic weak coupling to atmospheric pressure with a relaxation time of 0.5 ps.[21] Van der Waals and electrostatic interactions were calculated using a triple range cutoff scheme. Interactions within a short-range cutoff of 0.8 nm were calculated every time step from a pair list that was generated every five steps. At these time points, interactions between 0.8 and 1.4 nm were also calculated and kept constant between updates. A reaction-field contribution was added to the electrostatic interactions and forces to account for a homogeneous medium outside the long-range cutoff, using the relative permittivity of SPC water (61).[23] All interaction energies were calculated according to the GROMOS force field, parameter set 53A6.[24] Force field parameters used for GTP and 8-Br-GTP are listed in the supplementary material.[25] In the H-REMD simulations, all sugar-base interactions in GTP and 8-Br-GTP are treated using the soft-core interactions through Eqs. (1), (2), and (4).

REMD is implemented in GROMOS05 such that only switches between "neighboring" replicas are attempted. Replica exchanges are attempted every 2.5 ps (elementary period, t^{elem}). However, not all replica pairs are considered at the same time. After an odd number of elementary periods, exchanges between the $(2i+1)$th and the $(2(i+1))$th replicas are attempted (we call them replica exchanges of type I) and after an even number of t^{elem} exchanges between the $(2i)$th and the $(2i+1)$th replicas are attempted (exchanges of type II). This means that the effective time between two switches of the same type is twice the elementary switching period ($2 \times 2.5\ \text{ps} = 5\ \text{ps}$). During the initial 40 t^{elem} (100 ps) no replica exchanges were attempted in order to equilibrate the systems at the individual replicas. The conformations were written out every 2.5 ps, just before the replica exchange attempt. In order to see how effective REMD is in sampling the less stable conformations, we started all REMD runs from the most stable conformation, i.e., in the case of GTP from *anti* and in the case of 8-Br-GTP from *syn* conformations.

A novel concept of degenerate $\lambda_{\max}$ simulations was applied as presented in our previous study.[11] In this approach more replicas (n) are simulated at the highest λ-value, $\lambda_{\max}$, simultaneously and the time between switching attempts between these replicas is set such that every replica at $\lambda_{\max}$ spends a time $t^{\text{total}}_{\lambda_{\max}}$, during which the system can convert from one conformation to the other. Usually $t^{\text{total}}_{\lambda_{\max}} \gg t^{\text{elem}}$ because the conversion probability rather shows a sigmoidal than a linear time dependence, and the middle of the sigmoidal curve corresponds to a significantly longer time than t^{elem}. The values of $\lambda_{\max}$ and n as well as the number of λ-values between 0 and $\lambda_{\max}$ and their exact values were refined by the fast mimicking approach in Ref. 11. This approach maximizes the number of global conformational transitions (N_{gct}) per CPU. We defined the number of global conformational transitions as the number of times the conformational state changes for each particular replica when monitored at the lowest λ-value (λ_0). At which particular λ-value the conformational transition occurs is irrelevant for this measure. For more details, see Ref. 11. The λ-values at which REMD simulations were performed; $t^{\text{total}}_{\lambda_{\max}}$ and n are summarized in Table I.

III. RESULTS

A. Standard MD simulations

Theoretically, one can obtain the populations of *syn* and *anti* conformations for GTP/8-Br-GTP from sufficiently long MD simulations, during which many *anti*↔*syn* transitions occur. However, 25 ns of MD simulation of GTP and 8-Br-

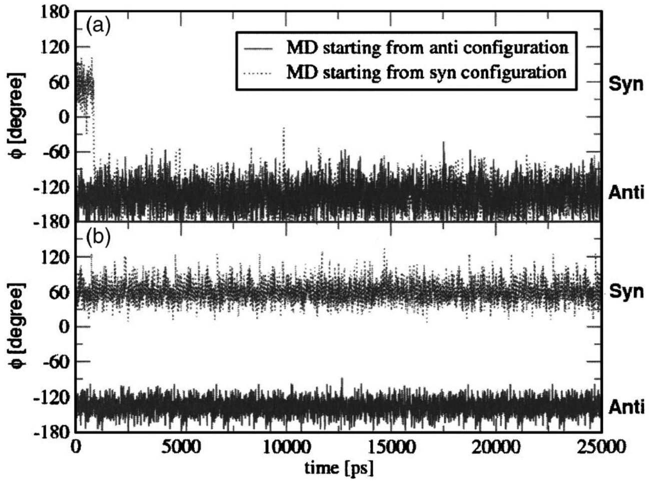


FIG. 3. Time dependence of dihedral angle (ϕ), around the glycosidic bond during two MD simulations (starting from *anti* and *syn* configurations) for (a) GTP and (b) 8-Br-GTP. *Anti* and *syn* conformational regions are indicated.

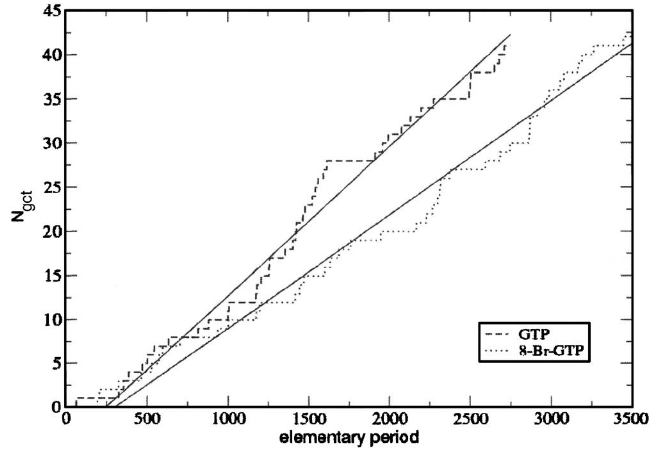


FIG. 4. Time dependence of the number of global transitions (N_{gct}) for REMD of GTP and 8-Br-GTP. Relaxation time calculated from the extrapolation of the linear part is for GTP 249 t^{elem} (elementary period, t^{elem} $=2.5$ ps) and for 8-Br-GTP 303 t^{elem}.

GTP shows that the *syn* and *anti* conformational states are separated by a quite high free energy barrier. For GTP we observe only one transition from the *syn* (less dominant) to the *anti* (more dominant) conformation and no transitions from *anti* to *syn* [Fig. 3(a)]. No transitions were observed at all for 8-Br-GTP, neither from *syn* to *anti* nor from *anti* to *syn* [Fig. 3(b)], indicating an even higher energy barrier between the *anti* and *syn* conformations for this molecule. This means that in order to obtain converged values of the *anti* and *syn* populations, we would have to run standard MD for an impractically long time. In such cases it may be very useful to use replica exchange MD, which can sample conformational space much more efficiently than standard MD.

B. REMD of GTP and 8-Br-GTP

All replicas within REMD were started from the same conformation (*anti* for GTP and *syn* for 8-Br-GTP). Therefore it takes some time to relax the complete set of REMD simulations. It is not straightforward to determine the relaxation time and time length of REMD simulation required to obtain sufficient statistics of measured quantities *a priori*. The REMD settings used here were chosen to maximize the number of global conformational transitions (N_{gct}) per CPU by fast mimicking.[11] Monitoring of N_{gct} is also useful for real REMD because when all replicas are in equilibrium, N_{gct} should show a linear time dependence. This allows us to determine the relaxation time of REMD by extrapolating the linear part of $N_{\text{gct}}(t)$ to 0 transition, as well as to estimate the length of REMD, which is needed to obtain a specific value of N_{gct}.

The time evolution of N_{gct} for GTP and 8-Br-GTP is shown in Fig. 4 from which the relaxation times are estimated to be 249 t^{elem} (622.5 ps) for GTP and 303 t^{elem} (757.5 ps) for 8-Br-GTP. This figure also shows that in order to reach at least 40 global conformational transitions, we perform REMD of GTP for 2750 t^{elem} (6875 ps) and REMD of 8-Br-GTP for 3500t^{elem} (8750 ps). Only the parts of the simulations after the relaxation times were used for quantitative analysis of these simulations.

The dihedral angle distributions of the glycosidic bond of GTP and 8-Br-GTP at different λ-values are shown in Figs. 5 and 6. The populations of *syn* and *anti* conformations obtained by integration of the distributions between bounds indicated in these figures are summarized in Tables II (GTP) and III (8-Br-GTP) together with the corresponding free energy difference between the *syn* and *anti* states as calculated by

$$\Delta G_{\text{syn-anti}} = -k_B T \ln\frac{[syn]}{[anti]}, \tag{19}$$

where k_B is again the Boltzmann constant and $T=298$ K. Usually we are most interested for values for unperturbed interactions ($\lambda=0.0$) which are for GTP,

$$[anti] = 95.6\,\% \;\pm 0.5\,\%\,,$$

$$[syn] = 4.4\,\% \;\pm 0.5\,\%\,,$$

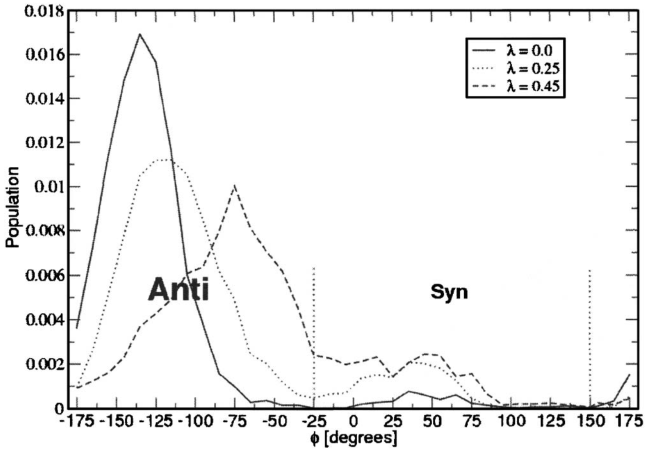


FIG. 5. Distribution of dihedral angle around the glycosidic bond of GTP at different levels of softness (λ). Populations were calculated using a bin width of 10°. *Anti* and *syn* conformational regions are indicated.

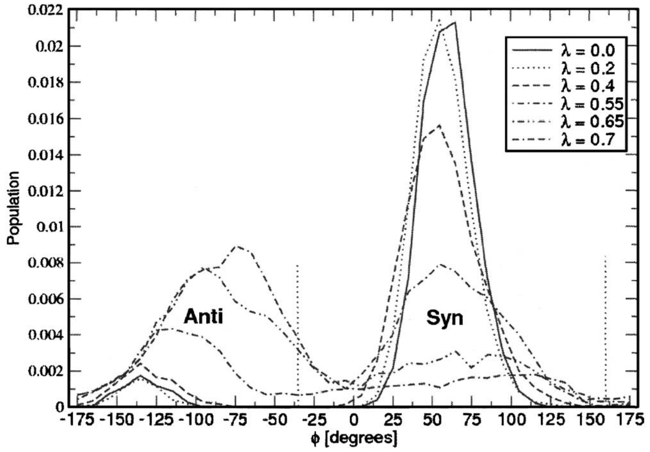


FIG. 6. Distribution of the dihedral angle around the glycosidic bond of 8-Br-GTP at different levels of softness (λ). Populations were calculated using a bin width of 10°. *Anti* and *syn* conformational regions are indicated.

$$\Delta G^{\mathrm{GTP}}_{\text{syn-anti}} = 7.6 \pm 0.3 \text{ kJ mol}^{-1},$$

and for 8-Br-GTP,

$$[anti] = 6.0\,\% \ \pm 1.8\,\% \,,$$

$$[syn] = 94.0\,\% \ \pm 1.8\,\% \,,$$

$$\Delta G^{\text{8-Br-GTP}}_{\text{syn-anti}} = -\,6.8 \pm 0.9 \text{ kJ mol}^{-1}.$$

Error estimates on the populations are calculated from block averages of occurrences of the conformational states followed by an extrapolation to infinite block length.[26]

C. Thermodynamic integration using a hidden dihedral angle restraint

The free energy difference between the *syn* and *anti* conformations can also be calculated by thermodynamic integration using a (hidden) dihedral angle restraint around the glycosidic bond. Theoretically, the obtained value for the free energy difference should not depend on the pathway by which the system is pulled from one stable conformation to another. However, in practice, this class of methods very often shows hysteresis resulting from inadequate relaxation of the systems at individual λ-values. In order to obtain reliable values, the free energy difference for both GTP and 8-Br-GTP was calculated through two different pathways $[-120°(anti) \leftrightarrow 60°(syn), 60°(syn) \leftrightarrow 240°(anti)]$ in both the forward and reverse directions (Figs. 7 and 8). This allows us to determine the convergence of the free energy difference and to monitor if sampling is sufficient for each point at the individual pathways. We note that similar values of Δ*G* calculated from two opposite directions do not automatically mean a complete convergence. Figure 7(b) shows, for example, that although the calculated values of Δ*G* are very similar (−9.2 ± 1.2 and −9.5 ± 1.3 kJ/mol) there are still four λ-values (0.333; 0.5; 0.556; 0.667) at which the values of $\langle \partial H/\partial \lambda \rangle_\lambda$ differ more than the statistical error estimate. The conformational sampling for these points was not converged even after 6 ns, and the small hysteresis is obtained as the result of a fortuitous cancellation of errors. Note that two simulations that are sampling conformations belonging to different narrow minima can easily result in averages that do not converge to the same value, while both show small statistical error estimates. Indeed, we observed for λ-values right before and after the very local and steep conformational barriers that the harmonic dihedral angle restraint results in a dihedral angle distribution with two distinct maxima, sometimes with insufficient transitions between them to obtain converged results. Moreover, the true maximum of the barrier is in these cases still not sampled. From the four transition pathways shown in Figs. 7 and 8, only the $-120°(anti) \leftrightarrow 60°(syn)$ pathway for GTP shows a small

TABLE II. Relative *syn* and *anti* conformational populations of GTP at different softness levels (λ). The corresponding free energy differences are also calculated.

λ	[*anti*]	[*syn*]	$\Delta G_{\text{syn-anti}}(\text{kJ mol}^{-1})$
0.0	0.954	0.044	7.62
0.25	0.860	0.140	4.49
0.45	0.768	0.232	2.96

TABLE III. Relative *syn* and *anti* conformational populations of 8-Br-GTP at different softness levels (λ). The corresponding free energy differences are also calculated.

λ	[*anti*]	[*syn*]	$\Delta G_{\text{syn-anti}}(\text{kJ mol}^{-1})$
0.0	0.060	0.940	−6.81
0.2	0.051	0.949	−7.24
0.4	0.187	0.813	−3.64
0.55	0.320	0.680	−1.67
0.65	0.619	0.381	1.20
0.7	0.722	0.278	2.36

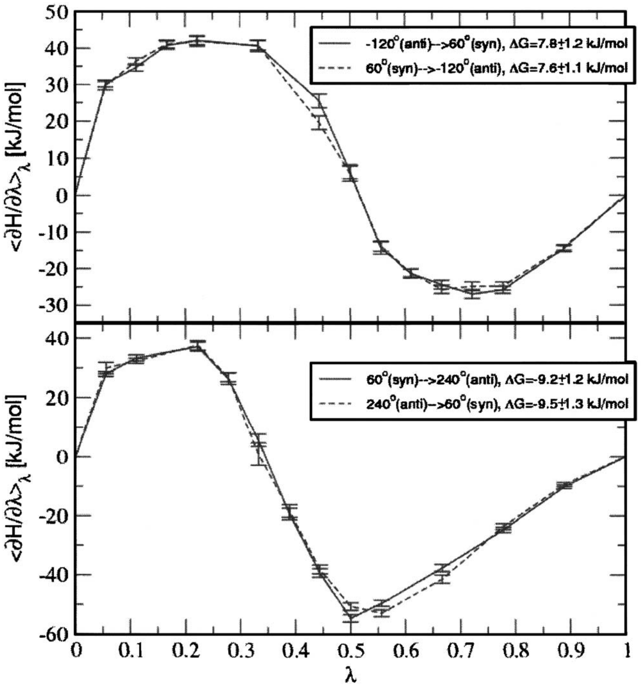


FIG. 7. Thermodynamic integration using hidden dihedral angle restraints for GTP. Error estimates for individual points result from block averaging and extrapolating the block length to infinity (Ref. 26).

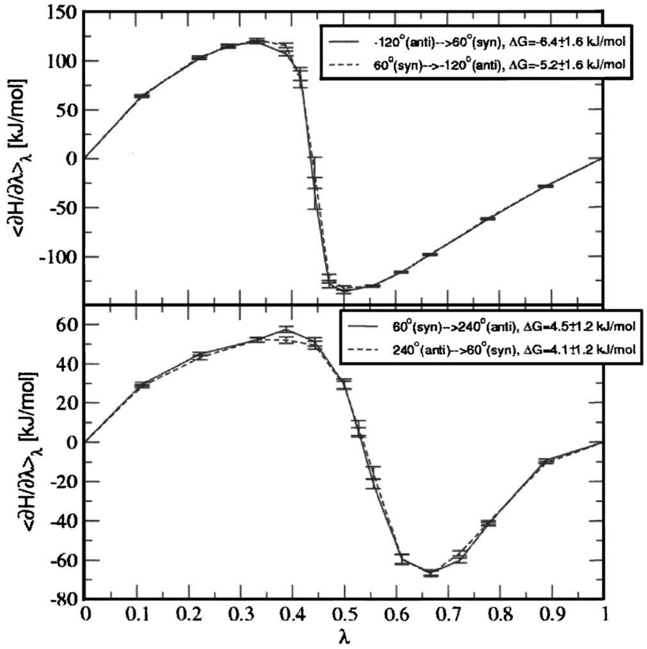


FIG. 8. Thermodynamic integration using hidden dihedral angle restraints for 8-Br-GTP. Error estimates for individual points result from block averaging and extrapolating the block length to infinity (Ref. 26).

hysteresis and the best convergence of $\langle \partial H/\partial\lambda\rangle_\lambda$ values in all λ-values. The steepness and curvature along this pathway also seems smallest. It leads to $\Delta G^{\text{GTP}}_{\text{syn-anti}}=7.7\pm 1.2$ kJ mol^{-1} which is very similar to the value obtained from the REMD simulations of GTP ($\Delta G^{\text{GTP}}_{\text{syn-anti}}=7.6\pm 0.3$ kJ mol^{-1}). Despite of the lack of full convergence in the other pathways, the calculated free energy differences for GTP ($\Delta G^{\text{GTP}}_{\text{syn-anti}}$ between 9.2 and 9.5 ± 1.3 kJ/mol) and for 8-Br-GTP ($\Delta G^{\text{8-Br-GTP}}_{\text{syn-anti}}$ between -6.4 and -4.1 ± 1.6 kJ/mol) are still within k_BT (2.5 kJ mol^{-1}) from the values obtained from REMD simulations ($\Delta G^{\text{GTP}}_{\text{syn-anti}}=7.6\pm 0.3$ kJ mol^{-1}, $\Delta G^{\text{8-Br-GTP}}_{\text{syn-anti}}=-6.8\pm 0.9$ kJ mol^{-1}).

IV. DISCUSSION

Any experimental observable that depends on the molecular structures (e.g., ^{3}J-coupling values) does not result from one single conformation but from an average over many (all) conformations that are present and over the experimental collection time. In order to compare such experimental values with values obtained from calculation, it is essential to generate the proper ensemble of conformations corresponding to the same boundary conditions (temperature, pressure) as at which the experiment was performed. According to the ergodic theory, such an ensemble of conformations can be generated by sufficiently long MD simulations. However, in many biological systems the energy barriers separating different conformations are so high that MD of even hundreds of nanoseconds does not produce converged conformational ensembles.

Examples of such biological systems are GTP and its eight-substituted analog 8-Br-GTP which both have a high energy barrier between their *syn* and *anti* conformations. This barrier results from nonbonded interactions only. The dihedral angle term around the glycosidic bond for these molecules is zero in our models (see the force field parameters in the supplementary material).[25] While *anti* is the more dominant conformation for GTP, *syn* is the more dominant conformation for 8-Br-GTP. Experiments indicate, however, that the less dominant conformations are still present for non-negligible amounts of time (estimates range between 5% and 30%).[12,27] The presented study shows for both systems that two 25 ns MD simulations (one started from *anti* and one from *syn*) do not show any transition from the more to the less stable conformation.

One possible way to sample such rare events in a more efficient way is to force the system to cross the conformational barrier. The application of hidden restraints[17] belongs to this class of methods and was shown to converge faster to the free energy difference between unrestrained, stable, conformations than, e.g., umbrella sampling.[28] We applied a hidden dihedral angle restraint around the glycosidic bond to force transitions between the *anti* and *syn* conformations. This approach does not directly yield the proper ensemble of conformations, but such an ensemble can be approximated by generating separate ensembles around two conformational minima (*anti*, *syn*) and subsequently weight these ensembles according to the free energy difference between them. Despite its efficiency this method still required a relatively large number of simulations (14–16) and many of these simulations needed to be sampled for 6 ns in order to obtain a reasonable convergence. The total simulation time for GTP was 78 ns per pathway and for 8-Br-GTP 84 ns per pathway. Note that in this approach, the actual end-points of unrestrained *anti* or *syn* conformations are not simulated, but the free energy derivative is approximated as zero. These would still need to be simulated in order to approximate the conformational ensemble as indicated above. In reality, these states suffer from an end-state problem in which the derivative fluctuates with unrestrained motion. Extrapolations towards $\lambda=0$ and $\lambda=1$ show that zero is a better approximation. As is explained in Sec. III C, the full convergence for individual $\langle \partial H/\partial\lambda\rangle_\lambda$ values was obtained only for GTP along the $-120°$ (*anti*) ↔ $60°$ (*syn*) pathway and the calculated value of $\Delta G^{\text{GTP}}_{\text{syn-anti}}=7.7\pm 1.2$ kJ mol^{-1} is in excellent agreement with $\Delta G^{\text{GTP}}_{\text{syn-anti}}=7.6\pm 0.3$ kJ mol^{-1} obtained from REMD of GTP. The $\langle \partial H/\partial\lambda\rangle_\lambda$ values are reasonably (but not fully) converged for the other three pathways shown in Figs. 7 and 8, and the calculated $\Delta G^{\text{GTP}}_{\text{syn-anti}}$ values are still within an acceptable range of 2.5 kJ mol^{-1} corresponding to thermal fluctuations when compared to values calculated from REMD.

The observed problems with the convergence of these simulations show that approaches which force the system to cross steep and high energy barriers are not necessarily computationally very efficient. Another important drawback of these methods is the requirement of a suitable reaction coordinate along which the system is forced to cross the energy barrier. This means that they can only be practically applied for simpler transition pathways.

REMD does not require prior knowledge about the transition path and can therefore be applied to more complex systems as well. In addition, REMD directly produces the proper ensemble of conformations. However, its efficiency can be much lower compared to methods forcing the molecule to cross a conformational barrier. As is explained in the Introduction, especially T-REMD becomes very inefficient for systems containing explicit solvent.[29] H-REMD involving a judiciously chosen perturbation of the Hamiltonian is less dependent on the overall system size and thus more generally applicable. On the other hand, it may not always be trivial to determine which parts of the Hamiltonian should be perturbed.

In the presented H-REMD scheme the Hamiltonian is modified over replicas by applying soft-core potentials for selected interactions. The efficiency is increased compared to other H-REMD approaches because soft-core potentials are very similar to unperturbed ones at longer distances. This means that only interactions between atoms that are close in distance (contributing most to the high energy barriers) contribute to the energy differences [Δ in Eq. (12)]. The switching probabilities are thus mainly governed by the interactions that really matter for enhancing the transitions over energy barriers. This is one of the main differences from other H-REMD schemes[5,6] or, e.g., the replica exchange with solute tempering[30] in which parts of the Hamiltonians are generally scaled rather than modified in their functional form (e.g., softened). For GTP and 8-Br-GTP only three (GTP) and six (8-Br-GTP) different levels of softness were required, thereby significantly reducing the time needed for the replicas to diffuse from the highest level of softness to the unperturbed Hamiltonian. In our previous study we showed that the overall efficiency can still be increased by simulating multiple replicas at the highest level of softness.[11] This leads to a total of six replicas for GTP and 12 replicas for 8-Br-GTP (Table I). The REMD simulations were performed such that after an initial equilibration period at least 40 global conformational transitions are observed. Both the equilibration time as well as the overall simulation time were obtained by monitoring the linear increase of the number of global transitions (N_{gct}). The total lengths of REMD simulation can then be calculated as $6\times 6.8=40.8$ ns for GTP and $12\times 8.7=104.4$ ns of 8-Br-GTP. When comparing these numbers with other methods (normal MD or thermodynamic integration with hidden restraints), one should keep in mind that REMD yields accurate free energy differences *and* the ensemble of relevant conformations for these molecules, which can be used for subsequent analysis.

In order to compare the efficiency of the presented H-REMD with standard T-REMD, we performed T-REMD for GTP using $T_0=298$ K and $T_{\max}=540$ K at which *anti*↔*syn* conversions are regularly observed. 50 replicas, separated by 4 K (in the range of 298–370 K), 5 K (370–450 K), and 6 K (450–540 K), were needed to obtain an average switching probability of 22.5%, close to the optimal of 20%.[31] This high number of replicas already implicates a low efficiency of T-REMD for GTP in explicit water. Indeed, T-REMD for 1 ns per replica, i.e., a total length of 50 ns (as compared to 40.8 ns used in H-REMD) did not reveal any global conformational transitions at all despite the occurrence of transitions at higher temperatures. The efficiency of T-REMD for 8-Br-GTP in explicit water can be expected to be even worse because MD up to 1500 K does not show regular *anti*↔*syn* conversions.

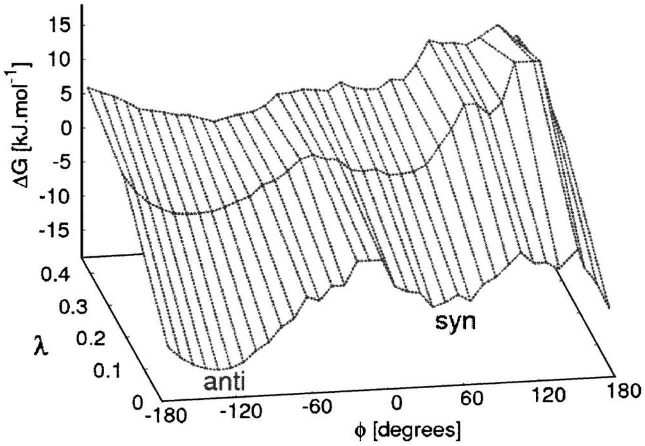


FIG. 9. Potential of mean force along the glycosidic bond of GTP as obtained from the populations in Fig. 5. Curves at individual λ-values were shifted in order to fulfill the perturbation formula (21) between different levels of softness for *anti* and *syn* conformations.

The dihedral angle distributions $P_\lambda(\varphi)$ in Figs. 5 and 6 allow us to generate the potential of mean force for GTP (Fig. 9) and for 8-Br-GTP (Fig. 10) using

$$G_\lambda(\varphi) = -k_BT\ln(P_\lambda(\varphi)) + C_\lambda, \tag{20}$$

where C_λ is a constant which is determined based on five φ points within the *anti* and *syn* regions such that the difference between $G_{\lambda j}(\varphi) - G_{\lambda i}(\varphi)$ is the same as calculated by the free energy perturbation formula,

$$\Delta G_{ij}(\varphi) = -k_BT\ln\langle e^{-(H_j-H_i)/k_BT}\rangle_i. \tag{21}$$

Figures 9 and 10 form the free energy landscapes on which the H-REMD simulations are performed and can be compared to the schematic representation in Fig. 2. Because of the nonmonotonous development of the free energy in the direction of λ for 8-Br-GTP, a three dimensional representation was very unclear. Rather, the individual curves are presented. The constant C_λ cancels exactly in the switching probability [Eqs. (9) and (12)] and is thus irrelevant for the REMD simulation. Figures 9 and 10 show that the soft-core

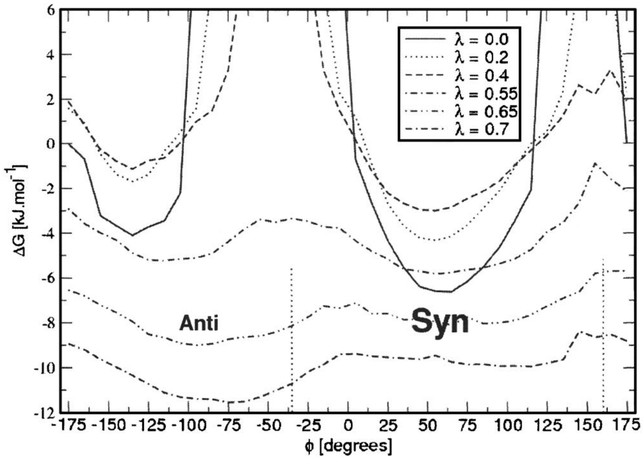


FIG. 10. Potential of mean force along the glycosidic bond of 8-Br-GTP as obtained from the populations in Fig. 6. Curves at individual λ-values were shifted in order to fulfill the perturbation formula (21) between different levels of softness for *anti* and *syn* conformations.

interactions at $\lambda_{\max}$ allow for a sufficient number of conformational transitions because (i) the conformational barrier is significantly decreased and (ii) the free energy difference between the stable conformations is reduced. This results in a higher population of the less dominant conformation and thus in a further increased occurrence of the conformational transitions. The free energy profiles also clearly show that the transitions between *anti* and *syn* conformations occur almost exclusively through the $-120°(\mathit{anti}) \leftrightarrow 60°(\mathit{syn})$ pathway because the barrier for the $60°(\mathit{syn}) \leftrightarrow 240°(\mathit{syn})$ pathway is significantly higher. Note that this is also the pathway in the thermodynamic integration simulations with hidden restraints that corresponds best to the REMD simulations. It is also interesting to observe how the interatomic interactions "lose" their local identity with increasing levels of softness. The nonsoft GTP and 8-Br-GTP have an inverse preference for the *syn* conformation (4.4% for GTP and 94% for 8-Br-GTP) but as can be seen from Tables II and III the *syn* populations become more similar with rising λ-values to a final 23.2% for GTP at $\lambda=0.45$ and 27.8% for 8-Br-GTP at $\lambda=0.7$. Furthermore, one may note that the position of the free energy minima is shifted at higher λ-values (most obvious for 8-Br-GTP at $\lambda=0.65$ and 0.7). This, of course, raises the question if the same dihedral angle intervals to define the *anti* and *syn* regions should be used for all λ-values and if the populations at these λ-values in Table III should not be revised. However, as these simulations involve unphysically perturbed molecules, the populations at these values have no direct meaning.

Even though GTP and 8-Br-GTP were used in this study mostly as model systems to show the application of our REMD simulations, there clearly is an interest in the conformational population of these molecules. Many six-or eight-substituted GTP and ATP analogs are being synthesized to act as specific inhibitors.[32–34] Bookser *et al.* correlated the binding affinities of such compounds to their *anti*/*syn* preference in water as obtained by ^{1}H NMR. Most adenosine analogs prefer the *syn* conformations, but the compounds with the highest adenosine kinase inhibitor potency all prefer the *anti* conformation. The methods described here may directly be applied to calculate the conformational preference prior to compound synthesis.

In the future, we think the H-REMD approach using soft-core interactions can be an asset to study (biomolecular) systems for which enhanced conformational sampling of a smaller part is required. Examples are loop regions of proteins (work in progress) or orientational sampling of small molecules bound to proteins (see, e.g., also our previous work involving binding free energy calculations using soft-core interactions[35,36]). The higher softness levels in these applications allow parts of the system to partially overlap and destabilize hydrogen bonds. Indeed, soft-core interactions allow us to see transitions which are not observed at high temperatures ($T=1000$ K) (unpublished results). This allows us to produce a wide variety of conformations of these parts of the system which are treated by soft-core interactions. Our approach may seem similar to two variants of T-REMD for local structure refinement: Partial and local replica exchange molecular dynamics.[37] These methods increase the temperature only in parts of the system. This will undoubtedly lead to a heat flow to the parts of the system that are still coupled to lower temperatures, but the assumption is made that this will not significantly modify the correct canonical ensemble. This is not guaranteed at all and can be different from system to system. Our approach does not induce a heat flow through the unperturbed system and does not require any such assumption. Moreover, it allows for even more flexibility by making it possible to select individual interactions (not only individual atoms or atom groups) which are treated by soft-core interactions, and each of these interaction can subsequently still respond to the same λ-value by specifying different values for α_{vdW} and α_{el} [Eqs. (1) and (2)].

As can be seen from the dihedral angle distributions in Figs. 5 and 6, the produced isothermic-isobaric ensemble does not only contain conformations from stable regions but also a few conformations from the energy barrier region. An approach involving only free energy estimates would not sample these conformations. We have calculated the populations of *syn* and *anti* conformations for GTP ($[\mathit{anti}]=0.956$, $[\mathit{syn}]=0.044$) and 8-Br-GTP ($[\mathit{anti}])=0.060$, $[\mathit{syn}]=0.940$). These values compare qualitatively with NMR experiments for guanosine and 8-Br-guanosine in dimethylsulfoxide (DMSO).[12,27] We stress that these experiments involve a different molecule in a different solvent and that several assumptions underlay the estimates of populations from the original NMR data. However, from the experiments it is clear that 8-Br-GTP has an inverse population preference with respect to GTP. Calculations to compare to original NMR data are underway. For now, we would like to point out that the quality of the obtained populations does not only depend on the convergence of values due to the enhanced sampling (main aim of our study) but also on the accuracy of the force field parameters used for GTP and 8-Br-GTP. More relevant is therefore the quantitative comparison (discussed above) of the calculated free energy differences by the two described methods as these both used the same force field parameters.

V. CONCLUSIONS

We present a novel Hamiltonian REMD scheme using soft-core potentials, which allows for effective REMD simulations with only a few replicas. The application of the method was demonstrated for two realistic biomolecules, GTP and 8-Br-GTP, in explicit water. By using soft-core interactions we only perturb those parts of the Hamiltonian that contribute most to a local free energy barrier. This results in efficient H-REMD simulations using only a few different levels of softness (three for GTP and six for 8-Br-GTP) and thus to a fast diffusion of the replicas between the lowest and the highest levels of softness. In order to give the systems time to undergo conformational transitions at the highest level of softness, we applied a degenerate highest softness level scheme. The optimal settings of the H-REMD schemes were obtained from an optimization procedure as described in our previous study.[11] The optimization procedure utilizes the number of global conformational transitions. The time dependence of this quantity was used to determine the relaxation and production times of H-REMD simulations. 40 global conformational transitions (20 from *anti* to *syn* and 20 from *syn* to *anti*) were observed within 6.8 ns of H-REMD simulation for GTP and 8.7 ns for 8-Br-GTP. No single transition from the more to the less dominant conformation was observed for either GTP or 8-Br-GTP within two 25 ns normal MD simulations (starting from initial *anti* and *syn* conformations) nor from T-REMD simulations of comparable length for GTP in explicit solvent. The obtained free energy differences between *anti* and *syn* conformations are in quantitative agreement with values calculated using thermodynamic integration with hidden dihedral angle restraints around the glycosidic bond and also in qualitative agreement with NMR estimates. The presented H-REMD approach using soft-core interactions allows for a softening of very specific, selected interactions which makes it a powerful technique to enhance the conformational sampling of smaller molecules in explicit solvent or of flexible parts of large biomacromolecules in explicit solvent.

ACKNOWLEDGMENTS

This work was supported by the Netherlands Genomics Initiative in the context of Horizon Breakthrough Grant No. 050-71-043 (J.H.) and by the Netherlands Organization for Scientific Research, VENI Grant No. 700.55.401 (C.O.).

[1] C. L. Brooks III, M. Karplus, and B. M. Pettitt, *Advances in Chemical Physics* (Wiley, New York, 1988), Vol. 71.
[2] M. Leitgeb, C. Schroder, and S. Boresch, J. Chem. Phys. **122**, 084109 (2005).
[3] J. P. Valleau and S. G. Whittington, in *Statistical Mechanics*, edited by B. J. Berne (Plenum, New York, 1977), Chap. 4, p. 145.
[4] U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).
[5] H. Fukunishi, O. Watanabe, and S. Takada, J. Chem. Phys. **116**, 9058 (2002).
[6] Y. Sugita, A. Kitao, and Y. Okamoto, J. Chem. Phys. **113**, 6042 (2000).
[7] T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber, and W. F. van Gunsteren, Chem. Phys. Lett. **222**, 529 (1994).
[8] D. D. Frantz, D. L. Freeman, and J. D. Doll, J. Chem. Phys. **93**, 2769 (1990).
[9] H. Li, G. Li, B. A. Berg, and W. Yang, J. Chem. Phys. **125**, 144902 (2006).
[10] A. Okur, D. R. Roe, C. Guanglei, V. Hornak, and C. Simmerling, J. Chem. Theory Comput. **3**, 557 (2007).
[11] J. Hritz and C. Oostenbrink, J. Chem. Phys. **127**, 204104 (2007).
[12] R. Stolarski, C. E. Hagberg, and D. Shugar, Eur. J. Biochem. **138**, 187 (1984).
[13] W. F. van Gunsteren, S. R. Billeter, A. A. Eising, P. H. Hünenberger, P. Krüger, A. E. Mark, W. R. P. Scott, and I. G. Tironi, *Biomolecular Simulation: The GROMOS96 Manual and User Guide* (Vdf Hochschulverlag AG an der ETH Zürich, Zürich, 1996).
[14] M. Christen, P. H. Hunenberger, D. Bakowies, R. Baron, R. Burgi, D. P. Geerke, T. N. Heinz, M. A. Kastenholz, V. Krautler, C. Oostenbrink, C. Peter, D. Trzesniak, and W. F. Van Gunsteren, J. Comput. Chem. **26**, 1719 (2005).
[15] R. H. Swendsen and J. S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
[16] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
[17] M. Christen, A.-P. E. Kunz, and W. F. Van Gunsteren, J. Phys. Chem. B **110**, 8488 (2006).
[18] J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).
[19] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977).
[20] R. W. Hockney, Methods Comput. Phys. **9**, 136 (1970).
[21] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).
[22] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans, in *Intermolecular Forces*, edited by B. Pullman (Reidel, Dordrecht, 1981), p. 331.
[23] I. G. Tironi, R. Sperb, P. E. Smith, and W. F. van Gunsteren, J. Chem. Phys. **102**, 5451 (1995).
[24] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. van Gunsteren, J. Comput. Chem. **25**, 1656 (2004).
[25] See EPAPS Document No. E-JCPSA6-128-509811 for force-field parameters of GTP and 8-Br-GTP. For more information on EPAPS, see http://www.aip.org/pubservs/epaps.html.
[26] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon, Oxford, 1987).
[27] H. Rosemeyer, G. Toth, B. Golankiewicz, Z. Kazimierczuk, W. Bourgeois, U. Kretschmer, H. P. Muth, and F. Seela, J. Org. Chem. **55**, 5784 (1990).
[28] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[29] J. W. Pitera and W. C. Swope, Proc. Natl. Acad. Sci. U.S.A. **100**, 7587 (2003).
[30] P. Liu, B. Kim, R. A. Friesner, and B. J. Berne, Proc. Natl. Acad. Sci. U.S.A. **102**, 13749 (2005).
[31] A. Kone and D. A. Kofke, J. Chem. Phys. **122**, 206101 (2005).
[32] T. Läppchen, A. F. Hartog, V. A. Pinas, G. J. Koomen, and T. den Blaauwen, Biochemistry **44**, 7879 (2005).
[33] F. Schwede, A. Christensen, S. Liaw, T. Hippe, R. Kopperud, B. Jastorff, and S. O. Doskeland, Biochemistry **39**, 8803 (2000).
[34] B. C. Bookser, M. C. Matelich, K. Ollis, and B. G. Ugarkar, J. Med. Chem. **48**, 3389 (2005).
[35] B. C. Oostenbrink, J. W. Pitera, M. M. H. Van Lipzig, J. H. N. Meerman, and W. F. van Gunsteren, J. Med. Chem. **43**, 4594 (2000).
[36] C. Oostenbrink and W. F. van Gunsteren, Proteins **54**, 237 (2004).
[37] X. Cheng, G. Cui, V. Hornak, and C. Simmerling, J. Phys. Chem. B **109**, 8220 (2005).

Paper 6

Hritz J, Oostenbrink C. Optimization of Replica Exchange Molecular Dynamics by Fast Mimicking. *J. Chem. Phys.* **2007,** 127, 204104

Optimization of replica exchange molecular dynamics by fast mimicking

Jozef Hritz and Chris Oostenbrink[a)]
Leiden Amsterdam Center for Drug Research (LACDR), Division of Molecular Toxicology, Vrije Universiteit, Amsterdam, The Netherlands

(Received 8 August 2007; accepted 5 September 2007; published online 27 November 2007)

We present an approach to mimic replica exchange molecular dynamics simulations (REMD) on a microsecond time scale within a few minutes rather than the years, which would be required for real REMD. The speed of mimicked REMD makes it a useful tool for "testing" the efficiency of different settings for REMD and then to select those settings, that give the highest efficiency. We present an optimization approach with the example of Hamiltonian REMD using soft-core interactions on two model systems, GTP and 8-Br-GTP. The optimization process using REMD mimicking is very fast. Optimization of Hamiltonian-REMD settings of GTP in explicit water took us less than one week. In our study we focus not only on finding the optimal distances between neighboring replicas, but also on finding the proper placement of the highest level of softness. In addition we suggest different REMD simulation settings at this softness level. We allow several replicas to be simulated at the same Hamiltonian simultaneously and reduce the frequency of switching attempts between them. This approach allows for more efficient conversions from one stable conformation to the other. © *2007 American Institute of Physics*. [DOI: 10.1063/1.2790427]

I. INTRODUCTION

Since the introduction of the replica exchange method (REM) using the Monte Carlo algorithm, replica exchange molecular dynamics (REMD) or parallel tempering simulations in the late 1990s,[1–5] there has been a steep increase in their popularity. In these simulations schemes the conformational sampling of a molecular system is greatly enhanced by connecting multiple simulations that are performed at different temperatures [temperature REMD (*T*-REMD)] or using different functional forms of the potential interaction energy [Hamiltonian REMD (*H*-REMD)].[6] By maintaining a detailed balance requirement when individual simulations are switched to a different temperature or Hamiltonian, the correct thermodynamic ensemble will be obtained for each of the simulation parameters. An intelligent choice of REMD settings allows for the swift generation of canonical ensembles of systems in which the potential energy barriers between stable conformations are too large to be crossed repeatedly in a normal molecular dynamics (MD) simulation. In this paper we will introduce an approach to efficiently optimize the settings for REMD simulation for systems with multiple stable conformations. Settings to optimize involve the number of simultaneous simulations (replicas) and the optimal simulation settings for each of these simulations (temperature, Hamiltonian).

Only researchers with access to extraordinary computational resources can afford a trial and error approach when searching for efficient REMD settings. Several studies describe approaches leading to efficient REMD simulations, mostly for *T*-REMDs. [Because one MD step requires the same amount of CPU time for any temperature or Hamiltonian, the allocation of replicas to CPUs can be trivial (one replica per CPU). However, in replica exchange Monte Carlo simulations, the average wall clock time to complete one Monte Carlo move varies with the temperature or the Hamiltonian, and the CPU allocation becomes an important issue.[7]] Many authors claim that the highest efficiency is obtained when the switching probability between neighboring replicas is constant at a value of approximately 20%.[8–13] This is still the most often used criterion in REMD applications despite the fact that in 2004 Trebst *et al.* presented the feedback-optimized parallel tempering approach.[14] It was shown that for the optimal temperature sets the switching probabilities between neighboring replicas are not constant but rather depend on the temperature.[14–16] A significant practical drawback of this approach, however, is the simulation time required to obtain the optimized settings. Especially for biomolecular simulations, this hampers the practical applicability of the method. The application of feedback-optimized parallel tempering for the 36-residue villin headpiece subdomain HP-36 required REMD simulations that covered 400 000 switching trials, which takes many years of CPU time.[16] It is probably for this reason that less applications of the method have been described.

For this reason we have developed a set of efficient tools for optimizing the settings of REMD (both *T*- and *H*-REMD) simulations for systems, of which multiple stable conformations are known. By generating the appropriate conformational ensemble of the system REMD gives insight into the relative populations of stable conformations. We present the practical application of the proposed optimization scheme for two biologically relevant systems: GTP and 8-Br-GTP (Fig. 1). GTP is an important component in, e.g., cellular signaling, while 8-Br-GTP is considered as promising anti-

[a)] Electronic mail: c.oostenbrink@few.vu.nl

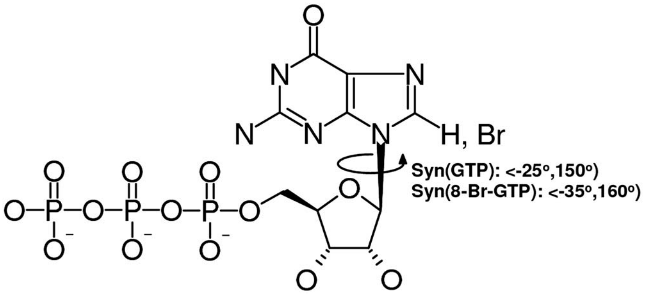


FIG. 1. Structure of GTP and 8-Br-GTP in syn conformation. Conformational transitions between the syn and anti states occur by rotation around the glycosidic bond (indicated).

bacterial agent. It inhibits FtsZ polymerization but does not affect tubulin polymerization.[17] GTP and 8-Br-GTP both have two stable orientations of the base toward the sugar: anti and syn (Fig. 1). Nuclear magnetic resonance (NMR) studies show that while GTP prefers to be in the anti conformation (with an estimated population of ~70% from NMR experiments), 8-Br-GTP shows preference for the syn conformation (population ~90%). Our MD study shows that there is a high energy barrier between the anti and syn conformations for both molecules. In two MD simulations of GTP starting from anti and syn conformation (each of 20 ns) only one single transition from syn to anti was observed (after ~1 ns) while no anti→syn transition was seen at all. Analogous simulations of 8-Br-GTP did not reveal any transition indicating an even higher potential energy barrier between the anti and syn conformations.[18] In these cases REMD can be used to enhance the conformational sampling and obtain the correct conformational ensemble with the proper populations of syn and anti conformations. From this ensemble, the free energy difference between the two states as well as a variety of molecular properties can be calculated.

II. METHODS

A. MD settings

All MD simulations were conducted using the GROMOS05 MD simulation package running on a linux cluster.[19] All bonds were constrained, using the SHAKE algorithm[20] with a relative geometric accuracy of 10^{-4}, allowing for a time step of 2 fs used in the leapfrog integration scheme.[21] Periodic boundary conditions, with a truncated octahedral box, were applied. After the steepest descent minimization to remove bad contacts between molecules, the initial velocities were randomly assigned from a Maxwell–Boltzmann distribution at 298 K, according to the atomic masses. The temperature was controlled using a weak coupling to a bath of 298 K with a time constant of 0.1 ps.[22] The solute molecules (GTP or 8-Br-GTP) and solvent (i.e., explicit water molecules and 3 Na^+ counterions) were independently coupled to the heat bath. The pressure was controlled using isotropic weak coupling to atmospheric pressure with a time constant 0.5 ps.[22] Van der Waals and electrostatic interactions were calculated using a triple range cutoff scheme. Interactions within a short-range cutoff of 0.8 nm were calculated every time step from a pair list that was generated every five steps.

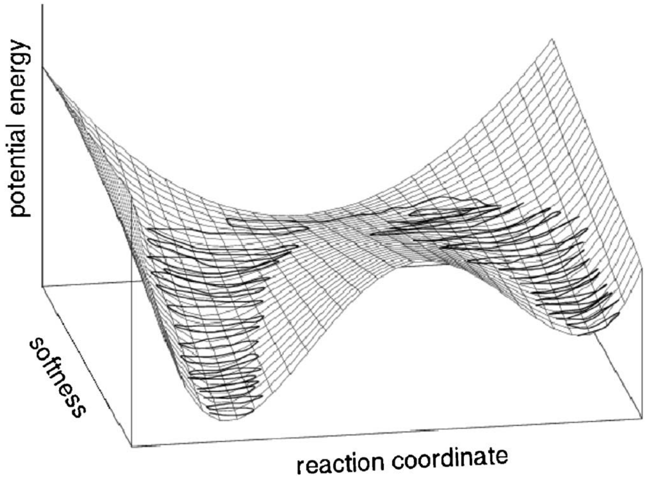


FIG. 2. Schematic figure of the energy landscape profile as function of the softness of nonbonded interactions. In the presented *H*-REMD using soft-core interactions, the replicas at higher levels of softness have a higher probability for the conformational transition between two stable conformational states. Still, this transition requires some time.

At these time points, interactions between 0.8 and 1.4 nm were also calculated and kept constant between updates. A reaction-field contribution was added to the electrostatic interactions and forces to account for a homogeneous medium outside the long-range cutoff, using the relative permittivity of SCP water (61).[23] All interaction energies were calculated according to the GROMOS force field, parameter set 53A6.[24] Force field parameters used for GTP and 8-Br-GTP are listed in the supplementary material of Ref. 18.

In this work we will focus on the optimization of a *H*-REMD approach in which the modification of the Hamiltonian consists of a softening of selected interactions. For this reason we have used the following form for Van der Waals and electrostatic soft-core interactions as a function of the interatomic distance r_{ij}:[25]

$$E_\lambda^{\mathrm{vdw}}(r_{ij}) = \left(\frac{C_{12}}{A_\lambda + r_{ij}^6} - C_6\right)\frac{1}{A_\lambda + r_{ij}^6}, \tag{1}$$

$$E_\lambda^{\mathrm{el}}(r_{ij}) = \frac{q_i q_j}{4\pi\varepsilon}\frac{1}{\sqrt{B_\lambda + r_{ij}^2}}, \tag{2}$$

where $A_\lambda = \alpha_{\mathrm{vdw}}(C_{12}/C_6)\lambda^2$ and $B_\lambda = \alpha_{\mathrm{el}}\lambda^2$. C_{12} and C_6 are the Lennard–Jones parameters, q_i and q_j are the partial charges of particles i and j, and α_{vdw} and α_{el} are the softness parameters that can be set for every pair of atoms. In the current study we used in all simulations $\alpha_{\mathrm{vdw}} = \alpha_{\mathrm{el}} = 1$ and the softness of the interactions was controlled through the λ parameter. It can be seen that at longer distances the soft-core interaction approximates the interaction for normal atoms and that they differ mostly at short distances between atoms. Potential energy barriers are mostly the result of short-ranged repulsion between atoms, which can strongly be reduced at higher levels of softness (Fig. 2). In this study, only interactions between the nucleotide base and sugar are treated using the soft-core interaction.

The systems were first equilibrated for 100 ps MD at constant pressure, where position restraints were applied on heavy atoms of GTP/8-Br-GTP, after which another 100 ps of equilibration at constant pressure without any restraints were carried out. Finally, the systems were simulated for 1 ns at different λ values, where coordinates of the whole system were recorded every 1 ps. Ten independent 1 ns MD simulations at different levels of softness were performed for GTP ($\lambda=0.0,0.05,\ldots,0.4,0.45$) and 15 for 8-Br-GTP ($\lambda=0.0,0.05,\ldots,0.65,0.7$).

B. REMD

In a REMD simulation, a number of noninteracting MD runs (called replicas) are simulated at different conditions. Let us label replicas as $0,1,\ldots,n$. After a given time (elementary period, t^{elem}), an exchange between two neighboring replicas is attempted, followed by another set of independent MD simulations. In the GROMOS05 implementation switches are first attempted between pairs $0\leftrightarrow 1, 2\leftrightarrow 3,\ldots$ (type I of replica exchange trials) and after the next t^{elem} of MD simulations switches are subsequently attempted between pairs $1\leftrightarrow 2, 3\leftrightarrow 4,\ldots$ (type II of replica exchange trials). This means that the effective time between switching attempts of the same type is twice the elementary period. In our study we used $t^{\mathrm{elem}}=2.5$ ps leading to an effective period of 5 ps between identical switching attempts.

In contrast to T-REMD, where the temperature is increased to facilitate the crossing of high energy barriers, we are modifying the Hamiltonian within H-REMD by making use of the soft-core interactions, Eqs. (1) and (2). This will lead to a decrease of the potential energy barrier between conformations and thus also facilitate the transition from one potential energy minimum into the other (see Fig. 2). The proper ensemble of conformations at every value of λ can be obtained by applying the Metropolis criterion for the exchange probability w_{λ_i,λ_j} between neighboring replicas running at Hamiltonians corresponding to the parameters λ_i and λ_j,

$$w_{\lambda_i,\lambda_j}=\min[1,\exp(-\Delta_{\lambda_i,\lambda_j})], \tag{3}$$

where

$$\Delta_{\lambda_i,\lambda_j}=\beta[E_{\lambda_j}(q_{\lambda_i})-E_{\lambda_i}(q_{\lambda_i})+E_{\lambda_i}(q_{\lambda_j})-E_{\lambda_j}(q_{\lambda_j})]. \tag{4}$$

The coordinates q_{λ_i} represent a conformation that was obtained from a simulation at the Hamiltonian corresponding to parameter λ_i and E_{λ_j} is the potential energy calculated according to the Hamiltonian corresponding to the parameter λ_j [Eqs. (1) and (2)]. $\beta=1/k_BT$ with k_B as the Boltzmann constant and T as the absolute temperature.

C. Concept of double/multiple replicas at the highest lambda/temperature

The elementary period, t^{elem}, in REMD should be long enough to relax the energy before the next switching attempt. Otherwise one observes many "reswitches" at the next switching attempt of the same type after a replica switch with low probability. Of course one can increase t^{elem}, but then the overall REMD efficiency decreases with decreasing number of replica switches. Therefore, a reasonable balance will need to be struck.

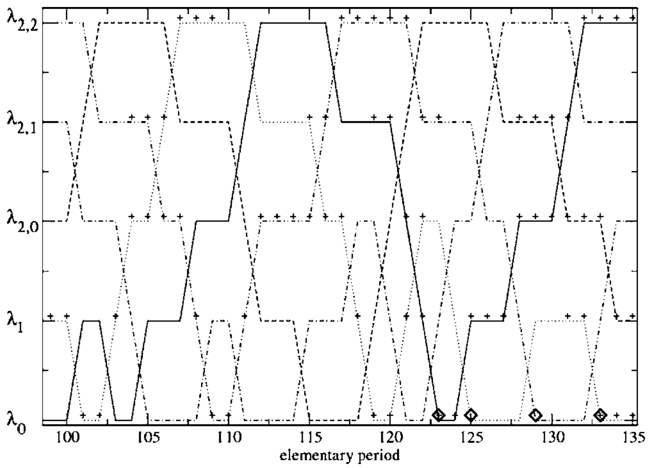


FIG. 3. Schematic example of real REMD with three replicas at $\lambda_{\max}$ ($\lambda_{2,0}=\lambda_{2,1}=\lambda_{2,2}=\lambda_{\max}$) and one middle λ_1 illustrating global conformational transitions, indicated by a diamond symbol. Crosses indicate the presence of syn conformation. Switches between replicas at $\lambda_{2,0}$ and $\lambda_{2,1}$ are attempted only at times, t, which are a multiple of 5 t^{elem} and between $\lambda_{2,1}$ and $\lambda_{2,2}$ if $t+t^{\mathrm{elem}}$ is a multiple of 5 t^{elem}. It ensures that a replica that switches from $\lambda_{2,0}$ to $\lambda_{2,1}$ will spend the required time, $t^{\mathrm{total}}_{\lambda_{\max}}=10t^{\mathrm{elem}}$ at $\lambda_{2,1}$ and $\lambda_{2,2}$ altogether.

However, there are more relaxation processes that play a role. Even at values of λ where the barrier has been removed or at temperatures where it is readily crossed, the system still needs time to move from one conformation to the other. (Note: Even for a very small conformational barrier the average transition time is significantly longer than times generally used in REMD between switching trials. This is mainly true for transitions to the less favorable conformational states.) The occurrence of syn states for a set of simulations starting at anti under the conditions of the highest λ value or temperature and the occurrence of anti states for simulations starting from a syn conformation as a function of time, will typically be represented by a sigmoidal curve. We define the transition time $t^{\mathrm{transit}}_{\lambda_{\max}}$ as the time at which this sigmoidal curve reaches saturation. The height of the curve allows us to estimate the relative populations of syn and anti at this λ value and from that the transition probabilities at times larger than $t^{\mathrm{transit}}_{\lambda_{\max}}$, $P^{\mathrm{anti}\to\mathrm{syn}}_{\lambda_{\max}}(t^{\mathrm{transit}}_{\lambda_{\max}})$, and $P^{\mathrm{syn}\to\mathrm{anti}}_{\lambda_{\max}}(t^{\mathrm{transit}}_{\lambda_{\max}})$.

The conformational transition time, $t^{\mathrm{transit}}_{\lambda_{\max}}$, is for the majority of systems much longer than t^{elem}. If t^{elem} would be extended to $t^{\mathrm{transit}}_{\lambda_{\max}}$, then the efficiency gain of REMD becomes very low. In order to combine frequent switching trials with long enough relaxation times to allow for anti↔syn conformational transitions to occur at $\lambda_{\max}$ we introduce a new scheme, called degenerated $\lambda_{\max}$ scheme, involving multiple (n) replicas at $\lambda_{\max}$ (or at the highest temperature for T-REMD).

This can be done by "degenerating" $\lambda_{\max}$ into $\lambda_{\max,0},\lambda_{\max,1},\ldots,\lambda_{\max,n-1}$. In the switching scheme all λ values are ordered as $\lambda_0,\lambda_1,\ldots,\lambda_{\max-1}$, $\lambda_{\max,0},\lambda_{\max,1},\ldots,\lambda_{\max,n-1}$, and only switches are attempted between neighboring λ values. The switching frequency between replicas at $\lambda_{\max}$ is reduced by only allowing switching attempts after a given multiple of the elementary period (Fig. 3). We define the time $t^{\mathrm{total}}_{\lambda_{\max}}$ as the total simulation time be-

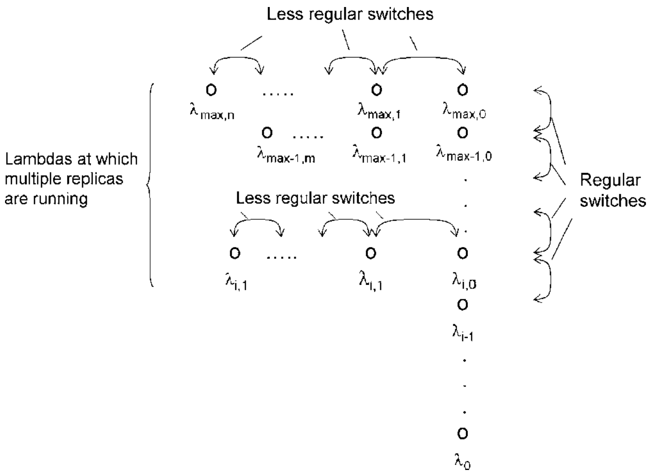


FIG. 4. Schematic figure showing the application of degenerate λ's. By limiting the number of switching attempts between replicas at the same λ value, the systems are allowed more time to relax toward different conformations at this value of λ.

tween the switch $\lambda_{max,0} \rightarrow \lambda_{max,1}$ until the return switch $\lambda_{max,1} \rightarrow \lambda_{max,0}$. Note that the switching probability between replicas at λ_{max} equals 1 because they correspond to the same Hamiltonian and Δ in Eq. (4) becomes 0 by definition. The procedure to find the optimal $t_{\lambda_{max}}^{total}$ is described in Sec. II D 2. The switching trial frequency between all other pairs of replicas is much higher (with an elementary period of 2.5 or 5 ps between two switching trials of the same type), which maintains an efficient diffusion of replicas between the lowest and highest λ value.

An increasing number of replicas at λ_{max} increases the convergence of conformational populations at λ_{max}, which subsequently leads to a faster convergence of conformational populations at $\lambda_{max-1}, \lambda_{max-2}, \ldots$ and finally to the faster convergence of populations over the whole REMD system. The REMD efficiency gain when using more replicas at λ_{max} comes from the fact that multiple conformational transition "attempts" are performed in parallel.

With n replicas at λ_{max}, a replica that has had sufficient time to show conformational transition becomes available at $\lambda_{max,0}$ with a period of $(1/n-1)t_{\lambda_{max}}^{total}$. This allows for values of n up to $n_{max}=t_{\lambda_{max}}^{total}/t^{elem}+1$, although the efficiency is limited in this case because replicas tend to stay for several periods of $t_{\lambda_{max}}^{total}$ at λ_{max} rather than get an opportunity to switch down to λ_{max-1}. For this reason, the maximal overall efficiency is obtained at $n \approx w_{\lambda_{max-1},\lambda_{max,0}} n_{max}$. Especially for complex systems the required time $t_{\lambda_{max}}^{total}$ at λ_{max} can be high, so a higher number of replicas at λ_{max} can significantly increase the overall REMD efficiency, almost by a factor $(n-1)$. For the examples described in this work, a one-dimensional setup with multiple replicas at λ_{max} seems sufficient, but the concept can easily be extended to having a degenerate set of replicas at a certain λ value or temperature by going to a second or third dimension of λ or T. See Fig. 4.

D. Searching for the optimal REMD settings

The efficiency of REMD to sample conformational space depends mainly on the efficiency of conformational transitions at λ_{max} and an efficient diffusion of replicas between the lowest and highest λ values. In this section, we describe the tools that are used to optimize the choice of λ values and the number of replicas at λ_{max}.

Section II D 1 describes the algorithm we propose for the fast mimicking of REMD without performing actual MD. This algorithm requires knowledge of the optimal value of λ_{max}, the optimal conformational conversion time at λ_{max}, $t_{\lambda_{max}}^{total}$, as well as estimates of the saturated conformational transition probabilities, $P_{\lambda_{max}}^{anti \rightarrow syn}$, $P_{\lambda_{max}}^{syn \rightarrow anti}$. In addition, the algorithm makes use of switching probabilities between replicas at different λ values. Section II D 2 describes the selection of λ_{max}, the calculation of $t_{\lambda_{max}}^{total}$, and conversion probabilities. Section II D 3 describes how switching probabilities are sampled from precalculated probability distributions and Sec. II D 4 finally describes the approach we take to use the mimicking algorithm to obtain the most efficient settings for REMD simulations.

1. Mimicked REMD

Mimicked REMD assumes a number of discreet stable conformational states exists, labeled, $c=0,1,\ldots,N_c-1$ (in our case anti and syn). In the present case these conformations are described as anti and syn. Instead of performing MD we estimate the probability of a conformational transition as well as the REMD switching probabilities and subsequently propagate the conformational states through time. This requires knowledge of the probability distributions of switching probabilities $\{w_{\lambda_i,\lambda_j}^{c_k,c_l}\}$ for all neighboring pairs of λ values: λ_i, λ_j as well as approximate probabilities of the conformational changes at each λ value from one stable conformational region to the other ones. We would like to note that if precise conformational transition probabilities would be known then no REMD mimicking is needed, as one could easily derive the individual conformational populations as well as the free energy difference between the stable conformations. The merit of the current approach comes from the fact that the transition probabilities can be estimated for λ_{max} fairly easily, while they are ~0 for all other λ values.

As explained in Sec. II C the transition between conformational states is a dynamic process in which the conformational transitions depend sigmoidally on time. From these curves at λ_{max}, one can estimate the probability of the transition $c_i \rightarrow c_j$, after the (prolonged) residence time, $P_{\lambda_{max}}^{c_i \rightarrow c_j}(t_{\lambda_{max}}^{transit})$. Because λ_{max} will be selected such that the corresponding $t_{\lambda_{max}}^{total}$ is shortest (see Sec. II D 2) and because for all other replicas the time between switching attempts $(2t^{elem})$ is much shorter than the corresponding t_{λ}^{total}, it is reasonable to assume that the probabilities of conformational transitions at any λ value other than λ_{max} approach zero $[P_{\lambda \neq \lambda_{max}}^{c_i \rightarrow c_j}(2t^{elem})=0]$.

In our REMD mimicking algorithm we begin by assigning starting conformational states to all replicas. This can be done either randomly, sampled from the correct ensemble, or biased toward one of the states.

Subsequently four steps make up the main cycle of our algorithm:

(1) Intralambda conformational transitions. We predict if the conformational state changes during the elementary period (in our case $t^{\text{elem}}=2.5$ ps) at each λ value by the given probabilities. As described above all conformational transition probabilities are set to zero for all λ values, except for $\lambda_{\max}$, where $P^{c_i\to c_j}_{\lambda_{\max}}(t^{\text{total}}_{\lambda_{\max}}) = P^{c_i\to c_j}_{\lambda_{\max}}(t^{\text{transit}}_{\lambda_{\max}})$ is nonzero once per number of cycles corresponding to the $t^{\text{total}}_{\lambda_{\max}}$ spent at $\lambda_{\max,1},\ldots,\lambda_{\max,n-1}$.

(2) Replica exchanges of type I (between $0\leftrightarrow 1, 2\leftrightarrow 3,\ldots$). According to the actual conformational state at each λ value we take the corresponding probability distributions of switching probabilities $\{w^{c_k,c_l}_{\lambda_i,\lambda_j}\}$ for all neighboring λ pairs of type I. A switching probability is randomly chosen from the distribution of switching probabilities $\{w^{c_k,c_l}_{\lambda_i,\lambda_j}\}$ and a second random number determines if the switch occurs.

(3) Intralambda conformational transitions. The same as the first step.

(4) Replica exchanges of type II (between $1\leftrightarrow 2$, $3\leftrightarrow 4,\ldots$). The same as the second step but now for the λ pairs of type II.

The length of the whole cycle corresponds to the double of elementary switching attempt period (in our case 2×2.5 ps$=5$ ps). Sampling of a half million cycles (corresponding to 10^6 t^{elem} of real REMD) takes typically 15 min (depending on the exact number of replicas) using an unoptimized python script.

2. Selection of $\lambda_{\max}$

The average conformational transition time at the unperturbed Hamiltonian ($\lambda=0.0$), $t^{\text{transit}}_{0.0}$ is by far too long to sample sufficient transitions reversibly. In *H*-REMD, the Hamiltonian is parameterized such that the potential energy barrier associated to the conformational transition is reduced for higher values of λ, thereby also reducing $t^{\text{transit}}_{\lambda_i}$.

To obtain the optimal value of $\lambda_{\max}$, we perform ten short MD simulations (200 ps) starting from anti and 10 short MD runs starting from syn conformation at different λ values (for GTP $\lambda=0.4;0.45;0.5$ and for 8-Br-GTP $\lambda=0.6;0.65;0.7;0.8;0.9$). (Note: $\lambda_{\max}$–values lower than 0.4 respectively. 0.6 were not considered based on the results of 1 ns runs at different λ values used for the calculation of switching probability distributions as described in the next paragraph.)

Our aim is to find the value of $t^{\text{total}}_{\lambda_{\max}}$ for which we would get the highest number of conformational transitions at $\lambda_{\max}$ during a given length of REMD simulation (see also Sec. II C). While the time dependency of the number of simulations where the conformation has changed with respect to the initial conformation has an S-curve profile, it is clear that the most efficient $t^{\text{total}}_{\lambda_{\max}}$ will be in the interval between the midpoint of the S curve and the saturated region. Quantitatively $t^{\text{total}}_{\lambda_{\max}}$ is obtained as the time corresponding to the maximum of the number of conformational changes divided by time. Because these maxima occur at different times for the anti → syn transition and the syn → anti transitions, and because we need to have a large enough number of both types of transitions, we select the longer time of both transitions at one $\lambda_{\max}$ value. We select the value of $\lambda_{\max}$ for which $t^{\text{total}}_{\lambda_{\max}}$ is shortest and where enough transitions occur in both directions. The set of ten MD runs starting from different conformations is subsequently prolonged for the selected $\lambda_{\max}$, in order to refine the converged values of the conformational conversion probabilities $P^{\text{anti}\to\text{syn}}_{\lambda_{\max}}(t^{\text{transit}}_{\lambda_{\max}})$ and $P^{\text{syn}\to\text{anti}}_{\lambda_{\max}}(t^{\text{transit}}_{\lambda_{\max}})$. Because the exact shape of the S curve is not properly converged from only ten simulations, it is not wise to directly read $P^{\text{transit}}_{\lambda_{\max}}(t^{\text{total}}_{\lambda_{\max}})$ from these curves. Rather, we set $P^{\text{transit}}_{\lambda_{\max}}(t^{\text{total}}_{\lambda_{\max}}) = P^{\text{transit}}_{\lambda_{\max}}(t^{\text{transit}}_{\lambda_{\max}})$, a value which converges also for a limited number of simulations. Using $P^{\text{transit}}_{\lambda_{\max}}(t^{\text{transit}}_{\lambda_{\max}})$ also ensures the generating the proper conformational populations at $\lambda_{\max}$ within the mimicked REMD.

3. Generation of probability distributions of switching probabilities

Let us consider the system, which has several different stable conformational sates (syn, anti), for which standard MD simulations do not produce enough transitions due to the conformational barriers between these states. We assume that the conformational space within each of these stable regions is sampled "reasonably" well by a relatively short MD simulation (~1 ns). The description below follows the schematic representation in Fig. 5.

To obtain the probability distributions of the switching probabilities needed for the REMD mimicking we performed a set of 1 ns MD simulations starting from different stable conformers at different values of λ (alternatively, one could use different temperatures) between 0.0 and $\lambda_{\max}$ usually at increments of 0.05 (in principle one can use also REMD at these λ values for this purpose). Every 1 ps we store one conformational configuration frame leading to 1000 different configurations. It is possible that during these MD simulations conformational transitions occur. For this reason, we collect conformations belonging to the same stable conformational region c and the same value of λ at which the simulation was performed into one set of conformations

$$\{q^c_\lambda\}, \lambda = 0, 0.05, 0.1, \ldots, \lambda_{\max};$$

$$c = 0, 1, \ldots, N_c - 1.$$

For every combination of λ and c we obtain about 1000 different structures. The rth structural frame from this "trajectory" is expressed as $q^c_\lambda(t_r)$. For each of these structures we calculate its energy not only using the Hamiltonian at which it was simulated (represented by λ_i), but also the energy according to the Hamiltonian corresponding to all other λ values (λ_j), and expressed as: $E_{\lambda_j}(q^c_{\lambda_i}(t_r))$.

For each pair of configurations $q^{c_k}_{\lambda_i}(t_r)$; $q^{c_l}_{\lambda_j}(t_s)$ we calculate switching energy difference

$$\Delta^{c_k,c_l}_{\lambda_i,\lambda_j}(t_r,t_s) = \beta[E_{\lambda_j}(q^{c_k}_{\lambda_i}(t_r)) - E_{\lambda_i}(q^{c_k}_{\lambda_i}(t_r)) + E_{\lambda_i}(q^{c_l}_{\lambda_j}(t_s)) - E_{\lambda_j}(q^{c_l}_{\lambda_j}(t_s))] \tag{5}$$

and the corresponding switching probability

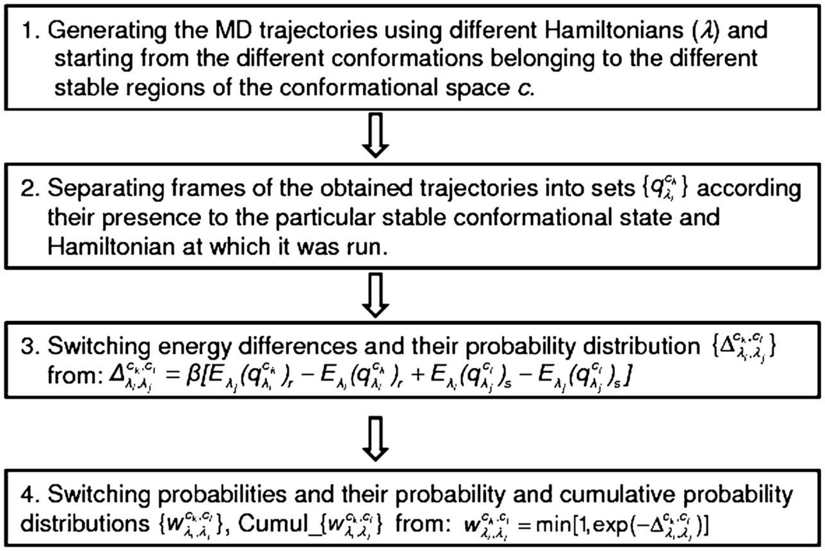

1. Generating the MD trajectories using different Hamiltonians (λ) and starting from the different conformations belonging to the different stable regions of the conformational space c.
2. Separating frames of the obtained trajectories into sets $\{q^{c_k}_{\lambda_i}\}$ according their presence to the particular stable conformational state and Hamiltonian at which it was run.
3. Switching energy differences and their probability distribution $\{\Delta^{c_k,c_l}_{\lambda_i,\lambda_j}\}$ from: $\Delta^{c_k,c_l}_{\lambda_i,\lambda_j} = \beta[E_{\lambda_j}(q^{c_k}_{\lambda_i})_r - E_{\lambda_i}(q^{c_k}_{\lambda_i})_r + E_{\lambda_i}(q^{c_l}_{\lambda_j})_s - E_{\lambda_j}(q^{c_l}_{\lambda_j})_s]$
4. Switching probabilities and their probability and cumulative probability distributions $\{w^{c_k,c_l}_{\lambda_i,\lambda_j}\}$, Cumul_$\{w^{c_k,c_l}_{\lambda_i,\lambda_j}\}$ from: $w^{c_k,c_l}_{\lambda_i,\lambda_j} = \min[1,\exp(-\Delta^{c_k,c_l}_{\lambda_i,\lambda_j})]$

FIG. 5. Scheme describing the steps for obtaining the probability distribution of the switching probabilities.

$$w^{c_k,c_l}_{\lambda_i,\lambda_j}(t_r,t_s) = \min[1,\exp(-\Delta^{c_k,c_l}_{\lambda_i,\lambda_j}(t_r,t_s))], \tag{6}$$

where $\beta = 1/k_BT$, k_B is the Boltzmann constant, and T is the absolute temperature.

The distributions of $\Delta^{c_k,c_l}_{\lambda_i,\lambda_j}$ and $w^{c_k,c_l}_{\lambda_i,\lambda_j}$ values are thus calculated from altogether $\sim 10^6$ values, which are marked as $\{\Delta^{c_k,c_l}_{\lambda_i,\lambda_j}\}$ and $w^{\text{syn,anti}}_{0.0,0.3}$ after normalization. This gives us the probability distribution for a randomly chosen pair of configurations, $q^{c_k}_{\lambda_i}$, $q^{c_l}_{\lambda_j}$ with the switching probability between them given by $w^{c_k,c_l}_{\lambda_i,\lambda_j}$. For the REMD mimicking we found it convenient to produce Cum_$\{w^{c_k,c_l}_{\lambda_i,\lambda_j}\}$, which contains the cumulants of the probability distribution of switching probabilities.

4. Searching algorithm for the parameters of the most efficient REMD

500 000 cycles of REMD mimicking (corresponding to 10^6 t^{elem} of real REMD, in our case 2.5 μs) takes ~ 15 min which allows us to perform "REMD" for hundreds of different combinations of λ values and numbers of replicas at $\lambda_{\max}$ producing time sequences of the occurrence of conformational states at different $w^{\text{syn,anti}}_{0.0,0.3} = \min[1,\exp(-\Delta^{\text{syn,anti}}_{0.0,0.3})]$ values. From these one can decide for different criteria from which to prefer one set of parameters over the other.

For cases where different conformational states can be defined (e.g., the anti and syn states), we propose to measure efficiency of REMD, through the "number of global conformational transitions," N_{gct}. We count the number of conformational transitions for each particular replica at the lowest λ value, λ_0.

The schematic REMD example shown in Fig. 3, contains three replicas at $\lambda_{\max}$ ($\lambda_{2,0} = \lambda_{2,1} = \lambda_{2,2} = \lambda_{\max}$) and one middle value of lambda, λ_1. The replica represented by a solid curve, is in the anti conformation at λ_0 after 103 t^{elem} and then exchanges to higher λ values. At $\lambda_{2,1}$ it makes the transition to the syn conformation (after 119 t^{elem}) and subsequently exchanges with replicas at lower λ values until it finally returns to λ_0, but now in the different conformational state (syn) as before. The combination of replica exchanges and a conformational transition is counted as one global anti→syn transition. From this moment onwards, we will monitor for the next syn→anti global conformational transition (for the same replica), etc. From this description one can see that for a global conformational transition to occur it is not necessary for the replica to reach $\lambda_{\max}$, because the conformational transition can occur also at lower λ values (see, e.g., in Fig. 3 the replica represented by a dotted line. At 131 t^{elem} an anti→syn transition occurs at λ_1 leading to the global conformational transition at 133 t^{elem}) or even at λ_0 within real REMD. (This is not the case for the mimicked REMD, where conformational transition can occur only at $\lambda_{\max}$.) On the other hand, the replica represented by the dotted curve is in a syn conformation at λ_0 at 102 t^{elem}, then went up to $\lambda_{\max}$ where it changes its conformation to anti and later back to syn again. When it exchanges back to λ_0 at 119 t^{elem}, this event is not counted as a global conformational transition, because the conformations at 102 t^{elem} and 119 t^{elem} are the same. In the case of MD (as a special case of REMD with only one replica at λ_0) N_{gct} is equal to the number of conformational transitions within one MD run. This allows for a direct comparison of the efficiency between very different REMDs or MDs performed under different conditions.

Our aim is to find the REMD settings which give us the maximum value of N_{gct} per CPU. (In the following, we assume that in real REMD simulations every replica is assigned one CPU.) Having decided on the optimal value of $\lambda_{\max}$ and $t^{\text{total}}_{\lambda_{\max}}$ we vary:

(1) the number and placement of middle λ's (between 0.0 and $\lambda_{\max}$) and
(2) the number of replicas at $\lambda_{\max}$.

Whether it is efficient to "invest" into more replicas at $\lambda_{\max}$ or not depends on the ratio of $t^{\text{total}}_{\lambda_{\max}}$ relative to the average time needed for global conformational transitions to occur. If more time is needed for a conformational change at $\lambda_{\max}$, it becomes more likely that an increased number of replicas at $\lambda_{\max}$ improves the overall efficiency per CPU.

In cases where only a few replicas are needed one can test various settings in a systematic manner and select the optimal combination. In spite of the efficiency of mimicked REMD, a larger number of required replicas will make the optimization problem more complex. We propose three useful approaches:

(1) The number and placement of λ values between 0.0 and $\lambda_{\max}$ (middle λ's) and the number of replicas at $\lambda_{\max}$ can be optimized independently. Middle λ's are essential for the diffusion of replicas between 0.0 and $\lambda_{\max}$, which will be the same for any number of replicas at $\lambda_{\max}$. The optimal setting of middle λ's will therefore be independent of the number of replicas at $\lambda_{\max}$. However, the gain in efficiency due to optimization of the middle λ's is more pronounced if the conformational change at $\lambda_{\max}$ is not the bottleneck of the simulation. Therefore, we propose to search for the optimal settings with a high number of replicas at $\lambda_{\max}$, initially 5 in the case of 8-Br-GTP. After this initial optimization, the number of optimal replicas at $\lambda_{\max}$ should be obtained after which one should check if having one more middle λ does not improve results even further. Once the optimal number of replicas at $\lambda_{\max}$ has been obtained for an optimal set of m middle λ values, this $n^{\text{opt}}(m)$ can be used to reduce the searching possibilities for different numbers of middle λ values by taking into account the following inequalities: $n^{\text{opt}}(k>m) \geq n^{\text{opt}}(m)$ and $n^{\text{opt}}(k<m) \leq n^{\text{opt}}(m)$.

(2) Searching for the optimal settings of middle λ's. When expanding or reducing the number of middle λ's in the optimization process, we do not need to consider the complete range $(0.0, \lambda_{\max})$ for the new λ values. If we denote the optimal λ values in a scheme with n middle λ's as ${}^{n}\lambda_i^{\text{opt}}$ (${}^{n}\lambda_0^{\text{opt}}=0.0$; ${}^{n}\lambda_{n+1}^{\text{opt}}=\lambda_{\max}$), then it is most likely that in a scheme with $n+1$ middle λ's, the optimal λ values are in the following range: ${}^{n+1}\lambda_i^{\text{opt}} \in \langle {}^{n}\lambda_{i-1}^{\text{opt}}, {}^{n}\lambda_i^{\text{opt}} \rangle$. Similarly, one can generally write that for a reduction of the number of middle λ's, the optimal λ values are restricted to ${}^{n}\lambda_i^{\text{opt}} \in \langle {}^{n+1}\lambda_i^{\text{opt}}, {}^{n+1}\lambda_{i+1}^{\text{opt}} \rangle$.

(3) In many cases, already a short simulation indicates if a certain set of settings is promising or not. In an approach we call "continuous filtering" we disregard possible settings on the fly, thereby only spending computer time on the most promising settings.

We perform a mimicked REMD simulation for all potentially relevant REMD settings for 30 000 cycles. We subsequently disregard all settings that show less than 10% of the (at that point) maximum value of N_{gct}. After 40 000 cycles this threshold is increased to 20% and by another 10% every 10 000 cycles until after 100 000 cycles only those settings are kept that lead to 80% of the maximum value of N_{gct}. We subsequently refine the search by increasing the threshold by 2.5% every 100 000 cycles and select the optimal setting after 500 000 cycles.

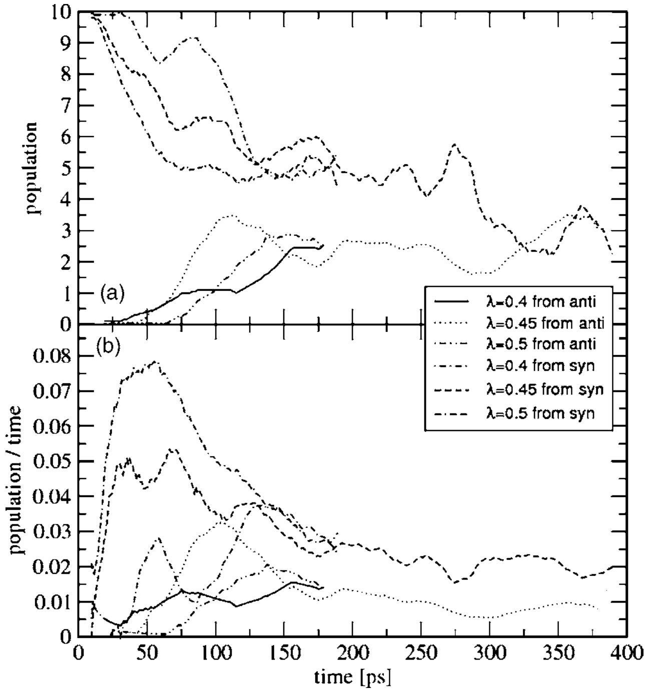


FIG. 6. (a) Number of syn conformations observed in a set of ten MD simulations of GTP as function of time at different λ values starting from ten different anti and syn conformations. Runs for λ=0.45 were prolonged up to 400 ps in order to reach convergence. (b) Population of converted conformations per time (syn → anti or anti → syn). For clarity, running averages over 20 time points have been taken.

III. RESULTS

Both GTP and 8-Br-GTP have two stable conformations: anti and syn, depending on the dihedral angle around the glycosidic bond (Fig. 1). Based on the distributions of this dihedral angle we define for GTP to be in the syn conformation, when its dihedral angle around the glycosidic bond is within the interval $<-25°, 150°)$, and otherwise to be in the anti conformation. For 8-Br-GTP the syn conformational interval is $<-35°, 160°)$.[18]

A. Selection of $\lambda_{\max}$

MD simulations of GTP in explicit water were performed for 200 ps at different values of λ (0.4; 0.45; 0.5). Ten simulations started from anti and ten simulations started from syn conformations. During the MD runs conformational transitions occur.

Figure 6(a) show the number of simulations in which GTP is in a syn conformation as function of time, when starting from anti and when starting from syn conformation. We obtained $t_{\lambda_{\max}}^{\text{total}}$ as the larger of two times corresponding to the maximas of the time dependence of the number of changed conformations divided by total time. From Fig. 6(b) follows: $t_{0.4}^{\text{total}}=155$ ps, $t_{0.45}^{\text{total}}=100$ ps, and $t_{0.5}^{\text{total}}=135$ ps. As the optimal value of $\lambda_{\max}$ we choose 0.45 because it gives the shortest $t_{\lambda_{\max}}^{\text{total}}$ (100 ps) together with a sufficient number of conformational transitions after this time. Another advantage of $\lambda_{\max}=0.45$ over 0.5 is the more efficient diffusion of replicas between λ_0 and $\lambda_{\max}$. For $\lambda_{\max}=0.45$, the set of ten MD simulations starting from syn and anti was prolonged leading

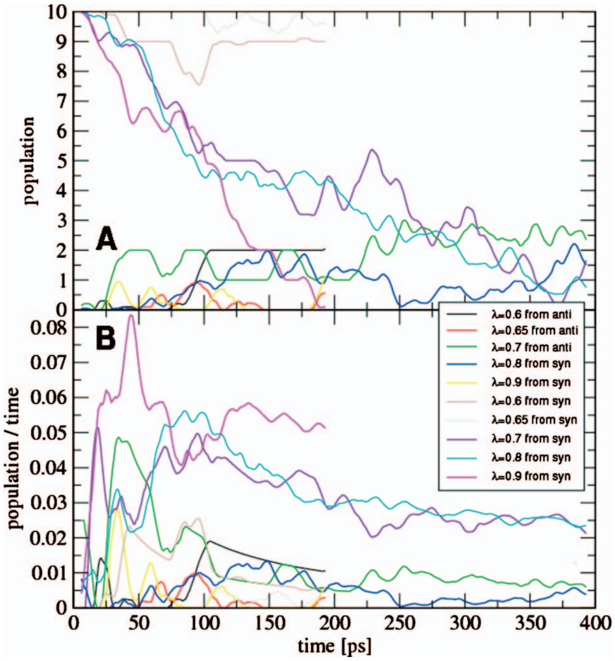


FIG. 7. (Color) (a) Number of syn conformations observed in a set of ten MD simulations of 8-Br-GTP as function of time at different λ values starting from ten different anti and syn conformations. Runs for λ=0.7 and λ=0.8 were prolonged up to 400 ps in order to reach convergence. (b) Population of converted conformations per time (syn→anti or anti→syn). For clarity, running averages over 20 time points have been taken.

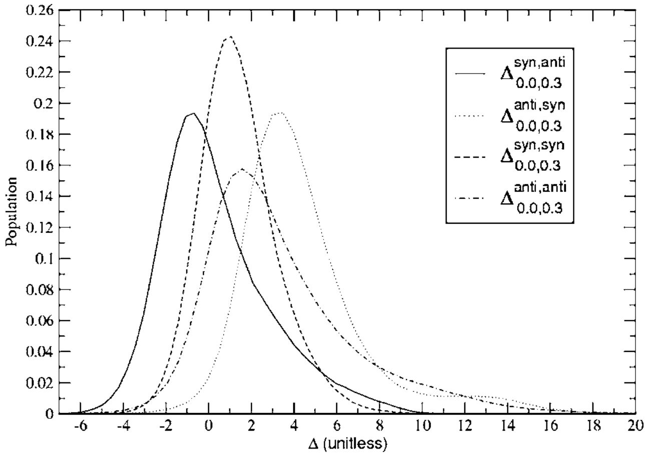


FIG. 8. Probability distributions of switching energy differences for GTP: $\Delta^{c_i,c_j}_{0.0,0.3}=\beta[E_{0.3}(q^{c_i}_{0.0})_r-E_{0.0}(q^{c_i}_{0.0})_r+E_{0.0}(q^{c_j}_{0.3})_s-E_{0.3}(q^{c_j}_{0.3})_s]$, where c_i, c_j are different conformational states: syn, anti and r, s are different configurations from MD trajectories.

to estimate of the converged conformational transition probabilities $P^{\text{anti}\rightarrow\text{syn}}_{\lambda=0.45}=0.3$ and $P^{\text{syn}\rightarrow\text{anti}}_{\lambda=0.45}=0.7$ [Fig. 6(a)].

The same kind of analysis was performed for 8-Br-GTP using λ=0.6;0.65;0.7;0.8;0.9 and leading to the population time dependences shown in Figs. 7(a) and 7(b). We considered as λ_{max} candidates $\lambda_{\text{max}}=0.7$ and $\lambda_{\text{max}}=0.8$ for which the prolonged MD simulations were performed. Based on these we decided to take $\lambda_{\text{max}}=0.7$ for 8-Br-GTP with the corresponding $t^{\text{total}}_{\lambda_{\text{max}}}=100$ ps and saturated transition probabilities $P^{\text{anti}\rightarrow\text{syn}}_{\lambda=0.7}=0.2$ and $P^{\text{syn}\rightarrow\text{anti}}_{\lambda=0.7}=0.8$. For λ=0.9 the transition syn→anti is quite efficient. However, the conversion probability from anti to syn has become very low which means that these transitions would occur only very rarely at λ=0.9.

B. Generation of probability distribution of switching probabilities

To obtain the probability distributions of switching probabilities we performed MD simulations for (8-Br-)GTP at λ's in the interval $\langle 0.0,\lambda_{\text{max}}\rangle$ with a λ increment of 0.05. For each λ two MD simulations of 1 ns were performed, one starting from an anti and the second from a syn conformation. Configurations of trajectories were stored every ps leading to 1000 frames for each trajectory. Configurations have been separated into sets according their conformational state ($\{q^{\text{anti}}_{\lambda}\}$, $\{q^{\text{syn}}_{\lambda}\}$).

The distributions for the switching energy differences $\{\Delta^{c_k,c_l}_{\lambda_i,\lambda_j}\}$ as well as for the switching probabilities $\{w^{c_k,c_l}_{\lambda_i,\lambda_j}\}$ and their cumulative values were obtained as described in Sec. II D 3 and schematically outlined in Fig. 5. Representative examples of the switching energy differences and switching probability distributions are given in Figs. 8 and 9.

C. REMD mimicking

REMD mimicking as described in the Sec. II D 1 produces as output the time evolutions of conformational states (anti and syn) at all λ values that are included. Because of its speed we can run mimicked REMD for a very long time and thus obtain converged populations of anti and syn conformations for GTP and 8-Br-GTP. Typically, we simulate 500 000 cycles of mimicked REMD corresponding to 10^6 t^{elem} in real REMD (~2.5 μs).

Syn populations at individual λ values of mimicked REMD for GTP (500 000 cycles) with two replicas at $\lambda_{\text{max}}=0.45$ combined with no or a single middle λ value are shown in Table I. Theoretically, syn populations at λ=0.0 should be the same for all mimicked REMDs. We do not suffer here from insufficient lengths of REMD simulations, but inaccuracies rather stem from the probability distributions of switching probabilities. Discrepancies in the syn

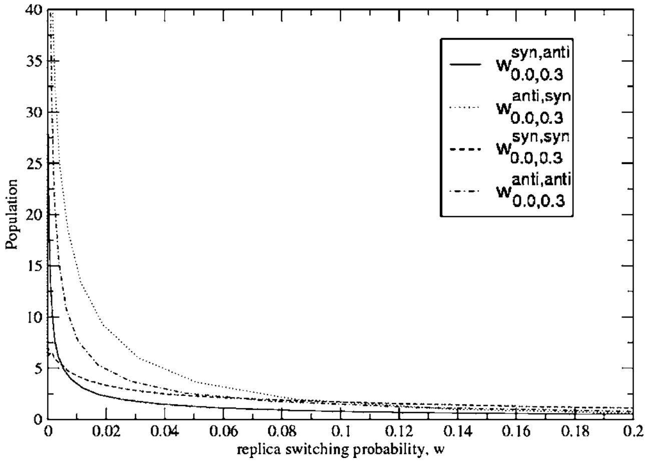


FIG. 9. Probability distribution of replica exchange switching probabilities for GTP: $w^{\text{syn,anti}}_{0.0,0.3}=\min[1,\exp(-\Delta^{\text{syn,anti}}_{0.0,0.3})]$ for GTP. Notice that probability $w^{\text{syn,anti}}_{0.0,0.3}$ is much lower than for the opposite replica switch $w^{\text{anti,syn}}_{0.0,0.3}$.

TABLE I. Syn populations at λ_0, ${}^1\lambda_1$, $\lambda_{\max,0}$, $\lambda_{\max,1}$ depending on the placement of ${}^1\lambda_1$ from mimicked REMD of GTP with two replicas at $\lambda_{\max}=0.45$ over 500 000 cycles.

${}^1\lambda_1$	None	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4
[syn](λ_0)	0.053	0.047	0.053	0.050	0.044	0.043	0.049	0.047	0.081
[syn](${}^1\lambda_1$)	⋯	0.058	0.095	0.133	0.163	0.228	0.287	0.295	0.298
[syn]($\lambda_{\max,0}$)	0.302	0.303	0.296	0.301	0.305	0.307	0.299	0.306	0.302
[syn]($\lambda_{\max,1}$)	0.300	0.301	0.295	0.299	0.303	0.304	0.298	0.304	0.300

populations at $\lambda=0.0$ can be explained from an insufficient sampling during the 1 ns MD simulation from which the probabilities are obtained. For λ values that are evident outliers compared to the other values, this simulation should probably be extended.

Largest deviations are expected for REMDs involving middle λ values, for which we multiply very small and very high probabilities (i.e., ${}^1\lambda_1=0.05$ and ${}^1\lambda_1=0.4$).

The root-mean-square deviation of [syn] at λ_0 divided by its average value when excluding the values obtained with ${}^1\lambda_1=0.05$ and ${}^1\lambda_1=0.4$ is calculated to be 0.068 indicating the relative inaccuracy in the distributions of switching probabilities.

The average of syn populations at λ_0 for runs with ${}^1\lambda_1$ between 0.1 and 0.35 amounts to $[\text{syn}]_{\text{aver}}=0.048$, which corresponds to a free energy difference $\Delta G^{\text{GTP}}_{\text{syn-anti}}=7.40$ kJ mol^{-1} at 298 K. As shown in Ref. 18 the obtained population of syn conformations from real *H*-REMD was 0.044 for GTP corresponding to a value of $\Delta G^{\text{GTP}}_{\text{syn-anti}}=7.63$ kJ mol^{-1}. 500 000 cycles of mimicking REMD using the same settings as the real REMD reported there ($\lambda=0.0;0.25,0.45,0.45,0.45,0.45$) gives [syn]=0.043, corresponding to $\Delta G^{\text{GTP}}_{\text{syn-anti}}=7.69$ kJ mol^{-1}. It shows that if most of the transitions can indeed occur only at $\lambda_{\max}$, then REMD mimicking can produce very reasonable values of populations of the individual conformational states.

Therefore, REMD mimicking is not only useful as a tool for the optimization of REMD settings but can also as be used in itself for calculating the populations of conformations. We can see an analogy between our method and a free energy method that makes use of alternative pathways to calculate free energy differences more efficiently along unphysical thermodynamic cycles.[26] We want to note, however, that mimicked REMD covers very many of such pathways simultaneously.

D. Finding the optimal settings for REMD

In the previous paragraph we demonstrated how long mimicked REMD can approximate syn and anti populations for GTP and 8-Br-GTP. These simulations would correspond to several μs of real REMD. In real REMD, however, we can afford to simulate only limited time lengths (~tens ns), therefore it is crucial to find settings which corresponds to the highest possible REMD efficiency. We have used mimicked REMD, because it allows us to attempt REMD for hundreds of different combinations of λ's. From the time sequences of conformational states at different λ values we can choose the combination of λ's that produces the highest number of global conformational transitions, N_{gct}, which will ensure the fastest convergence of populations at λ_0.

1. GTP

Mimicked REMD (500 000 cycles) without middle replica and two replicas at $\lambda_{\max}=0.45$ gives 5221 global conformational transitions, which is $5221/3=1740$ global conformational transitions per CPU. The results for mimicked REMD with an increasing number of replicas at $\lambda_{\max}=0.45$ are summarized in the Table II. It shows that the highest efficiency per CPU ($7506/4=1877$) is obtained for three replicas at $\lambda_{\max}$.

In the next step we performed mimicked REMD with different numbers of replicas at $\lambda_{\max}=0.45$ and one middle λ-value ${}^1\lambda_1\in\{0.05,0.1,0.15,0.2,0.25,0.3,0.35,0.4\}$ for 500 000 cycles [Fig. 10(a)]. With two replicas at $t^{\text{total}}_{\lambda_{\max}}=10t^{\text{elem}}$ the highest values of N_{gct} are obtained for ${}^1\lambda_1=0.15$ (6069), ${}^1\lambda_1=0.2$ (6071), and for ${}^1\lambda_1=0.25$ (6089), which are higher than the values of N_{gct} obtained from REMD without middle λ having the same two replicas at $\lambda_{\max}=0.45$. It means that putting one middle ${}^1\lambda_1$ can increase the efficiency of diffusion between $\lambda_0=0.0$ and $\lambda_{\max}=0.45$. The efficiency per CPU for mimicked REMD containing one middle ${}^1\lambda_1=0.25$, however, is lower (1522 compared to 1877). Still the efficiency per CPU may be higher for REMD containing one middle lambda in combination with a higher number of replicas at $\lambda_{\max}$. Figure 10 presents the efficiency of REMD containing one middle lambda with an increasing number of replicas at t^{elem}. It reveals that per CPU the most efficient set of λ's is [0.0, 0.25, 0.45, 0.45, 0.45, 0.45] with $12646/6=2108$ global conformational transitions per CPU. It also shows that the differences due to the placement of ${}^1\lambda_1$ are more pronounced for settings with a higher number of replicas at $\lambda_{\max}$.

Because the optimal REMD setting containing one ${}^1\lambda_1$ is more efficient per CPU than a set of λ's without any middle λ value, we continued to test REMD containing two middle λ values. As is explained in the methods Sec. II D 4 the optimal setting of REMD with two middle λ's will still have $t+t^{\text{elem}}$ replicas at $\lambda_{\max}=0.45$, because this is the optimal

TABLE II. Number of global conformational transitions, N_{gct}, during mimicked REMD of GTP with $n=2,3,4,5$ replicas at $\lambda_{\max}=0.45$ and no middle λ value after 500 000 cycles.

n replicas at $\lambda_{\max}=0.45$	2	3	4	5
N_{gct}	5221	7506	8954	9911
N_{gct}/CPU	1740	**1877**	1791	1652

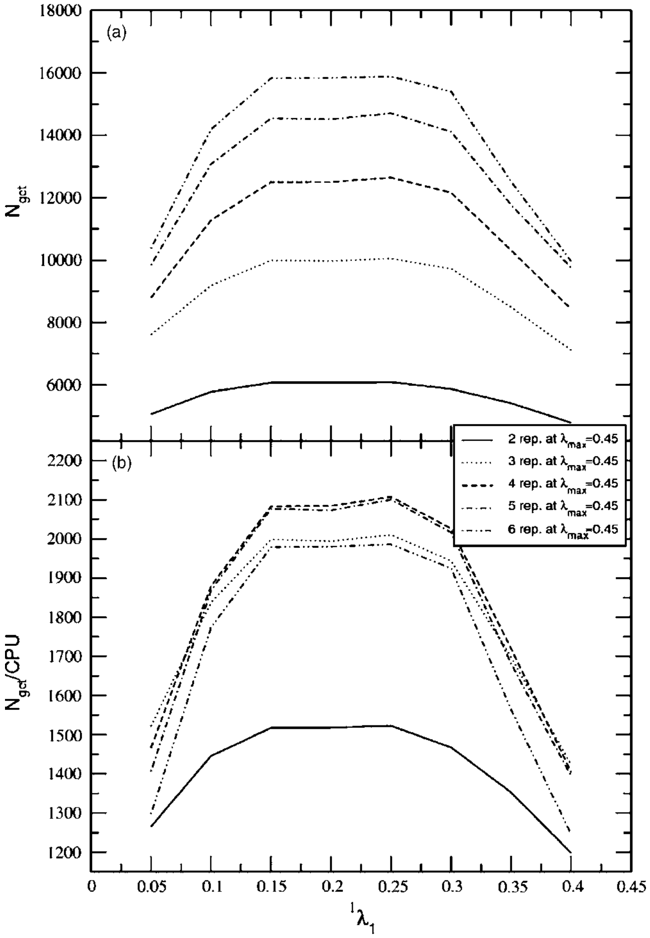


FIG. 10. Number of global conformational transitions, N_{gct} observed in 500 000 cycles of mimicked REMD of GTP with one middle replica $^1\lambda_1$ at different positions and with different numbers of replicas at λ_{max}=0.45: (a) absolute value and (b) value per CPU.

setting of REMD with one middle λ value. Table III shows N_{gct} per CPU for n=4,5,6 and two additional values $^2\lambda_1$ and $^2\lambda_2$. It shows that the optimal setting for REMD in this case is [0.0, 0.15, 0.3, 0.45, 0.45, 0.45, 0.45, 0.45]. However, its efficiency is slightly lower than for the optimal setting using only one middle $^1\lambda_1$ (2092 compared 2108 of global conformational transitions per CPU).

In conclusion the most effective setting predicted by mimicked REMD is [0.0, 0.25, 0.45, 0.45, 0.45, 0.45] with 12646/6=2108 global conformational transitions per CPU. This number corresponds to 500 000 REMD mimicking cycles or 10^6 elementary switching periods of real REMD. This means that one can expect ~13 global conformational transitions per 1000 t^{elem} of real REMD, which amounts to 2.5 ns with t^{elem}=2.5 ps. From an extrapolation of the linear increase of N_{gct} as a function of time, over ten individual 5000 cycles mimicked REMD simulations we estimate the REMD equilibration period to be about 200 t^{elem} (0.5 ns in real REMD). For comparison, in real REMD we reported that N_{gct} increases by ~17 N_{gct} per 1000 t^{elem} and that a REMD equilibration of 249 t^{elem} was required.[18]

2. 8-Br-GTP

8-Br-GTP has a higher conformational barrier between the anti and syn conformations than GTP, therefore we have found a higher value of λ_{max}=0.7 and the optimal REMD will probably require more middle replicas. For this reason we will not use a systematic search for the optimal settings as for GTP, but rather use a more efficient approach as described in Sec. II D 4.

We initiate the optimization of the middle λ's using five replicas at λ_{max}=0.7 and obtain $^1\lambda_1^{opt}$=0.5 as the optimal value if only one middle value of λ is taken into account, with N_{gct}=1079 per CPU (Table IV). When increasing the number of middle λ values to two, we restricted the searching ranges to $^2\lambda_1 \in \langle 0.05,0.5\rangle$ and $^2\lambda_2 \in \langle 0.5,0.65\rangle$. Similarly the optimizations were extended to three and four middle λ values. The N_{gct} per CPU for different numbers of middle λ values in combination with five replicas at λ_{max} are summarized in Table IV, where arrows indicate the steps in the optimization. As can be seen from this table, the optimal setting containing four middle λ values in combination with five replicas at λ_{max} is less efficient (1252 global conformational transitions per CPU) than the optimal three middle λ-values setting (1274). Because this may still be caused by too few conformational transitions at λ_{max}, the same optimal settings of four middle λ values [0.2; 0.4; 0.55; 0.65] was tested in combination with n=6,7,8 and indeed for n^{opt} $(m=4)=7$ we obtained a more efficient setting (1315 N_{gct} per CPU).

Taking into account the inequalities: $n^{opt}(m>4) \geq n^{opt}(m=4)=7$ and $n^{opt}(m<4) \leq n^{opt}(m=4)=7$ that are described in Sec. II D 4, the optimization procedure was continued by including five middle λ values in combination with n=7 and subsequently with n=8. It appears that $n^{opt}(m=5)=7$, but that its efficiency is lower (1305) than for the optimal four middle λ-values setting (1315 N_{gct} per CPU), meaning that the overall optimal scheme has ≤4 middle λ values with $n \leq 7$. Because the optimal three middle λ-values setting $[n^{opt}(m=3)=6]$ is also not giving a higher efficiency, the overall most efficient setting for 8-Br-GTP has thus been determined as $^4\lambda_i^{opt} \in [0.0;0.2;0.4;0.55;0.65;0.7;0.7;0.7;0.7;0.7;0.7;0.7]$.

N_{gct} per CPU (1315) is lower for 8-Br-GTP than for GTP (2108) indicating that REMD of 8-Br-GTP will be computationally more demanding. Using the optimal set for 8-Br-GTP of 12 replicas we estimate ~16 global conformational transitions during 1000 t^{elem} of real REMD (~2.5 ns) and the equilibration period to be ~180 t^{elem} (based on a linear extrapolation of time dependence of N_{gct} as function of time from ten individual runs of 5000 mimicked REMD cycles).

IV. DISCUSSION

The present study on the GTP/8-Br-GTP two state model systems deepens our understanding about REMD efficiency. For simplicity, we assumed that conformational transitions occur only at the highest lambda, λ_{max} in the case of H-REMD, or T_{max} for T-REMD. To obtain a different conformational state for a given replica at λ_0, the replica has to diffuse from λ_0 through the middle λ values to λ_{max}, where a conformational change has to occur. The system subsequently needs to diffuse from λ_{max} back to λ_0. Such process we named a global conformational transition. The advantage

TABLE III. Number of global conformational transitions, N_{gct} per CPU for mimicked REMD of GTP using $n=4,5,6$ replicas at $\lambda_{\max}=0.45$ and two additional λ-values $^2\lambda_1$, $^2\lambda_2$. The maximum is obtained for $^2\lambda_1=0.15$ and $^2\lambda_2=0.3$, at $n=5$.

$n=4$	$^2\lambda_1$				$^2\lambda_2$			
		0.1	0.15	0.2	0.25	0.3	0.35	0.4
	0.05	1614	1741	1726	1727	1697	1413	1269
	0.1	⋯	1787	1807	1879	1964	1692	1622
	0.15	⋯	⋯	1802	1898	**1990**	1807	1707
	0.2	⋯	⋯	⋯	1839	1988	1767	1761
	0.25	⋯	⋯	⋯	⋯	1841	1701	1719
	0.3	⋯	⋯	⋯	⋯	⋯	1678	1688
	0.35	⋯	⋯	⋯	⋯	⋯	⋯	1484
$n=5$	$^2\lambda_1$				$^2\lambda_2$			
		0.1	0.15	0.2	0.25	0.3	0.35	0.4
	0.05	1626	1814	1755	1766	1746	1446	1261
	0.1	⋯	1834	1871	1960	2030	1795	1644
	0.15	⋯	⋯	1813	1948	**2092**	1867	1803
	0.2	⋯	⋯	⋯	1945	2070	1845	1790
	0.25	⋯	⋯	⋯	⋯	1931	1776	1795
	0.3	⋯	⋯	⋯	⋯	⋯	1742	1740
	0.35	⋯	⋯	⋯	⋯	⋯	⋯	1470
$n=6$	$^2\lambda_1$				$^2\lambda_2$			
		0.1	0.15	0.2	0.25	0.3	0.35	0.4
	0.05	1562	1742	1742	1709	1702	1353	1167
	0.1	⋯	1781	1827	1920	2026	1722	1524
	0.15	⋯	⋯	1812	1917	**2075**	1830	1715
	0.2	⋯	⋯	⋯	1947	2069	1795	1648
	0.25	⋯	⋯	⋯	⋯	1906	1739	1714
	0.3	⋯	⋯	⋯	⋯	⋯	1715	1716
	0.35	⋯	⋯	⋯	⋯	⋯	⋯	1430

of monitoring the N_{gct} compared to the number of roundtrips between the lowest and highest temperature as proposed by Trebst *et al.*[14–16] comes from the fact that many roundtrips do not necessarily ensure any conformational changes at λ_0. For example for GTP, we have observed simulations in which there were many roundtrips thanks to relatively high values of $w^{\mathrm{anti,anti}}_{\lambda_i,\lambda_j}$, while $w^{\mathrm{anti,syn}}_{\lambda_i,\lambda_j}$ were very low. In such cases a syn conformation that occurs at $\lambda_{\max}$ will not be able to efficiently diffuse down to λ_0. Another advantage of counting N_{gct} is that if $P^{\mathrm{transit}}_{\lambda\neq\lambda_{\max}}(t^{\mathrm{elem}})\neq 0$ it does not require a replica to diffuse all the way up to $\lambda_{\max}$ because a conformational transition within real REMD can already occur at any λ value. The disadvantage of monitoring N_{gct} is that defined stable conformational states are required.

We described the approach for finding the optimal $\lambda_{\max}$ at which conformational transitions occur in sufficiently short time, which is, however, typically still much longer than the elementary switching period, t^{elem}, between switching attempts. The occurrence of conformational transitions shows a sigmoidal time dependency, from which a relatively long $t^{\mathrm{transit}}_{\lambda_{\max}}$, $t^{\mathrm{total}}_{\lambda_{\max}}$ can be estimated. The presented approach using n multiple replicas at $\lambda_{\max}$ allows several replicas to

TABLE IV. Number of global transitions, N_{gct} per CPU for the mimicked REMD of 8-Br-GTP, having optimal set of m middle lambdas $^m\lambda_i^{\mathrm{opt}}$ combined with n replicas at $\lambda_{\max}=0.7$. The maximum is obtained for $m=4$ and $n=7$. The arrows indicate the sequence of steps during the optimization procedure.

m	0		1		2		3		4		5
Set of $^m\lambda_i^{\mathrm{opt}}$	None		[0.5]		[0.4,0.55]		[0.35,0.5,0.65]		[0.2,0.4,0.55,0.65]		[0.2,0.4,0.45,0.55,0.65]
N_{gct}/CPU, $n=5$	159	→	1079	→	1263	→	1274	→	1252		
									↓		
N_{gct}/CPU, $n=6$							1301		1264		
							↑		↓		
N_{gct}/CPU, $n=7$							1297	←	1313	→	1305
									↓		↓
N_{gct}/CPU, $n=8$									1282		1266

spend the required time $t^{\mathrm{total}}_{\lambda_{\mathrm{max}}}$ at λ_{max} (in parallel). At the same time it allows for frequent switching attempts between $\lambda_0, \ldots, \lambda_i, \ldots \lambda_{\mathrm{max},0}$. In order to optimize the number of replicas at λ_{max} as well as the number and placement of λ values between λ_0 and λ_{max}, we present a REMD mimicking approach. As this approach propagates the conformations based on calculated probabilities, it is very fast and allows for simulations that correspond to μs time scales in real REMD.

This approach was inspired by the following analogy with real REMD. In real REMD "parallel" MD simulations are halted from time to time and switch their temperatures or Hamiltonians depending on the current energies. If we collect the switching probabilities into probability distributions then we can mimic the REMD simulation by sampling a switching probability from the probability distribution and subsequently perform the switch depending on this probability. Although real REMD switching probabilities converge to the same distribution, it builds up this distribution relatively inefficiently because for every structure at a particular time point, only one partner conformation is selected to attempt a replica switch. In our approach, all possible pairs of structures at a given set of λ_i, λ_j are used to estimate the distribution of switching probabilities. In many cases quite long REMD simulations are required to obtain a statistically sound probability distribution of switching probabilities. In this work we perform relatively short MD simulations at different λ's (H-REMD) or temperatures (T-REMD) starting from different conformational states and subsequently calculate the corresponding distributions of energy differences using different Hamiltonians or temperatures. These distributions are then used to calculate the probability distributions of switching probabilities. Together with estimates of the probability of conformational transitions at λ_{max} it allows us to mimic several μs of REMD simulations in a few minutes. We can then study the effect of different middle λ sets on the REMD efficiency, characterized by N_{gct} per CPU. For example, we clearly showed that increasing the number of replicas does not necessary increase the efficiency (per CPU) of REMD simulations. This means that "blind" brute force applications of REMD may be very inefficient.

A systematic search for the optimal λ settings that give the highest value of N_{gct} per CPU is affordable for simpler systems with a small number of replicas. However, the optimization process for complex systems requiring many replicas this can lead to a huge number of combinations. For such cases set we have suggested approaches to reduce the number of relevant combinations. The whole proposed optimization scheme is largely automated and fully parallelized.

Once the optimal settings have been found, REMD mimicking can make two more practical predictions for real REMD: it can estimate: (1) the equilibration time and (2) the required length of an REMD simulation to reach a given number of N_{gct}. For GTP this later time estimate showed an accuracy of about 20% in real REMD. This allows the user to estimate the feasibility of the real REMD simulation and to allocate the needed computational time in advance.

The most important advantage of the presented optimization scheme for REMD settings compared to the feedback-optimized parallel tempering approach[14–16] is its speed. While feedback optimization of the 36-residue villin headpiece subdomain HP-36 requires several years of CPU time[14–16] optimization by REMD mimicking can be performed within one week. This is a crucial factor for REMD simulations of complex system such as proteins. On the other hand, our approach requires some preliminary knowledge of stable conformational states, at least of the most dominant ones. Starting with an incomplete set of stable conformational states can lead to not completely optimal REMD settings. However, real REMD using partially optimized settings may suffice to reveal additional stable conformational states, which can then be used to refine the REMD settings. The whole process can be repeated iteratively until the real REMD leads to converged populations of the individual conformations. Note that one set of REMD generated conformations may yield part of the simulations that are used to calculate the probability distributions required for refining the λ settings.

We also showed that REMD mimicking is not only useful for the optimization of REMD settings, but also by itself can give us reasonable estimates of conformational populations. For GTP we obtained an excellent agreement for the syn population at $\lambda_0=0.0$ from mimicked REMD ($[\mathrm{syn}]_{\mathrm{aver}}=4.8\%$), as compared to real REMD ($[\mathrm{syn}]=4.4\%$) presented in our other work.[18] The advantage of mimicked REMD compared to thermodynamic cycle approaches is the fact that global conformational transition can occur through many different pathways, which are directly counted in a massively parallel manner in REMD mimicking using probability distributions of all combinations of switching probabilities. Usage of probability distributions, instead of average values takes into account the conformational variety of structures belonging to the same stable conformational region.

V. CONCLUSIONS

We have presented an algorithm that mimics replica exchange (REMD) simulations by a stochastic propagation in time of conformational states rather than explicit MD simulations. The approach was demonstrated for Hamiltonian REMD simulations on two model systems, GTP and 8-Br-GTP, for which two stable conformations are known, but the potential energy barrier separating them is too high to be crossed in normal MD. The method is, however, also readily applicable to temperature REMD.

We have shown that mimicked REMD can be used to serve three different purposes: (1) it can be used to obtain the optimal set of λ values by allowing an extremely fast attempt to try different REMD settings; (2) it gives the user an estimate of the simulation length in real REMD simulations, both for the equilibration of conformational states over the replicas and for the time required to obtain reasonably converged populations; and (3) it can make estimates of such populations directly which were shown to match remarkably well with real REMD simulations.

We are convinced that our method can contribute significantly to deepen our understanding of the processes governing REMD simulations. For many different applications, it will help to design more efficient REMD simulations for systems as are described in this work, but also for many more complex systems.

ACKNOWLEDGMENTS

This work was supported by the Netherlands Genomics Initiative in the context of Horizon Breakthrough Grant No. 050-71-043 (postdoc stay of J.H.).

[1] R. H. Swendsen and J. S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
[2] M. C. Tesi, E. J. J. van Rensburg, E. Orlandini, and S. G. Whittington, J. Stat. Phys. **82**, 155 (1996).
[3] K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604 (1996).
[4] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
[5] U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).
[6] Y. Sugita, A. Kitao, and Y. Okamoto, J. Chem. Phys. **113**, 6042 (2000).
[7] D. J. Earl and M. W. Deem, J. Phys. Chem. B **108**, 6844 (2004).
[8] D. A. Kofke, J. Chem. Phys. **117**, 6911 (2002).
[9] D. A. Kofke, J. Chem. Phys. **121**, 1167 (2004).
[10] A. Kone and D. A. Kofke, J. Chem. Phys. **122**, 206101 (2005).
[11] C. Predescu, M. Predescu, and C. Ciobanu, J. Chem. Phys. **120**, 4119 (2004).
[12] C. Predescu, M. Predescu, and C. Ciobanu, J. Phys. Chem. B **109**, 4189 (2005).
[13] N. Rathore, M. Chopra, and J. J. de Pablo, J. Chem. Phys. **122**, 024111 (2005).
[14] S. Trebst, D. A. Huse, and M. Troyer, Phys. Rev. E **70**, 046701 (2004).
[15] H. G. Katzgraber, S. Trebst, D. A. Huse, and M. Troyer, J. Stat. Mech.: Theory Exp. **2006**, P03018 (2006).
[16] S. Trebst, M. Troyer, and U. H. E. Hansmann, J. Chem. Phys. **124**, 174903 (2006).
[17] T. Läppchen, A. F. Hartog, V. A. Pinas, G. J. Koomen, and T. den Blaauwen, Biochemistry **44**, 7879 (2005).
[18] J. Hritz and C. Oostenbrink (in preparation).
[19] M. Christen, P. H. Hünenberger, D. Bakowies, R. Baron, R. Burgi, D. P. Geerke, T. N. Heinz, M. A. Kastenholz, V. Krautler, C. Oostenbrink, C. Peter, D. Trzesniak, and W. F. Van Gunsteren, J. Comput. Chem. **26**, 1719 (2005).
[20] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977).
[21] R. W. Hockney, Methods Comput. Phys. **9**, 136 (1970).
[22] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).
[23] I. G. Tironi, R. Sperb, P. E. Smith, and W. F. van Gunsteren, J. Chem. Phys. **102**, 5451 (1995).
[24] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. van Gunsteren, J. Comput. Chem. **25**, 1656 (2004).
[25] T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber, and W. F. van Gunsteren, Chem. Phys. Lett. **222**, 529 (1994).
[26] A. E. Mark, W. F. Van Gunsteren, and H. J. C. Berendsen, J. Chem. Phys. **94**, 3808 (1991).

Paper 7

Gabor Nagy, Chris Oostenbrink, Jozef Hritz*: Exploring the Binding Pathways of the 14-3-3ζ Protein: Structural and Free-Energy Profiles Revealed by Hamiltonian Replica Exchange Molecular Dynamics with Distance Field Distance Restraints. *PLoS ONE* **2017**,12(7), e0180633

RESEARCH ARTICLE

Exploring the binding pathways of the 14-3-3ζ protein: Structural and free-energy profiles revealed by Hamiltonian replica exchange molecular dynamics with distancefield distance restraints

Gabor Nagy[1¤], **Chris Oostenbrink**[2], **Jozef Hritz**[1]*

1 CEITEC-MU, Masaryk University, Brno, Czech Republic, **2** Institute for Molecular Modeling and Simulation, University of Natural Resources and Life Sciences, Vienna, Austria

¤ Current address: Department of Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany
* jozef.hritz@ceitec.muni.cz

Citation: Nagy G, Oostenbrink C, Hritz J (2017) Exploring the binding pathways of the 14-3-3ζ protein: Structural and free-energy profiles revealed by Hamiltonian replica exchange molecular dynamics with distancefield distance restraints. PLoS ONE 12(7): e0180633. https://doi.org/10.1371/journal.pone.0180633

Editor: Yaakov Koby Levy, Weizmann Institute of Science, ISRAEL

Received: September 27, 2016

Accepted: June 19, 2017

Published: July 20, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by Czech Science Foundation, no. GF15-34684L (J.H., G.N.). J.H. was also supported from the SoMoPro II Programme, co-financed by the European Union and the South Moravian Region (Czech Republic). The paper reflects only the author's views and the Union is not liable for any use that may be made of

Abstract

The 14-3-3 protein family performs regulatory functions in eukaryotic organisms by binding to a large number of phosphorylated protein partners. Whilst the binding mode of the phosphopeptides within the primary 14-3-3 binding site is well established based on the crystal structures of their complexes, little is known about the binding process itself. We present a computational study of the process by which phosphopeptides bind to the 14-3-3ζ protein. Applying a novel scheme combining Hamiltonian replica exchange molecular dynamics and distancefield restraints allowed us to map and compare the most likely phosphopeptide-binding pathways to the 14-3-3ζ protein. The most important structural changes to the protein and peptides involved in the binding process were identified. In order to bind phosphopeptides to the primary interaction site, the 14-3-3ζ adopted a newly found wide-opened conformation. Based on our findings we additionally propose a secondary interaction site on the inner surface of the 14-3-3ζ dimer, and a direct interference on the binding process by the flexible C-terminal tail. A minimalistic model was designed to allow for the efficient calculation of absolute binding affinities. Binding affinities calculated from the potential of mean force along the binding pathway are in line with the available experimental estimates for two of the studied systems.

Introduction

14-3-3 proteins are important regulatory factors found in all kingdoms of life and are vital for the survival of higher organisms. In mammals, the 14-3-3 family consists of seven isoforms that can be found in large abundance within the brain. The human 14-3-3ζ isoform was selected for this project because of its high biological relevance. 14-3-3 proteins function

the information contained therein. G.N. also acknowledges the BioTop doctoral exchange program, funded by the Austrian Science Fund (FWF; grant no. W1224). C.O. gratefully acknowledges the financial support of the Austrian Science Foundation (FWF; grant no. I 1999-N28) and the European Research Council (ERC; grant no. 240608). The computational simulations were realised in the National Supercomputing Center IT4Innovations, which was supported by The Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project–LM2015070. Further computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures." The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

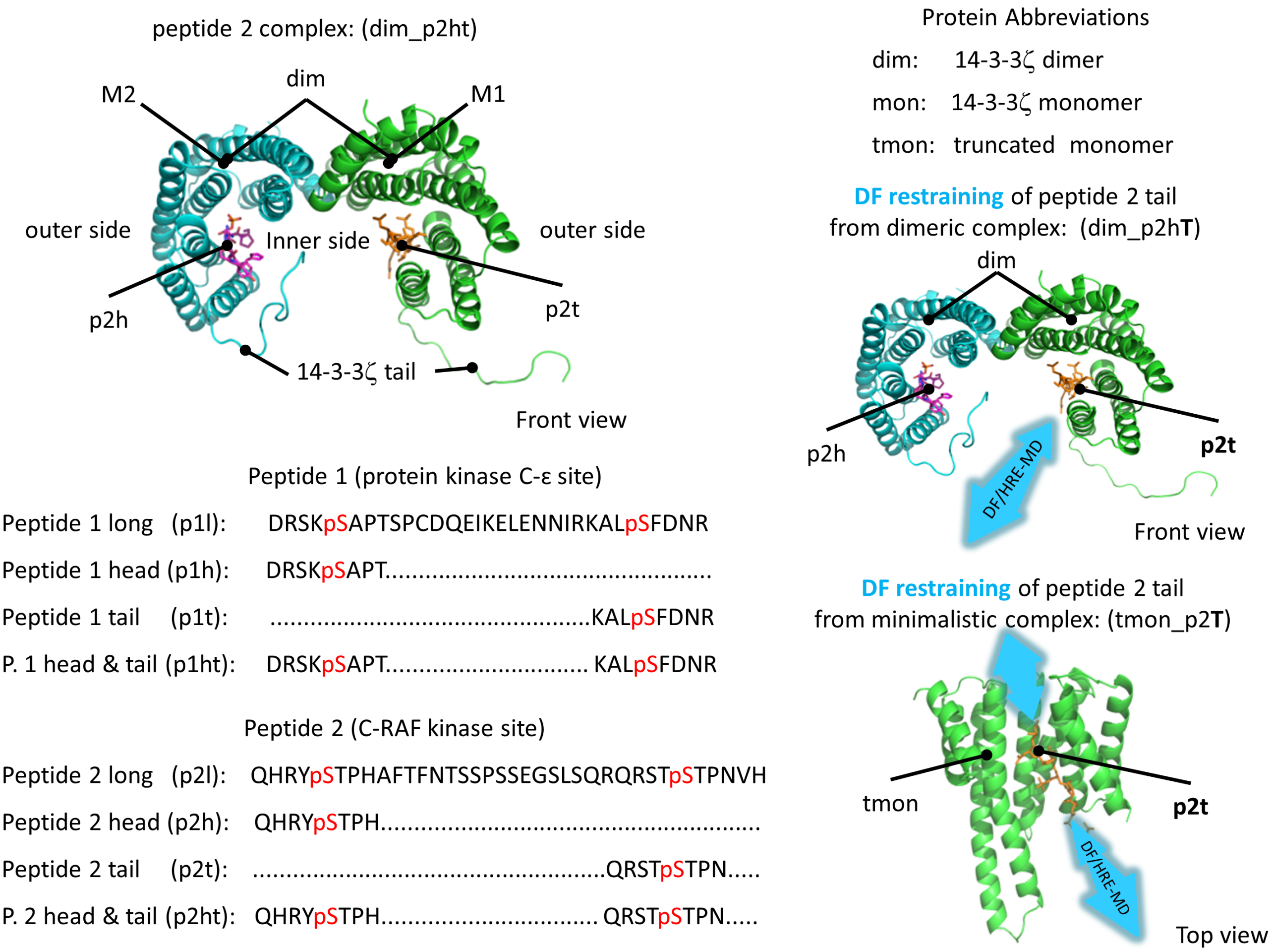
Competing interests: The authors have declared that no competing interests exist.

mainly as dimers, which are composed of two 28-kDa monomers that are both capable of binding phosphorylated serine (pS) and threonine (pT) motifs in other proteins. Crystal structures of all seven mammalian homodimers are now available and show that each monomer is composed of nine α-helices, arranged in an antiparallel fashion. The helices form an amphipathic groove that mediates pS and pT target binding [1]. Most 14-3-3 targets have two phosphoserine/threonine-containing motifs with a consensus sequence RSXpSXP (mode I) or RX[FY]XpSXP (mode II) [2], representing the optimal recognition sites for 14-3-3. Upon binding to these sites 14-3-3 proteins induce conformational changes in their target protein, (and 'finish the job') when phosphorylation alone may lack the power to drive the necessary allosteric changes for modulating the activity of an intracellular protein. Owing to their dimeric nature, 14-3-3 proteins are capable of distinguishing between non-, single- and double- phosphorylated binding protein partners; in this sense 14-3-3 proteins are considered to act as coupled binary devices [3]. Recently, we have presented an approach based on experimental ^{31}P NMR spectroscopy which revealed that a double-phosphorylated protein can be complexed with the 14-3-3ζ dimer in a much more dynamic fashion than was originally thought. In addition to the traditionally considered single partner with two phosphorylation sites occupying the individual binding cavities within the 14-3–3ζ dimer, two more major binding modes were confirmed [4]. All that is currently known about the structural features of the phosphopeptide binding to 14-3-3 proteins originates from the available crystal structures of 14-3-3 proteins in apo and holo state. Comparison of the various 14-3-3 crystal structures also revealed the different width of the main peptide binding groove, thus suggesting a dynamic opening process [5].

Here, we present the structural and energetic features of selected phosphopeptides along their binding/unbinding pathways to/from the 14-3-3ζ binding site, determined by enhanced sampling computational approaches. The main applied methodology is based on Hamiltonian replica exchange molecular dynamics (HRE-MD) combined with distancefield (DF) distance restraints [6], which has several advantages over the more conventional umbrella sampling and distance restraints approaches. HRE-MD is a highly parallel perturbed molecular dynamics (MD) technique, whereby each parallel simulation (replica) represents a discrete state along a thermodynamic pathway. The simulation of replicas are independent of each other, however, at given time periods the replicas can exchange their coordinates to allow the ensembles of each replica to include conformations derived from multiple starting coordinates. The application of HRE-MD instead of a set of individual MD simulations with different distance restraints significantly enhances sampling efficiency. HRE-MD could be combined with regular distance restraints, however, such an approach could lead to protein damage when coordinates at larger Cartesian distances switch to shorter distances [7]. Applying DF restraints avoids the protein damage whilst the ligand is pulled into the binding site by using an altered reaction coordinate, for which the distance to the binding site is defined by the shortest sterically possible pathway. Combining DF restraints and HRE-MD allows the simulation coordinates of the replicas to traverse reversibly between the various states of the binding pathway and represent conformations in structural ensembles restrained at different DF distances that would otherwise require much longer simulation times due to kinetic barriers. The set of thermodynamic ensembles that is generated is used to determine the structural features of the peptide-binding pathways as well as the corresponding potential-of-mean-force (PMF) profiles which can be used for the calculation of the absolute binding affinities.

Results and discussion

We studied the binding of phosphorylated peptides to the 14-3-3ζ protein by constructing two different models of 14-3-3ζ, one containing a full length 14-3-3ζ dimer (dim) and one



Peptide 1 (protein kinase C-ε site)

Peptide 1 long (p1l):	DRSKpSAPTSPCDQEIKELENNIRKALpSFDNR
Peptide 1 head (p1h):	DRSKpSAPT...
Peptide 1 tail (p1t):	....KALpSFDNR
P. 1 head & tail (p1ht):	DRSKpSAPT.............................. KALpSFDNR

Peptide 2 (C-RAF kinase site)

Peptide 2 long (p2l):	QHRYpSTPHAFTFNTSSPSSEGSLSQRQRSTpSTPNVH
Peptide 2 head (p2h):	QHRYpSTPH..
Peptide 2 tail (p2t):	....QRSTpSTPN.....
P. 2 head & tail (p2ht):	QHRYpSTPH................................ QRSTpSTPN.....

Fig 1. 3D representation, nomenclature and sequence of the model molecules used during the simulations. Dimer and monomer 14-3-3ζ systems (on the right) were simulated with fragments of phosphopeptide 1 and 2 (sequence shown on the left). The phosphoserine in the sequence is highlighted in red. 14-3-3ζ monomers are represented as cartoons in green and cyan, phosphopeptides are shown in stick representations in orange and purple. The phosphopeptide under DF restraining is labelled by an upper-case letter in the abbreviations.

https://doi.org/10.1371/journal.pone.0180633.g001

including only a truncated 14-3-3ζ monomer (tmon) without the flexible C-terminal stretch (tail, aa. 230–245). The proper simulation of a full length 14-3-3ζ dimer requires a large simulation box in order to avoid artificial periodic effects arising from the opening motion of 14-3-3ζ and the flexible C-terminal regions. Furthermore, the presence of the flexible C-terminal stretch also significantly slowed down the convergence of calculated free energies, as is demonstrated by the PMF calculations (see below). Four different phosphopeptide fragments are used as models for this study which were derived from the diphosphorylated PKC-ε and C-RAF kinase binding sites for 14-3-3ζ, and are referred to as the head and tail fragments of peptides 1 and 2, respectively (p1h, p1t, p2h, p2t), based on their location in the full protein sequence. These phosphopeptide fragments were chosen because their crystal structure bound to 14-3-3ζ was available and their binding affinities were previously measured [8–10].

Phosphopeptide sequences, abbreviations and naming conventions used in this work are given in Fig 1. The monomers of 14-3-3 proteins are horse-shoe shaped, and the first four α-helices form the dimerization interface. The two monomers of a 14-3-3ζ dimer are denoted as M1 and M2 when a distinction is necessary. The helices three, five, seven and nine of each monomer form the phosphopeptide binding grooves on the inner side of the 14-3-3ζ dimer. When referring to our simulations we denote our simulated complex by first defining the 14-3-3ζ model (dim or tmon) followed by peptide fragment present in the simulation (either p1h, p1t, p2h, p2t or nothing, in case of apo simulations). In the case of dimeric simulations, if two peptide fragments are present we will denote it by p1ht or p2ht representing either peptide 1 (region 342–372 of protein kinase C-ε) or peptide 2 (region 229–264 of C-RAF kinase) fragments. In addition, when DF distance restraints are applied to pull one of the phosphopeptide fragments from its respective binding groove, we mark this fragment with an upper-case letter (e.g. tmon_p2T or dim_p2hT).

Our results are divided into three sections. The first section describes phosphopeptide binding pathways elucidated from the DF/HRE-MD simulations. We characterized the binding/unbinding pathways of studied 14-3-3ζ/phosphopeptide complexes by determining the most important protein-peptide interactions between 14-3-3 and its binding partner, and identifying the regions most frequently populated by the restrained phosphopeptide. In the second section we focus on the structural changes along the binding pathways for both 14-3-3ζ and the phosphopeptide. Here we aim to identify the large-scale protein motions, which may be important for the phosphopeptide and protein binding. The third section presents the free energy changes along the binding pathways. The determined free-energy profiles allowed us to calculate estimates for the corresponding phosphopeptide binding affinities.

1. Major binding pathways of the 14-3-3ζ protein

In this section, we have elucidated the phosphopeptide binding pathways of 14-3-3ζ, using DF/HRE-MD simulations. We achieved this by gradually pulling one of the phosphopeptides bound to the 14-3-3ζ from its respective binding site in a reversible HRE-MD process. During the DF/HRE-MD simulations replicas with increasing index numbers restrained the phosphopeptide to regions with DF distances further away from the peptide binding groove (interaction site 1, IS1). Note that for every dimeric 14-3-3ζ complexed with two phosphopeptide fragments (e.g. p1h and p1t in dim_p1Ht) one phosphopeptide fragment (p1h in case of dim_p1Ht) per process was pulled from its monomer, whilst the other (p1t in case of dim_p1Ht) remained in its respective monomer binding site (Fig 1). Such a setup avoids complications arising from the two available binding sites (for p1h) within one 14-3-3ζ dimer.

1.1. Characterization of the phosphopeptide binding pathways. In order to characterize the phosphopeptide binding pathways of 14-3-3ζ, we monitored the position and interactions of the phosphopeptide during its binding/unbinding process. Fig 2 summarizes the results of such an analysis for simulation dim_p1hT (S1–S7 Figs corresponds to the other 7 DF/HRE-MD simulations). The position of the DF-restrained phosphopeptide was tracked on a three dimensional grid through its virtual atom (as described in Methods section) near the pSer residue during the DF/HRE-MD simulations (Fig 2A). The most occupied/preferred positions of the phosphopeptide at various DF distances were aligned along a dominant binding/unbinding pathway (as shown in Fig 2B). Similar dominant pathways were observed for seven out eight simulations (as shown in S1B–S7B Figs).

The analysis of phosphopeptide-protein interaction energies (Fig 2C) shows a local minimum around the DF distance of 3.5–6.0 nm, in addition to the global minimum of original binding site at the DF distance of about 0.2 nm. The local interaction energy minimum co-occurred with stable, specific interactions with certain 14-3-3ζ residues around amino acids 64–70 (Fig 2D). Similar characteristic interaction energy minima were also observed for all seven simulations with a dominant pathway, including stable interactions between the phosphopeptide and the same 14-3-3ζ residues (64–70, in S1D–S6D Figs), suggesting a previously

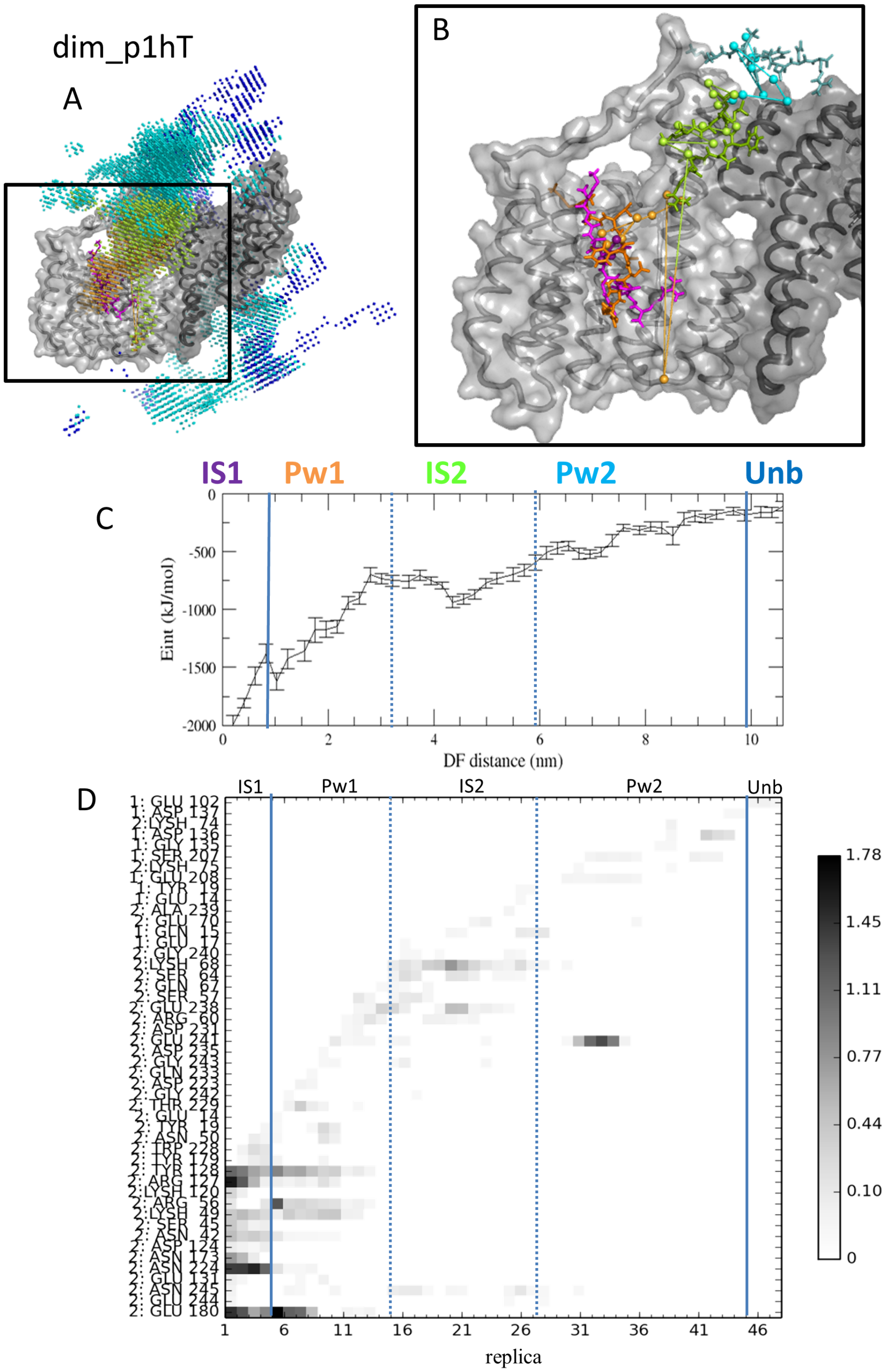


Fig 2. Pathway visualization of the DF/HRE-MD simulation dim_p1hT. A) The volume sampled by the p1t phosphopeptide during the simulation is shown by dots around the 14-3-3ζ protein coloured based on their position along the pathway. The most probable points in space to find the peptide in for each replica (probability density peaks) are represented by larger spheres. B) Density peaks and a few representative structures corresponding to the density peaks, coloured according to their position along the pathway. Replica density peaks are connected by lines to visualize the binding pathway. The 14-3-3ζ protein in panels A-B is shown as a surface representation, with the two monomers shown in light and dark grey. C) Average interaction energy between 14-3-3ζ and p1t. D) Interaction map between any atom of the pulled p1t peptide and the amino acids of the 14-3-3ζ protein, summarized for each replica (only amino acids with at least 0.1 hydrogen bond/salt bridge on average are shown). The scale indicates the average number of interactions.

https://doi.org/10.1371/journal.pone.0180633.g002

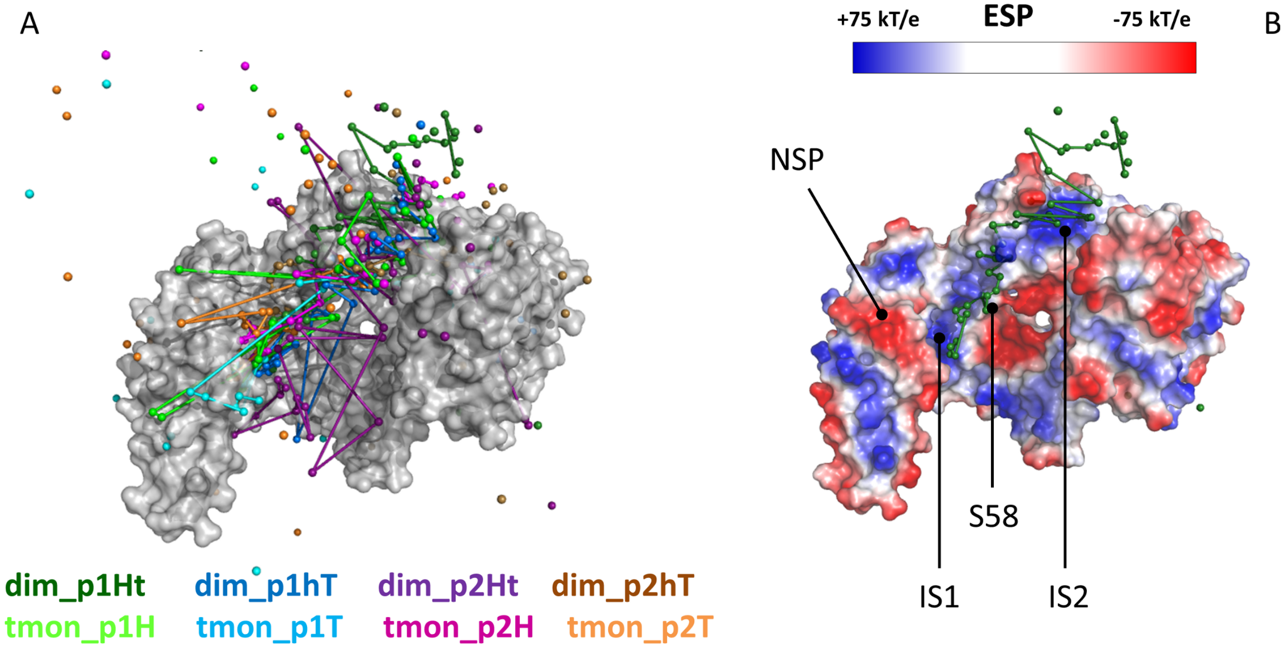


Fig 3. Comparison of binding pathways of different phosphopeptides. A) Binding pathways from DF/HRE-MD simulations depicted as connected dots referring to the probability density peaks along the respective pathway (See Table 1 for more information). B) Electrostatic surface potential (ESP) of 14-3-3ζ in blue, white and red for positive, neutral, and negative surface patches, respectively. The positively charged main interaction site (IS1), and secondary interaction site (IS2) are connected by a positive surface along the binding pathways. A negative surface patch (NSP) involved in the binding process is also indicated. Serine58 (S58), a phosphorylation site is located near the binding pathway. For clarity, the 14-3-3ζ C-terminal tail is not shown.

https://doi.org/10.1371/journal.pone.0180633.g003

unknown secondary interaction site. To better characterize our binding pathways, we divided the pathways into five parts; starting from the primary interaction site (IS1) within the main phosphopeptide binding groove, followed by the early pathway (Pw1) between the two interaction sites, the vicinity of the secondary interaction site (IS2), the late pathway (Pw2) after IS2, and the unbound state (Unb), where the pathway becomes diffuse and no significant interactions with 14-3-3ζ are observed for the phosphopeptide.

Experimental support for the existence of the IS2 site can possibly be found in previously published ^{31}P NMR titration data of the doubly phosphorylated (at positions 19 and 40) human tyrosine hydroxylase 1 peptide binding to 14-3-3ζ [4]. Fig 2A of the referred article shows an "unexpected" peak labelled as pS40*, which was different from pS40 in the free state or bound to IS1. One of the possible explanations is that while pS19 is bound to IS1, pS40 can interact with IS2—still being partially solvent exposed—resulting in the additional peak.

1.2. Comparison between phosphopeptide pathways. Comparing the binding pathways obtained for the four studied phosphopeptide fragments (with both 14-3-3ζ models), we found a remarkable similarity between their most probable pathways in 7 out of 8 simulations (with the exception of dim_p2Ht, Fig 3A). The dominant pathways generally followed helix three, "sliding" from IS1 in the primary binding groove to IS2, and becoming increasingly diffuse before the final detachment of phosphopeptide and reaching the Unb state. This general behaviour indicates the existence of one dominant pathway for all four phosphopeptides, observed both in the dimeric and truncated monomeric simulations. Examining the electrostatic surface potential (ESP) in multiple 14-3-3ζ conformations showed both IS1 and IS2 display highly positive surface potentials; and the connecting region along helix three also exhibits a mildly positive surface potential (Fig 3B). Thus, these areas can form favourable electrostatic interactions with a negatively charged phosphoserine side chain, which could explain the presence of a dominant binding pathway.

In addition, based on the ESP we identified a negative surface patch (NSP) which was often found to accommodate the positively charged side chain in position -3 or -2 relative to the phosphorylation site of the peptide fragment. The NSP is primarily composed of residues Y179, E180, N183, D223, N224, L227 and W228. The NSP provides an additional anchor point on the binding pathway and also concurs with the strong selection of positive amino acids in this position as reported by Jaffe *et. al.* [2].

It was previously reported [11], that phosphorylation and mutation of S58 into negative amino acids lead to dimer dissociation and a (usually negative) change in the 14-3-3 binding affinity, independently of the oligomeric state. S58 is located near the dimer interface of 14-3-3ζ and directly next to the binding pathway between IS1 and IS2. A negatively charged residue in position 58 could negate the mildly positive surface potential (Fig 3 and S13 Fig), and disrupt the observed binding pathway, which is in line with the observed experimental changes.

We identified the amino acids of 14-3-3ζ which were important phosphopeptide interaction partners at each of the five stages of the binding pathway. Most interactions between 14-3-3ζ and the phosphopeptides were polar hydrogen bonds or salt bridges. During our simulations, as the phosphopeptide sampled the available volumes along the dominant pathway, the overall probability to find a particular 14-3-3 amino acid interacting with the restrained phosphopeptide was calculated (shown in S8 Fig) as function of the replica number (restraining DF distance). In Fig 4 we display the table of all the 14-3-3ζ residues found significant (over 10% probability at a given replica) in this analysis. The residues are coloured according to the stage in which the residue was found most significant.

In dark purple we have depicted residues found most significant while the phosphopeptide is bound in IS1 including many residues previously identified as crucial interaction partners (such as Lys 49, and Arg 127) based on 14-3-3 crystal structures, and residues from the negative surface patch such as Asp 224, which mostly interacted with the positively charged residues of the phosphopeptide. The second group of amino acids (shown in orange) were found most prominent on the early pathway (Pw1) between IS1 and IS2. Many of these residues were probable interaction partners already when the peptide was in IS1 (such as Glu 180 and Arg 56), whilst other residues became significant only as the phosphopeptides left the primary binding site (e.g. Glu 113 and Asp 124).

The third group of residues were most prominent whilst the phosphopeptide visited IS2. Most of these residues (e.g. Ser 64, Lys 68 and Lys 75) are located in the positive surface patch at the end of helix three (marked with light green in Fig 4), near the dimerization interface. There were also three residues (Lys 11, Ser 207 and Glu 209, shown light blue) considered as important interaction partners only during our dimeric DF/HRE-MD simulations. In dimeric simulations these residues from the other, 14-3-3 monomer enhanced and expanded the positive surface patch at IS2, and prevented the phosphopeptide to diverge from a dominant pathway for ~1.0 nm longer in the distance field during the late pathway stage (Pw2) compared to monomeric simulations.

The final group of interacting residues, were located on the disordered C-terminal region (aa. 231–245) of the monomer from which the peptide is being pulled, shown in brown in Fig 4. The bulk of the tail interactions were observed in the PW1, IS2 and Pw2 stage of the binding pathway (including the salt bridges with Glu 241 and Glu 244), although some of the tail residues could interact with the phosphopeptide fragments while they were bound to IS1 (Thr 232, Gly 234 and Ala 237, most prominently with the p2h).

Fig 4 also highlights that the occurrence of phosphopeptide—tail interactions was very high in two of the DF/HRE-MD simulations (dim_p1hT and dim_p2Ht) and very low in the other

Position	residue		dim_p1Ht	dim_p1hT	dim_p2Ht	dim_p2hT	tmon_p1H	tmon_p1T	tmon_p2H	tmon_p2T
H3	ASN	38	+	-	-	++	+	+	+	-
H3	ASN	42	-	++	+	++	+	++	++	+
H3	SER	45	+	++	+	++	++	++	++	++
H3	LYS	49	+++	+++	+	+++	+++	+++	+++	+++
H3	ARG	56	++	++	+	+++	-	+	+	+
H3	SER	57	-	+	-	+	-	-	+	-
H3	ARG	60	+	++	-	+++	++	+	++	+
H3	SER	64	++	++	-	++	-	++	++	-
H3	GLN	67	+	+	-	+	-	+	++	+
H3	LYS	68	+++	+++	-	+++	+++	++	+++	++
H3	THR	69	++	-	-	-	++	-	+	++
H3	GLU	70	++	+	++	++	+	+	+++	+
H4	LYS	75	+	+	-	-	+	-	+	-
H5	GLU	113	-	-	+++	++	-	-	-	-
H5	LYS	120	+	+	+	++	+	+	++	+
H5	ASP	124	-	+	-	+	-	++	-	+
H5	ARG	127	-	+++	+	++	+	+++	++	++
H5	TYR	128	+	+++	+	+++	+++	+++	++	++
H5	GLU	131	+	+	-	++	-	+	+	++
H7	ASN	173	+	++	+	+++	++	+++	++	++
H7	TYR	179	++	+	+	++	++	++	++	+++
H7	GLU	180	+++	+++	+	+++	+++	++	+++	+++
H7	ASN	183	+	-	-	-	++	-	+	++
H9	GLN	219	-	-	++	+	-	-	-	-
H9	ASP	223	+	+	++	+	++	-	++	-
H9	ASN	224	+++	+++	+	++	+++	+++	+++	+++
H9	THR	226	-	-	+	-	-	+	+	-
H9	LEU	227	+	-	+	++	+	+	++	+
H9	TRP	228	+++	+	+	++	+++	++	+	+++
L9	THR	229	-	+	+	-	+	-	+	+
L9	SER	230	-	-	-	+	++	-	+	+
T	THR	232	-	-	++	-	n/a	n/a	n/a	n/a
T	GLN	233	-	+	+++	-	n/a	n/a	n/a	n/a
T	GLY	234	-	-	++	-	n/a	n/a	n/a	n/a
T	ASP	235	-	+	+++	-	n/a	n/a	n/a	n/a
T	ALA	237	-	-	+++	-	n/a	n/a	n/a	n/a
T	GLU	238	-	++	+	-	n/a	n/a	n/a	n/a
T	ALA	239	-	+	+	-	n/a	n/a	n/a	n/a
T	GLY	240	-	+	+	-	n/a	n/a	n/a	n/a
T	GLU	241	-	+++	+++	-	n/a	n/a	n/a	n/a
T	GLY	242	-	+	++	-	n/a	n/a	n/a	n/a
T	GLU	244	-	+	+++	-	n/a	n/a	n/a	n/a
T	ASN	245	-	++	+	+	n/a	n/a	n/a	n/a
H1	LYS	11	++	-	-	+	n/a	n/a	n/a	n/a
L8	SER	207	-	++	-	+	n/a	n/a	n/a	n/a
L8	GLU	209	+	-	-	++	n/a	n/a	n/a	n/a

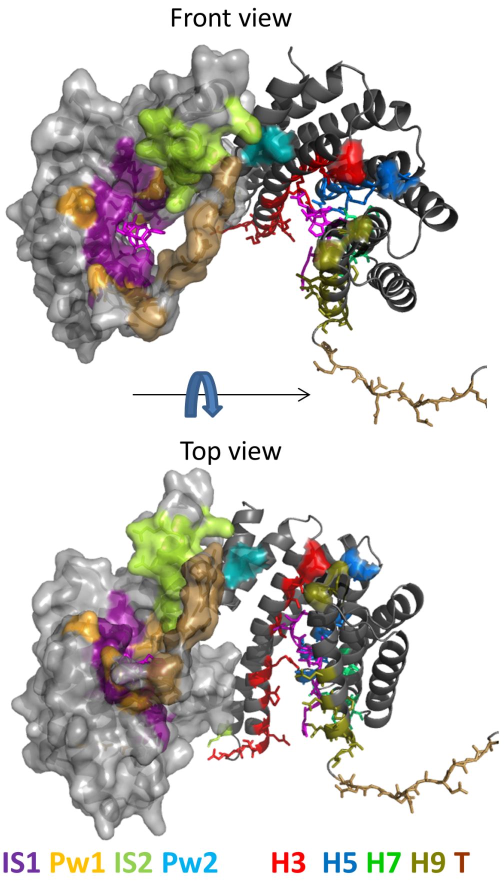


Fig 4. 14-3-3ζ residues involved in the phosphopeptide interactions. The table on the left side shows the 14-3-3ζ amino acids, which were found important based on the interaction map analysis (Fig 2, S1–S7 Figs) of the DF/HRE-MD simulations. Entries in the various simulations are marked as not applicable (n/a), not significant (-), significant (+), important (++), or major interaction partners (+++). The last three amino acids in the table were interaction partners from the other monomer. The right side shows the three dimensional structure of the 14-3-3ζ protein, where the monomers are depicted in cartoon and surface representations, respectively. The amino acids are depicted in stick representation, coloured differently for the two monomers. In the cartoon representation, peptide-interacting amino-acids are coloured according their position in the secondary structure, where helices 3, 5, 7 and 9 are shown in red, blue, green and brown, respectively, while the C-terminal stretch is depicted in light purple. For the surface representation, amino acids found significant only in the bound-state replicas (IS1) are shown as dark purple, amino acids along the binding pathway are marked in orange, light green and light blue, if they appeared prior, in, or after the secondary interaction site (IS2). See S8 Fig for details.

https://doi.org/10.1371/journal.pone.0180633.g004

two cases (dim_p1hT and dim_p2hT). The strongest interactions with the 14-3-3 tail region were observed for the simulation dim_p2Ht, which at the same time had the weakest interactions with IS2 and did not follow the dominant phosphopeptide binding pathway observed during the other seven DF/HRE-MD simulations (S7 Fig). These observations suggest that C-terminal tail may disrupt the dominant peptide binding pathway.

1.3. Comparing the pathways within 14-3-3ζ dimeric and truncated monomeric model. The summary of our DF/HRE-MD simulations indicates a dominant phosphopeptide binding pathway for 14-3-3ζ where the phosphopeptide in the process of unbinding "slides" from the IS1 towards IS2 along the described pathway shown in Figs 3 and 4. Here, we summarize the differences observed between our full dimeric systems (dim) and the minimalistic, truncated, monomeric systems (tmon) during the detailed analysis of preferred positions and phosphopeptide interactions (shown in Figs 2–4 and S1–S8 Figs).Our minimalistic systems lacked the neighbouring monomer and the C-terminal region of the 14-3-3ζ protein, which, in the case of full dimeric systems, restricted the accessible space for the peptides and necessitated longer DF distances to reach the unbound state.

The truncated monomer pathways were more similar to each other near the main peptide binding groove (IS1), and showed enhanced interactions with helices 7–9, which were partially blocked by the C-terminal region in the dimers. The monomeric pathways fanned out after leaving the IS2 around 5 nm in the DF and sampled the outer side of 14-3-3ζ monomer before final detachment from the 14-3-3ζ surface, and showed no significant interactions above 7 nm.

Full dimeric simulations lead to a more complete picture of 14-3-3ζ binding pathways and could yield additional, biologically relevant information. The pathway comparison revealed a more localized phosphopeptide presence for all dimeric systems near the dimerization interface, along with new interaction partners from the other monomer (K11, S207, and E209). During dimeric HRE-MD simulations which followed the dominant pathway, phosphopeptides occupied the secondary interaction site (IS2) for a wider range of distances than their monomeric counterparts and had non-negligible interactions between the restrained peptide and the 14-3-3ζ dimer for DF distances smaller than 9.5 nm.

The interactions between the phosphopeptide and the C-terminal tail of the restrained 14-3-3 monomer showed a large variation between the four full dimeric simulations. In case of the p2h peptide fragment, strong interactions with the tail seemed to prevent the phosphopeptide to follow the dominant pathway (S7 Fig), which is a major difference compared to the truncated simulation with the same peptide (S2 Fig). The signs of direct interference from the C-terminal tail, and the tail-associated discrepancies in the full dimeric phosphopeptide binding simulations necessitated further investigation, shown below.

2. Structural changes along the binding pathway

The structural ensembles generated at different replicas within the DF/HRE-MD allowed us to analyse structural changes as function of the DF distance from IS1. The results of this analysis allowed us to monitor the slow conformational degrees of freedom which may affect the binding/unbinding process of 14-3-3ζ. In the following section we identify these large scale motions, and compare our DF/HRE-MD simulations in the bound (IS1) and unbound (Unb) states with unrestrained MD simulations under the same conditions and—whenever possible—experimental observations. This analysis allowed us to explore the observed discrepancies, assess the effect of these large scale motions on the phosphopeptide binding process during the limited sampling of our simulations, as well as effects of the applied DF/HRE-MD restraints on the system.

2.1. The C-terminal region. The 14-3-3ζ proteins contain a 15 amino acid long C-terminal region, which is a highly flexible disordered segment of the protein (Fig 5). In our simulations two types of models were employed: full length dimeric (dim) models, and truncated monomeric (tmon) models lacking the C-terminal tail. The C-terminal regions in our full-length models visited both the inner (dimer interface formed by helices 3–4 of both monomers) and outer (helices 7–9) side of the 14-3-3ζ protein surface, as well as free conformations in which the tail was detached from the protein surface (Fig 5A). The position of the C-terminal tail in the simulations was determined based on the distances measured between the terminal carbonyl (of N245) and the Nζ atom of K68 (in IS2), Nζ of K120 (in IS1), and Cγ of D197 (on the outer side), with cut-off distances 2.5, 2.0, and 2.5 nm, respectively. The tail was considered to be on the inner side of 14-3-3ζ if the terminus was within the cut-off distance of IS1 or IS2, considered to be on the outer side if it was within the cut-off distance of D197 but not the other two residues, and was considered free otherwise. The detailed population

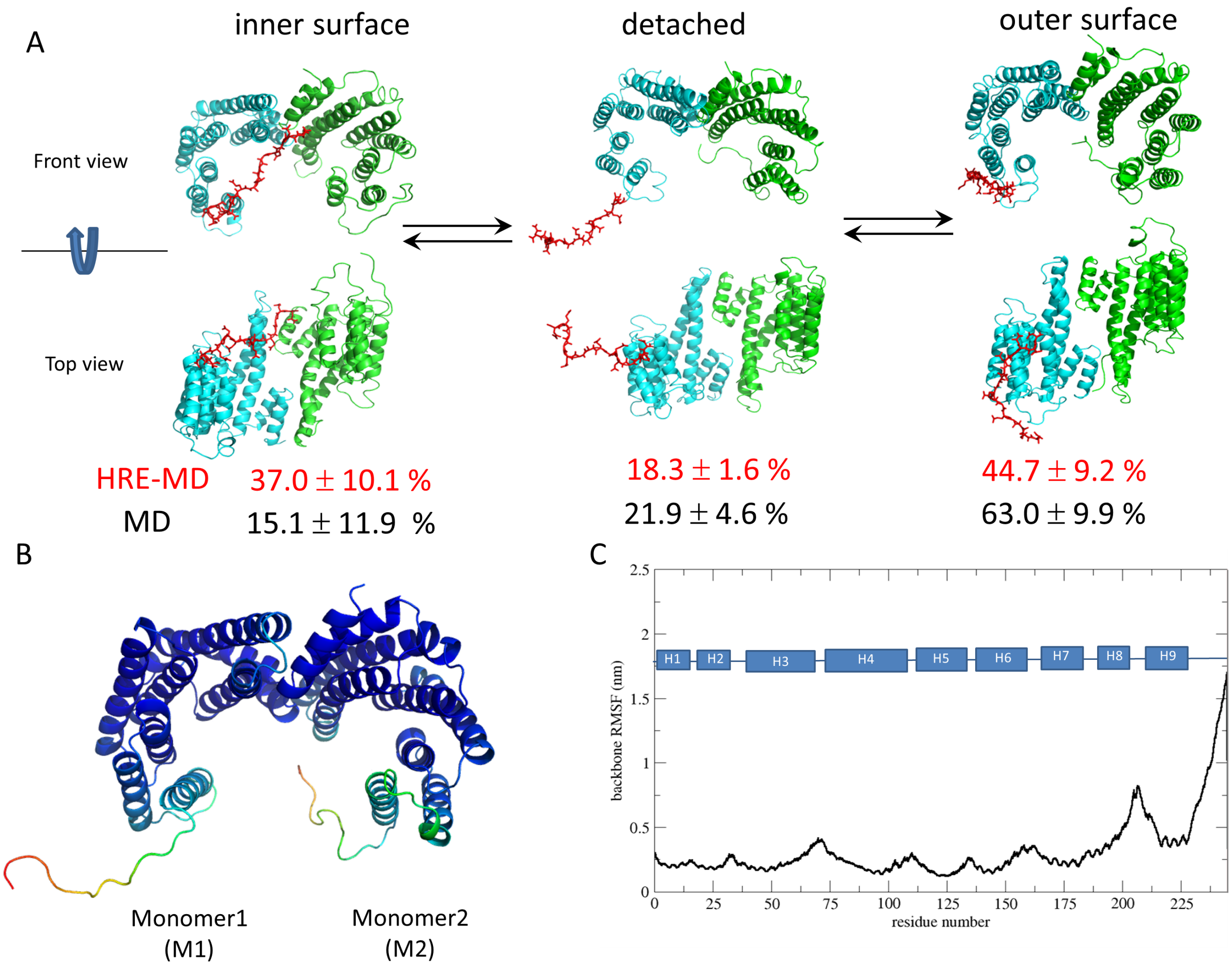


Fig 5. Flexibility of the 14-3-3ζ C-terminal tail. A) Representative structures of the conformational sub-states of the C-terminal stretch (marked in red) interacting with the inner or outer surface of 14-3-3ζ or being detached and exposed to the solvent. Populations of these conformational states as obtained from DF/HRE-MD (in red) and MD (in black) simulations are listed under the cartoon figures. B) 14-3-3 model starting conformation, coloured according to the backbone RSMF, where blue and red represent the least and most flexible amino-acids, respectively. C) The atom-positional root-mean-square fluctuations (RMSF) of the 14-3-3ζ backbone atoms.

https://doi.org/10.1371/journal.pone.0180633.g005

distributions and number of transitions between the four states are shown in S2 and S3 Tables for HRE-MD and MD simulations, respectively.

The exchange between the three conformational sub-states of the protein tail was a process observed, but not properly sampled on the time scales of our DF/HRE-MD simulations. Analysis of the tail residence during the DF/HRE-MD simulations (S2 Table) revealed a strong bias towards the conformation of the tail at the start of the DF/HRE-MD run (shown in Fig 5B). The tail of monomer 1 (M1) at the start of the simulations was facing outwards, while the tail of monomer 2 (M2) was facing towards the inner side of the 14-3-3 dimer (in both cases the starting conformation was detached from the surface). During the simulations each C-terminal region of M1 dominantly sampled free conformations and the outer surface of 14-3-3ζ, while tail of M2 sampled free conformations and the inner surface. A few transitions from the inner to the outer side were observed for the M2 tail, suggesting a slow conformational transition and preference towards the outer surface.

The bias of the preferred C-terminal tail conformation also explains the differences observed in the phosphopeptide interaction analysis. During the simulations dim_p1Ht and dim_p2hT the peptide fragment was pulled from M1, and since the C-terminal tail was mostly located on the outer side of the 14-3-3 dimer, no or very little interaction between the tail and the phosphopeptide was observed (Fig 4). On the other hand, for the simulations dim_p1hT and dim_p2Ht the peptide was pulled from M2 and the tail was close to the inner surface and the phosphopeptide binding pathway, allowing for strong tail-peptide interactions.

The explanation for the difference between the p2h binding pathways in tmon_p2H (S2 Fig) and dim_p2Ht (S7 Fig) is that the C-terminal tail partially buried and neutralized the positive surface potential of IS2 in dim_p2Ht. This prevented the p2h fragment in dim_p2Ht from following the dominant pathway, which consequently explored less likely, alternative binding/unbinding pathways. In the case of dim_p1hT (Fig 2), IS2 was more exposed, and the C-terminal tail formed salt bridges and remained attached to the phosphopeptide as it traversed through the dominant pathway (until the p1t fragment was pulled out of reach).

We compared the tail behaviour of the DF/HRE-MD simulations to four unrestrained MD simulations (two with and two without peptides) of 40 ns length (S3 Table). The unrestrained simulations showed a similar behaviour (black and red percentages in Fig 5A) as the DF/HRE-MD simulations. The tails showed a strong bias towards either the inner or outer side of the 14-3-3ζ dimer, and even fewer transitions between the two were observed. However once the C-terminal tail travelled from the inner surface to outer surface, we did not observe its return. Consequently the classical MD simulations have shown an even stronger preference towards the tail to be found near the outer surface. Despite this preference, the C-terminal tails in both DF/HRE-MD and unrestrained MD simulations remained the most flexible part of 14-3-3ζ (as shown in Fig 5C) with regular transitions between detached and surface bound conformations.

The biological function of the 14-3-3 C-terminal region is not fully understood. It was suggested previously [12] that the 14-3-3ζ C-terminal tail can have an auto-inhibitory effect by binding to the phosphopeptide binding groove of its respective 14-3-3ζ monomer (IS1). The C-terminal tail was not observed occupying the IS1 in any of our MD simulations, but this occurred transiently (<5) % in the HRE-MD simulations of dim complexes. Our results suggest that instead of occupying IS1, the tail rather interacts with amino acids on the outer surface of its 14-3-3 monomer, or amino acids from IS2. In addition, the C-terminal region also spent a considerable amount of time (~20%) being detached from the globular parts of 14-3-3ζ, and in some of the DF/HRE-MD simulations directly interacted with the phosphopeptides themselves during the binding process. Even though the sampling of the conformational substates is incomplete, our simulations are more in line with NMR studies that showed high flexibility and only transient interactions between the structured part of 14-3-3ζ and the C-terminal region [13].

2.2. 14-3-3ζ monomer opening. Structural changes in the 14-3-3ζ protein monomers leading to an opening or closing of the peptide-binding groove were previously reported using both computational and experimental methods [5,14]. This breathing motion was observed in our simulations as well, for both dim and tmon systems and their phosphopeptide complexes. We measured the groove width of a 14-3-3ζ monomer associated with the opening process, as the distance between the Cα atom of G53 (middle of helix 3) and L191 (middle of helix 8) (Fig 6A). Based on the measurement of the groove width distributions, we defined three sub-states (closed, open, wide-open) of the 14-3-3ζ monomer, depicted in Fig 6. We considered the 14-3-3ζ monomer closed, if the groove width was below 2.5 nm, wide-open above 2.9 nm and open in between.

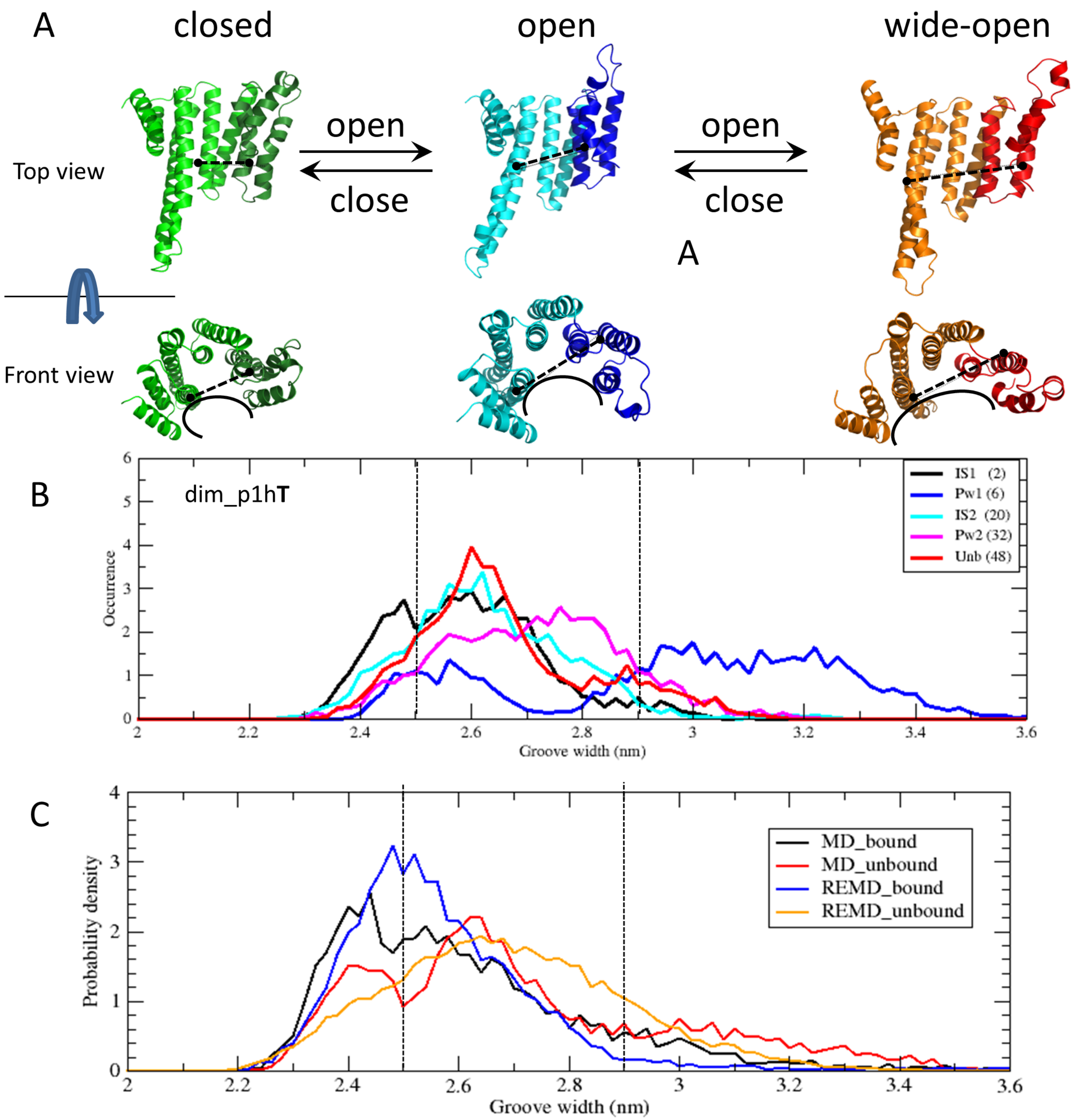


Fig 6. Changes in the 14-3-3ζ monomer groove width. A) Models of 14-3-3ζ at different levels of opening (the green, blue and orange colours mark closed, open and wide-open states, respectively), the last 3 helices are marked with a darker colour. B) Representative replicas of five different stages along the binding pathway (with the replica ID number shown in brackets) from DF/HRE-MD simulation dim_p1hT. The panel shows groove width distributions whilst the phosphopeptide moves from the binding site (IS1) to the unbound state (Unb). C) Average groove width distributions of 14-3-3ζ monomers as obtained from DF/HRE-MD bound (IS1) and unbound (Unb) states, and unrestrained MD simulations in holo (bound) and apo (unbound) states.

https://doi.org/10.1371/journal.pone.0180633.g006

We monitored the groove width distribution as the function of phosphopeptide location along the binding/unbinding pathway and found that the breathing motion in the restrained monomer was changed for all DF/HRE-MD simulations. Fig 6B presents the distribution of the groove widths for replicas representing the five stages along the binding pathway of the simulation dim_p1hT (replicas 2, 6, 20, 32 and 48 for IS1, Pw1, IS2, Pw2 and Unb, respectively). When the phosphopeptide was further away from the primary binding site (in the IS2, Pw2 or Unb stage) the groove width distributions were similar, centered on the open state (~2.6 nm), with a smaller probability to visit both the closed and wide-open states. When the phosphopeptide was bound to IS1, however, the interactions with residues from helices three, five and seven resulted in narrower groove width distributions which were also shifted towards the closed state. A similar observation was reported by Hu *et. al.* [14] (See S13 Fig). Interestingly, shortly after the phosphopeptide left IS1 (typically between replicas 6–11, Pw1) very wide grove width distributions were observed, with a high probability of the monomer from which the peptide was pulled adopting a wide-open conformation. The wide-open conformation in Pw1 is present for all our 14-3-3ζ/phosphopeptide models and is not dominant for any other part of the binding/unbinding pathway. These results (to our knowledge not reported before) indicate that the wide-open conformation of 14-3-3ζ may be necessary for the phosphopeptide ligand to detach from or enter into the primary binding site at IS1.

The distributions of the groove width for apo (unbound) and holo (bound) 14-3-3ζ in unrestrained MD simulations were also compared in Fig 6C with average groove width distributions for IS1 (bound) and Unb (unbound) DF/HRE-MD replicas. As the figure shows, the range of groove width distributions for the DF/HRE-MD and unrestrained MD simulations were very similar, and a shift towards the closed state was also observed between the bound and unbound monomers. The similarity between groove width distributions suggest, that the presence of a phosphopeptide strongly affected the breathing motion of 14-3-3ζ monomers in our simulations, but the applied DF restraints did not perturb this conformational degree of freedom directly.

2.3. Inter-monomer twist within 14-3-3ζ homo-dimer. Apart from the internal motion of the monomers, motion of monomers relative to each other was also observed in our (dim) simulations. This motion was monitored by measuring the dihedral angle of the Cα atoms of the residues L43(M1)-A54(M1)-A54(M2)-L43(M2) where M1 and M2 indicate different monomers within the 14-3-3ζ dimer (Fig 7A).

During DF/HRE-MD simulations an inter-monomer twist of 175 ± 15° was observed, with little variation with respect to the pulled phosphopeptide fragment or its DF restraint position (shown in Fig 7B). The similar distributions of the inter-monomer twist angles suggest that this conformational degree of freedom is not involved in the phosphopeptide binding process. We speculate the inter-monomers twist may still be important for protein-protein binding processes, as it can help 14-3-3 to better adapt the binding interface.

In some preliminary regular MD simulations at lower salt content, the inter-monomers twist angle was significantly reduced (see Methods section and S9 Fig). However, the inter-monomer twist angles in the DF/HRE-MD simulations were consistent with the twist angles observed in the available crystal structures of 14-3-3ζ dimers (~180 ± 3° calculated from the pdb entries 2WHO and 1A4O) and yielded a better agreement with SAXS measurements (S14 Fig).

2.4. Secondary structure changes of phosphopeptides along the binding pathway. The secondary structure of the 14-3-3ζ dimer remained mostly unchanged during our DF/HRE-MD and classical MD simulations, except for the occasional shortening of helix nine, and smaller conformational changes within the loops and disordered tails (e.g shown in S10 Fig). Secondary structure analyses of phosphopeptides for individual replicas of the

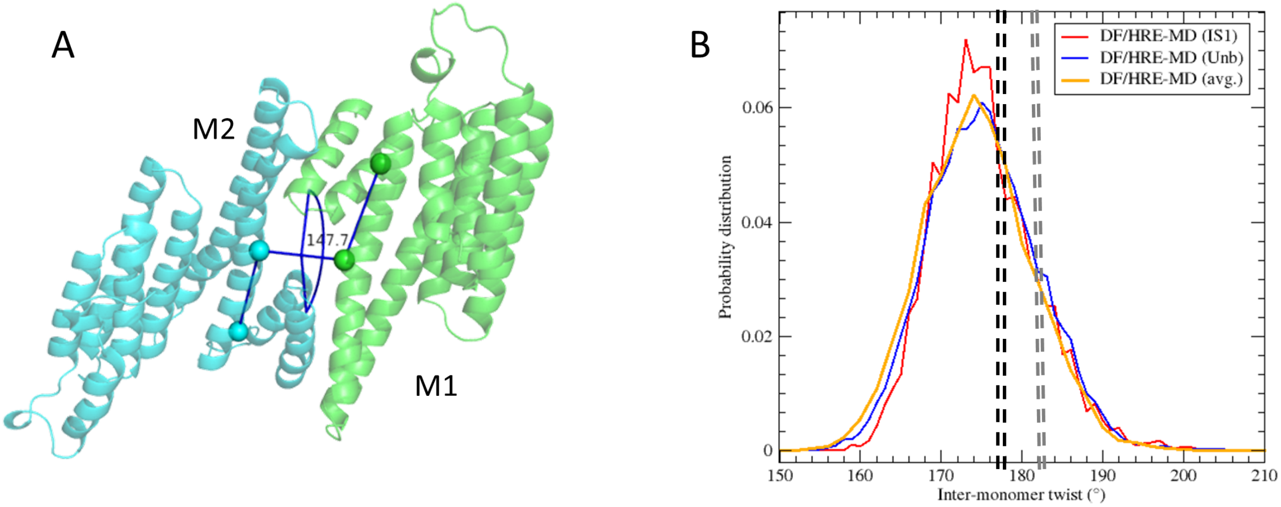


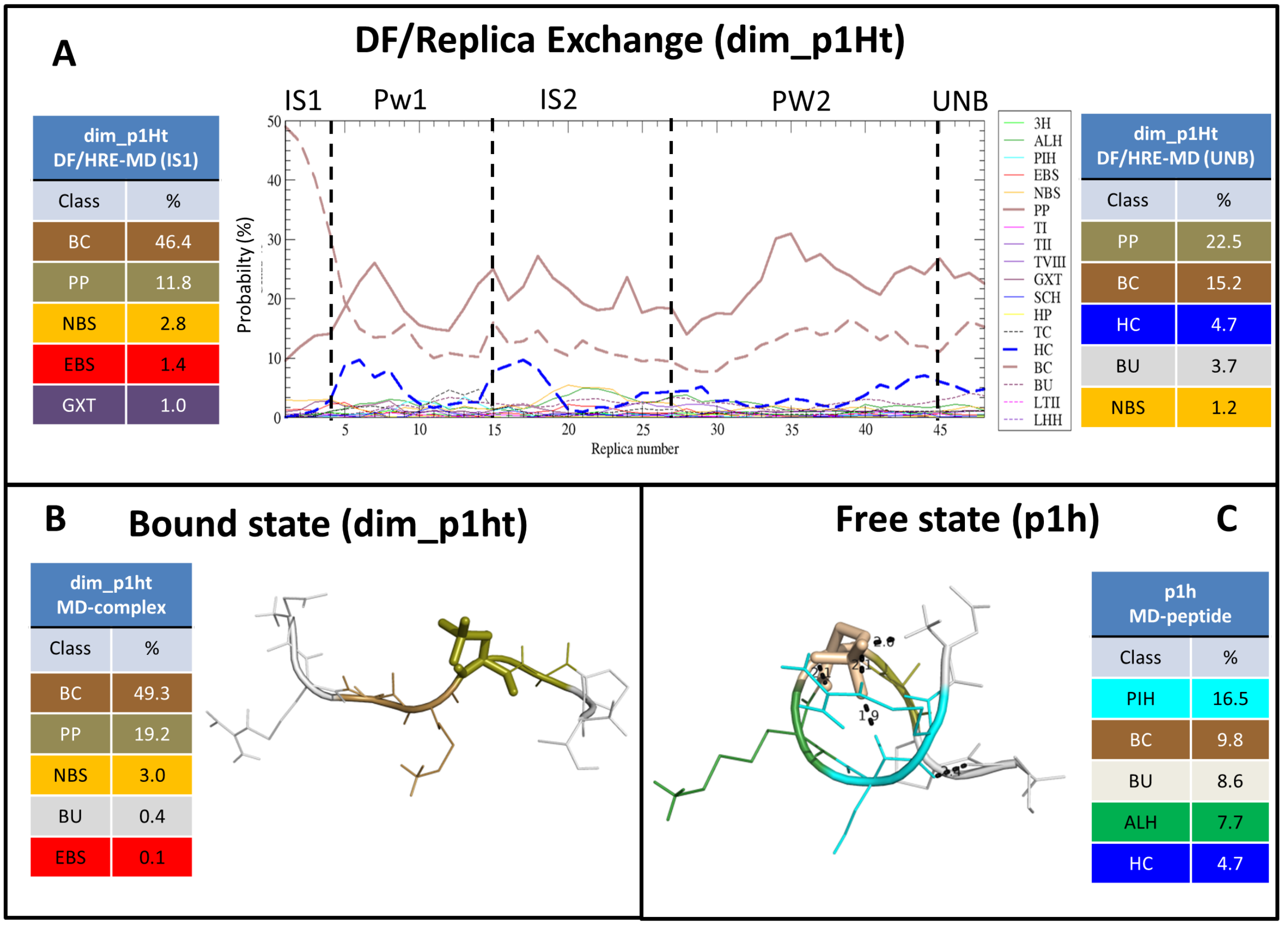
Fig 7. 14-3-3ζ inter-monomer twist. A) The top view of the protein dimer, with the twist angle between the monomers displayed in dark blue. B) Probability distribution of inter-monomers twist angles averaged over all DF/HRE-MD simulations in the IS1 (bound) and Unb (unbound) stages, and for the whole pathway.

https://doi.org/10.1371/journal.pone.0180633.g007

DF/HRE-MD along the binding pathway were also performed using the DISICL algorithm. The secondary structure (SS) of phosphopeptides along their binding pathway was summarized for all eight DF/HRE-MD simulations in S11 and S12 Figs. In Fig 8A we present the DISICL profile of the p1h phosphopeptide along the binding pathway during the simulation dim_p1Ht. The phosphopeptide in IS1 (replicas 1–4) adopts an extended conformation, dominated by the β-cap (BC, brown) and polyproline-like (PP, maroon) classes (~60% of all amino acids in the ensemble). After the phosphopeptide leaves IS1 it adopts less extended conformations, the PP and BC classes become less characteristic, and the population of the helix-cap (HC, blue) class rises slightly.

The analysis of the other phosphopeptides revealed a similar highly extended backbone structure for all four phosphopeptide fragments in the bound state, with a dominance of the polyproline-like (PP) and β-cap (BC) classes. This is also in agreement with the conformations observed in the crystal structures and unrestrained MD simulations (a representative conformation is shown in Fig 8B). Similar trends were also observed for later stages of the eight DF/HRE-MD simulations. The presence of helical classes, mainly the helix-cap, π-helical (PIH, cyan), and α-helical (ALH, green) population was gradually increased to 5–20% as the peptide reached the unbound state, whilst the population of the extended classes (PC and BC) decreased to ~30%). The population of the secondary structure classes in the 100 ns MD simulations of the free phosphopeptide fragments (e.g shown in Fig 8C), also showed an increased helical preference (~20%) compared to the extended 14-3-3ζ-bound peptide fragments. In these simulations, the average population of π-helix is ~10%, significantly higher than what was expected for a random coil peptide (~0.4%). We found that this conformation was stabilized by intra-molecular interactions with the pSer sidechain.

Despite the common trends, phosphopeptide-specific conformational changes were also observed during the DF/HREMD simulations, with the largest conformational changes in secondary structure occurring between replicas 5–12 (Pw1, DF: 0.9–2.2 nm) and 20–25 (IS2, DF: 4.4–5.5 nm) along the binding pathway. The simulation dim_p2Ht showed an unusually high π-helical content, similar to the free peptide simulations. Note that dim_p2Ht did not follow the dominant binding pathway. The distorted π-helical conformation of the backbone in this



dim_p1Ht DF/HRE-MD (IS1)	
Class	%
BC	46.4
PP	11.8
NBS	2.8
EBS	1.4
GXT	1.0

dim_p1Ht DF/HRE-MD (UNB)	
Class	%
PP	22.5
BC	15.2
HC	4.7
BU	3.7
NBS	1.2

dim_p1ht MD-complex	
Class	%
BC	49.3
PP	19.2
NBS	3.0
BU	0.4
EBS	0.1

p1h MD-peptide	
Class	%
PIH	16.5
BC	9.8
BU	8.6
ALH	7.7
HC	4.7

Fig 8. Phosphopeptide p1h secondary structure. Changes in the backbone structure are shown during DF/HRE-MD simulation dim_p1Ht (panel A) and unrestrained MD (panels B and C) simulations, analysed by the DISICL algorithm. The change in the average secondary structure content during the DF/HRE-MD simulation is shown in the middle of panel A, whilst the most dominant conformations in bound and unbound states are tabulated on the left and right side, respectively. Representative conformations for the bound and unbound states are depicted on the left and right sides, respectively, where the residues are coloured according to secondary structure classification. Intra-molecular hydrogen bonds are depicted as dashed lines. The tables besides the depictions show the 5 most populated secondary structure classes for the corresponding MD simulation and the appropriate (bound(IS1)/unbound) stage of the DF/HRE-MD simulation, respectively. The most populated DISICL classes are depicted in the following colours: π-helix (PIH)–cyan, Extended β-strand (EBS)–red, normal β-strand (NBS)–orange, polyproline-like (PP)–brown, turn type VIII (TVIII)–indigo, Gamma turn (GXT)–maroon, β-cap (BC)–gold, helix-cap (HC)–blue, turn-cap (TC)–black. DISICL secondary structure elements are listed in S1 Table.

https://doi.org/10.1371/journal.pone.0180633.g008

simulation increased (to ~10%) gradually while the phosphopeptide gets unbound, stabilised by an intra-molecular hydrogen-bond bridge between the pSer 5 side-chain and residues His 2 and Arg 3. These interactions appear early in the pathway, when the phosphopeptide is prevented to follow the positive surface patch of the 14-3-3 dimer, due to the intervention of the C-terminal tail.

3. Free-energy changes along the binding pathway

3.1. PMF profiles and binding affinity determination. The HRE-MD simulations were used to determine the Helmholtz free energy profiles ($A^{wham}(l)$) along the binding pathway of the four phosphopeptide fragments bound to both dim and tmon models of 14-3-3ζ. It is important to emphasise that the DF/HRE-MD methodology greatly enhances the sampling of phosphopeptide binding/unbinding but it does not enhance the sampling of slow conformational

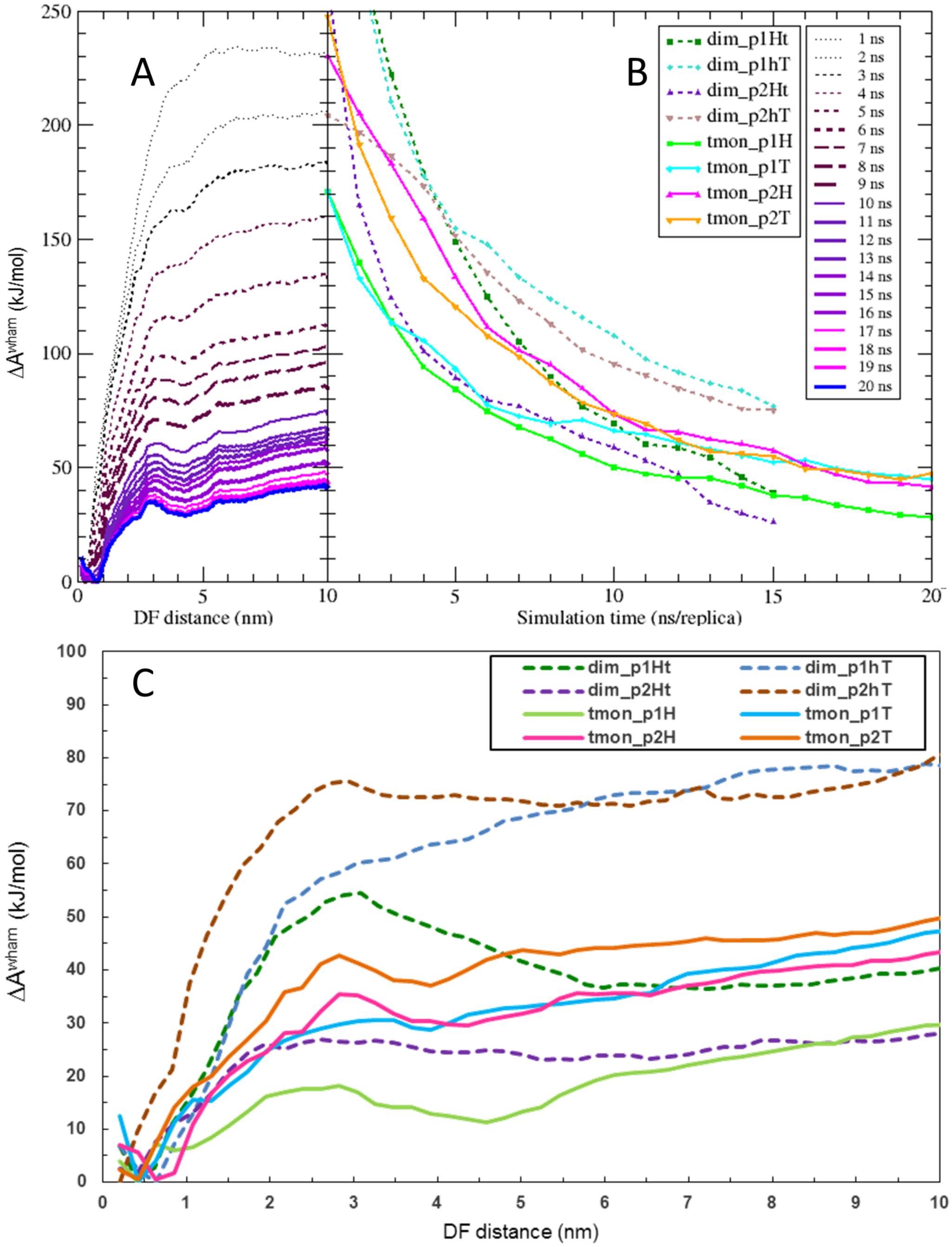


Fig 9. Free energy profiles of the DF/HRE-MD simulations. A) The raw free-energy profiles, derived from WHAM analysis, as function of simulation time per replica (every ns) for the simulation tmon_p2H. B) Convergence of the free energy difference between the unbound and bound state of all eight DF/HRE-MD simulations. C) Final free energy profiles (ΔA^{wham}) for the eight 14-3-3ζ peptide-binding simulations.

https://doi.org/10.1371/journal.pone.0180633.g009

transitions of the protein (e.g. C-terminal tail) to the same extent. Due to the slow transitions between different conformations, the simulations of dimeric complexes are not fully converged and should be considered only in a qualitative way (e.g. shape of curves $A^{wham}(l)$ in Fig 9). On the other hand, our minimalistic systems (tmon complexes) lack the C-terminal tail, sample a less complex conformational space and, therefore, the convergence of free energy profiles is much better. Fig 9A depicts the change of the free-energy profile ($A^{wham}(l)$) over simulation time, as obtained from the weighted histogram analysis method (WHAM) for simulation tmon_p2h. The physical meaning of the free energy profile ($A^{wham}(l)$) is that the probability ($\rho(l)$) of finding the phosphopeptide at the DF distance l is proportional to exp($-A^{wham}(l)$). The $A^{wham}(l)$ profiles for all systems were shifted by a constant such, that their minimal value is zero. Fig 9B presents the convergence of $A^{wham}(l)$ profiles as a function of DF/HRE-MD simulation time for the 8 studied complexes.

In order to calculate binding free energies that can be compared with the experimental binding affinities, a standard state correction has to be added to the raw WHAM profile as described in the Methods section (Eq 5). This correction relates the volume associated with the unbound state to the standard state volume. Table 1 shows the sampled and accessible volume (based on the restraining distance), as well as the approximate free energy from the WHAM calculations for every replica of simulation tmon_p2h, together with the zero-energy DF distance (l_0). Once the unbound replicas are identified, the unbound volume is calculated from the sampled volume of the unbound replicas. The free energy of binding is calculated by subtracting the free energy of the unbound state (Eq 4) from the free energy of the bound state (Eq 3) and adding the standard state correction in accordance with Eq 5.

Table 2 summarises the resulting binding free energies and draws a comparison with the available experimental binding affinities. The binding affinities of 14-3-3ζ to very similar peptide fragments with identical binding recognition sequences were recently documented using fluorescence spectroscopy and isothermal calorimetry (ITC) methods [8,15], resulting in binding affinities in the 100–10 μM range corresponding to binding free energies of -20 to -30 kJ/mol. Surface plasmon resonance (SPR) experiments [2,10] on identical or very similar peptides reported binding affinities of about 100 nM, corresponding to binding free energies of -45 kJ/mol (p2t peptide). Our computational absolute binding affinities in the range of -30 to -55 kJ/mol correspond roughly to the available experimental data taking into consideration the fact that no fitting parameters were used. Please note that the presented experimental binding free-energies were measured for longer peptides with an identical recognition sequence, thus the free energies of binding do not need to agree completely with the ones calculated from our simulations. The monomeric PMF profiles as well as two of the dimeric profiles (dim_p1Ht and dim_p2hT) have a second, local free-energy minimum at 3.5–5.0 nm in the distancefield, associated with IS2 and a free energy barrier at ~3.0 nm separating IS1 and IS2 (Fig 9C). For the other two complexes (dim_p1hT and dim_p2Ht) neither the barrier nor the local minimum was observed, this is likely due to interference from the C-terminal tail of 14-3-3ζ, which was—in both cases—mostly located on the inner side of the 14-3-3 dimer.

3.2. Impact of C-terminal tail on binding to the IS2. Comparison of free energy profiles between full length dimeric and truncated monomeric models of the 14-3-3ζ complexes allowed us to estimate the impact of the C-terminal tail on the phosphopeptide binding. The conformational transition of the C-terminal tail between the inner and outer side of 14-3-3ζ (Fig 2, S2 and S3 Tables) was a slow collective motion, poorly sampled in the dimeric simulations. However, two of DF/HRE-MD simulations (dim_p1Ht and dim_p2hT) thoroughly sampled the most probable binding pathways whilst the C-terminal tail remained attached to the outer protein surface. The other two simulations (dim_p1hT and dim_p2Ht) sampled pathways, during which the tail was attached to IS2 (more structural details in section 2.1).

The PMF profile of simulations where the C-terminal tail was for a considerable time present on the inner side of the 14-3-3ζ dimer (dim_p1hT and dim_p2Ht) does not exhibit a local free-energy minimum associated with IS2 (Fig 9C), or a free-energy barrier which would prevent the phosphopeptide to directly bind to IS1. In case of dim_p2Ht, the C-terminal tail was

Table 1. Details of the DF/HRE-MD simulation of tmon_p2H. The table contains quantities for every replica that are required for the free energy calculations. The columns show the state assignment, the replica number, the sampled volume ($V_{sampled}$), replica exchange probability (P_{ex}), reaction coordinate value (λ), ideal distancefield distance (l_0), accessible volume ($V(l)$) and raw free-energy according to the free-energy profile (A^{wham}). $V(l)$ and A^{wham} were calculated based on the restraining distance assigned to the replica. The last row contains the unbound volume (V_{unb}), calculated from the volume sampled by the phosphopeptide in all Unb state replicas, and the volume of simulation box (V_{box}).

State	Replica	$V_{sampled}$	P_{ex}	λ	l_0	$V(l)$	A^{wham}
tmon_p2H	Number	nm^3			Nm	nm^3	kJ/mol
IS1 (bound)	1	0.15	0.105	0.00	0.2	0.03	7
IS1 (bound)	2	0.27	0.223	0.02	0.4	0.09	5.6
IS1 (bound)	3	0.37	0.219	0.04	0.6	0.21	0.6
IS1 (bound)	4	0.44	0.222	0.06	0.9	0.60	1.7
IS1 (bound)	5	0.55	0.175	0.08	1.1	0.86	10.9
IS1/Pw1	6	0.75	0.099	0.11	1.3	0.77	16.7
IS1/Pw1	7	1.02	0.100	0.13	1.5	0.99	20.2
IS1/Pw1	8	1.38	0.130	0.15	1.7	1.21	22.9
Pw1	9	1.59	0.143	0.17	1.9	1.89	24.5
Pw1	10	1.60	0.123	0.19	2.2	2.19	28.1
Pw1	11	1.50	0.123	0.21	2.4	1.70	28.3
Pw1	12	1.78	0.118	0.23	2.6	2.02	32.2
IS2	13	2.05	0.089	0.25	2.8	2.47	35.4
IS2	14	2.65	0.102	0.27	3.0	2.99	35.2
IS2	15	2.71	0.121	0.29	3.3	5.61	33.9
IS2	16	2.66	0.130	0.32	3.5	7.00	31.8
IS2	17	2.78	0.113	0.34	3.7	5.84	30.3
IS2	18	2.72	0.098	0.36	3.9	6.93	30.3
IS2	19	3.19	0.137	0.38	4.1	10.94	29.8
IS2	20	2.78	0.170	0.40	4.4	13.14	29.5
IS2	21	3.35	0.156	0.42	4.6	10.47	30.5
IS2	22	4.12	0.120	0.44	4.8	12.05	31.1
IS2	23	4.91	0.119	0.46	5.0	13.82	31.7
IS2	24	5.61	0.150	0.48	5.2	15.69	32.6
IS2	25	6.76	0.128	0.51	5.5	27.42	34.7
IS2	26	7.37	0.114	0.53	5.7	31.08	35.7
IS2	27	7.69	0.142	0.55	5.9	22.56	35.4
Pw2	28	8.16	0.140	0.57	6.1	23.65	35.7
Pw2	29	7.67	0.132	0.59	6.3	24.40	35.7
Pw2/Unb	30	7.95	0.124	0.61	6.5	24.72	35.2
Pw2/Unb	31	9.47	0.125	0.63	6.8	24.76	36.2
Pw2/Unb	32	9.57	0.125	0.65	7.0	24.51	37
Pw2/Unb	33	9.75	0.111	0.67	7.2	23.97	37.5
Pw2/Unb	34	10.37	0.129	0.69	7.4	23.25	38.1
Unb (unbound)	35	10.69	0.138	0.72	7.6	22.52	39.1
Unb (unbound)	36	10.30	0.117	0.74	7.9	32.17	39.6
Unb (unbound)	37	9.83	0.114	0.76	8.1	30.47	39.9
Unb (unbound)	38	9.59	0.109	0.78	8.3	18.96	40.4
Unb (unbound)	39	10.17	0.112	0.80	8.5	17.80	40.7
Unb (unbound)	40	9.34	0.111	0.82	8.7	16.40	41
Unb (unbound)	41	8.84	0.107	0.84	9.0	15.00	40.9
Unb (unbound)	42	9.89	0.123	0.86	9.2	13.74	41.7
Unb (unbound)	43	9.90	0.124	0.88	9.4	12.47	41.8

(*Continued*)

Table 1. (*Continued*)

State	Replica	$V_{sampled}$	P_{ex}	λ	l_0	$V(l)$	A^{wham}
tmon_p2H	**Number**	**nm^3**			**Nm**	**nm^3**	**kJ/mol**
Unb (unbound)	44	9.04	0.109	0.91	9.6	11.20	42.1
Unb (unbound)	45	7.60	0.064	0.93	9.8	9.98	42.9
Unb (unbound)	46	6.96	0.090	0.96	10.2	8.82	43.7
Unb (unbound)	47	7.73	0.134	0.98	10.4	7.74	43.9
Unb (unbound)	48	6.87	0.059	1.00	10.6	6.73	45
	V_{unb}	74.8	nm^3		V_{box}	626.4	nm^3

https://doi.org/10.1371/journal.pone.0180633.t001

Table 2. Comparison of the experimental and calculated binding free energies. The ΔG_{exp} shows the experimental free energies calculated from dissociation constants [8–10]. ΔA_{bind}(mon) shows the binding free energy calculated from the tmon DF/HRE-MD simulations (20ns/replica).

Simulated	ΔG_{exp}	ΔA_{bind}(mon)
System	**kJ/mol**	**kJ/mol**
tmon_p1H	-28.8	-30.9
tmon_p1T	-23 *	-49.1
tmon_p2H	-21.6	-47.8
tmon_p2T	-27.4 / -44.2	-52.9

*: weak binding, estimate based on detection limit

https://doi.org/10.1371/journal.pone.0180633.t002

directly interacting with IS2 and prevented interactions between the p2h peptide and IS2. In case of dim_p1hT, IS2 was partially available for the p1t peptide, however, interactions between p1t and the C-terminal tail disrupted interactions with IS2. On the other hand, in simulations where the C-terminal tail was not present (all tmon simulations) or mostly present on the outer surfaces of 14-3-3ζ (dim_p1Ht and dim_p2hT) the drop in free-energy profile in the IS2 area and a free energy barrier (at ~3.0 nm separating IS1 and IS2) are present. These two features suggest an intermediate state (when the peptide is bound IS2) along the phosphopeptide binding pathway, which is diminished by the presence of the C-terminal tail near the inner side of the 14-3-3ζ dimer. This led us to the conclusion that one of the roles of the C-terminal tail may be to weaken the interaction between phosphopeptides and IS2 of the 14-3-3ζ protein.

Conclusions

We explored the phosphopeptide binding pathways of the 14-3-3ζ protein through molecular dynamics simulations of four phosphopeptide fragments derived from PKC-ε and C-RAF kinase. The pathways were explored by a novel Hamiltonian replica exchange molecular dynamics method with incorporated distancefield restraints (DF/HRE-MD). The eight DF/HRE-MD simulations (4 dimeric and 4 truncated monomeric complexes) combined corresponded to more than 6.7 μs of enhanced-sampling simulation time, allowing for the unbiased determination of the most probable binding/unbinding pathways, the corresponding structural changes of the phosphopeptides and 14-3-3ζ, as well as the PMF profiles along the binding pathway.

The determined binding pathways were very similar for 7 out of the 8 studied complexes, suggesting a dominant phosphopeptide pathway, which roughly followed helix 3 between the primary binding site (IS1) and a newly identified secondary interaction site (IS2), localized in the second half of helix 3 (residues 60–70). We found that the flexible C-terminal tail of 14-3-3ζ may interact with both IS2 and the phosphopeptide in our simulations. When the C-terminal region interfered with the phosphopeptide binding pathway the interactions between the phosphopeptide and 14-3-3ζ IS2 were changed significantly, and this change was also reflected in the corresponding PMF profiles.

We confirmed previous findings suggesting that 14-3-3 monomers in complex with a phosphopeptide are shifted towards a more closed conformation as compared to the apo state. Our DF/HRE-MD simulations revealed that the 14-3-3ζ monomer adopts a wide-opened conformation when a phosphopeptide is to enter or leave IS1. The phosphopeptide secondary structure during the DF/HRE-MD simulations also changed from an extended conformation at IS1 into a less ordered structure with ~20% helical content, with a high probability of π-helical conformations in the unbound state. Sequence specific rearrangements in the peptide structure were detected during the Pw1 and IS2 stages, followed by the gradual increase of helical content as the phosphopeptides detached from 14-3-3ζ.

The DF/HRE-MD simulations allowed effective pathway sampling of the phosphopeptide fragments within the 14-3-3ζ protein. Application of distancefield restraints prevented the 14-3-3ζ protein damage that was observed in cases when regular distance-restraints were applied. While the full length 14-3-3ζ WT dimer models complexed with the four phosphopeptides proved to be useful for exploring the structural properties, their size and slow convergence prevented effective free-energy calculation for these systems. Therefore minimalistic models based on a truncated 14-3-3ζ monomer were designed, which proved to be more efficient in calculating the potential-of-mean-force (PMF) profiles for the binding of phosphopeptide/14-3-3ζ complexes. The binding free energies derived from the calculated PMF profiles of minimalistic 14-3-3ζ models show a reasonable agreement with known experimental binding affinities between similar PKC-ε and C-RAF kinase fragments and the 14-3-3ζ protein.

Taken together, these findings deepen our understanding about the binding phenomena of phosphopeptides to 14-3-3ζ and the obtained results likely have a wider applicability for other human isoforms considering their high sequence identity.

Methods

All molecular dynamics (MD) and Hamiltonian replica exchange MD (HRE-MD) simulations were performed using the GROMOS11 software package [16]. The structure preparation and analysis of the simulation trajectories was based on the GROMOS++ analysis package [17]. The PyMol molecular graphics system version 1.5.0.3 [18] was used for visualization and calculation of the electrostatic surface potentials based on an adaptive Poisson-Boltzmann solver [19] as implemented in the PyMol software. Changes in the protein and peptide secondary structure were followed by the DISICL algorithm [20].

Generated starting coordinates

The starting coordinates of the presented 14-3-3ζ models were generated based on the same crystal structure (pdb code: 2WH0) in order to prevent structural changes due to different protein starting structures. The crystal structure 2WH0 represents the dimeric 14-3-3ζ in complex with protein kinase C ε peptide fragments [8] (p1h, p1t). The structure of the C-RAF proto oncogene peptide fragments [9] (p2h, p2t) bound to 14-3-3ζ was taken from the structure of their co-crystal (pdb code: 4FJ3), after aligning the 14-3-3ζ dimer structures with 2WH0. The missing parts of the 14-3-3ζ protein, including the C-terminal tail, were added using Modeller version 9v8 [21]. The phosphopeptide structure was energy minimized in vacuum to avoid clashes with the 2WH0 protein atoms. The 14-3-3ζ dimer model without ligands was constructed based on the 2WH0 crystal structure where all peptide atoms were removed. The sequence of the complete peptides, as well as the head and tail fragments are indicated in Fig 1. The head and tail peptide fragments were chosen on the basis that they contain the consensus binding motif and the phosphorylation site in the middle of the sequence, consist of 8 amino acids for all peptide fragments, and have a total charge of -1.

Minimalistic (truncated monomeric) models were constructed from the dimer complexes by extracting the corresponding 14-3-3ζ monomer (along with its phosphopeptide) and truncating the last 15 amino acids (C-terminal stretch).

The prepared PDB coordinate files were transformed into GROMOS configuration files compatible with the GROMOS 54a7 force field [22], with phosphorylation parameters taken from the Vienna PTM 54a7 extension [23]. The structures were then energy minimised, solvated in pre-equilibrated SPC water [24] to fill a rectangular box with a minimal solute-to-wall distance of 2.0, 2.0, and 2.5 nm in the X, Y and Z dimensions, respectively, to provide sufficient space to pull the phosphopeptide fragments out of the binding site. The solute models were rotated in such a way that the largest dimension of the solute complex was oriented along the Z axis. Sodium and chloride ions were added to all simulation boxes to neutralise the total charge and provide a NaCl concentration of 0.15 or 0.25 M for electrostatic screening. The systems of the monomers typically contained 57 000 atoms, whilst the dimer complexes amounted to roughly 121 000 atoms. After solvation, the simulation boxes were energy minimised, then heated up from 60 to 298 K while gradually reducing the position restraints on protein and peptide atoms (initial force constant of 2.5 x 10^4 kJ/mol/nm^2) in 5 discrete steps of 20 ps each, and subsequently equilibrated for 60 ps without position restraints.

Molecular dynamics simulations

Equilibration and all simulations were performed under periodic boundary conditions, using a 2-fs time step and the SHAKE algorithm [25] to constrain bond lengths and H–O–H bond angles. The weak-coupling algorithm [26] was used to maintain a stable temperature (300 K) and pressure (101 kPa) when required, with relaxation times of 0.1 and 0.5 ps respectively. For long-range interactions the reaction field method [27] was used with a 1.4 nm cut-off and 61 as dielectric permittivity [28]. During all of the production simulations roto-translational constraints [29] were kept on all protein atoms to prevent tumbling of the complexes in the rectangular simulation box.

During our simulations SHAKE errors occurred frequently at planar amide groups found in peptide bonds of the polypeptide chain and in the side chain amide groups with a delocalised character, such as arginines, glutamines and asparagines. The backbone N-H bond of arginine 18 appeared particularly regularly due to local structural strain. To treat this frequently occurring numerical problem, we increased the mass of this particular hydrogen by a factor of 5, in both our MD and HRE-MD production runs. MD simulations of all 14-3-3ζ dimer, monomer and complex systems were performed for 40 ns, and all phosphopeptide related MD simulations for 100 ns.

Preliminary simulations proved to be quite sensitive to the NaCl concentration. Our initial intention was to run all simulations at 0.15 M NaCl (physiological) concentration. This corresponds to the salt concentration at which most of experimental binding affinities were measured. However, preliminary simulations of the dimeric systems at 0.15 M showed significant instabilities in terms of inter-monomer twist angles, defined as the dihedral angle of the Cα atoms of the residues L43(M1)-A54(M1)-A54(M2)-L43(M2) where M1 and M2 indicate different monomers within the 14-3-3ζ dimer (Fig 7A). These problems were not observed at higher, 0.25 M NaCl concentration (S9A Fig). The comparison of experimental and calculated SAXS profiles of dim systems (S14 Fig) indicated that low values of inter-monomer twist (often observed at 0.15 M NaCl) do not agree with solution SAXS data. Therefore, we decided to perform all dimeric (dim) simulations (presented in this study) at 0.25 M NaCl and all tmon simulations at 0.15 M NaCl where inter-monomer twist instability cannot occur.

Distancefield replica-exchange simulations

DF distance restraints and HRE-MD were applied as they are implemented in the GROMOS11 version 1.3.0 [16]. To compute the binding free energies for the phosphopeptide fragments (p1h, p1t, p2h, p2t), a reaction coordinate was defined as the DF distance between a virtual atom at the binding site of the corresponding 14-3-3ζ monomer and a virtual atom at the phosphorylation site of the peptide. The virtual atoms for all peptides were defined by the centre of mass of the C_β carbon atoms of the phosphorylated serine and its two neighbouring amino acids. The virtual atom for the binding site was defined as the centre of mass of the C_α atoms in residues N50, A192 and the amide hydrogen of N224 of the appropriate monomer. Simulations were started along the reaction coordinates to pull one of the peptides out of its respective binding pocket. This was done using a harmonic distancefield [6] distance restraint in 20 discrete steps with a minimal-energy distance ranging from 0.2 to 10.6 nm, a force constant of 2500 kJ/mol/nm^2 and a simulation length of 250 ps at each step. The final configurations at each step were used to start 100 ps of HRE-MD [30] equilibration. During the HRE-MD equilibration with 48 replicas (Table 1) exchange events were prohibited and position restraints (25 kJ/mol/nm^2) were applied to the protein atoms. After this equilibration, the final coordinates were then used to start the DF/HRE-MD production simulations, where replica exchanges were permitted every 10 ps and position restraints on the proteins were replaced with roto-translational constraints to maintain the protein orientation. Each of the four HRE-MD simulations of dim complexes was run for 15 ns; and of tmon complexes for 20 ns, using 48 replicas to cover a DF distance of ~10 nm from the IS1. The average interaction energy, and hydrogen bonds were monitored between peptide and protein atoms at each replica to determine the minimal distance for the unbound state (Fig 2C).

HRE-MD production simulations for each peptide fragment uniformly used a harmonic DF restraining potential with a force constant of 350 kJ/mol/nm^2. The harmonic potential was linearised after deviations larger than 2 nm in order to avoid large forces over longer distances. The DF for the distance calculations was updated each 100 steps with a grid spacing and protein cut-off of 0.2 nm, and a single smoothing step to decrease repulsion at the protein surface [6]. HRE-MD simulations were kept at a constant temperature of 300 K by the weak coupling algorithm and were performed at a constant volume. The combined simulation time length of the DF/HRE-MD comes to a total of ~6.7 μs of enhanced sampling to study the 14-3-3ζ peptide binding.

Phosphopeptide pathway assignment

The phosphopeptide binding-pathways obtained from DF/HRE-MD simulations were divided into 5 parts (IS1, Pw1, IS2, Pw2, Unb) during the analysis. The replicas of the DF/HRE-MD simulations were assigned to IS1 or IS2 based on three factors:

1. the average protein-phosphopeptide interaction energy of the replica had to be near a local minimum.
2. interactions observed with 14-3-3ζ residues in the corresponding interaction sites (more details below).

3. the most probable locations to find the phosphopeptide for the replica had to be in the proximity of the interaction site.

An interaction partner was considered stable in the interaction analysis if on average at least 0.1 interaction was present in the particular replica. Interaction partners for IS1 were K49, R127, Y128 towards the pS of the peptide. Interaction partners for IS2 were K68 towards the pS and any interaction involving S64-G70. Proximity from an interaction site was determined by measuring the distance between density peak maxima of the phospopeptide virtual atom for that particular replica (as shown in Fig 8) and Cα of K49 and K68 for IS1 and IS2, respectively. The density peak was considered proximal within 1.5 nm. The replicas were assigned to unbound state (Unb) if the interaction energy was close to zero and no stable hydrogen-bonds were detected between the phosphopeptide and 14-3-3ζ. Replicas that were fulfilling only part of the criteria were assigned to Pw1 and Pw2 instead. An example of assignment is show in Table 1, and replica density peaks are coloured according to their assignment in Fig 2, S1–S7 Figs.

Free energy calculations

The free energy profile was calculated from the replica exchange simulations, using a weighted histogram analysis method (WHAM) [31]. The WHAM is an iterative method that uses the probability distribution of biased ensembles ($\rho_i^{(b)}$) and reconstructs the unbiased probability distribution (ρ) by fitting the free-energy of a given number of windows (Nw) along the reaction coordinate. Using the DF distance distributions of the replicas—biased by the harmonic DF restraining potentials (B_j)–the probability distribution along the peptide binding pathway is calculated iteratively according to Eqs 1 and 2:

$$\rho(l) = \frac{\sum_{i=1}^{Nw} n_i\, \rho_i^{(b)}(l)}{\sum_{j=1}^{Nw} n_j e^{-(B_j(l)-A_j^{wham})/RT}} \tag{1}$$

$$e^{-A_j^{lwham}/RT} = \int \rho(l)\, e^{-B_j(l)/RT} dl \tag{2}$$

Where *ρ(l)* is the probability of finding the peptide at the DF distance *l*, A_j^{wham} is the Helmholtz free energy at the DF window j, n_i is the number of data points in replica i, R is the ideal gas constant, and T is the temperature. The resulting free-energy profile($A^{wham}(l)$) was used to calculate the free energy of the bound (A_{bound}) and unbound (A_{unb}) states, by integrating the profile over the DF distances associated with the two states, as shown in Eqs 3 and 4, respectively:

$$A_{bound} = -RT \ln \int_{bound} e^{-A^{wham}(l)/RT}\, dl \tag{3}$$

$$A_{unb} = -RT \ln \int_{unb} e^{-A^{wham}(l)/RT}\, dl \tag{4}$$

Because of the shape of $A^{wham}(l)$ in the bound area the A_{bound} value is quite insensitive on the particular choice of boundaries of the integral in Eq 3. On the other hand, A_{unb} depends significantly on the particular choice of the *unb* region boundaries in Eq 4. This dependence is mainly due to entropic contributions from the increased available volume of the unbound state, and is compensated by the standard state correction (ΔA_{std}) in the final calculation of the binding free energy (ΔA_{bind}) by Eq (5):

$$\Delta A_{bind} = A_{bound} - A_{unb} + \Delta A_{std}; \Delta A_{std} = -RT \ln \frac{V_{unb}}{V_0} \tag{5}$$

where V_0 is the standard state volume (1.66 nm^3) corresponding to a 1 M concentration. The physical meaning of Eq 5 is that the binding free energy (ΔA_{bind}) corresponds to the free-energy difference of bringing a ligand from a bound to an unbound state and subsequently from the unbound to the standard state volume [32]. ΔA_{bind} can be directly compared with experimental values calculated from dissociation constants. The unbound volume (V_{unb}) was calculated based on the number of grid-points visited by the phosphopeptide virtual atom in any replica of unbound state (Table 1).

To calculate the free-energy profiles of the peptide binding processes, we used the DF distributions of the DF distances collected over the 48 replicas of the HRE-MD simulations and 100 DF widows for the iterative WHAM process, which was continued for 10 000 steps or until the energy change was less than 10^{-5} kJ/mol. Calculations of the available volume were approximated from the DF grid based on the number of gridpoints assigned to a given distance. The

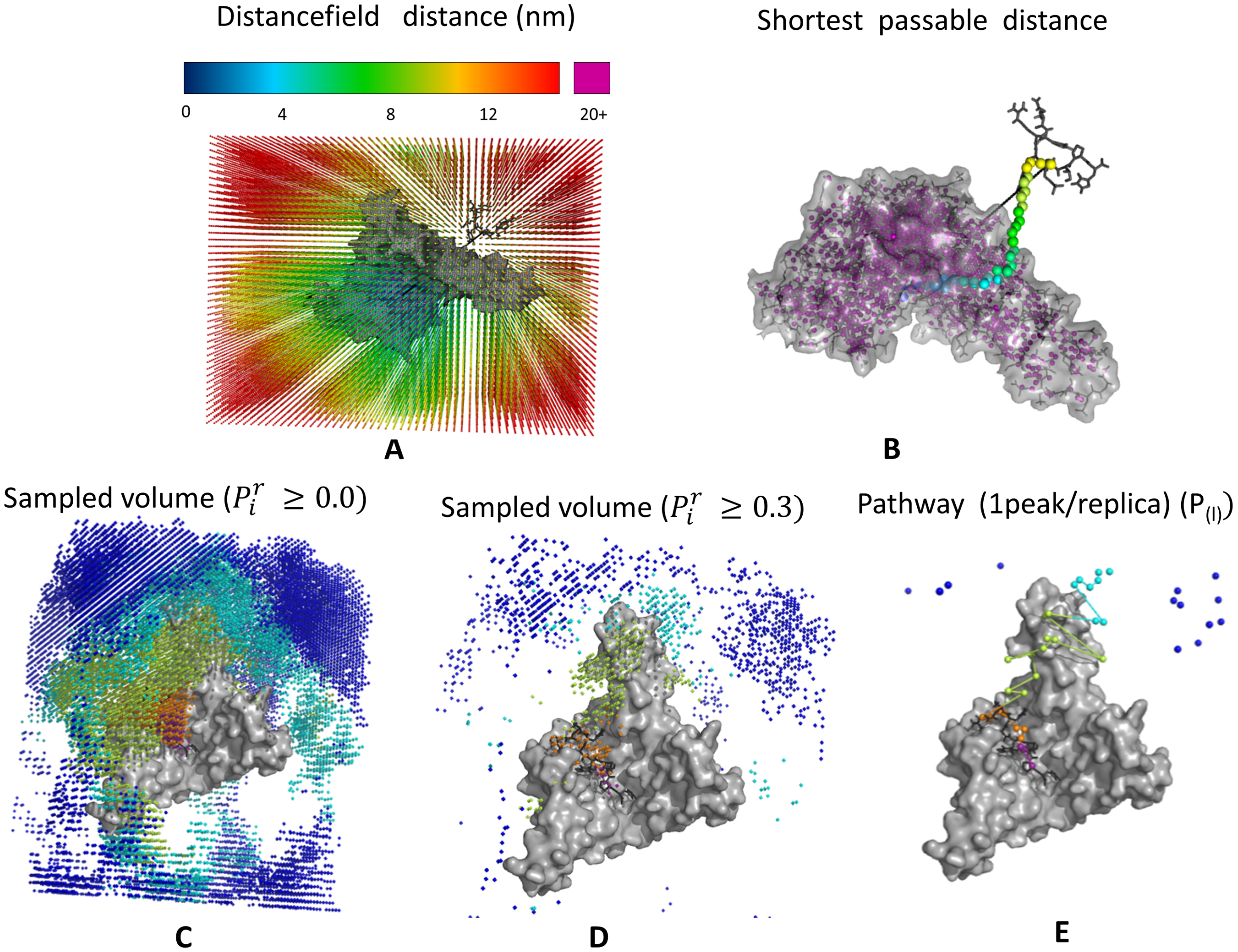

Distancefield distance (nm)
0
4
8
12
20+
A
Shortest passable distance
B
Sampled volume ($P_i^r \geq 0.0$)
C
Sampled volume ($P_i^r \geq 0.3$)
D
Pathway (1peak/replica) ($P_{(l)}$)
E

Fig 10. Graphical representation of pathway mapping in simulations. A) Distancefield grid, where grid points are coloured according to distancefield (DF) distance. B) Direct distance (as a black dashed line), and DF distance (along the coloured grid points) between the ligand and the protein binding side. Grid points located within the protein (which should be avoided) are shown in purple. C-E) Grid points sampled during a replica exchange simulation, coloured according to their position along the pathway. Panels C and D are showing grid points at different levels of relative probability (P_i^r) per replica; with all visited grid points in Panel C, and only the often-visited grid points per replica in Panel D. Panel E shows the derived peptide-binding pathway, with one peak maximum per replica. The surface of the protein is shown in grey.

https://doi.org/10.1371/journal.pone.0180633.g010

volumes sampled for the replica exchange simulations were calculated based on a similar grid with a mesh size identical to DF grids (0.2 nm) where the movements of the peptide virtual atom were tracked. By recording which gridpoint was the closest to the peptide virtual atom position at each frame (after every 5 ps of simulation), we could determine the probability density of the given gridpoint (P_i). To find the conformations most often visited by the peptide, the relative probability density (P_i^r) of the gridpoint i was calculated according to Eq 6:

$$P_i^r = \frac{P_i - P_{min}}{P_{max} - P_{min}} \quad P_i = \frac{O_i}{N} \tag{6}$$

Where P_{max} and P_{min} are the probabilities of visiting the least and most often visited gridpoints, O_i is the number of times the gridpoint was visited and N is the number of simulation frames used in the calculation (an example is shown in Fig 10).

Supporting information

S1 Table. DISICL secondary structure elements. Structure elements and abbreviations are listed below. For a detailed description of the DISICL classes see [20].
(DOCX)

S2 Table. C-terminal tail distributions in DF/HRE-MD simulations. The table displays the name of the simulation, the identity of the monomer (mon), the probability to find the C-terminal tail near the primary binding site (IS1), secondary binding site (IS2), on the outer protein surface (out), and free in solution (sol) over the entire length of the simulation, along the total number of transitions between the listed sub-states (observed transitions) Note that the phosphopeptide fragments p1h and p2t were binding to M1, whilst the p1t and p2h were binding to M2.
(DOCX)

S3 Table. C-terminal tail distribution in unrestrained MD simulations. The table displays the name of the simulation, the identity of the monomer (mon), the probability to find the C-terminal tail near the primary binding site (IS1), secondary binding site (IS2), on the outer protein surface (out), and free in solution (sol) over the entire length of the simulation, along with the total number of transitions between the listed sub-states (Observed transitions) Note that simulations were started with the M1 tail close to the outer surface and the M2 tail close to the inner surface (near IS2). Dim-2 denotes dimeric 14-3-3ζ- simulation started from an alternative set of starting coordinates.
(DOCX)

S1 Fig. Pathway visualization of the simulation tmon_p1H. A) The volume sampled by the p1h phosphopeptide during the simulation is shown by dots around the 14-3-3ζ protein coloured based on their position along the pathway. The most probable points in space to find the peptide in for each replica (probability density peaks) are represented by larger spheres. B) Density peaks and a few representative structures corresponding to the density peaks, coloured according to their position along the pathway. Replica density peaks are connected by lines to visualize the binding pathway. The 14-3-3ζ protein in panels A-B is shown as a surface representation, with the two monomers shown in light and dark grey. C) Average interaction energy between 14-3-3ζ and the p1h. D) Interaction map between any atom of the pulled p1h peptide and the amino acids of the protein, summarized for each replica (only amino acids with at least 0.1 hydrogen bond/salt bridge on average are shown). The scale indicates the number of interactions.
(TIF)

S2 Fig. Pathway visualization of the simulation tmon_p2H. A) The volume sampled by the p2h phosphopeptide during the simulation is shown by dots around the 14-3-3ζ protein coloured based on their position along the pathway. The most probable points in space to find the peptide in for each replica (probability density peaks) are represented by larger spheres. B) Density peaks and a few representative structures corresponding to the density peaks, coloured according to their position along the pathway. Replica density peaks are connected by lines to visualize the binding pathway. The 14-3-3ζ protein in panels A-B is shown as a surface representation, with the two monomers shown in light and dark grey. C) Average interaction energy between 14-3-3ζ and the p2h. D) Interaction map between any atom of the pulled p2h peptide and the amino acids of the protein, summarized for each replica (only amino acids with at least 0.1 hydrogen bond/salt bridge on average are shown). The scale indicates the number of interactions.
(TIF)

S3 Fig. Pathway visualization of the simulation tmon_p1T. A) The volume sampled by the p1t phosphopeptide during the simulation is shown by dots around the 14-3-3ζ protein coloured based on their position along the pathway. The most probable points in space to find the peptide in for each replica (probability density peaks) are represented by larger spheres. B) Density peaks and a few representative structures corresponding to the density peaks, coloured according to their position along the pathway. Replica density peaks are connected by lines to visualize the binding pathway. The 14-3-3ζ protein in panels A-B is shown as a surface representation, with the two monomers shown in light and dark grey. C) Average interaction energy between 14-3-3ζ and the p1t. D) Interaction map between any atom of the pulled p1t peptide and the amino acids of the protein, summarized for each replica (only amino acids with at least 0.1 hydrogen bond/salt bridge on average are shown). The scale indicates the number of interactions.
(TIF)

S4 Fig. Pathway visualization of the simulation tmon_p2T. A) The volume sampled by the p2t phosphopeptide during the simulation is shown by dots around the 14-3-3ζ protein coloured based on their position along the pathway. The most probable points in space to find the peptide in for each replica (probability density peaks) are represented by larger spheres. B) Density peaks and a few representative structures corresponding to the density peaks, coloured according to their position along the pathway. Replica density peaks are connected by lines to visualize the binding pathway. The 14-3-3ζ protein in panels A-B is shown as a surface representation, with the two monomers shown in light and dark grey. C) Average interaction energy between 14-3-3ζ and the p2t. D) Interaction map between any atom of the pulled p2t peptide and the amino acids of the protein, summarized for each replica (only amino acids with at least 0.1 hydrogen bond/salt bridge on average are shown). The scale indicates the number of interactions.
(TIF)

S5 Fig. Pathway visualization of the simulation dim_p1Ht. A) The volume sampled by the p1h phosphopeptide during the simulation is shown by dots around the 14-3-3ζ protein coloured based on their position along the pathway. The most probable points in space to find the peptide in for each replica (probability density peaks) are represented by larger spheres. B) Density peaks and a few representative structures corresponding to the density peaks, coloured according to their position along the pathway. Replica density peaks are connected by lines to visualize the binding pathway. The 14-3-3ζ protein in panels A-B is shown as a surface representation, with the two monomers shown in light and dark grey. C) Average interaction energy

between 14-3-3ζ and the p1h. D) Interaction map between any atom of the pulled p1h peptide and the amino acids of the protein, summarized for each replica (only amino acids with at least 0.1 hydrogen bond/salt bridge on average are shown). The scale indicates the number of interactions.
(TIF)

S6 Fig. Pathway visualization of the simulation dim_p2hT. A) The volume sampled by the p2t phosphopeptide during the simulation is shown by dots around the 14-3-3ζ protein coloured based on their position along the pathway. The most probable points in space to find the peptide in for each replica (probability density peaks) are represented by larger spheres. B) Density peaks and a few representative structures corresponding to the density peaks, coloured according to their position along the pathway. Replica density peaks are connected by lines to visualize the binding pathway. The 14-3-3ζ protein in panels A-B is shown as a surface representation, with the two monomers shown in light and dark grey. C) Average interaction energy between 14-3-3ζ and the p2t. D) Interaction map between any atom of the pulled p2t peptide and the amino acids of the protein, summarized for each replica (only amino acids with at least 0.1 hydrogen bond/salt bridge on average are shown). The scale indicates the number of interactions.
(TIF)

S7 Fig. Pathway visualization of the simulation dim_p2Ht. A) The volume sampled by the p2h phosphopeptide during the simulation is shown by dots around the 14-3-3ζ protein coloured based on their position along the pathway. The most probable points in space to find the peptide in for each replica (probability density peaks) are represented by larger spheres. B) Density peaks and a few representative structures corresponding to the density peaks, coloured according to their position along the pathway. Replica density peaks are connected by lines to visualize the binding pathway. The 14-3-3ζ protein in panels A-B is shown as a surface representation, with the two monomers shown in light and dark grey. C) Average interaction energy between 14-3-3ζ and the p2h. D) Interaction map between any atom of the pulled p2h peptide and the amino acids of the protein, summarized for each replica (only amino acids with at least 0.1 hydrogen bond/salt bridge on average are shown). The scale indicates the number of interactions. Note that dim_p2Ht did not follow the general pathway observed for other DF/HRE-MD simulations, and IS2 residues were not significant interaction partners.
(TIF)

S8 Fig. Interaction map between the perturbed phosphopeptides and amino acids of 14-3-3ζ. The map shows an average over the 8 HRE-MD simulations, where the average number of hydrogen bonds observed during the simulations are shown as the function of replica number. Replica 1–3 refers to the bound state in interaction site1 (IS1), and larger replica numbers correspond to higher DF distance from IS1.
(TIF)

S9 Fig. The salt concentration dependence of the 14-3-3ζ inter-monomer twist. A) The drift of the monomer twist angle during MD simulations. B) Atom-positional root-mean-square deviation (RMSD) from the original conformation (derived from the crystal structure) during the various simulations. The structural deviation calculated for the individual monomers of simulation X are marked as X_M1 and X_M2, respectively.
(TIF)

S10 Fig. 14-3-3ζ secondary structure analysis. DISICL backbone analysis shows a stable secondary structure during the MD simulation dim. The left side of the Figure shows a schematic representation of the 14-3-3ζ helices. The most populated DISICL classes are depicted in the

following colours: α-helix (ALH)–green, π-helix (PIH)–cyan, helix-cap (HC)–blue, turn-cap (TC)–black, polyproline-like (PP)–brown, turn type I (TI)–magenta, Turn type II (TII)–purple. For a description of abbreviations of DISICL secondary structure elements see S1 Table.
(TIF)

S11 Fig. Pathway dependence of phosphopeptide secondary structure. The population of secondary structure elements is dependent on the replica ID (and DF distance) within the dim DF/HRE-MD simulations. The change of the backbone secondary structure content was analysed by the DISICL algorithm. For a description of abbreviations of DISICL secondary structure elements see S1 Table.
(TIF)

S12 Fig. Pathway dependence of phosphopeptide secondary structure. The population of secondary structure elements is dependent on the replica ID (and DF distance) within the tmon DF/HRE-MD simulations. The change of the backbone secondary structure content was analysed by the DISICL algorithm. For a description of abbreviations of DISICL secondary structure elements see S1 Table.
(TIF)

S13 Fig. The effect of phosphorylation of Ser58 on the electrostatic surface potential (ESP) of 14-3-3ζ. The IS1 and IS2 in A) 14-3-3ζ WT are connected by a mildly positive surface potential patch in contrary to B) 14-3-3ζ phosphorylated at S58, where the connection is disrupted by the negatively charged phosphoserine. Blue, white and red colour represents the positive, neutral, and negative surface patches, respectively. For the clarity, the 14-3-3ζ C-terminal tails are not shown.
(TIF)

S14 Fig. SAXS comparison. Small angle X-ray **scattering c**alculated from the crystal structure (cryst, pdb code 2WHO), the IS1 and UnB stages dim_p1hT, and the unrestrained MD simulations with a low intermonomer twist angles (dim_0.15 dim_p1ht_0.15). The_SAXS curves were calculated using the software Crysol (1), and compared to the experimental SAXS curve (exp) after ensemble averaging. The full length 14-3-3ζ protein sample used for SAXS measurements was expressed and purified as described in Hritz et al.[4]. SAXS data were collected on the beamline BM29 BioSAXS ESFR in Grenoble, France. The concentrations of the 14-3-3-WT during the measurement were 1.18; 2.35 and 4.70 mg/ml in the 50 mM Tris buffer, pH = 8.0. The data were recorded at 20.12˚C using the pixel 1M PILATUS detector at a sample-detector distance of 2.867 m, and a wavelength (λ) of 0.099 nm, covering the range of momentum transfer $0.025\ nm^{-1} < s < 5\ nm^{-1}$ (s = 4π sin(θ)/λ, where 2θ is the scattering angle). No radiation damage was observed during the data collection.The data were processed using standard procedures with the PRIMUS software (2). Solvent contributions (buffer backgrounds collected before and after the protein sample) were averaged and subtracted from the associated protein sample. A slight concentration dependency was noticeable. Therefore, the scattering curves collected at different concentrations were used to obtain a final zero concentration scattering curve through extrapolation according to the guidelines provided by (3).
(TIF)

Acknowledgments

We kindly acknowledge Dr. Anita de Ruiter, for support on the calculation of WHAM profiles, Vojtěch Zapletal for SAXS measurements, Erik Župa for help with 14-3-3ζ sample purification and Jamie McDonald for critical reading of the manuscript.

Author Contributions

Conceptualization: Gabor Nagy, Chris Oostenbrink, Jozef Hritz.

Data curation: Gabor Nagy.

Formal analysis: Gabor Nagy.

Funding acquisition: Chris Oostenbrink, Jozef Hritz.

Investigation: Gabor Nagy.

Methodology: Gabor Nagy, Chris Oostenbrink, Jozef Hritz.

Project administration: Jozef Hritz.

Resources: Jozef Hritz.

Supervision: Jozef Hritz.

Validation: Gabor Nagy, Jozef Hritz.

Visualization: Gabor Nagy.

Writing – original draft: Gabor Nagy, Chris Oostenbrink, Jozef Hritz.

Writing – review & editing: Gabor Nagy, Chris Oostenbrink, Jozef Hritz.

References

1. Gardino AK, Smerdon SJ, Yaffe MB. Structural determinants of 14-3-3 binding specificities and regulation of subcellular localization of 14-3-3-ligand complexes: A comparison of the X-ray crystal structures of all human 14-3-3 isoforms. Semin Cancer Biol. 2006; 16: 173–182. https://doi.org/10.1016/j.semcancer.2006.03.007 PMID: 16678437

2. Yaffe MB, Rittinger K, Volinia S, Caron PR, Aitken A, Leffers H, et al. The Structural Basis for 14-3-3: Phosphopeptide Binding Specificity. Cell. 1997; 91: 961–971. https://doi.org/10.1016/S0092-8674(00)80487-0 PMID: 9428519

3. Johnson C, Crowther S, Stafford MJ, Campbell DG, Toth R, MacKintosh C. Bioinformatic and experimental survey of 14-3-3-binding sites. Biochem J. 2010; 427: 69–78. https://doi.org/10.1042/BJ20091834 PMID: 20141511

4. Hritz J, Byeon I-JL, Krzysiak T, Martinez A, Sklenar V, Gronenborn AM. Dissection of Binding between a Phosphorylated Tyrosine Hydroxylase Peptide and 14-3-3 zeta: A Complex Story Elucidated by NMR. Biophys J. 2014; 107: 2185–2194. https://doi.org/10.1016/j.bpj.2014.08.039 PMID: 25418103

5. Yang X, Lee WH, Sobott F, Papagrigoriou E, Robinson CV, Grossmann JG, et al. Structural basis for protein—protein interactions in the 14-3-3 protein family. Proceedings of the National Academy of Sciences. 2006; 103: 17237–17242.

6. de Ruiter A, Oostenbrink C. Protein-Ligand Binding from Distancefield Distances and Hamiltonian Replica Exchange Simulations. J Chem Theory Comput. 2013; 9: 883–892. https://doi.org/10.1021/ct300967a PMID: 26588732

7. Oostenbrink C, de Ruiter A, Hritz J, Vermeulen N. Malleability and Versatility of Cytochrome P450 Active Sites Studied by Molecular Simulations. Curr Drug Metab. 2012; 13: 190–196. PMID: 22208533

8. Kostelecky B, Saurin AT, Purkiss A, Parker PJ, McDonald NQ. Recognition of an intra-chain tandem 14-3-3 binding site within PKCε. EMBO reports. 2009; 10: 983–989. https://doi.org/10.1038/embor.2009.150 PMID: 19662078

9. Molzan M, Kasper S, Roeglin L, Skwarczynska M, Sassa T, Inoue T, et al. Stabilization of Physical RAF/14-3-3 Interaction by Cotylenin A as Treatment Strategy for RAS Mutant Cancers. ACS Chem Biol. 2013; 8: 1869–1875. https://doi.org/10.1021/cb4003464 PMID: 23808890

10. Muslin AJ, Tanner JW, Allen PM, Shaw AS. Interaction of 14-3-3 with signaling proteins is mediated by the recognition of phosphoserine. Cell. 1996; 84: 889–897. PMID: 8601312

11. Truong AB, Masters SC, Yang HZ, Fu HA. Role of the 14-3-3 C-terminal loop in ligand interaction. Proteins. 2002; 49: 321–325. https://doi.org/10.1002/prot.10210 PMID: 12360521

12. Williams DM, Ecroyd H, Goodwin KL, Dai H, Fu H, Woodcock JM, et al. NMR spectroscopy of 14-3-3ζ reveals a flexible C-terminal extension: differentiation of the chaperone and phosphoserine-binding

activities of 14-3-3ζ. Biochemical Journal. 2011; 437: 493–503. https://doi.org/10.1042/BJ20102178 PMID: 21554249

13. Hu G, Li H, Liu J-Y, Wang J. Insight into Conformational Change for 14-3-3σ Protein by Molecular Dynamics Simulation. International Journal of Molecular Sciences. 2014; 15: 2794–2810. https://doi.org/10.3390/ijms15022794 PMID: 24552877

14. Sluchanko NN, Gusev NB. Oligomeric structure of 14-3-3 protein: What do we know about monomers? FEBS Letters. 2012; 586: 4249–4256. https://doi.org/10.1016/j.febslet.2012.10.048 PMID: 23159940

15. Molzan M, Ottmann C. Synergistic Binding of the Phosphorylated S233- and S259-Binding Sites of C-RAF to One 14-3-3ζ Dimer. Journal of Molecular Biology. 2012; 423: 486–495. https://doi.org/10.1016/j.jmb.2012.08.009 PMID: 22922483

16. Schmid N, Christ CD, Christen M, Eichenberger AP, van Gunsteren WF. Architecture, implementation and parallelisation of the GROMOS software for biomolecular simulation. Comput Phys Commun. 2012; 183: 890–903. https://doi.org/10.1016/j.cpc.2011.12.014

17. Eichenberger AP, Allison JR, Dolenc J, Geerke DP, Horta BAC, Meier K, et al. GROMOS plus plus Software for the Analysis of Biomolecular Simulation Trajectories. J Chem Theory Comput. 2011; 7: 3379–3390. https://doi.org/10.1021/ct2003622 PMID: 26598168

18. Schrödinger, LLC. The PyMOL Molecular Graphics System.

19. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: application to microtubules and the ribosome. Proceedings of the National Academy of Sciences. 2001; 98: 10037–10041.

20. Nagy G, Oostenbrink C. Dihedral-Based Segment Identification and Classification of Biopolymers I: Proteins. J Chem Inf Model. 2014; 54: 266–277. https://doi.org/10.1021/ci400541d PMID: 24364820

21. Sali A, Blundell T. Comparative Protein Modeling by Satisfaction of Spatial Restraints. J Mol Biol. 1993; 234: 779–815. https://doi.org/10.1006/jmbi.1993.1626 PMID: 8254673

22. Schmid N, Eichenberger AP, Choutko A, Riniker S, Winger M, Mark AE, et al. Definition and testing of the GROMOS force-field versions 54A7 and 54B7. European Biophysics Journal. 2011; 40: 843–856. https://doi.org/10.1007/s00249-011-0700-9 PMID: 21533652

23. Petrov D, Margreitter C, Grandits M, Oostenbrink C, Zagrovic B. A Systematic Framework for Molecular Dynamics Simulations of Protein Post-Translational Modifications. Wei G, editor. PLoS Computational Biology. 2013; 9: e1003154. https://doi.org/10.1371/journal.pcbi.1003154 PMID: 23874192

24. Berendsen HJC, Postma JPM, Van Gunsteren WF, Hermans J. Interaction models for water in relation to protein hydration. Intermolecular Forces. 1981; 11: 331–338.

25. Ryckaert J, Ciccotti G, Berendsen H. Numerical-Integration of Cartesian Equations of Motion of a System with Constraints—Molecular-Dynamics of N-Alkanes. J Comput Phys. 1977; 23: 327–341. https://doi.org/10.1016/0021-9991(77)90098-5

26. Berendsen H, Postma J, Van Gunsteren W, Dinola A, Haak J. Molecular-Dynamics with Coupling to an External Bath. J Chem Phys. 1984; 81: 3684–3690. https://doi.org/10.1063/1.448118

27. Tironi I, Sperb R, Smith P, Van Gunsteren W. A generalized reaction field method for molecular-dynamics simulations. J Chem Phys. 1995; 102: 5451–5459. https://doi.org/10.1063/1.469273

28. Heinz T, van Gunsteren W, Hunenberger P. Comparison of four methods to compute the dielectric permittivity of liquids from molecular dynamics simulations. J Chem Phys. 2001; 115: 1125–1136. https://doi.org/10.1063/1.1379764

29. Amadei A, Chillemi G, Ceruso MA, Grottesi A, Di Nola A. Molecular dynamics simulations with constrained roto-translational motions: Theoretical basis and statistical mechanical consistency. J Chem Phys. 2000; 112: 9–23. https://doi.org/10.1063/1.480557

30. Sugita Y, Kitao A, Okamoto Y. Multidimensional replica-exchange method for free-energy calculations. J Chem Phys. 2000; 113: 6042–6051. https://doi.org/10.1063/1.1308516

31. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. Journal of computational chemistry. 1992; 13: 1011–1021.

32. Doudou S, Burton NA, Henchman RH. Standard free energy of binding from a one-dimensional potential of mean force. J Chem Theory Comput. 2009; 5: 909–918. https://doi.org/10.1021/ct8002354 PMID: 26609600

5 Alchemical (relative) Free energy calculations

Paper 8: Hritz, J.; Lappchen T.; Oostenbrink, C. Calculations of binding affinity between C8-substituted GTP analogs and the bacterial cell-division protein FtsZ. *Eur.Biophys. J.* **2010**, 39, 1573-1580

Paper 9: Hritz, J.; Oostenbrink, C. Efficient free energy calculations for compounds with multiple stable conformations separated by high energy barriers. *J. Phys. Chem. B* **2009**, 113, 12711-12720

Paper 10: Jandova, Z.; Trosanova, Z.; Weisova, V.; Oostenbrink, C.*, Hritz, J.*:Free energy calculations on the stability of the 14-3-3ζ protein. *BBA - Proteins and Proteomics*, **2018**, 1866, 442-450.

The previous chapter describes the challenges of achieving sufficient sampling of conformational space. Applications also included calculations of binding free energies by applying H-REMD using distance-field distance restraints. Although these approaches enhance sampling along the binding pathways by several orders of magnitude with respect to traditional MD – they are computationally still very demanding.

However, in many cases binding free energy differences between the similar type of ligands need to be calculated, i.e. relative rather than absolute binding affinities. This is a typical scenario faced by pharma companies where there is a lead compound regarding a particular protein target and in the following steps, the chemical substitutions are sought that should lead to an increase in binding affinity. For this type of case, instead of calculating individual (absolute) binding affinities, it is possible to calculate the difference between them by calculating the alchemical free energy difference in the following way:

$$\Delta\Delta G_{bind} = \Delta G_{bind}(Y) - \Delta G_{bind}(X) = \Delta G_{YX}(free) - \Delta G_{YX}(bound) \qquad \textbf{(5.1)}$$

The $\mathbf{\Delta G_{YX}}$ in the free or bound state means the free energy difference between two chemically different compounds in the corresponding environment. Rarely, can it be calculated from a single simulation of one of two compounds. More often it is calculated

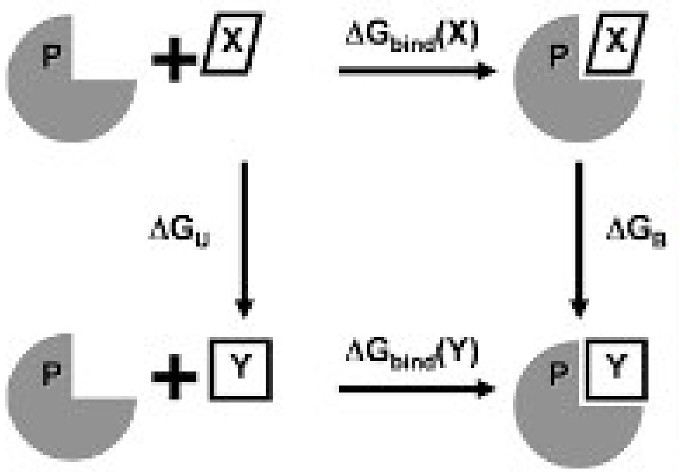


Figure 6: Thermodynamic cycle used for the computation of the difference in binding free energies between compound X and Y into the same protein target (P).

through several intermediate steps –e.g. by alchemical change/perturbation of one that is a direct consequence of thermodynamic cycle in Fig. 6.[28]

It is usually done by a set of MD simulations at different Hamiltonians, defined as, e.g.: $H(\lambda_i) = (1- \lambda_i)H^X + \lambda_i H^Y$ controlled by the λ_i parameter, linking the Hamiltonians H^X representing compound X and H^Y representing compound Y. The free energy difference is then calculated by the free energy perturbation formula:[29]

$$\Delta G_{YX} = \sum_{\lambda_{i=0}}^{1} \Delta G_{\lambda_i}\,;\ \Delta G_{\lambda_i} = -k_B T \ln \langle e^{-\frac{H(\lambda_i)-H(\lambda_{i+1})}{k_B T}} \rangle_{H(\lambda_i)} \tag{5.2}$$

where $k_B T$ is the Boltzmann constant, multiplied by the absolute temperature. The bottleneck of traditional free energy calculation methods is achieving sufficient conformational sampling for systems with high energy barriers.

5.1 Efficient calculations of relative free energies

High energy barriers within biomolecules restrict the efficiency of MD conformational sampling and naturally also prohibit the use of traditional free energy calculation schemes such as alchemical free energy calculation.[30–32] The applicant developed the enhanced sampling – one-step perturbation method (ES-OS) to tackle the calculation of free energy differences between similar C8-substituted GTP analogs in a highly efficient way.[P8] **A single MD simulation of a judiciously chosen reference state (using two sets of soft-core interactions) was sufficient to determine conformational distributions of chemically similar compounds and the free energy differences between them.** The ES-OS method was applied to a set of five biologically relevant 8-substituted GTP analogs having high energy barriers between the anti and syn conformations of the guanine with respect to the ribose part.[P9] The reliability of ES-OS was verified by comparing the results to Hamiltonian replica exchange simulations of GTP and 8-Br-GTP and the experimentally determined ^{3}J(C4,H1') coupling constant for GMP in water. Additional simulations in vacuum, octanol and in the binding site of the FtsZ protein allowed us to calculate differences in the solvation free energies, lipophilicities (log P) and relative binding affinities for FtsZ.[P8,P9] **The applicant also derived the theoretical basis describing how free energy contributions from individual conformational regions could be calculated and have their relationship with the overall free energy derived, leading to a set of multi-conformational free energy formulas.[P9]** The equation for the solvation free energy, i.e. relative free energy difference between compounds A and B in aqueous (aq) with respect to the vacuum (v) environment was derived as:

$$\Delta\Delta G_{AB}(solv) = -k_B T \ln\left(\sum_{i}^{n} \frac{[i]_{(aq)}^{A}[i]_{(v)}^{B}}{[i]_{(v)}^{A}} e^{-\frac{\Delta\Delta G_{AB}^{i}(solv)}{k_B T}}\right) \tag{5.3}$$

where [i] stands for the population of particular conformation state of particular compound (e.g. A) in a given environment. For 8-substituted GTP, the conformational states can be [anti] and [syn]. The usefulness of this equation for the community lies in the fact that when faced with compounds with a high intramolecular energy barrier one does not have to cross it a sufficient number of times in the free energy perturbation calculations. The calculations of $\Delta\Delta G_{AB}^{i}(solv)$ for individual conformational states has fast convergence because there is no need to cross the high energy barrier.

The calculation of relative free energy differences by alchemical perturbation has general validity and thus can be applied not only for the presented analogs of organic compounds but also for the predicting of the impact of site-directed mutagenesis on the overall thermodynamics of the protein. We presented such practical application for the particular mutations of 14-3-3ζ protein.[P10] The predicted thermodynamic stability changes were also compared with the experimental data. [P10]

Paper 8

Hritz, J.; Lappchen T.; Oostenbrink, C. Calculations of binding affinity between C8-substituted GTP analogs and the bacterial cell-division protein FtsZ.
Eur.Biophys. J. **2010**, 39, 1573-1580

Eur Biophys J (2010) 39:1573–1580
DOI 10.1007/s00249-010-0614-y

ORIGINAL PAPER

Calculations of binding affinity between C8-substituted GTP analogs and the bacterial cell-division protein FtsZ

Jozef Hritz · Tilman Läppchen · Chris Oostenbrink

Received: 1 April 2010 / Revised: 21 May 2010 / Accepted: 26 May 2010 / Published online: 18 June 2010

Abstract The FtsZ protein is a self-polymerizing GTPase that plays a central role in bacterial cell division. Several C8-substituted GTP analogs are known to inhibit the polymerization of FtsZ by competing for the same binding site as its endogenous activating ligand GTP. Free energy calculations of the relative binding affinities to FtsZ for a set of five C8-substituted GTP analogs were performed. The calculated values agree well with the available experimental data, and the main contribution to the free energy differences is determined to be the conformational restriction of the ligands. The dihedral angle distributions around the glycosidic bond of these compounds in water are known to vary considerably depending on the physicochemical properties of the substituent at C8. However, within the FtsZ protein, this substitution has a negligible influence on the dihedral angle distributions, which fall within the narrow range of −140° to −90° for all investigated compounds. The corresponding ensemble average of the coupling constants $^{3}J(\mathrm{C4,H1'})$ is calculated to be 2.95 ± 0.1 Hz. The contribution of the conformational selection of the GTP analogs upon binding was quantified from the corresponding populations. The obtained restraining free energy values follow the same trend as the relative binding affinities to FtsZ, indicating their dominant contribution.

Keywords GTP analogs · FtsZ · One-step free energy perturbation · Conformational selection · Restraining free energy · Ensemble average · Heteronuclear coupling constant

Introduction

The cell division protein FtsZ is considered a promising antibacterial target (Vollmer 2006; Huang et al. 2007; Paradis-Bleau et al. 2007; Lock and Harry 2008; Kapoor and Panda 2009), and the recent discovery of a small synthetic FtsZ inhibitor with potent in vitro and in vivo bactericidal activity against multidrug-resistant *Staphylococcus aureus* suggests that these high expectations are justified (Haydon et al. 2008; Czaplewski et al. 2009).

In the presence of guanosine 5′-triphosphate (GTP), FtsZ assembles into a variety of polymeric structures, the nature of which is very much dependent on the exact experimental conditions employed (reviewed by Adams and Errington 2009). Linear protofilaments and protofilament bundles arising from lateral association are among the more frequently studied polymeric species of FtsZ. Polymerization of FtsZ activates its GTPase activity by insertion of acidic residues from the synergy loop into the nucleotide binding pocket of the preceding monomer in the protofilament (Oliva et al. 2004). Despite considerable

J. Hritz · C. Oostenbrink (✉)
Leiden-Amsterdam Center for Drug Research,
Section of Molecular Toxicology,
Department of Chemistry and Pharmacochemistry,
Vrije Universiteit, De Boelelaan 1083,
1081 HV Amsterdam, The Netherlands
e-mail: c.oostenbrink@few.vu.nl

J. Hritz
e-mail: jozef.hritz@gmail.com

T. Läppchen
Department of Biomolecular Engineering,
Philips Research, High Tech Campus 11,
M/S WBC02 P263, 5656 AE Eindhoven, The Netherlands
e-mail: tilman.lappchen@philips.com

C. Oostenbrink
Institute of Molecular Modeling and Simulation,
University of Natural Resources and Applied Life Sciences,
Muthgasse 18, 1190 Vienna, Austria

efforts, FtsZ polymer dynamics, the associated GTPase reaction kinetics, and the modulation of both by pH, nature and concentration of cations, GTP, guanosine 5′-diphosphate (GDP), FtsZ, and certainly regulation by accessory proteins is still not fully understood at a molecular level (Löwe and Amos 1998; Michie and Lowe 2006; Mendieta et al. 2009). In particular, the relationship between GTP hydrolysis and FtsZ polymer dynamics remains controversial. While earlier studies suggested direct exchange of nucleotide in protofilaments (Romberg and Mitchison 2004; Tadros et al. 2006), recent data show that terminal FtsZ subunit exchange is independent of nucleotide state and faster than GTP hydrolysis, supporting the hypothesis that nucleotide exchange occurs only on recycling terminal subunits (Chen and Erickson 2009). In addition, the previously accepted view that the GTP-bound form of the FtsZ protofilament is intrinsically straight while the GDP-bound form is curved has recently been challenged by the finding that FtsZ structures in various crystal forms and nucleotide states did not show evidence of a conformational switch in the FtsZ monomer involving domain movement (Oliva et al. 2007), although it should be taken into account that the crystal structures might not be representative for the GTP- or GDP-bound state but in fact could correspond to a transition state. Strikingly, even 8-morpholino-GTP, one of a series of C8-substituted GTP analogs acting as competitive inhibitors of GTP-driven FtsZ polymerization and GTP hydrolysis, was found to bind to *Aquifex aeolicus* FtsZ in essentially the same way as GDP without inducing any significant conformational changes in the protein (Läppchen et al. 2008). Until now, the molecular basis of the observed inhibitory action of the investigated C8-substituted GTP derivatives has not been completely resolved. Although the C8-morpholino substituent protrudes from the surface of the monomer, the currently available FtsZ protofilament structures (Oliva et al. 2004, 2007) suggest that the inhibitory action cannot be simply attributed to direct steric clashes between the C8 substituent and the next FtsZ monomer in a growing protofilament. It is important to note, however, that stabilization of intersubunit contacts and the rate of GTPase activity are also dependent on the presence of divalent and monovalent cations and pH (Mendieta et al. 2009), suggesting that the C8-substituted GTP derivatives might act by interfering with vital hydrogen-bonding interactions via rearrangement of water molecules and cations in the active site.

In a series of C8-substituted GTP analogs, inhibitory potencies were found to correlate with the corresponding binding affinities to the FtsZ monomer and with the Sterimol parameters of their C8 substituents (Läppchen et al. 2008). Intrigued by this observation, we set out to rationalize these results in terms of binding free energies. C8-substituted GTP analogs with two stable conformations (*anti*, *syn*) separated by high energy barriers belong to a challenging class of compounds for binding affinity calculations (Hritz and Oostenbrink 2007, 2008). Recently we have developed the enhanced sampling one-step free energy perturbation method (ES-OS) that allows for efficient free energy calculations for GTP analogs in explicit solvent based on sufficient sampling of both relevant conformations (Hritz and Oostenbrink 2009).

This paper presents calculations of relative free energies of binding to the FtsZ protein for a set of five C8-substituted GTP analogs in which H8 is replaced by halogen atoms or a methyl group (Fig. 1). Molecular docking simulations of the C8-substituted GTP analogs well reproduced the phosphate and ribose groups of the molecules, while the base was found to be in both the *syn* and the *anti* conformation (Läppchen 2007). The recent high-resolution crystal structure of the *Aquifex aeolicus* FtsZ protein, co-crystallized with 8-morpholino-GTP, shows that also compounds with a bulky substituent at the C8 position bind to the protein in the *anti* conformation (Läppchen et al. 2008), even though in solution the *syn* conformation is expected to be dominant (Davies 1978; Stolarski et al. 1984; Cho and Evans 1991). This observation significantly simplifies the free energy calculations of the compounds in the binding site of the FtsZ protein, because the simulation can be restricted to a single ligand conformation and no high energy barriers need to be crossed. For this reason we calculate the free energy difference between the various compounds bound to the FtsZ protein using the one-step (OS) perturbation method (Liu et al. 1996; Oostenbrink and van Gunsteren 2005). The corresponding values in solution (where both *anti* and *syn* conformations contribute) were calculated earlier using enhanced sampling OS (ES-OS) (Hritz and Oostenbrink 2009).

The computationally predicted values are compared with the available experimental binding affinities to nucleotide-free *Methanococcus jannaschii* FtsZ protein (Läppchen et al. 2008). The main contributions to the

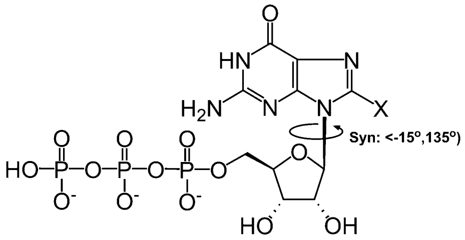


Fig. 1 Structure of C8-substituted analogs of GTP in *syn* conformation, where X = H, F, Cl, Br, CH_3. Conformational transitions between the *syn* and *anti* conformations occur by rotation around the glycosidic bond indicated by the *arrow*. The glycosidic dihedral angle (χ) is defined over atoms: C4-N9-C1′-O4′

relative binding free energies are analyzed and provide an explanation for the empirically observed correlation between Sterimol parameters of C8 substituents and binding affinities (Läppchen et al. 2008).

Materials and methods

One-step free energy perturbation (OS) (Liu et al. 1996)

The aim of OS is to efficiently determine the free energy differences between chemically similar compounds from a single molecular dynamics (MD) simulation of a designed reference compound, *S*. The free energy between the reference compound (*S*) and the real compounds (*R*), can be calculated from a simulation using the reference Hamiltonian and applying Zwanzig's perturbation formula (Zwanzig 1954):

$$\Delta G_{SR} = G_R - G_S = -k_{\mathrm{B}}T\ln\left\langle \mathrm{e}^{\frac{-(H_R(\mathbf{q},\mathbf{p})-H_S(\mathbf{q},\mathbf{p}))}{k_{\mathrm{B}}T}}\right\rangle_S \qquad (1)$$

where k_{B} is the Boltzmann constant, T is absolute temperature, and H_S and H_R are the Hamiltonians for the (soft) reference compound and one of the real compounds, respectively. The angular brackets indicate an ensemble average over the positions, **q**, and momenta, **p**, obtained from a simulation of the reference state.

The ensemble average $\langle A\rangle^R$ of a property A for the real compound (*R*) can be estimated by reweighting the individual values corresponding to the particular configuration ($\mathbf{q}_i$, $\mathbf{p}_i$) for the reference compound, A_i^S by a Boltzmann factor p_i^{SR}:

$$\langle A\rangle^R = \sum_i p_i^{SR}A_i^S;\ \text{with}\ p_i^{SR} = \frac{\mathrm{e}^{\frac{-(H_R(\mathbf{q}_i,\mathbf{p}_i)-H_S(\mathbf{q}_i,\mathbf{p}_i))}{k_BT}}}{\sum_i \mathrm{e}^{\frac{-(H_R(\mathbf{q}_i,\mathbf{p}_i)-H_S(\mathbf{q}_i,\mathbf{p}_i))}{k_BT}}}. \qquad (2)$$

Simulation setup

The crystal structure of *Aquifex aeolicus* FtsZ complexed with 8-morpholino-GTP (Läppchen et al. 2008) was downloaded from the protein databank (Berman et al. 2003) (www.pdb.org; PDB ID: 2R75, chain B). The morpholino substituent was replaced by a single bromine (Br) atom. Molecular dynamics (MD) simulations of the FtsZ protein in complex with 8-Br-GTP were performed using the GROMOS05 simulation package (Christen et al. 2005) in combination with the GROMOS 53A6 force field (Oostenbrink et al. 2004). Force field parameters for five C8-substituted GTP analogs are available in the supplementary material of (Hritz and Oostenbrink 2009). The magnesium cation and all 283 crystallographic water oxygens were kept and position restrained during initial equilibration steps.

Rectangular periodic boundary conditions were used with an additional 24,544 water molecules; 13 of them were replaced by 13 sodium cations in order to electroneutralize the whole system (note: crystallographic waters were not considered for replacement by sodium cations). The system finally contained 24,814 explicit simple point charge (SPC) water molecules (Berendsen et al. 1981). All bonds were constrained, using the SHAKE algorithm (Ryckaert et al. 1977), with relative geometric accuracy of 10^{-4}, allowing for a time step of 2 fs in the leapfrog integration scheme (Hockney 1970). After a steepest-descent minimization to remove bad contacts between molecules, initial velocities were randomly assigned from a Maxwell–Boltzmann distribution at 298 K, according to the atomic masses. The temperature was kept constant using weak coupling (Berendsen et al. 1984) to a bath of 298 K with a relaxation time of 0.1 ps. The solute molecule and solvent were independently coupled to the heat bath. The pressure was controlled using isotropic weak coupling to atmospheric pressure (Berendsen et al. 1984) with a relaxation time of 0.5 ps. van der Waals and electrostatic interactions were calculated using a triple range cutoff scheme. Interactions within a short-range cutoff of 0.8 nm were calculated every time step from a pair list that was generated every five steps. At these time points, interactions between 0.8 and 1.4 nm were also calculated and kept constant between updates. A reaction-field contribution was added to the electrostatic interactions and forces to account for a homogeneous medium outside the long-range cutoff, using the relative permittivity (61) of SPC water (Tironi et al. 1995). Selected interactions were calculated using a soft-core van der Waals and electrostatic interaction between atoms *i* and *j* (Beutler et al. 1994):

$$E_{ij}^{\mathrm{vdw}}(r_{ij},\lambda_{\mathrm{vdw}}) = \left(\frac{C12_{ij}}{A_{ij}(\lambda_{\mathrm{vdw}})+r_{ij}^6} - C6_{ij}\right)\frac{1}{A_{ij}(\lambda_{\mathrm{vdw}})+r_{ij}^6}, \qquad (3)$$

$$E_{ij}^{\mathrm{el}}(r_{ij},\lambda_{\mathrm{el}}) = \frac{q_iq_j}{4\pi\varepsilon}\frac{1}{\sqrt{B_{ij}(\lambda_{\mathrm{el}})+r_{ij}^2}}, \qquad (4)$$

with r_{ij} being the interatomic distance; $A_{ij}(\lambda_{\mathrm{vdw}}) = \alpha_{\mathrm{vdw}}\frac{C12_{ij}}{C6_{ij}}\lambda_{\mathrm{vdw}}^2$ and $B_{ij}(\lambda_{\mathrm{el}}) = \alpha_{\mathrm{el}}\lambda_{\mathrm{el}}^2$. $C12_{ij}$ and $C6_{ij}$ are the Lennard–Jones parameters for atom pair i and j, q_i and q_j are the partial charges of particles i and j, and α_{vdw} and α_{el} are the softness constants. In the current study we used in all simulations $\alpha_{\mathrm{vdw}}\lambda_{\mathrm{vdw}}^2 = \alpha_{\mathrm{el}}\lambda_{\mathrm{el}}^2 = 0.3775$, the value empirically known to work well in standard OS (Schäfer et al. 1999; Oostenbrink and van Gunsteren 2004). It can be seen that at longer distances [$r_{ij} \gg A(\lambda_{\mathrm{vdw}})$ and $r_{ij} \gg B(\lambda_{\mathrm{el}})$] the soft-core interaction approximates the interaction for normal atoms and that they differ mostly at short distances between the atoms [$r_{ij} \le A(\lambda_{\mathrm{vdw}})$ or $r_{ij} \le B(\lambda_{\mathrm{el}})$]. The conformational space of the C8-substituted GTP analogs was adequately represented within the FtsZ protein by the reference state 8-soft_Br-GTP. Here, a bromine substituent was placed at position 8, for which all nonbonded interactions with the rest of the system (including protein and solvent) were evaluated as soft-core interactions (Eq. 3, 4). A single MD simulation of the reference state FtsZ:8-soft_Br-GTP was performed for 2 ns, and system coordinates were saved every 0.2 ps.

Results

Figure 2 presents the normalized dihedral angle (χ) distributions for the five C8-substituted analogs of GTP in complex with the FtsZ protein as calculated by reweighting the probabilities of individual configurations of the MD trajectory of the reference state (8-soft_Br-GTP) using Eq. (2). The distributions indicate that all five studied GTP analogs occupy a very similar conformational range, $\chi \in [-140°, -90°]$, when bound to FtsZ. This range is only about half of the *anti* conformational range observed for these compounds free in solution. The rest of the *anti* range is strongly prohibited by steric repulsion of Phe175 and hydrogen bonding with Asp179 in the FtsZ active site. The different glycosidic dihedral angle distributions in the bound and free state have a direct influence on the ^3J coupling constant values, calculated as ensemble averages $\left\langle {}^3J(\mathrm{C4,H1'})\right\rangle^R$ using Eq. 2. $^3J(\mathrm{C4,H1'})_i^S$ values for the individual configurations of the reference compound (S) were calculated using the Karplus equation (Karplus 1959):

$$^3J(\mathrm{C4,H1'})_i^S = A\cos^2(\chi_i + 120°) + B\cos(\chi_i + 120°) + C \qquad (5)$$

with the Karplus coefficients, $A = 4.4$ Hz, $B = -1.4$ Hz, and $C = 0.1$ Hz (Trantirek et al. 2002). The ^3J values for both states are listed in the caption of Fig. 2. While there is a large difference between the calculated $\left\langle {}^3J(\mathrm{C4,H1'})\right\rangle^R$ values of individual compounds in water, the values are almost identical when bound to the FtsZ protein and fall in the very narrow range of 2.95 ± 0.1 Hz. A separate simulation of real 8-Br-GTP bound to the protein yielded an average ^3J-value of 3.01 Hz.

Free energy differences of the five compounds relative to the reference 8-soft_Br-GTP bound to FtsZ as calculated by OS are listed on the left side of the thermodynamic cycle presented in Fig. 3. The free energies relative to GTP ($\Delta G^{\mathrm{OS}}_{\mathrm{GTP,R}}(\mathrm{FtsZ})$) obtained by cycle closure (right side) are listed in the second column of Table 1. Relative differences in binding affinities, $\Delta\Delta G^{\mathrm{calc}}_{\mathrm{GTP,R}}(\mathrm{bind})$, are calculated using Eq. 6.

$$\Delta\Delta G^{\mathrm{calc}}_{\mathrm{GTP,R}}(\mathrm{bind}) = \Delta G^{\mathrm{OS}}_{\mathrm{GTP,R}}(\mathrm{FtsZ}) - \Delta G^{\mathrm{ES-OS}}_{\mathrm{GTP,R}}(\mathrm{aq}), \qquad (6)$$

where $\Delta G^{\mathrm{ES-OS}}_{\mathrm{GTP,R}}(\mathrm{aq})$ are the free energy values relative to GTP in a water environment as calculated by the ES-OS method in our previous study (Hritz and Oostenbrink 2009). It is important to note, despite the different soft reference states used in water and for the FtsZ bound state, the relative free energy differences between real compounds remain valid for the calculation of relative binding

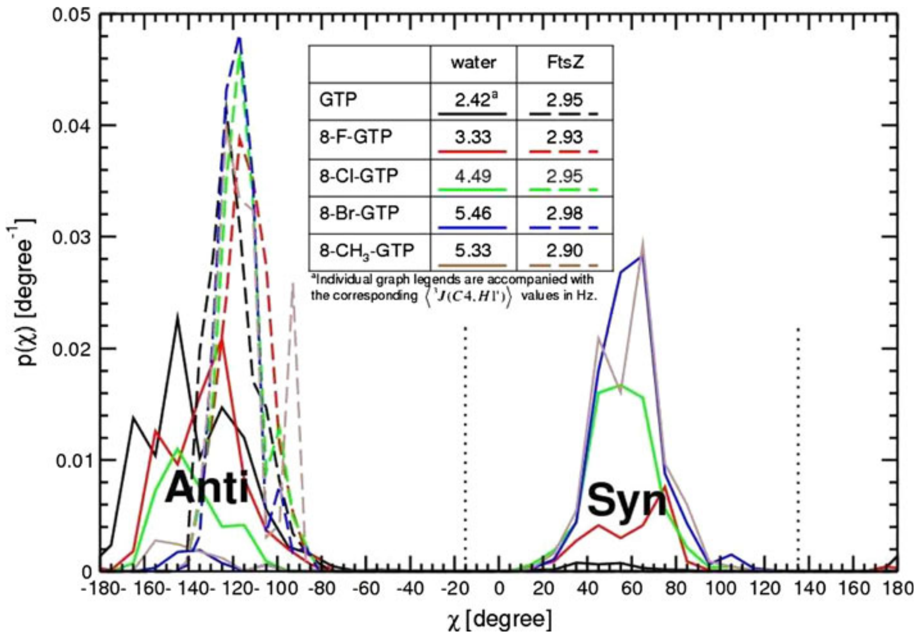


Fig. 2 Normalized dihedral angle (χ) distributions for C8-substituted analogs of GTP as calculated by ES-OS in water environment [*solid lines* (Hritz and Oostenbrink 2009)] and by OS in bound state to FtsZ protein (*dashed lines*). Ensemble averages of $\left\langle {}^3J(\mathrm{C4,H1'})\right\rangle$ values in Hz are listed in the legend above individual line symbols. *Anti* and *syn* conformational ranges are indicated by *dotted lines*. Note that the dihedral angle in the protein simulations remains in the range $[-140°, -90°]$.

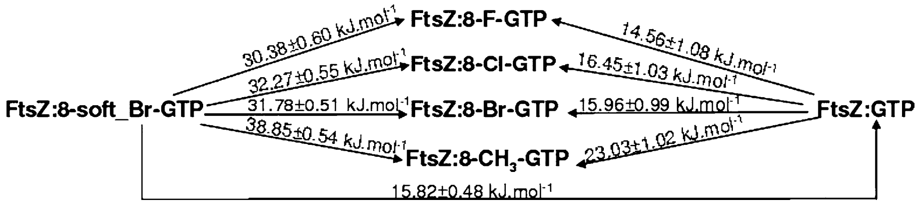


Fig. 3 Thermodynamic cycles used to calculate the free energy difference between compounds in complex with FtsZ protein. The free energy differences between the reference compound (8-soft_Br-GTP) and the real compounds were calculated by OS using the perturbation formula (Zwanzig 1954). The free energies relative to GTP were obtained by cycle closure

Table 1 Comparison of binding affinities to FtsZ protein for C8-substituted GTP analogs with respect to GTP as obtained from computational simulations, $\Delta\Delta G^{\mathrm{calc}}_{\mathrm{GTP,R}}(\mathrm{bind})$, and corresponding experimental values, $\Delta\Delta G^{\mathrm{exp}}_{\mathrm{GTP,R}}(\mathrm{bind})$ to *Methanococcus jannaschii* FtsZ (Läppchen et al. 2005, 2008)

	$\Delta G^{\mathrm{OS}}_{\mathrm{GTP,R}}(\mathrm{FtsZ})$ (kJ mol^{-1})	$\Delta G^{\mathrm{ES-OS}}_{\mathrm{GTP,R}}(\mathrm{aq})$ (kJ mol^{-1})	$\Delta\Delta G^{\mathrm{calc}}_{\mathrm{GTP,R}}(\mathrm{bind})$ (kJ mol^{-1})	$\Delta\Delta G^{\mathrm{exp}}_{\mathrm{GTP,R}}(\mathrm{bind})$ (kJ mol^{-1})
GTP	0	0	0	0
8-F-GTP	14.6 ± 1.1	15.0 ± 1.4	−0.8 ± 2.5	–
8-Cl-GTP	16.5 ± 1.0	12.8 ± 1.3	3.7 ± 2.3	8.0 ± 4.4
8-Br-GTP	16.0 ± 1.0	5.9 ± 1.3	10.1 ± 2.3	9.2 ± 1.8
8-CH$_3$-GTP	23.0 ± 1.0	12.6 ± 1.4	10.5 ± 2.4	~8.7[a]

Values of free energy differences in complex with FtsZ relative to GTP ($\Delta G^{\mathrm{OS}}_{\mathrm{GTP,R}}(\mathrm{FtsZ})$) were derived from the thermodynamic cycles shown in Fig. 3 based on the calculated values by OS free energy perturbation calculations. Relative free energy values in explicit water, $\Delta G^{\mathrm{ES-OS}}_{\mathrm{GTP,R}}(\mathrm{aq})$, were calculated by the ES-OS method (Hritz and Oostenbrink 2009)

[a] Value for $\Delta\Delta G^{\mathrm{exp}}_{\mathrm{GTP,8-CH_3-GTP}}(\mathrm{bind})$ was estimated from an experimental 50% inhibition concentration (IC$_{50}$) value of GTPase activity

affinities. The computationally predicted values are compared with the last column in Table 1, which lists the only available experimental binding affinities, to *Methanococcus jannaschii* FtsZ (Läppchen et al. 2008). We do not expect that the binding affinities will deviate significantly between *M. jannaschi* FtsZ and *A. aeolicus* FtsZ, since X-ray structures of FtsZ from both species show very similar monomer interfaces, which comprise the nucleotide binding pocket (Oliva et al. 2007). Moreover, for both proteins it was observed that different ligands do not lead to different conformation of the binding site (Oliva et al. 2007).

Discussion

The calculated relative binding affinities of C8-substituted GTP analogs to the FtsZ protein compare well to the experimental values, with root-mean-square error of 2.7 kJ mol^{-1} for 8-Cl-GTP, 8-Br-GTP, and 8-CH$_3$-GTP. Note that no empirical parameters, other than the force field to calculate the interactions, were used to obtain these values. Table 1 nicely illustrates that the relative free energies in both environments are equally important for the final free energies. The bromine and methyl substituent are both predicted to be much weaker binders with respect to chloride, by roughly ~6.5 kJ mol^{-1}. However, while the dominant contribution to this difference comes from the water environment for 8-Br-GTP, it comes from the bound state for 8-Me-GTP. It is also interesting to note that the GTP analog that is predicted to have the highest affinity is 8-F-GTP, resulting from similar contributions in both environments. No experimental data is available for this compound, as difficulties concerning its synthesis were only recently resolved (Liu et al. 2006; Ghosh et al. 2007).

It is usually considered that the *anti* conformation corresponds to a low and the *syn* conformation to a high value of the coupling constant, ^3J(C4,H1’) (Stolarski et al. 1984; Cho and Evans 1991; Ippel et al. 1996; Trantirek et al. 2002). Therefore it may seem surprising that the ensemble average $\left\langle {}^3J(\mathrm{C4,H1'})\right\rangle$ of GTP in water (to which the *syn* conformation contributes with ~3%) is calculated to be lower (2.4 Hz) than the value obtained for the bound state of GTP (2.95 Hz) in which only the *anti* conformation is observed (Fig. 2). This finding follows directly from the fact that within the FtsZ binding site the χ dihedral angle is restricted within tighter bounds for all five C8-substituted GTP analogs as compared with in aqueous solution.

The paradigm of conformational selection describes the binding process between ligand and protein by taking multiple conformations of the protein into account. It states that the correct conformation is selected from the complete ensemble of possible conformations, which are all populated to a given extent (Carlson 2002). Here, we apply this model to the GTP analogs and quantify the contribution of

Table 2 The free energy of restraining the conformation of C8-substituted GTP analogs to a conformational range $\chi \in [-140^\circ, -90^\circ]$ in water

	$[\langle -140^\circ, -90^\circ \rangle]^R_{(\mathrm{aq})}$	$\Delta G_R^{\mathrm{rest}}(\mathrm{aq})$ (kJ mol^{-1})	$\Delta\Delta G_{\mathrm{GTP,R}}^{\mathrm{rest}}(\mathrm{aq})$ (kJ mol^{-1})
GTP	50%	1.7	0.0
8-F-GTP	48%	1.8	0.1
8-Cl-GTP	18%	4.2	2.5
8-Br-GTP	3.3%	8.4	6.7
8-CH$_3$-GTP	4.0%	8.0	6.2

The free energy values $\Delta G_R^{\mathrm{rest}}(\mathrm{aq})$ are calculated from the populations $[\langle -140^\circ, -90^\circ \rangle]^R_{(\mathrm{aq})}$. A comparison of the relative free energies with respect to GTP, $\Delta\Delta G_{\mathrm{GTP,R}}^{\mathrm{rest}}(\mathrm{aq})$, indicates a significant contribution to the total relative binding affinity, $\Delta\Delta G_{\mathrm{GTP,R}}^{\mathrm{calc}}(\mathrm{bind})$ (Table 1)

the conformational restriction to the relative binding affinities. For this, we calculate the free energy that is needed to restrict the GTP analogs from their unbound state to a conformation that is possible in the protein, using Eq. 7.

$$\Delta G_R^{\mathrm{rest}}(\mathrm{aq}) = -kT \ln[\langle -140^\circ, -90^\circ \rangle]^R_{(\mathrm{aq})}, \tag{7}$$

where the unitless populations $[\langle -140^\circ, -90^\circ \rangle]^R_{(\mathrm{aq})}$ were obtained from a simple integration over the selected range of a normalized dihedral angle distribution of compound R in water (solid lines in Fig. 2). The relative values with respect to GTP in water, $\Delta\Delta G_{\mathrm{GTP,R}}^{\mathrm{rest}}(\mathrm{aq})$, are listed in the fourth column of Table 2. It is interesting to note that these follow the same trend as the $\Delta\Delta G_{\mathrm{GTP,R}}^{\mathrm{calc}}(\mathrm{bind})$ values (Table 1, fourth column). It seems that the conformational restriction of the GTP analogs accounts for roughly 65% of the difference in affinity for the FtsZ protein. We emphasize that, while the calculation of $\Delta\Delta G_{\mathrm{GTP,R}}^{\mathrm{calc}}(\mathrm{bind})$ requires extensive simulations in water and in the binding site of the FtsZ protein, the restraining free energy values were calculated from a simple integration of dihedral angle distributions obtained from the water simulation only. Our results strongly indicate that the previously reported correlation between the binding affinities and Sterimol B1 parameter of the substituents (Läppchen et al. 2008) follows from the differences in the restraining free energies of the GTP analogs, which in turn stem from a different *syn–anti* balance of the compounds, when free in solution.

In agreement with the experimental observations, all compounds adopt roughly the same conformation when bound to the protein, while the various substituents lead to different conformational ensembles when the compounds are free in solution. The conformational restriction upon binding to the FtsZ protein accounts for roughly 65% of the differences in binding affinity. As the binding affinity seems to be for a large part dependent on the conformational ensemble of the studied C8-substituted GTP analogs in water, a low specificity may be expected against other GTPases in which the binding site restricts the conformational freedom of C8-substituted GTP analogs in a similar manner. The inhibitory activities for such proteins will presumably display the same trend. A notable exception is tubulin, the eukaryotic homolog of FtsZ, where GTP analogs with small C8 substituents promoted assembly more than GTP itself.

Conclusions

Relative free energy differences of five C8-substituted GTP analogs in complex with the bacterial cell-division protein FtsZ were calculated using the one-step perturbation (OS) method. Combined with previous values for the water environment as obtained from enhanced sampling OS we calculated the relative binding free energies for these compounds. The results are in good agreement with the available experimental binding affinities. The dihedral angle distributions within the FtsZ binding site are much narrower as compared with those obtained in water. This results in significantly different ensemble averages of the ^3J coupling constants.

The contribution of conformational selection for the C8-substituted GTP analogs was quantified by calculating the restraining free energy in water that is needed to restrain the dihedral angle to the conformational range that is accessible within the binding site of the FtsZ protein. The restraining free energies follow the same trend as the binding free energies, accounting for about 65% of the differences in affinity. This suggests low specificity towards the FtsZ protein, because the same trend can be expected for any GTPases in which the binding site restricts the conformational freedom of C8-substituted GTP analogs in a similar manner. Our results also suggest an explanation for the empirically observed correlation between the Sterimol parameters and the binding affinity to the FtsZ protein.

Acknowledgments We gratefully acknowledge financial support from the Netherlands Organization for Scientific Research (NWO), Horizon Breakthrough grant 935.18.018 (J.H.) and VENI grant 700.55.401 (C.O.).

References

Adams DW, Errington J (2009) Bacterial cell division: assembly, maintenance and disassembly of the Z-ring. Nat Rev Microbiol 7:642–653

Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J (1981) Interaction models for water in relation to protein hydration. In: Pullman B (ed) Intermolecular forces. Reidel, Dordrecht, pp 331–342

Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular-dynamics with coupling to an external bath. J Chem Phys 81:3684–3690

Berman HM, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. Nat Struct Biol 10:980

Beutler TC, Mark AE, van Schaik RC, Gerber PR, van Gunsteren WF (1994) Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. Chem Phys Lett 222:529–539

Carlson HA (2002) Protein flexibility and drug discovery: how to hit a moving target. Curr Opin Chem Biol 6:447–452

Chen YD, Erickson HP (2009) FtsZ filament dynamics at steady state: subunit exchange with and without nucleotide exchange. Biochemistry 48:6664–6673

Cho BP, Evans FE (1991) Correlation between nmr spectral parameters of nucleosides and its implication to the conformation about the glycosyl bond. Biochem Biophys Res Commun 180:273–278

Christen M, Hunenberger PH, Bakowies D, Baron R, Burgi R, Geerke DP, Heinz TN, Kastenholz MA, Krautler V, Oostenbrink C, Peter C, Trzesniak D, Van Gunsteren WF (2005) The GROMOS software for biomolecular simulation: GROMOS05. J Comput Chem 26:1719–1751

Czaplewski LG, Collins I, Boyd EA, Brown D, East SP, Gardiner M, Fletcher R, Haydon DJ, Henstock V, Ingram P, Jones C, Noula C, Kennison L, Rockley C, Rose V, Thomaides-Brears HB, Ure R, Whittaker M, Stokes NR (2009) Antibacterial alkoxybenzamide inhibitors of the essential bacterial cell division protein FtsZ. Bioorg Med Chem Lett 19:524–527

Davies DB (1978) Conformations of nucleosides and nucleotides. Progress Nucl Magn Res Spectr 12:135–225

Ghosh A, Lagisetty P, Zajc B (2007) Direct synthesis of 8-fluoro purine nucleosides via metalation-fluorination. J Org Chem 72:8222–8226

Haydon DJ, Stokes NR, Ure R, Galbraith G, Bennett JM, Brown DR, Baker PJ, Barynin VV, Rice DW, Sedelnikova SE, Heal JR, Sheridan JM, Aiwale ST, Chauhan PK, Srivastava A, Taneja A, Errington J, Czaplewski LG (2008) An inhibitor of FtsZ with potent and selective anti-staphylococcal activity. Science 321:1673–1675

Hockney RW (1970) The potential calculations and some applications. Meth Comput Phys 9:136–211

Hritz J, Oostenbrink C (2007) Optimization of replica exchange molecular dynamics by fast mimicking. J Chem Phys 127:204104

Hritz J, Oostenbrink C (2008) Hamiltonian replica exchange molecular dynamics using soft-core interactions. J Chem Phys 128:144121

Hritz J, Oostenbrink C (2009) Efficient free energy calculations for compounds with multiple stable conformations separated by high energy barriers. J Phys Chem B 113:12711–12720

Huang Q, Tonge PJ, Slayden RA, Kirikae J, Ojima I (2007) FtsZ: a novel target for tuberculosis drug discovery. Curr Top Med Chem 7:527–543

Ippel JH, Wijmenga SS, de Jong R, Heus HA, Hilbers CW, de Vroom E, van der Marel GA, van Boom JH (1996) Heteronuclear scalar couplings in the bases and sugar rings of nucleic acids: their determination and application in assignment and conformational analysis. Magn Reson Chem 34:S156–S176

Kapoor S, Panda D (2009) Targeting FtsZ for antibacterial therapy: a promising avenue. Exp Opin Therap Targets 13:1037–1051

Karplus M (1959) Contact electron-spin coupling of nuclear magnetic moments. J Chem Phys 30:11–15

Läppchen T (2007) Synthesis of GTP analogues and evaluation of their effect on the antibiotic target FtsZ and its eukaryotic homologue tubulin. PhD thesis, University of Amsterdam (available online http://dare.uva.nl/en/record/211017); Chapter 3

Läppchen T, Hartog AF, Pinas VA, Koomen GJ, den Blaauwen T (2005) GTP analogue inhibits polymerization and GTPase activity of the bacterial protein FtsZ without affecting its eukaryotic homologue tubulin. Biochemistry 44:7879–7884

Läppchen T, Pinas VA, Hartog AF, Koomen GJ, Schaffner-Barbero C, Andreu JM, Trambaiolo D, Löwe J, Juhem A, Popov AV, den Blaauwen T (2008) Probing FtsZ and tubulin with C8-substituted GTP analogs reveals differences in their nucleotide binding sites. Chem Biol 15:189–199

Liu HY, Mark AE, van Gunsteren WF (1996) Estimating the relative free energy of different molecular states with respect to a single reference state. J Phys Chem 100:9485–9494

Liu J, Barrio J, Satyamurthy N (2006) Kinetics and mechanism of the defluorination of 8-fluoropurine nucleosides in basic and acidic media. J Fluor Chem 127:1175–1187

Lock RL, Harry EJ (2008) Cell-division inhibitors: new insights for future antibiotics. Nat Rev Drug Discov 7:324–338

Löwe J, Amos LA (1998) Crystal structure of the bacterial cell-division protein FtsZ. Nature 391:203–206

Mendieta J, Rico AI, Lopez-Vinas E, Vicente M, Mingorance J, Gomez-Puertas P (2009) Structural and functional model for ionic (K+/Na+) and pH dependence of GTPase activity and polymerization of FtsZ, the prokaryotic ortholog of Tubulin. J Mol Biol 390:17–25

Michie KA, Lowe J (2006) Dynamic filaments of the bacterial cytoskeleton. Ann Rev Biochem 75:467–492

Oliva MA, Cordell SC, Löwe J (2004) Structural insights into FtsZ protofilament formation. Nat Struct Mol Biol 11:1243–1250

Oliva MA, Trambaiolo D, Löwe J (2007) Structural insights into the conformational variability of FtsZ. J Mol Biol 273:1229–1242

Oostenbrink C, van Gunsteren WF (2004) Free energies of binding of polychlorinated biphenyls to the estrogen receptor from a single simulation. Proteins 54:237–246

Oostenbrink C, van Gunsteren WF (2005) Free energies of ligand binding for structurally diverse compounds. Proc Natl Acad Sci U S A 102:6750–6754

Oostenbrink C, Villa A, Mark AE, van Gunsteren WF (2004) A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J Comput Chem 25:1656–1676

Paradis-Bleau C, Beaumont M, Sanschagrin F, Voyer N, Levesque RC (2007) Parallel solid synthesis of inhibitors of the essential cell division FtsZ enzyme as a new potential class of antibacterials. Bioorg Med Chem 15:1330–1340

Romberg L, Mitchison TJ (2004) Rate-limiting guanosine 5′-triphosphate hydrolysis during nucleotide turnover by FtsZ, a prokaryotic tubulin homologue involved in bacterial cell division. Biochemistry 43:282–288

Ryckaert J-P, Ciccotti G, Berendsen HJC (1977) Numerical integration of cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 23:327–341

Schäfer H, van Gunsteren WF, Mark AE (1999) Estimating relative free energies from a single ensemble: hydration free energies. J Comput Chem 20:1604–1617

Stolarski R, Hagberg CE, Shugar D (1984) Studies on the dynamic syn-anti equilibrium in purine nucleosides and nucleotides with the aid of 1H and 13C NMR spectroscopy. Eur J Biochem 138:187–192

Tadros M, Gonzalez JM, Rivas G, Vicente M, Mingorance J (2006) Activation of the *E. coli* cell division protein FtsZ by low affinity interaction with monovalent cations. FEBS Lett 580:4941–4946

Tironi IG, Sperb R, Smith PE, van Gunsteren WF (1995) A generalized reaction field method for molecular-dynamics simulations. J Chem Phys 102:5451–5459

Trantirek L, Stefl R, Masse JE, Feigon J, Sklenar V (2002) Determination of the glycosidic torsion angles in uniformly C-13-labeled nucleic acids from vicinal coupling constants 3 J(C2/4–H1′) and 3 J(C6/8–H1′). J Biomol NMR 23:1–12

Vollmer W (2006) The prokaryotic cytoskeleton: a putative target for inhibitors and antibiotics? Appl Microbiol Biotech 73:37–47

Zwanzig RW (1954) High-temperature equation of state by a perturbation method. I. Nonpolar gases. J Chem Phys 22:1420–1426

Paper 9

Hritz, J.; Oostenbrink, C. Efficient free energy calculations for compounds with multiple stable conformations separated by high energy barriers.
J. Phys. Chem. B **2009**, 113, 12711-12720

Efficient Free Energy Calculations for Compounds with Multiple Stable Conformations Separated by High Energy Barriers

Jozef Hritz and Chris Oostenbrink*

Leiden Amsterdam Center for Drug Research, Division of Molecular Toxicology, VU University Amsterdam, The Netherlands

Received: April 1, 2009; Revised Manuscript Received: August 4, 2009

Compounds with high intramolecular energy barriers represent challenging targets for free energy calculations because of the difficulty to obtain sufficient conformational sampling. Existing approaches are therefore computationally very demanding, thus preventing practical applications for such compounds. We present an enhanced sampling-one step perturbation method (ES-OS) to tackle this problem in a highly efficient way. A single molecular dynamics simulation of a judiciously chosen reference state (using two sets of soft-core interactions) is sufficient to determine conformational distributions of chemically similar compounds and the free energy differences between them. The ES-OS method is applied to a set of five biologically relevant 8-substituted GTP analogs having high energy barriers between the anti and the syn conformations of the base with respect to the ribose part. The reliability of ES-OS is verified by comparing the results to Hamiltonian replica exchange simulations of GTP and 8-Br-GTP and the experimentally determined ^{3}J(C4,H1′) coupling constant for GMP in water. Additional simulations in vacuum and octanol allow us to calculate differences in the solvation free energies and in lipophilicities (log P). Free energy contributions from individual conformational regions are also calculated, and their relationship with the overall free energy is derived leading to a set of multiconformational free energy formulas. These relationships are of general applicability and can be used in free energy calculations for a more diverse set of compounds.

Introduction

The calculation of free-energy differences from molecular simulations using free energy perturbation (FEP) or thermodynamic integration (TI) is becoming more and more popular, not in the least thanks to increasing computer power and more accurate force fields.[1–5] In these approaches, the free energy difference is typically calculated from a set of simulations, connecting the two compounds or states between which the free energy is calculated. Convergence becomes difficult when applied to compounds that have a high energy barrier between multiple relevant conformations.[6–9] In this manuscript, a method is presented that efficiently calculates the free energy difference between multiple conformations of several molecules and simultaneously yields the free energy difference between these compounds. The problem of multiple conformations in free energy calculations was addressed earlier by the conformational space decomposition approach and the confine-and-release method.[6,10,11] In these approaches, the free energy difference between two compounds is calculated from the populations of individual conformation for both the compounds and the free energy difference between the compounds in specified conformations. In this work, we derive multiconformational free energy formulas which, in contrast to the conformational space decomposition approach, require the conformational populations of only one compound. This is highly advantageous, because calculating the conformational populations by applying an external force to overcome high energy barriers is known to be slowly converging.[12,13]

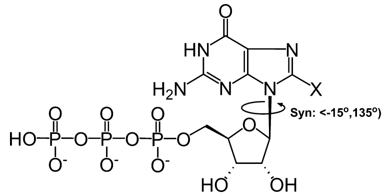


Figure 1. Structure of 8-substituted analogs of GTP in syn conformation, where X = H, F, Cl, Br, CH_3. Conformational transitions between the syn and anti conformations occur by rotation around the glycosidic bond indicated by the arrow. The glycosidic dihedral angle (χ) is defined over atoms: C4−N9−C1′−O4′.

If the occupancy of the conformational states changes during the process for which the free energy (ΔA) is calculated, sufficient conformational sampling is of utmost importance. 8-substituted GTP analogs (Figure 1) represent a group of biologically relevant compounds to which this applies. The base of GTP can adopt two stable conformations with respect to the sugar moiety by rotation around the glycosidic bond, χ. The preference for the anti ($\chi \approx -130°$) or syn ($\chi \approx 60°$) conformation is strongly dependent on the substituent at the C8 atom of the base. In a previous study, we showed that a high energy barrier between these two conformations in both GTP and 8-Br-GTP prevents sufficient conformational sampling by common computational techniques such as molecular dynamics (MD). No single anti → syn transition is seen within two independent 25 ns MD simulations of GTP or 8-Br-GTP in water at room temperature preventing the determination of their relative populations.[13] There are two possible transition pathways between the anti and the syn conformations, the one with the lowest energy barrier corresponding to the region with $\chi \approx -20°$. We determined that the height of this barrier for GTP is ~15 kJ·mol^{-1} with respect to the anti and ~7 kJ·mol^{-1} with respect to the syn conformation of GTP in water. In the

* Corresponding author. E-mail: c.oostenbrink@few.vu.nl.

10.1021/jp902968m CCC: $40.75 © 2009 American Chemical Society
Published on Web 09/01/2009

case of 8-Br-GTP, it is higher than 25 kJ·mol^{-1} with respect to both anti/syn conformations which prevents adequate sampling. The consequence of these high intramolecular barriers is the fact that during 25 ns of standard MD starting from both anti and syn conformations the only observed transition was for GTP, from syn to anti. In the same paper, we show a very poor convergence when calculating the free energy difference between the anti and the syn conformations using techniques with an external force to overcome the energy barrier. Significantly better results were obtained when, instead of forcing to overcome a high energy barrier, the barrier itself was decreased by applying soft-core interactions between the base and the sugar atoms of GTP/8-Br-GTP as part of a Hamiltonian replica exchange MD (H-REMD) scheme. This technique revealed the inverse preference for the anti and syn conformations of GTP and 8-Br-GTP, in agreement with NMR data.[14] High intramolecular energy barriers within the molecules not only hamper the determination of their conformational ensembles but also make it difficult to calculate the free energy differences between them. Bitetti-Putzer et al. proposed to perturb the Hamiltonian within generalized ensembles between two compounds with inverse conformational populations in order to enhance the conformational sampling and to calculate the free energy difference between them.[15] Unfortunately, such an approach does not work for 8-substituted GTP analogs because they all contain the energy barrier in the same conformational region and, for all linear combinations of the Hamiltonians of individual ligands, the barrier remains high. This problem can be overcome by applying “dual topology alchemical REMD”.[16] In this REMD scheme, every replica contains coordinates for both the starting- and the end-state of a perturbation, while a coupling parameter λ determines how the Hamiltionians for these states are mixed in every replica. The authors point out that a soft-core potential[17] may also be used at intermediate values of λ, potentially lowering barriers. In ref 13, we suggested an alternative approach to combine H-REMD between compounds with variations in softness. By using a single topology approach, in which only the softness is modified at intermediate replica’s to obtain sufficient conformational transitions, the free energy difference between compounds and the conformational preferences could be obtained. In order to obtain sufficient conformational transitions, it is not needed to decouple the system completely from its surroundings, but a high enough level of softness is sufficient. GTP and 8-Br-GTP could be connected by a number of replica’s in which the sugar−base interaction is made soft, such that conformational transitions become possible. This approach would suffer from large computational requirements as about 20 replicas would be needed. In addition, we note that finding the optimal settings of such a scheme is very demanding.[18] In order to calculate the free energy difference between any pair of GTP analogs by any of the mentioned H-REMD schemes, we would have to perform such an extensive H-REMD simulation for each of these pairs.

In contrast, this paper presents a very efficient enhanced sampling-one step perturbation (ES-OS) approach, combining the construction of an artificial reference Hamiltonian allowing for enhanced conformational sampling and simultaneous application of the one-step perturbation method.[19,20] In the one-step perturbation method, all interactions involving atoms which are chemically modified within a series of compounds are described using soft-core interactions. For many systems, such as the 8-substituted GTP analogs, this is not sufficient to enhance the conformational sampling. Therefore, we extend the set of interactions treated using a soft-core by the interactions between the atom pairs that are responsible for the high energy barrier (base-sugar in case of 8-substituted GTP analogs in water), thereby allowing for much faster conformational transitions.[13]

The set of 8-substituted analogs of GTP was chosen not only out of methodological interest but also because of the biological relevance of these compounds, which are considered as promising antibacterial agents.[21,22] 8-Br-guanosine (with an enforced syn conformation) was experimentally used to probe the syn conformation in a G quartet,[23] UNCG motifs,[24] and lead-dependent ribozyme.[25] Many studies involving 8-substituted analogs of cyclic adenosine and guanosine are reviewed in ref 26. Much less is known about the biological effects of 8-F-guanosines, probably because of the fact that these became synthetically feasible only very recently.[27,28]

This paper is organized as follows; first ES-OS is described and applied to a set of 8-substituted analogs of GTP in vacuum, water, and 1-octanol. From this, we obtain free energy differences between the compounds and the dihedral angle distributions around the glycosidic bond for each of them. The method is tested by comparing the distributions of the dihedral angle and energies for GTP and 8-Br-GTP as calculated by ES-OS to the values obtained from extensive H-REMD simulations using soft-core interactions.[13] From the free energy differences in different media, we subsequently calculated differences in log P and solvation free energies between the different compounds. Differences in free energies between compounds, solvation free energies, and lipophilicities were calculated over the complete conformational space as well as over ensembles for the anti and syn conformations individually. Theoretical relationships are derived showing how these individual conformational terms contribute to the overall values and how these equations can be used for free energy calculations for more diverse compounds.

Methods

Enhanced Sampling-One Step Perturbation (ES-OS). The aim of ES-OS is to determine simultaneously the free energy differences between multiple conformations separated by high energy barriers and between chemically similar compounds over the complete conformational space. We show its application for a set of five 8-substituted GTP analogs in which H8 is replaced by halogen atoms or a methyl group (Figure 1).

The ES-OS approach consists of two steps; the first is the (1) construction of an (artificial) reference compound. Typically, all interactions involving the atom or group of atoms that are different in the series of compounds are treated by soft-core interactions (see below). For practical reasons, we usually select the largest atom from the set; in our application, we used a soft Br atom at position 8 of the base. Often, such a modification of the Hamiltonian is insufficient to overcome high intramolecular energy barriers. Therefore, additional interactions are altered. Here, soft-core interactions were additionally applied between atoms of the base and the sugar in the reference molecule, as described in ref 13. (2) The second step is the projection to the Hamiltonian of the real compounds. The conformational sampling is strongly affected by the modifications to the Hamiltonian in the previous step. Following the idea of the one-step perturbation method,[20] the free energy between the reference compound (S) and the real compounds (R), can be calculated from a simulation using the reference Hamiltonian and applying Zwanzigs perturbation formula:[29]

$$\Delta G_{SR} = G_R - G_S = -k_BT \ln\langle \mathrm{e}^{-(H_R(\mathbf{q},\mathbf{p}) - H_S(\mathbf{q},\mathbf{p}))/k_BT}\rangle_S \tag{1}$$

where k_B is the Boltzmann constant, T is the absolute temperature, H_S and H_R are the Hamiltonian for the (soft) reference compound and one of the real compounds, respectively. The angular brackets indicate an ensemble average over the positions, $\mathbf{q}$ and momenta $\mathbf{p}$ obtained from a simulation of the reference state.

Any particular configuration ($\mathbf{q}_i$,$\mathbf{p}_i$) can be reweighted for the real compounds by calculating the normalized Boltzmann factor:

$$p_i^R(\mathbf{q}_i, \mathbf{p}_i) = \frac{e^{-(H_R(\mathbf{q}_i,\mathbf{p}_i)-H_S(\mathbf{q}_i,\mathbf{p}_i))/k_BT}}{\sum_i e^{-(H_R(\mathbf{q}_i,\mathbf{p}_i)-H_S(\mathbf{q}_i,\mathbf{p}_i))/k_BT}} \qquad (2)$$

The ES-OS approach has general applicability for any set of compounds. The high efficiency comes from the fact that the conformational characteristics for all compounds in the studied set as well as the free energy differences between them can be calculated from a single MD simulation of the reference compound. If a judiciously chosen reference state is used, which samples relevant conformations for the real compounds often enough, eqs 1 and 2 may be calculated with sufficient accuracy to obtain relevant results for the real compounds.[19,20,30,31]

Simulation Setup. All MD simulations and most of the analyses were conducted using the GROMOS05 simulation package running on a linux cluster.[32] All bonds were constrained, using the SHAKE algorithm[33] with a relative geometric accuracy of 10^{-4}, allowing for a time step of 2 fs used in the leapfrog integration scheme.[34] Periodic boundary conditions with a truncated octahedral box were applied. System coordinates were saved every 1 ps, and simulations of the reference compound were performed for 50 ns in water, 200 ns in vacuum, and 100 ns in octanol, depending on the convergence of the calculated free energy values and the speed of the simulations. Simulation of the reference compound in vacuum was much faster than simulation in a box containing 1926 SPC water molecules[35] or 324 flexible octanol molecules, respectively. After a steepest descent minimization to remove bad contacts between molecules, initial velocities were randomly assigned from a Maxwell−Boltzmann distribution at 298 K, according to the atomic masses. The temperature was kept constant using weak coupling to a bath of 298 K with a relaxation time of 0.1 ps.[36] The solute molecule (reference compound 8-Br-GTP using soft core interactions) and solvent were independently coupled to the heat bath. The pressure was controlled using isotropic weak coupling to atmospheric pressure with a relaxation time of 0.5 ps.[36] van der Waals and electrostatic interactions were calculated using a triple range cutoff scheme. Interactions within a short-range cutoff of 0.8 nm were calculated every time step from a pair-list that was generated every five steps. At these time points, interactions between 0.8 and 1.4 nm were also calculated and kept constant between updates. A reaction-field contribution was added to the electrostatic interactions and forces to account for a homogeneous medium outside the long-range cutoff, using the relative permittivity of SPC water (61),[37] octanol (10),[38] and vacuum (1, no reaction field). All interaction energies were calculated according to the GROMOS force field, parameter set 53A6.[39] Force field parameters for all compounds are listed in Supporting Information. Note that different partial charges of the phosphate groups in different environments result in an overall neutral reference compound in vacuum and octanol while it has a net charge of −3e in water which was electroneutralized by the addition of 3 Na$^+$ counterions. The functional form of the soft-core van der Waals and electrostatic interactions between atoms i and j was the following:[17]

$$E_{ij}^{vdw}(r_{ij}, \lambda_{vdw}) = \left(\frac{C12_{ij}}{A_{ij}(\lambda_{vdw}) + r_{ij}^6} - C6_{ij}\right)\frac{1}{A_{ij}(\lambda_{vdw}) + r_{ij}^6} \qquad (3)$$

$$E_{ij}^{el}(r_{ij}, \lambda_{el}) = \frac{q_iq_j}{4\pi\varepsilon}\frac{1}{\sqrt{B_{ij}(\lambda_{el}) + r_{ij}^2}} \qquad (4)$$

with r_{ij} the interatomic distance; $A_{ij}(\lambda_{vdw}) = \alpha_{vdw}\lambda_{vdw}^2(C12_{ij})/(C6_{ij})$, and $B_{ij}(\lambda_{el}) = \alpha_{el}\lambda_{el}^2$. $C12_{ij}$, $C6_{ij}$ are the Lennard-Jones parameters for atom pair i and j; q_i and q_j are the partial charges of particles i and j; α_{vdw} and α_{el} are the softness constants. In the current study, we used in all simulations $\alpha_{vdw} = \alpha_{el} = 1$, and the softness of the interactions was controlled through the parameters λ_{vdw}, λ_{el}. It can be seen that at longer distances ($r_{ij} \gg A(\lambda_{vdw})$ and $r_{ij} \gg B(\lambda_{el})$) the soft-core interaction approximates the interaction for normal atoms and that they differ mostly at short distances between the atoms ($r_{ij} \leq A(\lambda_{vdw})$ or $r_{ij} \leq B(\lambda_{el})$). Potential energy barriers are mostly the result of a short-ranged repulsion between atoms, which can strongly be reduced by increasing the levels of softness, thereby allowing us to make very specific, local modifications to the Hamiltonian.

The procedure for finding the optimal softness of selected interactions, such that enhanced conformational sampling is observed, is described in our previous paper.[18] In water, it led to softness parameters $\lambda_{\mathrm{vdw}} = 0.7$ and $\lambda_{\mathrm{el}} = 0.7$ for all nonbonded interactions between the base and the sugar atoms of 8-Br-GTP. In the ES-OS application for 8-substituted analogs of GTP, the set of soft-core interactions was further extended to all nonbonded interactions between bromine and the rest of the system (including water) using the same softness parameters. In vacuum, the procedure led to the same interactions with the same softness parameters ($\lambda_{\mathrm{vdw}} = 0.7$; $\lambda_{\mathrm{el}} = 0.7$) extended by soft-core interactions between the base and the phosphates with $\lambda_{\mathrm{vdw}} = 0$ and $\lambda_{\mathrm{el}} = 0.22$. The optimal softness parameters obtained for GTP in octanol were refined as $\lambda_{\mathrm{vdw}} = 0.7$ and $\lambda_{\mathrm{el}} = 0.6$ for interactions between base and sugar atoms as well as between bromine and the rest of system (including octanol). For interactions between the base and the phosphates, softness parameters $\lambda_{\mathrm{vdw}} = 0$ and $\lambda_{\mathrm{el}} = 0.27$ were used. We stress that, in order to calculate the free energy of transfer between two media, the reference molecules do not need to be identical. As long as the end-states of the perturbation (the real molecules) are identical in the different media, they can theoretically be connected through any arbitrary reference molecule in order to calculate the free energy difference between the real molecules. The most efficient reference molecule can be different in different media.

Results

During 50 ns of MD simulation of the reference state in water ∼150 anti ↔ syn transitions occurred ensuring sufficient convergence of the χ dihedral angle distribution (orange curve in Figure 2) and a precise determination of the conformational populations, $[\mathrm{anti}]_{(aq)}^S = 41.5\% \pm 3.0\%$, $[\mathrm{syn}]_{(aq)}^S = 58.5\% \pm 3.0\%$, and the free energy between them $\Delta G_S^{anti,syn}(\mathrm{aq}) = -k_BT \ln([\mathrm{syn}]_{(aq)}^S/[\mathrm{anti}]_{(aq)}^S) = -0.85 \pm 0.31$ kJ·mol^{-1}. Similar values were obtained in vacuum ($[\mathrm{anti}]_{(v)}^S = 40.8\% \pm 1.1\%$, $[\mathrm{syn}]_{(v)}^S = 59.2\% \pm 1.1\%$, $\Delta G_S^{anti,syn}(v) = -0.92 \pm 0.11$ kJ·mol^{-1}) and in octanol ($[\mathrm{anti}]_{(o)}^S = 49.3 \pm 4.5\%$, $[\mathrm{syn}]_{(o)}^S = 50.7\% \pm 4.5\%$, $\Delta G_S^{anti,syn}(o) = -0.069 \pm 0.45$ kJ·mol^{-1}). Error estimates on ensemble averages are calculated from block averages followed by an extrapolation to infinite block length.[40] Populations

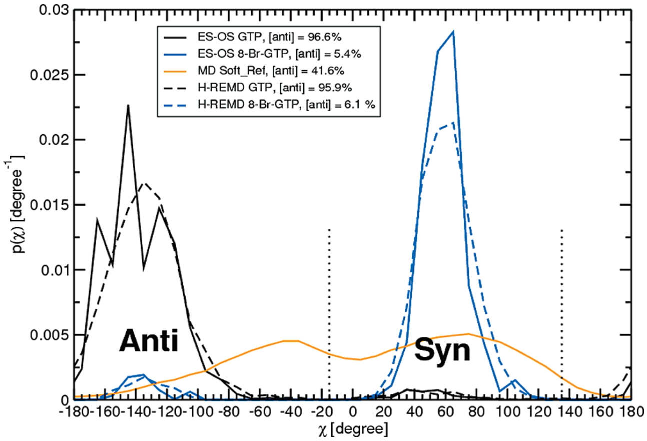


Figure 2. Normalized dihedral angle (χ) distributions for GTP and 8-Br-GTP calculated by ES-OS and H-REMD and for the reference state (8-Br-GTP using soft-core interactions) as obtained from MD. Anti and syn conformational regions are indicated by dotted lines.

(including their error estimates) are calculated from the ensemble average of occurrences of the conformational states. From the simulations in the reference state, the free energy differences to the real compounds, ΔG_{SR} were calculated, see Table 1. By selecting from the ensemble only structures that are in anti and syn conformation, free energy differences can be obtained for the molecules in these conformations ΔG_{SR}^{anti} and ΔG_{SR}^{syn}. From these values and the free energy difference $\Delta G_{S}^{anti,syn}$ for the reference state, we can calculate the free energy difference between the anti and the syn conformations for the real compounds, $\Delta G_{R}^{anti,syn}$ (see below) and subsequently the relative populations $[\text{anti}]^R$ and $[\text{syn}]^R$. All of these values are summarized in Table 1. From H-REMD calculations, we have earlier calculated $\Delta G_{GTP}^{anti,syn} = 7.6 \pm 0.3$ kJ·mol^{-1} and $\Delta G_{8-Br-GTP}^{anti,syn} = -6.8 \pm 0.9$ kJ·mol^{-1},[13] which fit remarkably well with the data in Table 1.

The reliability of the free energy differences between the compounds will be mostly dependent on the overlap in phase space sampled by the reference and real compounds (Figure 2). The sampling efficiency can be calculated as $S_{eff} = 2N_{\Delta H \le \Delta G}/N$, being twice the number of conformations in which the change in Hamiltonian in eq 1 is less than the overall change in free energy, divided by the total number of conformations sampled for the reference state.[41] The most efficient sampling will occur with a symmetric distribution of energies around $\Delta H = \Delta G$, leading to $S_{eff} = 1$. Sampling efficiencies were calculated to be between 3.4% (for 8-Br-GTP in water) and 6.1% (for 8-F-GTP in water) indicating that about 2000 relevant conformations are sampled for each of the real compounds.

In Figure 3, we compare distributions of nonbonded energies between (A) the base and water and (B) the base and the sugar of GTP and 8-Br-GTP as obtained from ES-OS and from previous extensive H-REMD simulations.[13] The distributions for the ES-OS approach were obtained by reweighting the probabilities of individual MD structures using eq 2. A good agreement can be observed. Figure 2 presents a similar comparison in terms of the χ dihedral angle distribution for GTP and 8-Br-GTP. The distributions in water obtained from ES-OS (solid curves) show good agreement with the same distributions as obtained from H-REMD simulations (dashed curves). The solid curves in Figure 2 and Figure 3 are less smooth than the dashed ones, which is due to the fact that only few (~50) MD structures have a significant (>0.1) probability for the real compounds. We note that the $[\text{anti}]_{(aq)}^R$ and $[\text{syn}]_{(aq)}^R$ populations show faster convergence. A reasonably accurate number can already be obtained from 10 ns of ES-OS simulation.

^3J(C4,H1′) values for individual frames of the reference compounds were calculated using the Karplus equation:

$$^3J(\text{C4,H1}') = A\cos^2(\chi + 120°) + B\cos(\chi + 120°) + C \tag{5}$$

where the most recent Karplus coefficients parametrized for guanosine containing systems were applied: $A = 4.4$ Hz, $B = -1.4$ Hz, $C = 0.1$ Hz.[42]

Distributions of ^3J(C4,H1′) were generated in the same way as for the glycosidic dihedral angle and subsequently used for the calculation of the average values $\langle ^3J(\text{C4,H1}')\rangle$, listed in the last column of Table 1. These values are directly comparable with experimental NMR data. Schwalbe et al. reported for different magnetization transfer delays, four values of ^3J(C4,H1′) for 5′-GMP in D_2O: 2.5 Hz, 3.5 Hz, 2.6 Hz, and 2.6 Hz.[43] Our value for GTP in H_2O as obtained from ES-OS, $\langle ^3J(\text{C4,H1}')\rangle = 2.42$ Hz is in very good agreement with these data. Unfortunately, we did not find experimental NMR ^3J(C4,H1′) data for other molecules in our set of 8-substituted GTP analogs.

Figure 4 shows the χ dihedral angle distributions for all real compounds in Table 1. The values for $[\text{anti}]^R$ and $[\text{syn}]^R$ that can be obtained by integrating these distributions correspond within numerical precision to the values obtained from the free energy differences in Table 1. In order to analyze the free energy differences under the restriction of a given conformation, it is useful to use the same definition of the anti and syn regions (anti: $\chi \in <-180°,-15°) \cup <135°,180°>$; syn: $<\chi \in <-15°,135°)$). In reality, the distributions are shifted within these regions for different compounds in different environments (Figure 4A (in water), Figure 4B (in vacuum), Figure 4C (in octanol)). The population of the anti conformation is lower in vacuum as compared with the water environment for all studied compounds. This is most probably due to hydrogen bonding and electrostatic interactions between the phosphate and the NH_2 group of the base which are more pronounced in vacuum than in water. A more detailed comparison of these effects requires additional simulations of guanosine in water and vacuum which is beyond the scope of this work. By using the thermodynamic cycles in Figure 5, the free energy differences between the compounds (Figure 5A) or between syn and anti states of the real compounds can be calculated (Figure 5B).

From the free energy differences between individual compounds and the soft reference state in different environments, we calculate the differences in solvation free energies and lipophilicities by applying the following equations:

$$\Delta\Delta G_{SR}(solv) = \Delta G_{SR}(aq) - \Delta G_{SR}(v) \tag{6}$$

$$\Delta\Delta G_{SR}(o/w) = \Delta G_{SR}(aq) - \Delta G_{SR}(o) \tag{7}$$

$$\Delta\log P^{o/w} = \frac{\Delta\Delta G(o/w)}{2.3k_BT} \tag{8}$$

The calculations were performed for the complete conformational space (Table 2A) and for the anti (Table 2B) or syn (Table 2C) conformations. We note that, since we have simulated the reference molecules in different phosphate protonation states in the different media, a contribution for the deprotonation of the phosphate tail should be included in the

TABLE 1: Free Energy Differences for 8-Substituted GTP Analogs with Respect to the Reference State (8-Br-GTP Using Soft-Core Interactions) As Calculated by ES-OS and Free Energy Differences between Their Anti and Syn Conformations in (A) Water, (B) Vacuum, and (C) Octanol[a]

(A)

In water	ΔG_{SR}(aq)[kJ·mol^{-1}]	ΔG_{SR}^{anti}(aq)[kJ·mol^{-1}]	ΔG_{SR}^{syn}(aq)[kJ·mol^{-1}]	$\Delta G_{R}^{anti,syn}$(aq)[kJ·mol^{-1}]	[anti]$^{R}_{(aq)}$	[syn]$^{R}_{(aq)}$	$\langle ^3J(C4,H1')\rangle$[Hz]
GTP	6.82 ± 0.87	4.72 ± 0.85	13.97 ± 0.62	8.40 ± 1.78	96.7%	3.3%	2.42 ± 0.60
8-F-GTP	21.85 ± 0.51	20.35 ± 0.62	24.04 ± 1.00	2.84 ± 1.93	75.9%	24.1%	3.33 ± 0.61
8-Cl-GTP	19.57 ± 0.38	19.93 ± 0.71	19.33 ± 0.53	−1.45 ± 1.55	35.8%	64.2%	4.49 ± 0.69
8-Br-GTP	12.67 ± 0.47	17.71 ± 0.76	11.48 ± 0.53	−7.08 ± 1.60	5.4%	94.6%	5.46 ± 1.00
8-CH_3-GTP	19.37 ± 0.54	23.15 ± 0.85	18.27 ± 0.61	−5.73 ± 1.77	9.0%	91.0%	5.33 ± 1.12

(B)

In vacuum	ΔG_{SR} (v)[kJ·mol^{-1}]	$\Delta G_{SR}^{anti}(v)$[kJ·mol^{-1}]	$\Delta G_{SR}^{syn}(v)$[kJ·mol^{-1}]	$\Delta GR^{anti,syn}(v)$[kJ·mol^{-1}]	$[anti]^{R}_{(v)}$	$[syn]^{R}_{(v)}$	$\langle ^3J(C4,H1')\rangle$[Hz]
GTP	−22.47 ± 0.70	−24.28 ± 0.79	−19.09 ± 0.67	4.27 ± 1.57	84.9%	15.1%	1.86 ± 0.47
8-F-GTP	−13.19 ± 0.33	−12.62 ± 0.52	−13.51 ± 0.48	−1.81 ± 1.11	32.5%	67.5%	4.23 ± 0.71
8-Cl-GTP	−17.03 ± 0.44	−13.40 ± 0.81	−18.08 ± 0.49	−5.60 ± 1.41	9.4%	90.6%	4.69 ± 0.57
8-Br-GTP	−20.69 ± 0.45	−12.71 ± 0.75	−21.95 ± 0.46	−10.16 ± 1.32	1.6%	98.4%	4.96 ± 0.74
8-CH_3-GTP	−18.14 ± 0.50	−9.88 ± 0.73	−19.41 ± 0.51	−10.45 ± 1.35	1.4%	98.6%	5.10 ± 0.94

(C)

In octanol	ΔG_{SR} (o)[kJ·mol^{-1}]	$\Delta G_{SR}^{anti}(o)$[kJ·mol^{-1}]	$\Delta G_{SR}^{syn}(o)$[kJ·mol^{-1}]	$\Delta GR^{anti,syn}(o)$[kJ·mol^{-1}]	$[anti]^{R}_{(o)}$	$[syn]^{R}_{(o)}$	$\langle ^3J(C4,H1')\rangle$[Hz]
GTP	2.12 ± 0.86	0.42 ± 0.95	10.26 ± 0.54	9.77 ± 1.94	98.1%	1.9%	1.91 ± 0.42
8-F-GTP	7.42 ± 0.71	6.26 ± 0.90	9.55 ± 0.70	3.22 ± 2.05	78.6%	21.4%	3.24 ± 0.63
8-Cl-GTP	1.21 ± 0.43	1.04 ± 0.77	1.38 ± 0.40	0.27 ± 1.62	52.7%	47.3%	3.96 ± 0.77
8-Br-GTP	−7.06 ± 0.37	−4.10 ± 0.79	−8.35 ± 0.35	−4.32 ± 1.59	14.9%	85.1%	4.99 ± 0.83
8-CH_3-GTP	3.83 ± 0.47	6.59 ± 0.75	2.59 ± 0.50	−4.07 ± 1.7	16.2%	83.8%	5.01 ± 0.98

[a] Values for $\Delta GR^{anti,syn}$ (5th column) were derived from the thermodynamic cycles shown in Figure 5B.

calculations. However, we assume this contribution to be constant for the various compounds and conformations, such that it cancels in the relative free energies.

For comparison, Δlog $P^{o/w}$ was also calculated using a widely used atomic contribution model[44] based on the two-dimensional (2D) structure of molecules as implemented in the program MOE.[45] These values are listed in the last column of Table 2.

Discussion

We want to emphasize that all data for the set of five 8-substituted GTP analogs as summarized in Tables 1 and 2 and Figures 4 and 5 were refined from only three MD simulations of the reference compound in different media. This demonstrates the big potential of the ES-OS approach to efficiently characterize chemically similar compounds with high intramolecular energy barriers. An obvious application is a lead optimization process during which one searches for the chemical modification which has certain desired physicochemical properties. We observe relatively rough dihedral angle distributions mostly in regions of conformational space that are rarely sampled by the reference compound and for compounds showing a large free energy difference with respect to the reference compound. However, even rough distributions provide very good overview for all compounds and therefore allow for an efficient identification of the ones having the desired properties. For the compounds of interest, additional MD simulations can be performed within the local minima and smooth dihedral angle distributions can be obtained by reweighting the distributions for each of these simulations with the population obtained with ES-OS. We recall that the relative populations converge much faster in ES-OS than the distributions. In the present application, this can be applied to ^3J value distributions and the resulting average value. The drawback obviously is the need for additional MD simulations.

The unphysical reference compound described here allows for enhanced conformational sampling either using ES-OS or

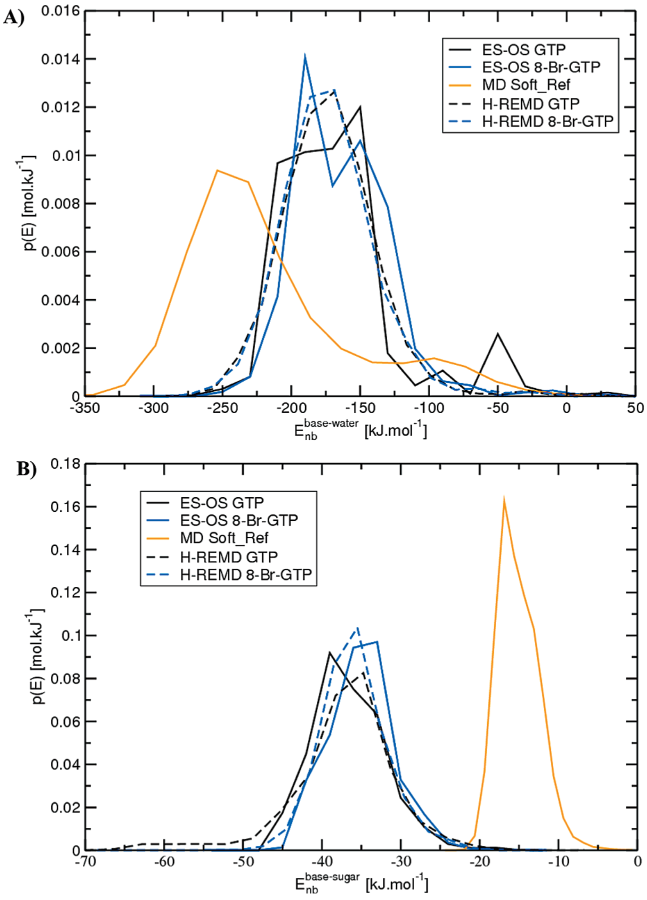


Figure 3. Normalized energy distributions for GTP and 8-Br-GTP calculated by ES-OS and H-REMD and for the reference state (8-Br-GTP using soft-core interactions) as obtained from MD. Nonbonded energies between (A) base and water and (B) base and sugar, which are modified in the reference state.

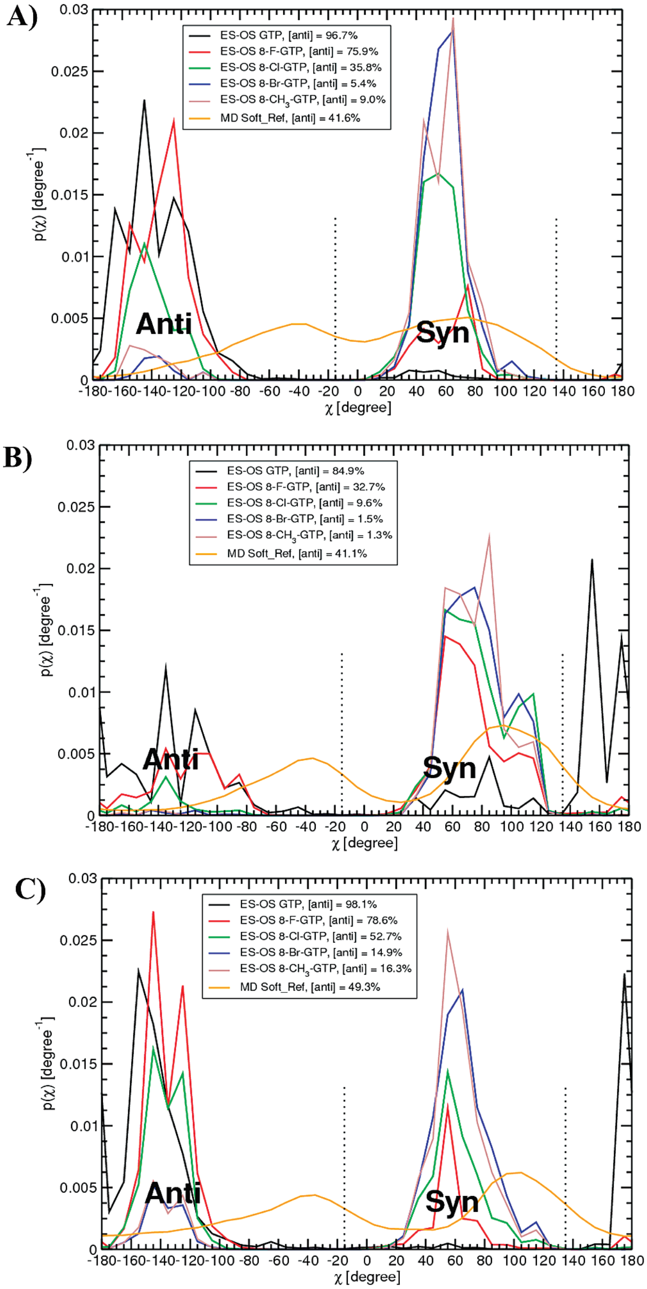


Figure 4. Normalized dihedral angle (χ) distributions for 8-substituted analogs of GTP as calculated by ES-OS and for the reference state (8-Br-GTP using soft-core interactions) as obtained from MD in (A) water, (B) vacuum, (C) octanol. Anti and syn conformational regions are indicated by dotted lines.

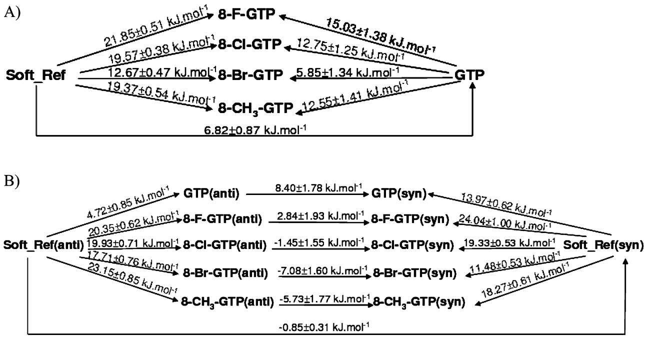


Figure 5. Thermodynamic cycles used to calculate the free energy difference between compounds (A) and between different conformations (B) in water. The free energy differences between the reference compound (Soft_Ref) and the real compounds where calculated by ES-OS. All other values were obtained by cycle closure.

H-REMD. In this application, the interactions that were softened (between sugar and base) were determined mostly by chemical intuition. Determining the optimal reference molecule may be the most crucial and least straightforward factor of this method, which will be the topic of continuing research. Now that a reference molecule for GTP is determined, it can be applied further in any guanosine containing systems, such as RNA loops. This would be useful not only for a structural refinement including the populations of conformations with guanosine in the anti or syn conformation but also for a consistent interpretation of NMR relaxation rates depending on the chemical shielding anisotropy (CSA). Recently, Ferner et al. presented a significant dependence of ^{13}C CSA on the dihedral angle around the glycosidic bond (−133 ppm for guanine in anti and −122 in syn). Only by taking the different conformations into account could they obtain fully consistent data for the guanine residue.[46]

Two applications that require the accurate calculation of free energy differences between 8-substituted GTP analogs in different media are the solvation free energy and lipophilicity differences, which are summarized in Table 2. The free energy differences in water that are summarized in Figure 5A and in the second column of Table 1A will prove to be useful in other applications as well, such as the calculation of differences in binding affinities of 8-substituted GTPs toward biomacromolecules. Despite the fact that many proteins will bind these compounds only in one conformation, it is important to note that for the unbound compounds both conformations (anti/syn) are contributing to the final affinity.

We emphasize that values calculated for complete conformational space (Table 2A) differ significantly from the same values calculated only considering the anti (Table 2B) or syn (Table 2C) conformation for several compounds. The reason is that the free energy differences for individual conformations are different and their contributions to the overall values depend on the conformational populations. Both of these effects are completely ignored by any 2D prediction method of solvation and lipophilicity[44] (often used as descriptor in e.g. medicinal chemistry). Therefore, 2D predictions in principle can not be expected to lead to accurate predictions for compounds with different conformational states. Please note in Table 1 that the value of ΔG_{SR} for the complete conformational space is always in between the values for the individual conformations ΔG_{SR}^{anti} and ΔG_{SR}^{syn}. This is not necessarily true for the relative free energy differences between different environments, $\Delta\Delta G_{SR}$. For example, differences in the solvation free energy of 8-CH$_3$-GTP and GTP for individual conformations are $\Delta\Delta G_{GTP,8-CH_3-GTP}^{anti}(solv) = 4.03 \pm 3.05$ kJ·mol^{-1} and $\Delta\Delta G_{GTP,8-CH_3-GTP}^{syn}(solv) = 4.62 \pm 2.41$ kJ·mol^{-1} while the value for the complete ensemble is $\Delta\Delta G_{GTP,8-CH_3-GTP}\ (solv) = 8.22 \pm 2.61$ kJ·mol^{-1} (Table 2).

In order to understand how ΔG and $\Delta\Delta G$ values for individual conformations contribute to values for the complete conformational space, we derive explicit relationships between them. The free energy for compound A is defined as

$$G_A = -k_B T \ln Z_A \tag{9}$$

where Z_A is the isothermal−isobaric partition function of compound A:

$$Z_A = \frac{1}{h^{3N}N!}\iiint e^{-(H_A+pV)/k_BT}(\mathrm{d}V\,\mathrm{d}\mathbf{p}^N\,\mathrm{d}\mathbf{q}^N) = e^{-G_A/k_BT} \tag{10}$$

Considering that the complete conformational space of compound A is divided into n^A distinct conformations (i^A), the partition function can be written as sum of partition functions of individual subspaces, or conformations:

$$Z_A = \sum_{i^A}^{n^A} Z_A^{i^A} = \sum_{i^A}^{n^A} e^{-G_A^{i^A}/k_BT} \tag{11}$$

which defines $G_A^{i^A}$ as the free energy of compound A, confined to conformation i^A.[6] The relative population of this conformation is defined as

$$[i^A]^A = \frac{Z_A^{i^A}}{Z_A} \tag{12}$$

Within this framework, Straatsma and McCammon address the free energy difference between compound A and compound B in the presence of multiple conformations by writing[6]

$$\Delta G_{AB} = G_B - G_A = (G_B - G_B^{i^B}) - (G_A - G_A^{i^A}) + (G_B^{i^B} - G_A^{i^A}) \tag{13}$$

The last term in this equation, the free energy difference between compound B in conformation i_B and compound A in conformation i_A, is typically obtained by standard TI or FEP calculations, during which the compounds occupy only one conformational state. The first two terms represent the free energy difference between sampling the complete conformational space and sampling only region i_A. These terms are directly linked to the population of the conformation, $G_A - G_A^{i^A} = k_BT \ln [i^A]^A$. Straatsma and McCammon point out that a free energy calculation in which the compounds occupy only one conformation should be corrected by these terms, both for compound A and for compound B.[6] However, in many cases obtaining an accurate estimate of the relative populations is much more difficult than obtaining a free energy difference between compounds confined to well-defined local minima. Also, in the current application, we have precise population values only for the reference compound. We therefore derive formulas that only depend on populations of one compound in order to evaluate ΔG_{AB}.

Using eqs 9−11, we can write

$$\Delta G_{AB} = G_B - G_A = -k_BT\ln\frac{Z_B}{Z_A} = -k_BT\ln\left(\sum_{i^B}^{n^B}\frac{Z_B^{i^B}}{Z_A}\right) \tag{14}$$

in which we multiply the individual terms by unity in the form $(1)/(n^A)\sum_{i^A}^{n^A}(Z_A^{i^A})/(Z_A^{i^A})$ to obtain

$$\begin{aligned}\Delta G_{AB} &= -k_BT\ln\left(\sum_{i^B}^{n^B}\frac{Z_B^{i^B}}{Z_A}\frac{1}{n^A}\sum_{i^A}^{n^A}\frac{Z_A^{i^A}}{Z_A^{i^A}}\right)\\ &= -k_BT\ln\left(\frac{1}{n^A}\sum_{i^B}^{n^B}\sum_{i^A}^{n^A}\frac{Z_B^{i^B}}{Z_A}\frac{Z_A^{i^A}}{Z_A^{i^A}}\right)\end{aligned} \tag{15}$$

and using eq 12 then yields

TABLE 2: Differences in Solvation Free Energies and Lipophilicities for 8-Substited GTP Analogs with Respect to the Reference State (8-Br-GTP Using Soft-Core Interactions; 2nd, 4th Column) and with Respect to GTP (3rd, 5th Column) As Calculated by ES-OS over (A) the Complete Conformational Space, (B) the Anti Conformation, and (C) the Syn Conformation[a]

(A)

	$\Delta\Delta G_{\mathrm{GTP,R}}$(solv)[kJ·mol^{-1}]	$\Delta\Delta G_{\mathrm{GTP,R}}$(solv)[kJ·mol^{-1}]	$\Delta\Delta G_{\mathrm{SR}}$(o/w)[kJ·mol^{-1}]	$\Delta\Delta G_{\mathrm{GTP,R}}$(o/w)[kJ·mol^{-1}]	Δlog $\mathrm{P}_{\mathrm{ES-OS}}^{\mathrm{o/w}}$	Δlog $\mathrm{P}_{\mathrm{2D}}^{\mathrm{o/w}}$
GTP	29.29 ± 1.57	0	4.70 ± 1.73	0	0	0
8-F-GTP	35.04 ± 0.84	5.75 ± 2.41	14.43 ± 1.22	9.73 ± 2.95	1.71 ± 0.52	0.58
8-CL-GTP	36.6 ± 0.82	7.31 ± 2.39	18.36 ± 0.81	13.66 ± 2.54	2.40 ± 0.45	2.73
8-BR-GTP	33.36 ± 0.92	4.07 ± 2.49	19.73 ± 0.84	15.03 ± 2.57	2.64 ± 0.45	3.19
8-CH$_3$-GTP	37.51 ± 1.04	8.22 ± 2.61	15.44 ± 1.01	10.74 ± 2.74	1.89 ± 0.48	1.29

(B)

anti	$\Delta\Delta G_{SR}^{anti}(solv)$[kJ·mol^{-1}]	$\Delta\Delta G_{GTP,R}^{anti}(solv)$[kJ·mol^{-1}]	$\Delta\Delta G_{SR}^{anti}(o/w)$[kJ·mol^{-1}]	$\Delta\Delta G_{GTP.R}^{anti}(o/w)$[kJ·mol^{-1}]	Δlog $P_{ES-OS}^{o/w,anti}$	Δlog $P_{2D}^{o/w,anti}$
GTP	29.00 ± 1.64	0	4.30 ± 1.80	0	0	0
8-F-GTP	32.97 ± 1.14	3.97 ± 2.61	14.09 ± 1.52	9.79 ± 3.32	1.72 ± 0.58	0.58
8-CL-GTP	33.33 ± 1.52	4.33 ± 2.99	18.89 ± 1.48	14.59 ± 3.28	2.56 ± 0.58	2.73
8-BR-GTP	30.42 ± 1.51	1.42 ± 2.98	21.81 ± 1.55	17.51 ± 3.35	3.07 ± 0.59	3.19
8-CH$_3$-GTP	33.03 ± 1.58	4.03 ± 3.05	16.56 ± 1.60	12.26 ± 3.40	2.15 ± 0.60	1.29

(C)

syn	$\Delta\Delta G_{SR}^{syn}(solv)$ [kJ·mol^{-1}]	$\Delta\Delta G_{GTP,R}^{syn}(solv)$ [kJ·mol^{-1}]	$\Delta\Delta G_{SR}^{syn}(o/w)$ [kJ·mol^{-1}]	$\Delta\Delta G_{GTP,R}^{syn}(o/w)$ [kJ·mol^{-1}]	Δlog $P_{ES-OS}^{o/w,syn}$	Δlog $P_{2D}^{o/w,syn}$
GTP	33.06 ± 1.29	0	3.71 ± 1.16	0	0	0
8-F-GTP	37.55 ± 1.48	4.49 ± 2.77	14.49 ± 1.7	10.78 ± 2.86	1.89 ± 0.50	0.58
8-CL-GTP	37.41 ± 1.02	4.35 ± 2.31	17.95 ± 0.93	14.24 ± 2.09	2.50 ± 0.37	2.73
8-BR-GTP	33.43 ± 0.99	0.37 ± 2.28	19.83 ± 0.88	16.12 ± 2.04	2.83 ± 0.36	3.19
8-CH$_3$-GTP	37.68 ± 1.12	4.62 ± 2.41	15.68 ± 1.11	11.97 ± 2.27	2.10 ± 0.40	1.29

[a] Values of Δlog $P_{ES-OS}^{o/w}$ with respect to GTP calculated by ES-OS (6th column) are compared with values calculated in MOE using an atomic contribution model (7th column).[44]

$$\Delta G_{AB} = -k_BT\ln\left(\frac{1}{n^A}\sum_{i^B}^{n^B}\sum_{i^A}^{n^A}[i^A]^A\ \mathrm{e}^{-(G_B^{i^B}-G_A^{i^A})/k_BT}\right) \quad (16)$$

In many practical applications, the different compounds occupy the same conformations. Here, we consider the same definition of the anti and syn conformations for all compounds. By imposing $i^A = i^B = i$ and multiplying the individual terms in eq 14 by unity in the form $(Z_A^i)/(Z_A^i)$, eq 16 reduces to

$$\begin{aligned}\Delta G_{AB} &= -k_BT\ln(\sum_i^n [i]^A\ \mathrm{e}^{-(G_B^i-G_A^i)/k_BT}) \\ &= -k_BT\ln(\sum_i^n [i]^A\ \mathrm{e}^{-\Delta G_{AB}^i/k_BT})\end{aligned} \quad (17)$$

It can be seen from eq 17 that the value of ΔG_{AB} is always in between the maximum and minimum of the individual ΔG_{AB}^i values. The requirement that $\sum_i^n [i]^A = 1$ ensures that $\Delta G_{AB} = \Delta G_{AB}^i$ if all individual ΔG_{AB}^i values are equal. Application of eq 17 to the values for ΔG_{SR}^{anti} and ΔG_{SR}^{syn} yields estimates that correspond up to the second decimal to the values of ΔG_{SR} in Table 1.

We also derive similar relationships for relative free energy differences, such as the ones reported for differences in solvation or lipophilicity. The relative solvation free energy $\Delta\Delta G_{AB}(solv)$ can be obtained using eq 6:

$$\begin{aligned}\Delta\Delta G_{AB}(solv) &= \Delta G_{AB}(aq) - \Delta G_{AB}(v) \\ &= -k_BT\ln\frac{Z_B(aq)Z_A(v)}{Z_A(aq)Z_B(v)}\end{aligned} \quad (18)$$

In the same notation as above, we can express $\Delta\Delta G_{AB}(solv)$ in terms of the differences in free energies over the individual conformations in water ($\Delta G_{AB}^{i^A}(aq)$) and in vacuum ($\Delta G_{AB}^{i^A}(v)$) using eq 16:

$$\Delta\Delta G_{AB}(solv) = \\ -k_BT\ln\left(\frac{\dfrac{1}{n^A(v)}\displaystyle\sum_{i^B(v)}^{n^B(v)}\sum_{i^A(v)}^{n^A(v)}[i^A]_{(v)}^A e^{-(G_B^{i^B}(v)-G_A^{i^A}(v))/k_BT}}{\dfrac{1}{n^A(aq)}\displaystyle\sum_{i^B(aq)}^{n^B(aq)}\sum_{i^A(aq)}^{n^A(aq)}[i^A]_{(aq)}^A e^{-(G_B^{i^B}(aq)-G_A^{i^A}(aq))/k_BT}}\right) \quad (19)$$

Note that conformation i^A may be defined differently in different media, but it only makes sense to consider it for compounds in the same medium. Therefore, we will indicate the solvation state by using (aq) or (v) only once for every symbol. $G_B^{i^B}(v)$ then represents the free energy of compound B in vacuum, confined to conformation i^B as it is defined in vacuum. Again, this equation becomes simpler if the definition of the conformational states is the same for both compounds (still allowing for different conformations in different media). That is, if $i^A(aq) = i^B(aq) = i(aq)$ and $i^A(v) = i^B(v) = i(v)$, then we obtain the following from eq 17:

$$\Delta\Delta G_{AB}(solv) = -k_BT\ln\left(\frac{\displaystyle\sum_{i(v)}^{n(v)}[i]_{(v)}^A e^{-\Delta G_{AB}^i(v)/k_BT}}{\displaystyle\sum_{i(aq)}^{n(aq)}[i]_{(aq)}^A e^{-\Delta G_{AB}^i(aq)/k_BT}}\right) \quad (20)$$

An alternative relationship between the absolute and the relative free energy differences can be obtained from eqs 11 and 18 and multiplying the individual terms by unity in the form $(Z_A^i(aq)Z_B^j(v))/(Z_A^i(aq)Z_B^j(v))$:

$$\begin{aligned}\Delta\Delta G_{AB}(solv) &= -k_BT\ln\frac{\displaystyle\sum_{i(aq)}^{n(aq)}Z_B^i(aq)\sum_{j(v)}^{n(v)}Z_A^j(v)}{Z_A(aq)Z_B(v)} \\ &= -k_BT\ln\left(\sum_{i(aq)}^{n(aq)}\sum_{j(v)}^{n(v)}\frac{Z_B^i(aq)Z_A^j(v)}{Z_A(aq)Z_B(v)}\frac{Z_A^i(aq)Z_B^j(v)}{Z_A^i(aq)Z_B^j(v)}\right)\end{aligned} \quad (21)$$

Using eq 9 and 12, we can write this as

$$\Delta\Delta G_{AB}(solv) = \\ -k_BT\ln(\sum_{i(aq)}^{n(aq)}\sum_{j(v)}^{n(v)}[i]_{(aq)}^A[j]_{(v)}^B e^{-(\Delta G_{AB}^i(aq)-\Delta G_{AB}^j(v))/k_BT}) \quad (22)$$

Equation 22 is similar to eq 17 in the sense that the value of $\Delta\Delta G_{AB}(solv)$ will be in between the maximum and minimum of the individual $(\Delta G_{AB}^i(aq) - \Delta G_{AB}^j(v))$ values. Again, the normalization condition $\sum_{i(aq)}^{n(aq)}\sum_{j(v)}^{n(v)}[i]_{(aq)}^A[j]_{(v)}^B = \sum_{i(aq)}^{n(aq)}[i]_{(aq)}^A\sum_{j(v)}^{n(v)}[j]_{(v)}^B = 1$ ensures that if all differences $(\Delta G_{AB}^i(aq) - \Delta G_{AB}^j(v))$ are equal, they will also be equal to the relative solvation free energy difference over the complete conformational space ($\Delta\Delta G_{AB}(solv)$).

In case the individual conformations are the same for compounds A and B in both media $i^A(aq) = i^B(aq) = i^A(v) = i^B(v) = i$, we can write eq 18 as

$$\begin{aligned}\Delta\Delta G_{AB}(solv) &= -k_BT\ln\frac{Z_A(v)\displaystyle\sum_i^n Z_B^i(aq)}{Z_A(aq)Z_B(v)} \\ &= -k_BT\ln\left(\sum_i^n\frac{Z_A(v)Z_B^i(aq)}{Z_A(aq)Z_B(v)}\right)\end{aligned} \quad (23)$$

Multiplication of the individual terms by unity in the form $(Z_A^i(v)Z_A^i(aq)Z_B^i(v))/(Z_A^i(v)Z_A^i(aq)Z_B^i(v))$ and applying eqs 6, 9, and 12 then yields

$$\Delta\Delta G_{AB}(solv) = -k_BT\ln\left(\sum_i^n\frac{[i]_{(aq)}^A[i]_{(v)}^B}{[i]_{(v)}^A}e^{-\Delta\Delta G_{AB}^i(solv)/k_BT}\right) \quad (24)$$

Note that in most cases $\sum_i^n([i]_{(aq)}^A[i]_{(v)}^B)/([i]_{(v)}^A)$ will not add up to 1. Therefore, even if all individual $\Delta\Delta G_{AB}^i(solv)$ are equal, eq 24 does not necessarily lead to the same value for the overall

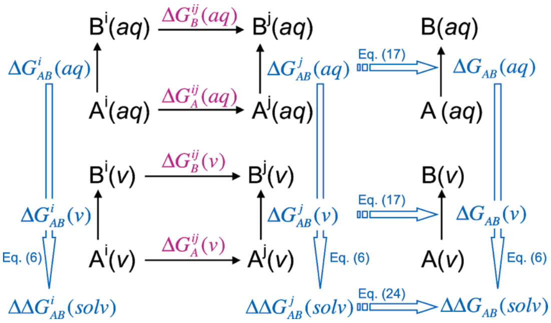


Figure 6. Graphical representation of multiconformational free energy formulas (eqs 17 and 24) relating absolute and relative free energy differences for compounds A and B in conformations *i* and *j* in aqueous environment and vacuum.

$\Delta\Delta G_{AB}(solv)$. This also explains why $\Delta\Delta G_{AB}$ $(solv)$ can be larger than both $\Delta\Delta G_{AB}^{anti}(solv)$ and $\Delta\Delta G_{AB}^{syn}(solv)$ as was observed for compound 8-CH$_3$-GTP relative to GTP (Table 2).

Excellent agreement is observed between eqs 20, 22, and 24 and the differences in solvation free energies and lipophilicities listed in Table 2. The set of eqs 16, 17, 19, 20, 22, and 24 relates the absolute and relative free energy differences for individual conformational states to values for the complete ensemble. We will refer to them as the multiconformational free energy formulas. Equations 17 and 24 are graphically represented in Figure 6 for the free energy differences in water and vacuum resulting in solvation free energies. Of course, the multiconformational free energy formulas have general validity for any environment. When comparing free energy differences in water and octanol, differences in lipophilicities can be calculated, and when comparing a water and a protein environment, differences in binding affinities are obtained. In this context, the different conformational states i^A(protein), i^B(protein) in the multiconformational free energy formulas may also refer to different relevant binding modes.

It is important to understand that here we calculated the free energy differences over the complete conformational space by two independent approaches. The first one, was a direct calculation by ES-OS using a reference state which samples the complete conformational space sufficiently. The second approach builds on ideas of configuration space decomposition[6] uses of the multiconformational free energy formulas which require only free energy differences for individual conformations and conformational populations of one compound (for eqs 16, 17, 19, and 20) in certain environments. While the first approach is more efficient, it is applicable only to sets of very similar compounds. On the other hand, the second approach has a much wider range of applicability. For example, the calculation of free energy differences between GTP analogs with larger and more diversely substituted groups cannot be addressed by OS perturbation approaches.[47] A large change in the Hamiltonian requires more steps of perturbation (e.g., within FEP[29] or TI[48]) between individual compounds, and for each of these steps, sufficient conformational sampling is required. One possible solution by using H-REMD is discussed in the introduction. On the other hand, the multiconformational free energy formulas allow for a much easier solution, in which only the conformational populations for one selected compound (soft or real) and the free energy differences for individual conformational states are needed. The populations can be obtained from a specifically designed soft reference state or by using enhanced sampling methods like REMD,[49,50] umbrella sampling,[51] local elevation,[52] hidden restraints,[12] and so forth. The free energy differences within one conformational space are generally converging very fast because there is no need to cross high energy barriers.

Conclusions

This paper presents the enhanced sampling-one step (ES-OS) method, which characterizes a set of chemically similar compounds with high intramolecular energy barriers structurally and energetically. The method was applied to a set of five 8-substituted GTP analogs. An unphysical reference state was constructed based on 8-Br-GTP where the Br atom was treated as a "soft atom"; that is, all nonbonded interactions involving Br are treated as soft-core interactions. This allowed us to perturb the reference state into the real states of 8-substituents of GTP. Another set of soft-core interactions was applied between the atoms of the base and the ribose allowing the reference state to cross high energy barriers. The final step of ES-OS is a one-step perturbation to the real compounds using Zwanzigs formula, thus obtaining free energy differences between the reference and the real compounds. Combining free energy differences between 8-substituted GTPs in water, vacuum, and octanol enabled us to calculate differences in solvation and lipophilicity. Using the probability of individual structures of the MD simulation, we constructed distributions of the dihedral angle around the glycosidic bond, the ^3J(C4,H1′) values, and the base−water and base−sugar nonbonded energy. The comparison of these distributions for GTP and 8-Br-GTP in water shows a good agreement with the ones calculated by H-REMD using soft-core interactions. Also, the ensemble average $\langle {}^3J(C4,H1')\rangle$ of GTP in water is in very good agreement with the experimental NMR value for GMP.

All free energy differences were calculated over the complete conformational space as well as separately for the anti and syn conformations. In analogy to the conformational space decomposition approach, the relationship between them was derived, which deepens our understanding of how free energy differences from individual conformations contribute to the overall values. The relationships have a significant practical potential also for more diverse compounds because the calculation of free energy differences within one conformation is in many cases much simpler as there is no need to cross high energy barriers.

Acknowledgment. We gratefully acknowledge financial support from The Netherlands Organization for Scientific Research (NWO), VENI Grant 700.55.401 (CO) and Horizon Breakthrough Grant 935.18.018 (JH).

Supporting Information Available: Force field parameters for 8-substituted analogs of GTP. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Simonson, T.; Archontis, G.; Karplus, M. *Acc. Chem. Res.* **2002**, *35*, 430.

(2) Brandsdal, B. O.; Osterberg, F.; Almlof, M.; Feierberg, I.; Luzhkov, V. B.; Aqvist, J. *Adv. Protein Chem.* **2003**, *66*, 123.

(3) Raha, K.; Kenneth, M.; Merz, J. *Ann. Rep. Comput. Chem.* **2005**, *1*, 113.

(4) Rodinger, T.; Pomes, R. *Curr. Opin. Struct. Biol.* **2005**, *15*, 164.

(5) Shirts, M. R.; Mobley, D. L.; Chodera, J. D. *Ann. Rep. Comput. Chem.* **2007**, *3*, 41.

(6) Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1989**, *90*, 3300.

(7) Straatsma, T. P.; McCammon, J. A. *J. Chem. Phys.* **1989**, *91*, 3631.

(8) Tobias, D. J.; Brooks, C. L.; Fleischman, S. H. *Chem. Phys. Lett.* **1989**, *156*, 256.

(9) Hermans, J.; Yun, R. H.; Anderson, A. G. *J. Comp. Phys.* **1992**, *13*, 429.

(10) Leitgeb, M.; Schroder, C.; Boresch, S. *J. Chem. Phys.* **2005**, *122*, 084109.

(11) Mobley, D. L.; Chodera, J. D.; Dill, K. A. *J. Chem. Theory & Comput.* **2007**, *3*, 1231.

(12) Christen, M.; Kunz, A.-P. E.; Van Gunsteren, W. F. *J. Phys. Chem. B* **2006**, *110*, 8488.

(13) Hritz, J.; Oostenbrink, C. *J. Chem. Phys.* **2008**, *128*, 144121.

(14) Stolarski, R.; Hagberg, C. E.; Shugar, D. *Eur. J. Biochem.* **1984**, *138*, 187.

(15) Bitetti-Putzer, R.; Yang, W.; Karplus, M. *Chem. Phys. Lett.* **2003**, *377*, 633.

(16) Min, D. H.; Li, H. Z.; Li, G. H.; Bitetti-Putzer, R.; Yang, W. *J. Chem. Phys.* **2007**, *126*, 144109.

(17) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1994**, *222*, 529.

(18) Hritz, J.; Oostenbrink, C. *J. Chem. Phys.* **2007**, *127*, 204104.

(19) Liu, H. Y.; Mark, A. E.; van Gunsteren, W. F. *J. Phys. Chem.* **1996**, *100*, 9485.

(20) Oostenbrink, C.; van Gunsteren, W. F. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6750.

(21) Läpp chen, T.; Hartog, A. F.; Pinas, V. A.; Koomen, G. J.; den Blaauwen, T. *Biochemistry* **2005**, *44*, 7879.

(22) Lappchen, T.; Pinas, V. A.; Hartog, A. F.; Koomen, G. J.; Schaffner-Barbero, C.; Andreu, J. M.; Trambaiolo, D.; Lowe, J.; Juhem, A.; Popov, A. V.; den Blaauwen, T. *Chem. Biol.* **2008**, *15*, 1.

(23) Dias, E.; Battiste, J. L.; Williamson, J. R. *J. Am. Chem. Soc.* **1994**, *116*, 4479.

(24) Proctor, D.; Kierzek, E.; Kierzek, R.; Bevilacqua, P. *J. Am. Chem. Soc.* **2003**, *125*, 2390.

(25) Yajima, R.; Proctor, D. J.; Kierzek, R.; Kierzek, E.; Bevilacqua, P. C. *Chem. & Biol.* **2007**, *14*, 23.

(26) Schwede, F.; Maronde, E.; Genieser, H. G.; Jastorff, B. *Pharmacol. & Ther.* **2000**, *87*, 199.

(27) Ghosh, A.; Lagisetty, P.; Zajc, B. *J. Org. Chem.* **2007**, *72*, 8222.

(28) Liu, J.; Barrio, J.; Satyamurthy, N. *J. Fluor. Chem.* **2006**, *127*, 1175.

(29) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420.

(30) Mark, A. E.; Xu, Y. W.; Liu, H. Y.; van Gunsteren, W. F. *Acta Biochimica Polonica* **1995**, *42*, 525.

(31) Oostenbrink, C.; van Gunsteren, W. F. Chem.—Eur. J., in press.

(32) Christen, M.; Hunenberger, P. H.; Bakowies, D.; Baron, R.; Burgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Krautler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; Van Gunsteren, W. F. *J. Comput. Chem.* **2005**, *26*, 1719.

(33) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.

(34) Hockney, R. W. *Meth. Comput. Phys.* **1970**, *9*, 136.

(35) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction models for water in relation to protein hydration. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, The Netherlands, 1981; pp 331.

(36) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.

(37) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451.

(38) Komooka, H. *Bull. Chem. Soc. Jpn.* **1972**, *45*, 1696.

(39) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656.

(40) Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids*; Clarendon Press: Oxford, 1987.

(41) Schäfer, H.; van Gunsteren, W. F.; Mark, A. E. *J. Comput. Chem.* **1999**, *20*, 1604.

(42) Trantirek, L.; Stefl, R.; Masse, J. E.; Feigon, J.; Sklenar, V. *J. Biomol. NMR* **2002**, *23*, 1.

(43) Schwalbe, H.; Marino, J. P.; King, G. C.; Wechselberger, R.; Bermel, W.; Griesinger, C. *J. Biomol. NMR* **1994**, *4*, 631.

(44) Wildman, S. A.; Crippen, G. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868.

(45) Molecular Operating Environment, v., Chemical Computing Group, Montreal, Canada.

(46) Ferner, J.; Villa, A.; Duchardt, E.; Widjajakusuma, E.; Wohnert, J.; Stock, G.; Schwalbe, H. *Nucleic Acids Res.* **2008**, *36*, 1928.

(47) Oostenbrink, C.; van Gunsteren, W. F. *J. Comput. Chem.* **2003**, *24*, 1730.

(48) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300.

(49) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140.

(50) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141.

(51) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187.

(52) Huber, P.; Torda, A. E.; Van Gunsteren, W. F. *J. Comput.-Aided Mol. Design* **1994**, *8*, 695.

JP902968M

Paper 10

Jandova, Z.; Trosanova, Z.; Weisova, V.; Oostenbrink, C.*, Hritz, J.*:Free energy calculations on the stability of the 14-3-3ζ protein. *BBA - Proteins and Proteomics*, **2018**, 1866, 442-450.

Contents lists available at ScienceDirect

BBA - Proteins and Proteomics

journal homepage: www.elsevier.com/locate/bbapap

Free energy calculations on the stability of the 14-3-3ζ protein

Zuzana Jandova[a], Zuzana Trosanova[b], Veronika Weisova[b], Chris Oostenbrink[a,*], Jozef Hritz[b,*]

[a] Institute of Molecular Modeling and Simulation, University of Natural Resources and Life Sciences, Vienna, Austria
[b] CEITEC-MU, Masaryk University, Kamenice 753/5, Bohunice, Brno, Czech Republic

ARTICLE INFO

Keywords:
14-3-3 protein
Protein stability
Molecular dynamics simulation
Differential scanning calorimetry
Free energy calculation
Thermodynamic integration

ABSTRACT

Mutations of cysteine are often introduced to e.g. avoid formation of non-physiological inter-molecular disulfide bridges in in-vitro experiments, or to maintain specificity in labeling experiments. Alanine or serine is typically preferred, which usually do not alter the overall protein stability, when the original cysteine was surface exposed. However, selecting the optimal mutation for cysteines in the hydrophobic core of the protein is more challenging. In this work, the stability of selected Cys mutants of 14-3-3ζ was predicted by free-energy calculations and the obtained data were compared with experimentally determined stabilities. Both the computational predictions as well as the experimental validation point at a significant destabilization of mutants C94A and C94S. This destabilization could be attributed to the formation of hydrophobic cavities and a polar solvation of a hydrophilic side chain. A L12E, M78K double mutant was further studied in terms of its reduced dimerization propensity. In contrast to naïve expectations, this double mutant did not lead to the formation of strong salt bridges, which was rationalized in terms of a preferred solvation of the ionic species. Again, experiments agreed with the calculations by confirming the monomerization of the double mutants. Overall, the simulation data is in good agreement with experiments and offers additional insight into the stability and dimerization of this important family of regulatory proteins.

1. Introduction

The 14-3-3 proteins form a family of acidic proteins with molecular weight of approximately 30 kD [1] ubiquitous in all eukaryotic organisms. There are seven highly conserved isoforms of the 14-3-3 family in mammals, two in yeast and fifteen in plants [2–4]. The 14-3-3 proteins function mostly as hetero and homo dimers [5,6], with each of the monomers consisting of nine antiparallel helices. The complete structure of the protein is shown in Fig. 1. The first four helices, starting from the N-terminus, create a hydrophobic dimer interface and the floor of the binding channel. The next 5 helices create the amphipathic interior of the peptide-binding channel [7,8]. Residues aligning the channel belong to the most highly conserved residues among all isoforms, while the sequence of the surface and flexible C-terminal residues is more diverse and responsible for isoform-substrate specificity [1,5,9].

14-3-3 proteins play a key role in regulating intracellular signaling pathways, common cellular processes, primary metabolism, anti-apoptotic pathways and cellular proliferation [10,11]. They act as adapters for > 200 binding partners by either allosteric modification of enzymes or controlling the assembly of protein complexes [12–16]. Among the best known 14-3-3 binding partners are tyrosine hydroxylase, Raf-1 (RAF proto-oncogene ser/thr-protein kinase), BCR (B-cell receptor), PKC (protein kinase C), MLK (mixed lineage ser/thr kinase), Tau protein or p53 (cellular tumor antigen p53) [16–18]. Their protein binding partners usually contain intrinsically disordered regions, typically containing motifs involving phosphoserine or phosphothreonine with a consensus sequence RSXpSXP (mode I) or RX[FY]XpSXP (mode II) [19]. Enhanced sampling molecular dynamics (MD) simulations of the phosphopeptides binding to 14-3-3ζ showed the presence of a dominant binding pathway, roughly following helix 3. In the process of phosphopeptide binding, 14-3-3ζ was observed to adopt a so-called “wide-opened” conformation that is not usually present in the apo or holo states [20]. 14-3-3 dimers contain two binding sites and are therefore able to accommodate also doubly-phosphorylated binding partners. In addition to the traditionally considered single partner with two phosphorylation sites occupying the individual binding cavities within the 14-3-3ζ dimer [21,22], two more major binding modes were confirmed by ^{31}P NMR titration experiments [23].

In this paper, we focus on computational predictions of the stability of the 14-3-3ζ isoform upon mutations, using MD simulations and thermodynamic integration (TI) to compute free-energy differences. Our computational predictions are compared to experimentally determined melting temperatures (T_M) as obtained from differential scanning calorimetry and to native gel electrophoresis experiments, as

[*] Corresponding authors.
E-mail addresses: chris.oostenbrink@boku.ac.at (C. Oostenbrink), jozef.hritz@ceitec.muni.cz (J. Hritz).

https://doi.org/10.1016/j.bbapap.2017.11.012
Received 29 August 2017; Received in revised form 31 October 2017; Accepted 25 November 2017
Available online 05 December 2017

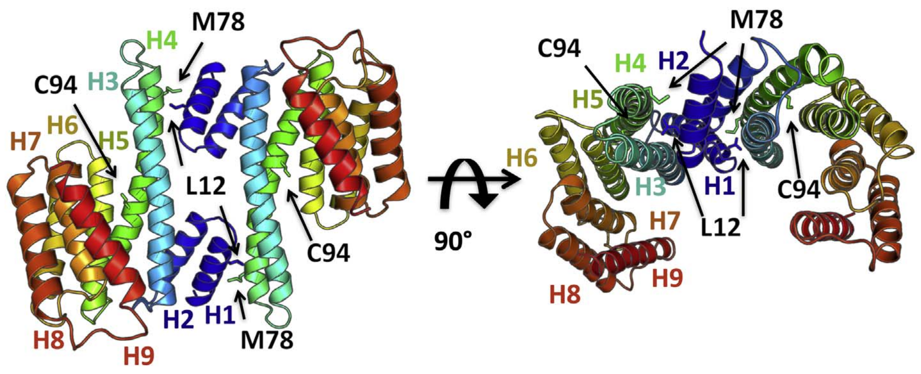


Fig. 1. 14-3-3ζ protein with helices H1-H9 marked in one monomer and the mutated residues L12, M78 and C94 indicated in stick representation. The unstructured C-terminus (residues 229–245) was removed because of its high flexibility.

indirect measures of protein stability and dimerization [24]. While the thermodynamic stability as computed from our simulations is not expected to quantitatively correlate with T_M, both are measures of stability of the protein [25]. A comparison of homologous proteins from mesophillic and thermophilic organisms e.g. reveals that for the majority of cases, the melting temperature is increased as the result of an increase of the unfolding free energy at the entire temperature range [26].

Mutations of cysteine may be introduced for multiple practical reasons. Cysteine contains a reactive sulfhydryl group, allowing for the formation of disulfide bridges. However, cysteines that are not involved in intramolecular disulfide bridges may be prone to unwanted oxidation or crosslinking under experimental conditions. On the other hand, various experimental methods make use of specific labels that are anchored to the protein (e.g. chromopohoric or spin labels) to gain insight into the function and dynamics of proteins. To maintain the specificity in labeling, other cysteines, naturally occurring in the sequence need to be replaced by other amino acids preferably without altering the overall protein stability. This is commonly done by a replacement by alanine or serine. While mutating the surface exposed single cysteines for alanines or serines usually does not alter the overall protein stability, it is much more complicated when mutating the cysteins in the hydrophobic core of the protein. The 14-3-3ζ protein contains three cysteines, of which two (Cys25 and Cys189) are located on the protein surface and their mutagenesis for the Ala has negligible impact on the protein stability [23]. The third, Cys94, is located in the central helix H4 and buried in a hydrophobic pocket, therefore its mutation influences the overall stability of the protein. Here, we investigate the possibility to propose protein mutants by MD simulations that would sustain or improve protein stability upon cysteine mutation and verify our results by experimental melting temperatures. Apart from the most common choices C94A and C94S, we have further considered three more hydrophobic mutants, C94I, C94L and C94V, based on the location of the residue in the hydrophobic core of the protein.

Further, we investigate a double mutation of Leu12 and Met78, which are located at the dimer interface and influence protein dimerization. In the 14-3-3ζ homodimer several salt bridges as well as buried polar and hydrophobic residues are engaged in the formation of a stable dimer interface [8]. The double mutant Leu12Glu/Met78Lys was identified to destabilise this interface and thereby hinder dimer formation while at the same time it does not change the overall charge of the 14-3-3ζ monomeric unit. As the mutations are in a position that would allow for stable salt bridges, a destabilization of the dimer interface seems somewhat counter intuitive. Here, our aim is to rationalize the experimental results by molecular simulations.

For the calculation of free energy differences, we make use of thermodynamic cycles. The one we use for the mutations of Cys94 can be seen in Fig. 2. In this figure, the dimeric 14-3-3ζ is indicated by two roughly L-shaped figures with the wildtype residue (Cys94) indicated with a triangle. A hypothetical unfolded protein is indicated by a wavy bar at the bottom of the cycle. Mutations are implied by the white squares.

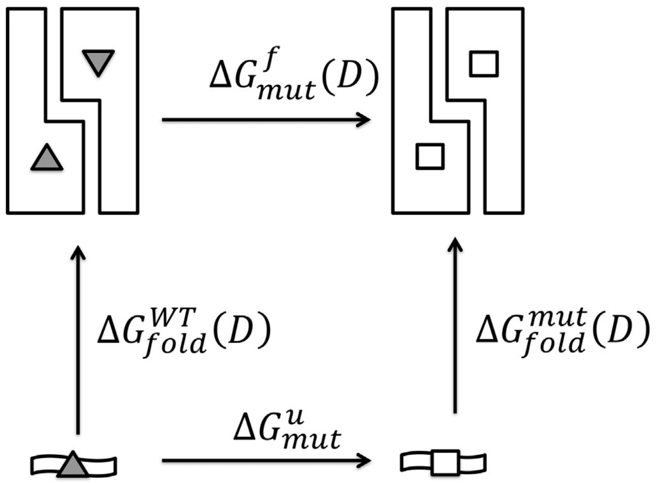


Fig. 2. Thermodynamic cycle for the mutation of Cys94. The upper arrow corresponds to the free energy of mutation in the dimer, the lower arrow in the unfolded state. Grey triangles represent WT Cys94, white squares the mutants.

Considering that the Gibbs free energy is a state function [27] the value around the cycle is equal to zero. This allows us to write:

$$\begin{aligned}\Delta\Delta G_{mut}(D) &= 2\,\Delta G^{u}_{mut} - \Delta G^{f}_{mut}(D) = \Delta G^{WT}_{fold}(D) - \Delta G^{mut}_{fold}(D)\\ &= \Delta\Delta G_{fold}(D)\end{aligned} \tag{1}$$

where $\Delta\Delta G_{mut}(D)$ represents the relative free energy difference of mutation of one aminoacid into another, which can be computed as the difference of the mutation free energy in the dimer, $\Delta G_{mut}{}^{f}(D)$ and in the unfolded state $\Delta G_{mut}{}^{u}$. The same quantity can be written as $\Delta\Delta G_{fold}(D)$, which is the relative free energy of folding and is the difference of the folding free energy of the wildtype, $\Delta G_{fold}{}^{WT}(D)$, and of the mutant, $\Delta G_{fold}{}^{mut}(D)$. The addition (D) explicitly refers to the dimeric state. Current computational resources do not allow us to simulate the full process of protein folding and unfolding to determine $\Delta\Delta G_{fold}(D)$ directly. However, we are able to estimate this value from free energy perturbation in a folded and unfolded state [28,29]. Selection of the right representation of the unfolded state is crucial for free energy calculations [30]. We have decided to use a tripeptide with its central residue being mutated and two neighbouring residues resembling the ones in protein context, as was previously shown to agree well with experimental results [31]. For the folded state, we simulated both the monomer and the dimer to obtain a better understanding of the impact of dimerization on the stability of the protein.

2. Methods

2.1. Experimental

2.1.1. Cloning, expression, and purification of 14-3-3ζ variants

A codon-optimized gene for 14-3-3ζ (GenScript, Piscataway Township, NJ) was inserted into pET15b and two mutations C25A and C189A were inserted. Due to the fact that these mutations have negligible impact on the protein stability (T_M is slightly increased from 59.5 °C to the 60.3 °C) and avoid the formation of any intermolecular artificial disulfide bonds, we used this construct as a 14-3-3ζ pseudo-WT. The 14-3-3ζ gene with the C25A and C189A mutations was therefore used as the "parental" construct for all other studied mutations, which were introduced by using the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent Technologies, Santa Clara, CA, USA). The DNA sequence for all constructs was confirmed by sequencing (Macrogene, Seoul, Rep. of Korea). All 14-3-3ζ variants were expressed and purified by procedure described in Hritz et al. [23]. Final purity of the prepared protein variants was verified by MALDI-TOF-MS spectroscopy.

2.1.2. Differential scanning calorimetry (DSC)

Melting temperatures (T_M) were obtained from DSC measurements carried out at a VP-DSC instrument (MicroCal Inc., Northampton, MA) at a heating scan rate of 1 °C per minute from 25 to 80 °C. Samples of 14-3-3ζ C94 mutants were prepared at 1 mg/ml concentration and dialyzed against a degassed 20 mM sodium phosphate buffer (pH 6.0). DSC data were analyzed using the Microcal Origin 7.0 software (MicroCal Inc., Northampton, MA).

2.1.3. Native polyacrylamide gel electrophoresis (Native-PAGE)

10 μl of 2 μM protein sample were mixed with 10 μl of 2 × loading buffer containing 2% glycerol, 2% Bromophenol Blue, and 160 mM TRIS-HCl (pH = 6.8). Samples were then loaded into a freshly prepared 12.5% native minigel. Electrophoresis was conducted for approximately 200 min using constant voltage (95 V) in a native electrophoretic buffer (2.5 mM TRIS-HCl, 19.2 mM Glycine, pH 8.3). The apparatus was cooled down on ice during whole procedure in order to prevent the thermal denaturation of the studied proteins. Finally, the gel was stained by Coomassie Brilliant Blue R-250 (AppliChem, Darmstadt, Germany).

2.2. MD simulations

As starting configuration for molecular dynamics simulations we used an apo 14-3-3ζ crystal structure (PDB ID:1A4O [8]) at a 2.8 Å resolution. Missing loops and residues were remodelled from a holo crystal structure (PDB ID:4HKC [32]). C-terminal residues 229–245 were removed from the structure to avoid interactions of the highly flexible and unstructured C-terminus with the rest of the protein. In preliminary simulations, a high strain in the backbone of Arg18 was observed, leading to failures of the SHAKE algorithm to constrain bond lengths. It appeared that the ϕ and ψ angles of Glu17 were at the very bottom of the left-handed helical configuration in the Ramachandran plot ($\phi = 73°$, $\psi = -16°$), leading to strain in the loop. By turning ψ of A16 and ϕ of E17 by approximately 180°, the strain was released and no further SHAKE failures were observed. All arginines, lysines and cysteines were protonated, while aspartates and glutamates were deprotonated. His164 was protonated on N_ε, which was based on better hydrogen bonding possibilities in its surrounding. All MD simulations were carried out using the GROMOS11 software simulation package [33], employing the 54a8 forcefield [34]. Proteins were energy-minimized in vacuum using the steepest-descent algorithm and subsequently solvated in a rectangular, periodic and pre-equilibrated box of single point charge (SPC) water [35]. Minimum solute to box-wall distances were set to 2, 2, and 2.5 nm for dimer and 1, 1 and 2.5 nm for monomer in the x,y and z-dimensions, respectively, after alignment of the largest intramolecular distance with the z-axis. This led to systems containing about 50 to 60 thousand atoms for the monomer and 110 to 117 thousand atoms for the dimer. Another minimization in water was performed using the steepest descent algorithm. To achieve electroneutrality of the system with monomer and dimer 8 and 16 sodium ions were added to the system, respectively. Because previous simulations on very similar systems were shown to have increased structural stability at higher salt concentrations [20], simulations were additionally performed at a concentration of 300 mM NaCl. To achieve this, additional 94–127 sodium and chloride ions were added to the box of monomer and 209–219 sodium and chloride ions were added to the box of dimer. For the equilibration, the following protocol was used: initial velocities were randomly assigned according to a Maxwell–Boltzmann distribution at 60 K. All solute atoms were positionally restrained with a harmonic potential using a force constant of 2.5×10^4 kJ/mol nm^{-2}. In each of the four subsequent 20 ps MD simulations, the force constant of the positional restraints was reduced by one order of magnitude and the temperature was increased by 60 K. Subsequently, the positional restraints were removed and rototranslational constraints were introduced on all solute atoms [36]. The last step of equilibration was performed at a constant pressure of 1 atm for 300 ps. After equilibration production runs of 60 ns were performed with constant number of particles, constant temperature (300 K) and constant pressure (1 atm). To sustain a constant temperature, we used the weak-coupling thermostat [37] with a coupling time of 0.1 ps. The pressure was maintained using a weak coupling barostat with a coupling time of 0.5 ps and an isothermal compressibility of 4.575×10^{-4} kJ^{-1} mol nm^{-3}. Solute and solvent were coupled to separate temperature baths. Implementation of the SHAKE algorithm [38] to constrain bond lengths of solute and solvent to their optimal values allowed for a 2-fs time-step. Nonbonded interactions were calculated using a triple range scheme. Interactions within a short-range cutoff of 0.8 nm were calculated at every time step from a pair list that was updated every fifth step. At these points, interactions between 0.8 and 1.4 nm were also calculated explicitly and kept constant between updates. A reaction field [39] contribution was added to the electrostatic interactions and forces to account for a homogenous medium outside the long-range cutoff using a relative dielectric constant of 61 as appropriate for the SPC water model [40]. Coordinate and energy trajectories were stored every 0.5 ps for subsequent analysis. To study the overall stability of the WT protein, a total of eight independent simulations were performed: Two independent simulations of both monomer and dimer in both 0 and 300 mM NaCl, further noted as MD1 and MD2.

2.3. Thermodynamic integration (TI)

Using thermodynamic integration, the free energy difference between two states A and B can be computed via multiple discrete intermediate steps using a coupling parameter λ. If λ is equal to 0, the system is in state A and, on the other hand the system is in state B when λ is equal to 0. At each intermediate step the average of the derivative of the Hamiltonian with respect to λ is calculated. By integrating over the derivatives along the path we obtain the total free energy difference (ΔG_{BA}).

$$\Delta G_{BA} = \int_0^1 \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{2}$$

TI calculations for perturbations of selected amino acid were initially performed at 11 evenly spaced λ steps. Starting from initial structures taken after above mentioned equilibration with low salt content, 20 ps of equilibration at each λ value were followed by 1 ns of production run. If necessary, the simulations were prolonged to maximally 5 ns per λ point or additional λ points were added to decrease the overall error estimate below 3 kJ/mol. For perturbed atoms we used soft-core parameters of 0.5 for the van der Waals and 0.5 nm^2 for electrostatic interactions [41]. A single topology approach was used, by introducing dummy atoms where necessary. Dummy atoms have all nonbonded interactions set to zero while the bonded interactions and the masses of individual atoms remain the same as for real atoms. This way we can selectively convert real atoms into dummy atoms and vice versa to switch between particular residues. Fig. 3 shows the complete list of structural formulas of perturbed residues, including the dummy atoms, for both sets of mutations.

3. Results and discussion

3.1. Overall protein stability

First, we investigated the overall stability of 14-3-3ζ as WT protein. We performed plain MD simulations of monomers and dimers in 0 and 300 mM NaCl for 60 ns. For each of the simulation two independent simulations were performed, noted as MD1 and MD2. The atom-positional root-mean-square deviation of C_α atoms per monomer increased up to a maximum of 1 nm after 30 ns of the dimer simulation MD1. Time series of the root-mean-square deviations of all simulations can be found in the Supplementary Material (Fig. S1). The root-mean-square fluctuations of C_α atoms (Fig. 4) as well as the secondary structure analysis (Fig. 5) shows, that the monomer simulations are characterized by a lower stability, especially in the first three helices H1 to H3 (residues 1–70), than the dimer simulations. In contrast, the last three helices (residues 162–228) seem to be less stable in simulation of dimers, which is in agreement with previous studies [20].

For the secondary structure analysis we used the DSSP algorithm (Define Secondary Structure of Proteins) [42]. Fig. 5 shows the average secondary structure propensity per residue for all simulations, i.e. MD1 and MD2 in 0 and 300 mM NaCl for the monomer simulations and averaged over both monomers for the dimer simulations. In the dimer simulation, these helices keep their α-helical secondary structure for almost the entire simulation, whereas in the monomer simulation we observe unfolding of the first three helices. Exactly these three helices are necessary for maintaining the dimer interface. After monomer separation, the interface residues become unstable which, after some time, results in a disruption of the secondary structure.

Apart from the changes in the local secondary structure of the protein a global opening and closing of the binding channel occurs. As previously described [20,43], the distance between the third and eighth helix fluctuates significantly. Fig. 6 shows the distance between Cα carbons of Gly53 and Leu191. This distance was previously suggested to describe the openness of 14-3-3ζ. While the opening distance of monomers in the dimer simulations seems to yield different distributions for every simulation, all four monomer simulations consistently show the maximum in the distance distribution at 2.8 nm. In particular the simulations of dimers at a low salt concentration seem to be more diverse in the openness of the binding channel.

Furthermore, similarly to the previous work of Nagy et al. [20], an inter-monomer twist of 14-3-3ζ was observed. It means that one of the monomers is rotating with respect to the other by up to 80°. The time series of a dihedral angle showing this twist, represented by the dihedral angle Leu43(M1)-Ala54(M1)-Ala54(M2)-Leu43(M2), where M1 and M2 represents monomer 1 and 2 respectively, can be found in the Supplementary material (Fig. S2). Our analysis shows, that a large variability is observed in simulations with salt concentration of 0 and 300 mM NaCl, with intermonomer twists ranging between 15 and 80°.

3.2. Mutation of Cys94

Cysteine 94 is located in a hydrophobic environment in the middle of helix 3. We have perturbed this residue into 5 different amino acids: Alanine, Isoleucine, Leucine, Valine and Serine. The structural formulas and details of these mutations are shown in Fig. 3. The free energy differences for TI perturbations in Ile-Cys-Asn tripeptide as well as in protein were evaluated based of the thermodynamic cycle, shown in Fig. 2. Perturbation in the protein was carried out in both, monomer and dimer. Fig. S3 shows the TI curves of the perturbations in tripeptide, monomer and dimer for all five mutations. Perturbation free energies for tripeptide ($\Delta G_{mut}{}^{u}$) as well as monomer ($\Delta G_{mut}{}^{f}(M)$) were

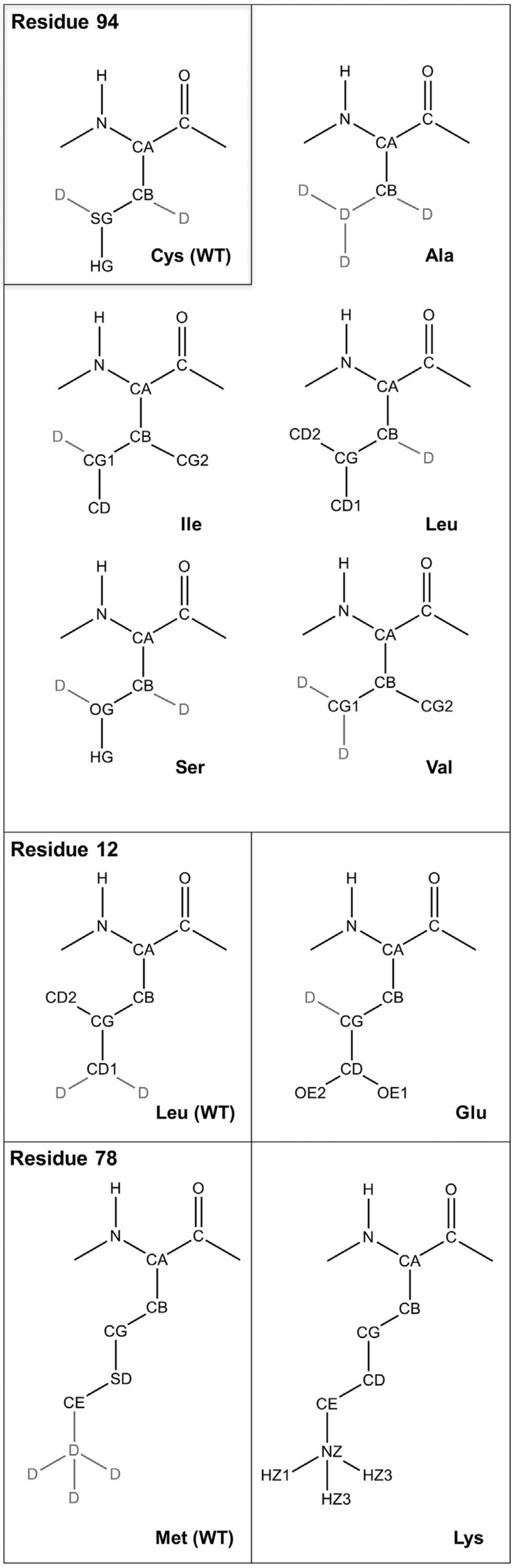


Fig. 3. Structural formulas of perturbed residues. D stands for dummy atoms.

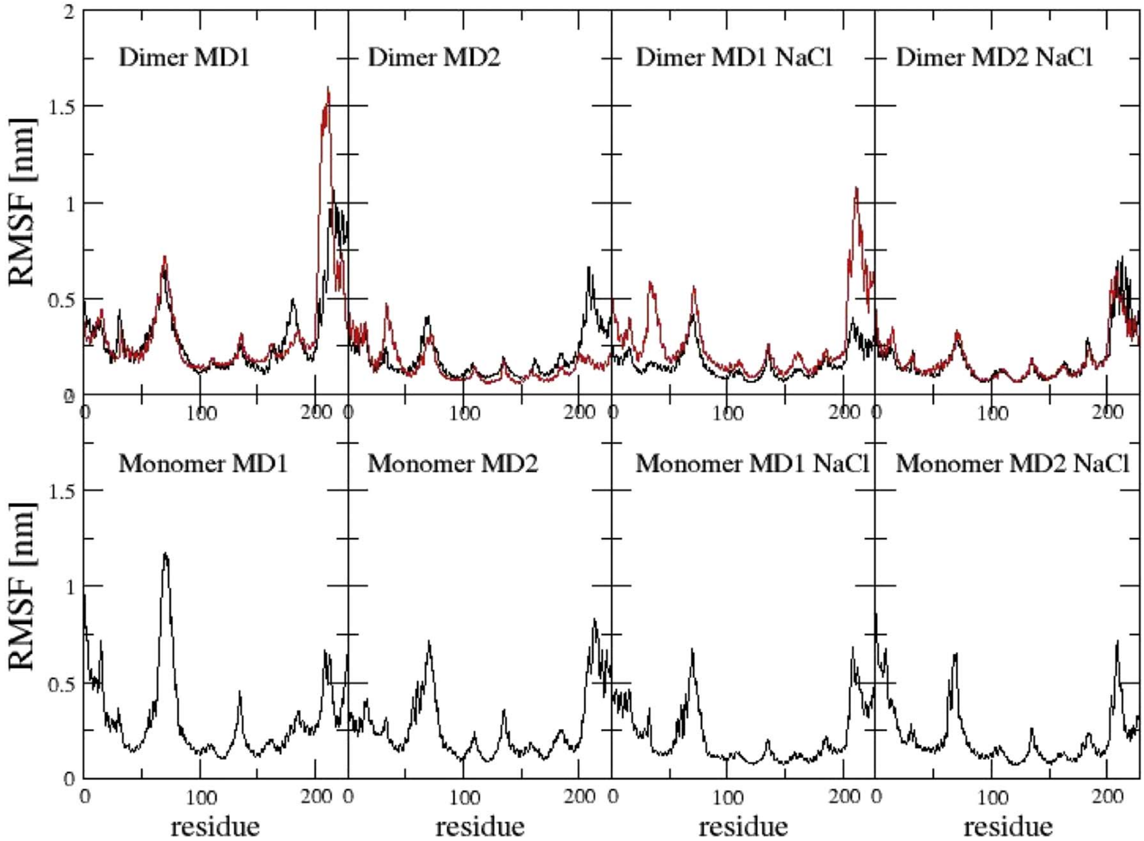


Fig. 4. RMSF of Cα atoms of dimers and monomers. First monomer of the dimer in red, second in black. NaCl indicates the simulations that were performed at 300 mM NaCl concentration.

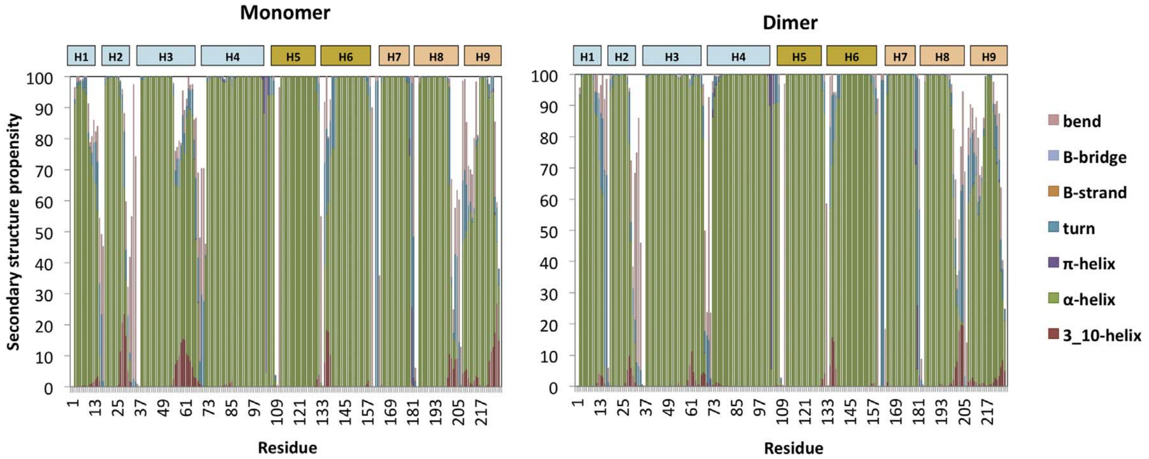


Fig. 5. Secondary structure propensities as calculated by the DSSP program [42] for 14-3-3ζ monomer and dimer simulations, averaged over all plain WT simulations.

multiplied by two, to be comparable with the double perturbation in dimer ($\Delta G_{mut}^{f}(D)$). Table 1 summarizes the free energies obtained from the simulations. The free energies as obtained for the monomers agree very well with the stability predicted for the dimers. Both sets of simulations suggest that the Ala and Ser mutations lead to a loss of stability ($\Delta\Delta G_{mut}^{f} > 0$) while the Val mutant shows comparable stability as the wildtype and Leu and Ile increase the stability at the simulated temperature. For comparison, the experimental melting temperatures as determined by DSC (experimental details described in the Methods) are also included in Table 1.

While $\Delta\Delta G_{mut}^{f}$ is a measure of stability at a given thermodynamic state (here at a temperature of 300 K), the melting temperature depends on the behavior of the protein at a range of different temperatures. Therefore, and because the thermal unfolding of the 14-3-3 constructs was observed to be fully irreversible, $\Delta\Delta G_{mut}^{f}$ and T_M cannot be expected to correlate exactly. Still, both quantities are indicative of protein stability. In fact, a common strategy by which thermophilic proteins increase the melting temperature is to decrease the free energy of folding over the entire temperature range [26]. Arguably, a single point mutation is more likely to crease the folding free energy at the entire temperature range, than to have a significant effect on the protein heat capacity, leading to alternative temperature dependencies.

Our simulation data agrees with the experimental stability estimates by identifying Ser and Ala as the two least stable mutants, with $\Delta\Delta G_{mut}(D)$ of 54 kJ/mol and 18 kJ/mol, respectively. However, the simulations predict Ile and Leu to be more stable than the WT, while the

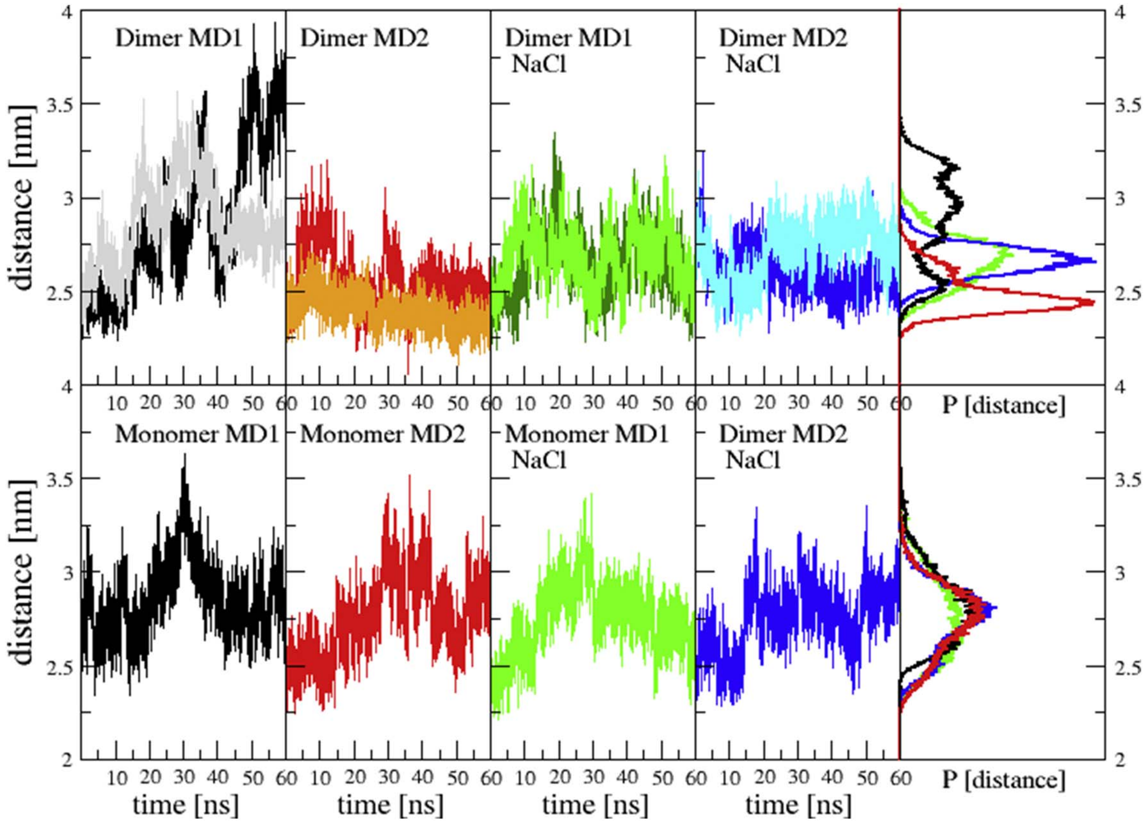


Fig. 6. Monomer opening: Distances between Cα carbons of Glu53 and Leu191. Different colours in dimer graphs correspond to monomer 1 and monomer 2. NaCl indicates the simulations that were performed at 300 mM NaCl concentration.

Table 1
Predicted free-energy differences, ΔG, for monomer, dimer and tripeptide simulation and their differences ΔΔG in kJ/mol. T_M corresponds to the experimentally determined melting temperatures in °C.

	$2\Delta G_{mut}^{f}(M)$	$\Delta G_{mut}^{f}(D)$	$2\Delta G_{mut}^{u}$	$2\Delta\Delta G_{mut}^{f}(M)$	$\Delta\Delta G_{mut}^{f}(D)$	T_M
	[kJ/mol]					[°C]
Ala	55 ± 1	59 ± 1	41 ± 2	14 ± 2.0	18 ± 2	54.0
Ile	24 ± 3	29 ± 4	51 ± 3	−26 ± 4	−22 ± 5	55.6
Leu	21 ± 2	23 ± 2	34 ± 2	−13 ± 3	−11 ± 2	57.0
Ser	2 ± 1	9 ± 1	−46 ± 1	47 ± 1	55 ± 1	52.5
Val	56 ± 1	61 ± 2	60 ± 1	−4 ± 2	1 ± 2	55.0
Cys (WT)	0	0	0	0	0	60.3

experimentally determined T_M is slightly lower than for the wildtype. The large instability of Ala and Ser is striking, as these are typically the first mutations suggested to replace Cys in experiments. Both simulations and the experimentally determined melting temperatures show that these particular mutations are rather unfortunate choices to replace Cys94 in 14-3-3ζ.

Our rationalization of the destabilizing effect of the C94A and C94S mutants is that these mutations disrupt the hydrophobic core in that region [44]. Considering that position 94 is located in a very hydrophobic region among three helices, any disturbance of this hydrophobicity might lead to a decrease of protein stability. Detailed snapshots from the last step of simulations are shown in Fig. 7. When a hydrophobic cysteine residue [45] is replaced by a hydrophilic serine, there are only a few possibilities for serine to create hydrogen bonds within its surrounding. Hydrogen bonding between the sidechain oxygen of Ser94 and the backbone oxygen of Leu90 was observed in 45% of simulation time in the last simulation of the TI simulations (λ = 1.00). Interestingly, in one monomer we could see a hydrogen bond between a water molecule and Ser94. While hydrogen bonding is energetically favourable, introducing water molecules into a hydrophobic core might lead to structural reorganization and destabilization. On the other hand, after insertion of alanine into a densely packed core, a voluminous hydrophobic cavity is created. It has been previously shown, that substitution of bulky hydrophobic amino acids by smaller, cavity-creating ones is unfavourable and leads to a loss of stability [46]. The other branched hydrophobic aminoacids, on the contrary, fill the cavity and contribute favourably to the hydrophobic interactions in the core and, thus, stabilise the protein.

3.3. *Mutation of Leu12 and Met78*

3.3.1. *Experimental*

Under normal circumstances, without any posttranslational modifications the monomer/dimer equilibrium of the 14-3-3ζ isoform is

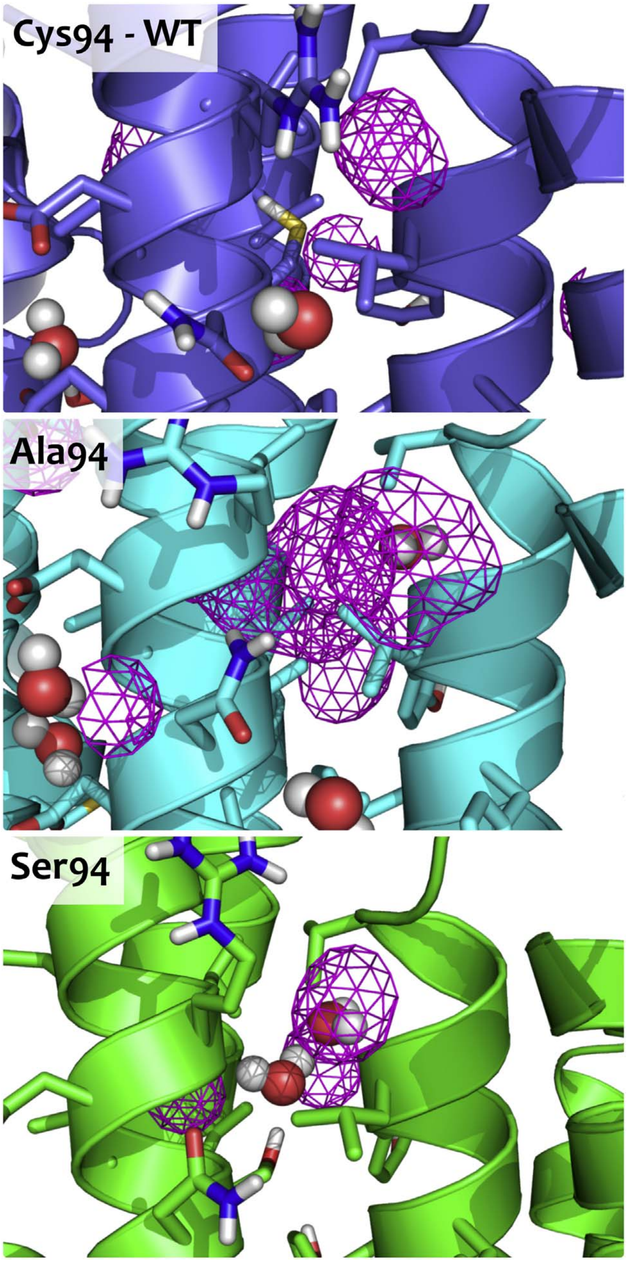


Fig. 7. Snapshots of the most the WT and the two most destabilizing mutants of Cys94. The cavities inside of the protein as computed by Pymol [47] are shown in pink.

strongly shifted towards the dimer [48]. A variety of 14-3-3 dimer-incapable forms were reported over last two decades and they are nicely summarized in Table 1 of the review of Sluchanko et al. [48] The listed mutations shifting this equilibrium towards monomer change the overall charge of 14-3-3 monomer with respect to the WT. Here, we designed a double mutation Leu12Glu and Met78Lys which does not change the overall charge of the 14-3-3ζ monomer and at the same time is almost exclusively in monomeric state at a μM concentration range. Fig. 8B shows a native PAGE gel of WT and the double mutant at 1 μM concentration. The L12E/M78K double mutant is propagating faster through the PAGE than WT. Considering that both of these proteins have the same overall charge per monomer our interpretation is that the L12E/M78K double mutant is in the monomeric form in contrast to the dimeric form of the 14-3-3ζ WT at the given conditions.

3.3.2. *Simulation*

To rationalize the experimental results computationally, we have performed in silico mutations L12E and M78K in three different systems. One of them was, again, a tripeptide representing the unfolded state within a protein context, Lys11-Leu12-Ala13 and Gln77-Met78-Ala79. The other two were in dimer and monomer simulations. In contrast to the mutations applied to Cys94, the individual mutations involve full charge changes. The L12E mutation leads to negative charges, while the M78K mutations leads to a positive charge (see Fig. 3). Thus, the overall charge of the system stays neutral and there is no need for additional corrections [49]. The total scheme of the resulting thermodynamic cycles can be seen in Fig. 8A. Here, we compute the relative dimerization free energies, $\Delta\Delta G_{dim}$ as the difference between the dimerization free energies of the mutant and the wildtype, which is the same as the difference in the mutation free energy in the mutant and in the monomers,

$$\Delta\Delta G_{dim} = \Delta G_{dim}^{mut} - \Delta G_{dim}^{WT} = \Delta G_{mut}^{f}(D) - 2\Delta G_{mut}^{f}(M) \tag{3}$$

Similarly to the Cys94 mutations, we can also define the relative folding free energy of the monomers as

$$\Delta\Delta G_{fold}(M) = \Delta G_{fold}^{mut}(M) - \Delta G_{fold}^{WT}(M) = \Delta G_{mut}^{f}(M) - \Delta G_{mut}^{u} \tag{4}$$

where in this case the mutational free energy of the unfolded state is the sum of two independent simulations of the tripeptides

$$\Delta G_{mut}^{u} = \Delta G_{mut}^{u}(L12E) + \Delta G_{mut}^{u}(M78K) \tag{5}$$

The folding free energy of the complete dimers is subsequently related to the previous properties as

$$\Delta\Delta G_{fold}(D) = \Delta\Delta G_{dim} + 2\Delta\Delta G_{fold}(M) \tag{6}$$

Table 2 shows the resulting free energy differences and TI curves for the mutations in all three systems can be found in Fig. S4. The free-

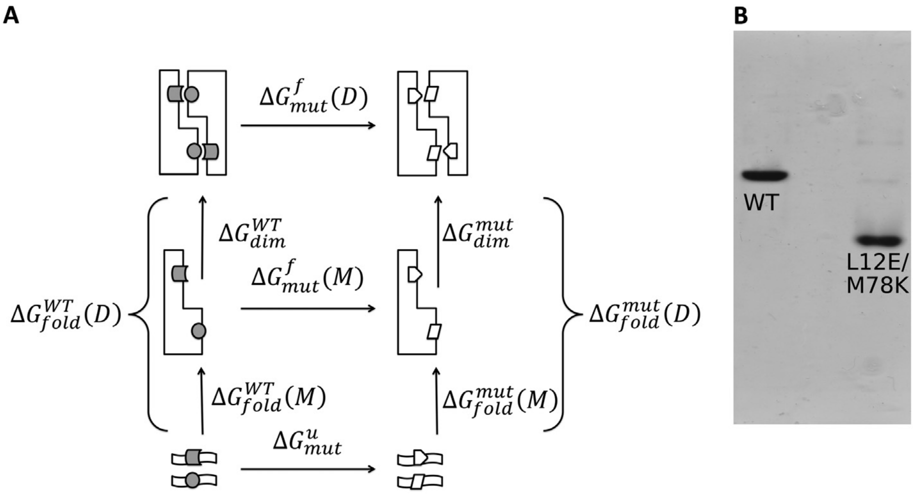


Fig. 8. A) Thermodynamic cycle of mutations at the 14-3-3ζ dimeric interface. WT Leu12 and Met78 are in grey cylinder and circle and Glu12 and Lys78 in white pentagon and parallelogram. B) Native PAGE of 14-3-3ζ (pseudo) WT and the L12E/M78K double mutant at 1 μM concentration showing the higher oligomeric state of WT.

Table 2
Free energy of mutation in different systems, and their differences, ΔΔG, in kJ/mol.

ΔG	[kJ/mol]
$\Delta G_{mut}{}^{f}(D)$	-1169 ± 9
$2\Delta G_{mut}{}^{f}(M)$	-1259 ± 11
$2\Delta G_{mut}{}^{u}$	-1296 ± 8
$2\Delta\Delta G_{fold}(M)$	37 ± 14
$2\Delta\Delta G_{dim}$	90 ± 14
$2\Delta\Delta G_{fold}(D)$	127 ± 12

energy calculations show that the difference of free energy between the mutant and the WT is 90 kJ/mol for dimer, i.e. 45 kJ/mol per monomer in dimer. Despite the fact, that these values are unusually high, they show that the 14-3-3ζ WT is much more stable in its dimeric form than the L12E-M78K double mutant. Our predictions agree with the experimental assay, where at 1 μM concentration no trace of a monomeric state was found for the WT, while the monomeric state was dominant for the L12K/M78E mutant (Fig. 8B).

The exact role of salt bridges in stabilising of protein complexes has been a source of dispute for a long time [50]. The driving forces for salt bridge formation are primarily energetically favourable Coulombic charge-charge interactions between two ionized amino acids. In order to build a stable salt bridge, however, two additional factors play a key role: the orientation of interacting sidechains and the local context [51]. The sidechain orientation is crucial for keeping charged moieties in close vicinity, which is necessary for electrostatic interactions. This might, however, lead to an unfavourable loss in entropy. If the local context is rich in other possible interaction partners, the probability of salt bridge formation might be decreased as well. Therefore, due to the immobilization and desolvation of the interacting residues, creating a salt bridge in solvent may be unfavourable. The time series of the distance between Leu12 and Met78 throughout the TI simulation resembles this case (Fig. S5). In the beginning of the simulation, the residue distance varies between 0.2 and 1.0 nm. In this phase, these neutral residues interact via hydrophobic interactions. After introducing small charges to the perturbing residues, the residues start to create a stable salt bridge with a length of about 0.3 nm. In the last perturbation step, nevertheless, the distance increases again. This may be ascribed to the high charge of both residues and preferential interaction of Lys78 with solvent. A snapshot of the final state can be seen in Fig. 9A. The simulation of the double mutant L12E, M78K was prolonged up to 20 ns. In monomer 1, the salt bridge was formed for the first 3 ns, then lost for a host time and reformed for another 4 ns, before it was lost completely. In monomer 2, the salt bridge was immediately lost, only to be formed transiently again after 15 ns (Fig. 9B). This indicates that, while the salt bridge could be formed theoretically, the residues prefer to interact favourably with the solvent.

4. Conclusion

Two types of mutation of the human regulatory protein 14-3-3ζ were studied by free energy calculations and experimental stability measurements. Both computational and experimental approaches agreed that the common Ala and Ser replacements of Cys94 leads to considerable protein destabilization. The folding free energies were significantly reduced, as was confirmed by a lower melting temperature. These findings could be rationalized at a molecular level by the hydrophobic nature of the particular position and the formation of small cavities upon mutation to Ala.

Similarly, free energy calculations on a double mutant (L12K/M78E) that was found to show a significantly reduced tendency to dimerize were performed and found to be in agreement with the experimental findings. The putative salt bridge that was introduced in the double mutant was only observed for a limited amount of time, while the ionic species of the sidechains, in particular Lys78, are more favourably solvated individually in the individual monomers.

Overall, our work nicely demonstrates the strength of molecular simulations and free energy calculations as a complementary tool to experimental observation. The simulations offer explanations at a molecular level that are not easily accessible experimentally. This type of computational tools can be also used for the more efficient rational design of mutations having desired physico-chemical properties.

Transparency document

The Transparency document associated with this article can be found, in online version.

Acknowledgments

Financial support of the Austrian Science Fund (grant numbers W1224 and I 1999-N28) is gratefully acknowledged. J.H., Z.T. and V.W. acknowledge the financial support of the Czech Science Foundation (GF15-34684L). CIISB research infrastructure project LM2015043, funded by Ministry of Education, Youth and Sports of the Czech Republic (MEYS CR), is gratefully acknowledged for partial financial support of the measurements at the Proteomics and Biomolecular Interactions Core Facilities, CEITEC – Masaryk University.

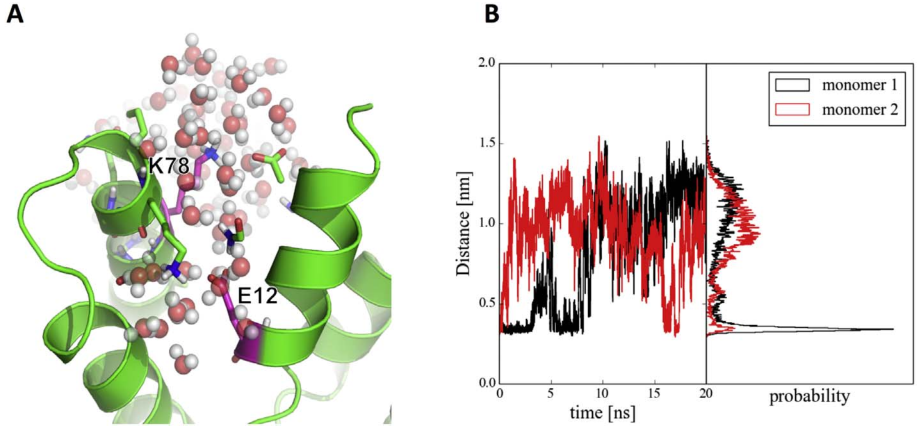


Fig. 9. A) A snapshot from the simulation at the last perturbation step (λ = 1.00), corresponding to the double mutant, K78 is solvated. B) Distance between NZ in K78 and CD in Glu12 from prolonged simulations at λ = 1.00.

References

[1] Aitken A, Collinge DB, van Heusden BPH, Isobe T, Roseboom PH, Rosenfeld G, Soll J. 14-3-3 proteins: a highly conserved, widespread family of eukaryotic proteins. Trends Biochem. Sci., 17: 498–501.

[2] W. Wang, D.C. Shakes, Molecular evolution of the 14-3-3 protein family, J. Mol. Evol. 43 (1996) 384–398.

[3] M. Rosenquist, P. Sehnke, R.J. Ferl, M. Sommarin, C. Larsson, Evolution of the 14-3-3 protein family: does the large number of isoforms in multicellular organisms reflect functional specificity? J. Mol. Evol. 51 (2000) 446–458.

[4] A. Aitken, 14-3-3 proteins: a historic overview, Semin. Cancer Biol. 16 (2006) 162–172.

[5] D.H. Jones, S. Ley, A. Aitken, Isoforms of 14-3-3 protein can form homo- and heterodimers in vivo and in vitro: implications for function as adapter proteins, FEBS Lett. 368 (1995) 55–58.

[6] G. Messaritou, S. Grammenoudi, E.M.C. Skoulakis, Dimerization is essential for 14-3-3 zeta stability and function in vivo, J. Biol. Chem. 285 (2010) 1692–1700.

[7] X.W. Yang, W.H. Lee, F. Sobott, E. Papagrigoriou, C.V. Robinson, J.G. Grossmann, M. Sundstrom, et al., Structural basis for protein-protein interactions in the 14-3-3 protein family, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 17237–17242.

[8] D. Liu, J. Bienkowska, C. Petosa, R.J. Collier, H. Fu, R. Liddington, Crystal structure of the zeta isoform of the 14-3-3 protein, Nature 376 (1995) 191–194.

[9] A.K. Gardino, S.J. Smerdon, M.B. Yaffe, Structural determinants of 14-3-3 binding specificities and regulation of subcellular localization of 14-3-3-ligand complexes: a comparison of the X-ray crystal structures of all human 14-3-3 isoforms, Semin. Cancer Biol. 16 (2006) 173–182.

[10] P.D. Burbelo, A. Hall, 14-3-3 proteins. Hot numbers in signal transduction, Curr. Biol. 5 (1995) 95–96.

[11] M.P. Rubio, K.M. Geraghty, B.H.C. Wong, N.T. Wood, D.G. Campbell, N. Morrice, C. Mackintosh, 14-3-3-affinity purification of over 200 human phosphoproteins reveals new links to regulation of cellular metabolism, proliferation and trafficking, Biochem. J. 379 (2004) 395–408.

[12] L.N. Johnson, D. Barford, The effects of phosphorylation on the structure and function of proteins, Annu. Rev. Biophys. Biomol. Struct. 22 (1993) 199–232.

[13] M.J. Hubbard, P. Cohen, On target with a new mechanism for the regulation of protein phosphorylation, Trends Biochem. Sci. 18 (1993) 172–177.

[14] T. Pawson, J.D. Scott, Signaling through scaffold, anchoring, and adaptor proteins, Science 278 (1997) 2075–2080.

[15] T. Pawson, M. Raina, P. Nash, Interaction domains: from simple binding events to complex cellular behavior, FEBS Lett. 513 (2002) 2–10.

[16] C. Mackintosh, Dynamic interactions between 14-3-3 proteins and phosphoproteins regulate diverse cellular processes, Biochem. J. 381 (2004) 329–342.

[17] K. Nagata, A. Puls, C. Futter, P. Aspenstrom, E. Schaefer, T. Nakata, N. Hirokawa, et al., The MAP kinase kinase kinase MLK2 co-localizes with activated JNK along microtubules and associates with kinesin superfamily motor KIF3, EMBO J. 17 (1998) 149–158.

[18] G. Tzivion, Y.H. Shen, J. Zhu, 14-3-3 proteins; bringing new definitions to scaffolding, Oncogene 20 (2001) 6331–6338.

[19] M.B. Yaffe, K. Rittinger, S. Volinia, P.R. Caron, A. Aitken, H. Leffers, S.J. Gamblin, et al., The structural basis for 14-3-3:phosphopeptide binding specificity, Cell 91 (1997) 961–971.

[20] G. Nagy, C. Oostenbrink, J. Hritz, Exploring the binding pathways of the 14-3-3zeta protein: structural and free-energy profiles revealed by Hamiltonian replica exchange molecular dynamics with distancefield distance restraints, PLoS One 12 (2017) e0180633.

[21] M.B. Yaffe, How do 14-3-3 proteins work?— gatekeeper phosphorylation and the molecular anvil hypothesis, FEBS Lett. 513 (2002) 53–57.

[22] C. Johnson, S. Crowther, M.J. Stafford, D.G. Campbell, R. Toth, C. MacKintosh, Bioinformatic and experimental survey of 14-3-3-binding sites, Biochem. J. 427 (2010) 69–78.

[23] J. Hritz, L. Byeon Ij, T. Krzysiak, A. Martinez, V. Sklenar, A.M. Gronenborn, Dissection of binding between a phosphorylated tyrosine hydroxylase peptide and 14-3-3ζ: a complex story elucidated by NMR, Biophys. J. (2014) 2185–2194.

[24] D.C. Rees, A.D. Robertson, Some thermodynamic implications for the thermostability of proteins, Protein Sci. 10 (2001) 1187–1194.

[25] W.J. Becktel, J.A. Schellman, Protein stability curves, Biopolymers 26 (1987) 1859–1877.

[26] A. Razvi, J.M. Scholtz, Lessons in stability from thermophilic proteins, Protein Soc. 15 (2006) 1569–1578.

[27] R. Cohen, Tom, J.G. Frey, B. Holström, IUPAC, Quantities, Units and Symbols in Physical Chemistry, 2nd ed., Blackwell Scientific Publications, Oxford, 1993.

[28] Y.Y. Shi, A.E. Mark, C.X. Wang, F. Huang, H.J. Berendsen, W.F. van Gunsteren, Can the stability of protein mutants be predicted by free energy calculations? Protein Eng. 6 (1993) 289–295.

[29] C.D. Christ, A.E. Mark, W.F. van Gunsteren, Feature article basic ingredients of free energy calculations: a review, J. Comput. Chem. 31 (2010) 1569–1582.

[30] Y. Sugita, A. Kitao, Dependence of protein stability on the structure of the denatured state: free energy calculations of I56V mutation in human lysozyme, Biophys. J. 75 (1998) 2178–2187.

[31] A.P. Eichenberger, W.F. van Gunsteren, S. Riniker, L. von Ziegler, N. Hansen, The key to predicting the stability of protein mutants lies in an accurate description and proper configurational sampling of the folded and denatured states, Biochim. Biophys. Acta Gen. Subj. 2015 (1850) 983–995.

[32] R. Bonet, I. Vakonakis, I.D. Campbell, Characterization of 14-3-3-zeta interactions with integrin tails, J. Mol. Biol. 425 (2013) 3060–3072.

[33] N. Schmid, C.D. Christ, M. Christen, A.P. Eichenberger, W.F. van Gunsteren, Architecture, implementation and parallelisation of the GROMOS software for biomolecular simulation, Comput. Phys. Commun. 183 (2012) 890–903.

[34] M.M. Reif, M. Winger, C. Oostenbrink, Testing of the GROMOS force-field parameter set 54A8: structural properties of electrolyte solutions, lipid bilayers, and proteins, J. Chem. Theory Comput. 9 (2013) 1247–1264.

[35] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, J. Hermans, Interaction models for water in relation to protein hydration, in: B. Pullman (Ed.), Intermolecular Forces: Proceedings of the Fourteenth Jerusalem Symposium on Quantum Chemistry and Biochemistry Held in Jerusalem, Israel, April 13–16, 1981, Springer Netherlands, Dordrecht, 1981, pp. 331–342.

[36] A. Amadei, G. Chillemi, M.A. Ceruso, A. Grottesi, A. Di Nola, Molecular dynamics simulations with constrained roto-translational motions: theoretical basis and statistical mechanical consistency, J. Chem. Phys. 112 (2000) 9–23.

[37] H.J.C. Berendsen, J.P.M. Postma, W.F. Vangunsteren, A. Dinola, J.R. Haak, Molecular dynamics with coupling to an external bath, J. Chem. Phys. 81 (1984) 3684–3690.

[38] J.P. Ryckaert, G. Ciccotti, H.J.C. Berendsen, Numerical integration of cartesian equations of motion of a system with constraints - molecular dynamics of N-alkanes, J. Comput. Phys. 23 (1977) 327–341.

[39] I.G. Tironi, R. Sperb, P.E. Smith, W.F. van Gunsteren, A generalized reaction field method for molecular dynamics simulations, J. Chem. Phys. 102 (1995) 5451–5459.

[40] T.N. Heinz, W.F. van Gunsteren, P.H. Hünenberger, Comparison of four methods to compute the dielectric permittivity of liquids from molecular dynamics simulations, J. Chem. Phys. 115 (2001) 1125–1136.

[41] T.C. Beutler, A.E. Mark, R.C. van Schaik, P.R. Gerber, W.F. van Gunsteren, Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations, Chem. Phys. Lett. 222 (1994) 529–539.

[42] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.

[43] G. Hu, H. Li, J.Y. Liu, J. Wang, Insight into conformational change for 14-3-3sigma protein by molecular dynamics simulation, Int. J. Mol. Sci. 15 (2014) 2794–2810.

[44] K.A. Dill, Dominant forces in protein folding, Biochemistry 29 (1990) 7133–7155.

[45] N. Nagano, M. Ota, K. Nishikawa, Strong hydrophobic nature of cysteine residues in proteins, FEBS Lett. 458 (1999) 69–71.

[46] A.E. Eriksson, W.A. Baase, X.J. Zhang, D.W. Heinz, M. Blaber, E.P. Baldwin, B.W. Matthews, Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect, Science 255 (1992) 178–183.

[47] L. Schrödinger, The PyMOL Molecular Graphics System. Version ~ 1, 8 ed., (2015).

[48] N.N. Sluchanko, N.B. Gusev, Oligomeric structure of 14-3-3 protein: what do we know about monomers? FEBS Lett. 586 (2012) 4249–4256.

[49] M.M. Reif, C. Oostenbrink, Net charge changes in the calculation of relative ligand-binding free energies via classical atomistic molecular dynamics simulation, J. Comput. Chem. 35 (2014) 227–243.

[50] P. Strop, S.L. Mayo, Contribution of surface salt bridges to protein stability, Biochemistry 39 (2000) 1251–1255.

[51] G.I. Makhatadze, V.V. Loladze, D.N. Ermolenko, X. Chen, S.T. Thomas, Contribution of surface salt bridges to protein stability: guidelines for protein engineering, J. Mol. Biol. 327 (2003) 1135–1148.

6 Structural and interaction properties of intrinsically disordered proteins (IDPs) determined by experimental NMR and computational studies

Paper 11: Hritz J.; Byeon I-J.; Krzysiak T.; Martinez A.; Sklenář V.; Gronenborn A.M. Dissection of binding between a phosphorylated tyrosine hydroxylase peptide and 14-3-3ζ: a complex story elucidated by NMR. *Biophys. J.* **2014**, 107, 2185-2194

Paper 12: Louša, P.; Nedozrálová, H.; Župa, E.; Nováček, J.; Hritz, J.*: Phosphorylation of the regulatory domain of human tyrosine hydroxylase 1 monitored using non-uniformly sampled NMR. *Biophysical Chemistry* **2017**, 223, 25-29

Paper 13: Zapletal, V.; Mládek, A.; Melková, K.; Louša, P.; Nomilner, E.; Jaseňáková, Z.; Kubáň, V.; Makovická, M.; Laníková, A.; Žídek L.; Hritz, J.* Choice of force field for proteins containing structured and intrinsically disordered regions. *Biophys. J.* **2020**, 118, 1621–1633

Paper 14: Pavlíková Přecechtělová, J.; Mládek, A.; Zapletal, V.; Hritz, J. Quantum Chemical Calculations of NMR Chemical Shifts in Phosphorylated Intrinsically Disordered Proteins, JCTC **2019**, 15, 5642-5658

Paper 15: Jansen, S.;Melková, K.; Trošanová, Z.; Hanáková, K.; Zachrdla, M.; Nováček, J.; Župa, E.; Zdráhal, Z.; Hritz, J.*; Žídek, L.*: Quantitative Mapping of MAP2c Phosphorylation and 14-3-3ζ Binding Sites Reveals Key Differences Between MAP2c and Tau. *J. Biol. Chem.* **2017**, 292, 6715-6727

6.1 Intrinsically disordered proteins (IDPs)

The structure-function paradigm is probably mentioned in every structural biology or biochemistry textbook. It is important to pay attention to the cases in which proteins seem to violate this quite general rule. Certain proteins do not adopt a unique 3D structure but they still play a well-defined physiological role.[33,34] Such proteins are usually referred to as intrinsically disordered proteins in the literature. Since the automated analysis of sequences predicted that one-third of the eukaryotic proteins would contain disordered regions longer than 30 amino acids, IDPs have recently attracted significant attention.[35,36] However, the biophysical characterization of IDPs is much more difficult than in the case of well-structured proteins. The lack of a unique structure represents a fundamental problem for crystallographic and cryoEM studies. Rather, computational simulations and nuclear magnetic resonance (NMR) play a crucial role in the atomic-resolution studies of IDPs.

However, even for these techniques, the characterization of IDPs is far from trivial: Computational approaches suffer from insufficient sampling, in a manner that is much more severe than for structured proteins, because of the necessity to sample a much wider conformational space. The resonance frequencies observed in NMR spectra are time- and ensemble-averaged. Because IDPs take up multiple rapidly inter-converting conformations, the averaging makes the distribution of measured frequencies more uniform and thus more difficult to distinguish. Peak overlap in spectra of large or highly repetitive IDPs is often so severe that it is impossible to assign frequencies to individual atoms using standard approaches. The experimental data need to be interpreted differently than results obtained for well-ordered proteins because the obtained values are averages of large ensembles of very diverse protein conformations. For example, the propensity of individual molecules to form transient secondary structures is estimated, instead of a determination of well-defined secondary structure elements in terms of atomic coordinates.

6.2 Structural ensembles of IDPs and IDRs verified by NMR data

MD generally can generate the canonical ensemble of any system if sufficient sampling is achieved. More details about the sampling aspect are discussed in the second scientific chapter. However, there is another important aspect – the quality of force-field (FF) parameters that has not been discussed yet because there is a wide range of biomolecular FFs that work quite well for globular proteins. However, recent publications indicate that many of them fail when simulating IDPs and new sets of FF parameter sets were designed for this class of proteins.[37] At the same time, these FF parameters have not been providing reliable data for globular proteins. This introduces several complications when trying to simulate proteins containing both globular parts as well as disordered regions.

This issue was addressed by the applicant for three different proteins: the δ subunit of RNA polymerase (δRNAP), the regulatory domain of human tyrosine hydroxylase 1 (RD-hTH1) and the selected region of microtubule-associated protein 2 c (MAP2c).[P13]

The structural characterization of the δ subunit- RNAP was complicated mainly by the physical properties of its C-terminal domain. This region of the sequence is extremely acidic and highly repetitive, which was demonstrated to be necessary for regulating the RNAP affinity towards nucleic acids. To proceed towards structural characterization of the native, full-length protein, we applied and further developed NMR techniques based on nonuniform sampling (NUS), which allowed us to overcome a severe overlap of peaks of the disordered C-terminal domain in the NMR spectra.

RD-hTH1 has a similar architecture, i.e. region 1-69 is unstructured and the rest (70-169) is structured. In the case of MAP2c, region 159-254 was chosen, containing transient helical elements. MD simulations in the microsecond range were generated for those proteins by applying a variety of popular biomolecular force fields. Primary NMR data were used to verify generated MD trajectories by using Amber99, Charmm22 and Charmm36 force-field parameters in combination with TIP3P and TIP4P-D explicit water models.[37,38] We noticed distinct differences in the predicted cross-relaxation steady-state NOE relaxation data, exemplified in Fig. 7. The dynamic behavior of the disordered region is very unrealistic in all simulations using TIP3P (original or modified) when compared to the TIP4P-D water model. We observed the same pattern also for other protein systems.[P13] This is an interesting observation, as the TIP4P-D water model was not parameterized using NMR relaxation data. Although simulations with TIP4P-D water model yield a quite realistic steady-state nuclear Overhauser effect (ssNOE) profile, there is still space for improvement (e.g. region 90-130 in Fig. 9). This motivated us to validate simulations of IDRs, in addition to chemical shifts (CSs),

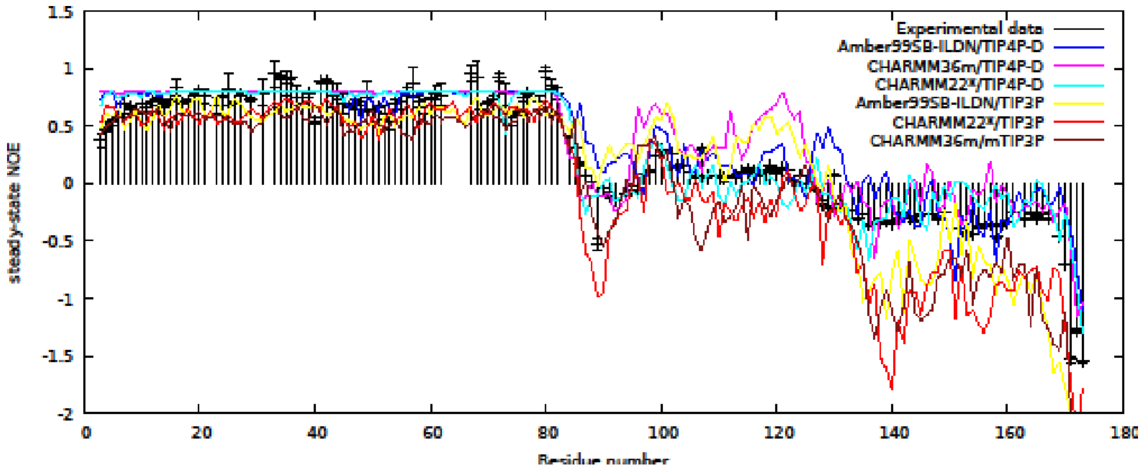


Figure 7: NMR relaxation data for δ subunit of RNA polymerase.

Comparison of predicted and measured dipolar cross-relaxation NMR rates (^{15}N-^{1}H steady-state NOE) for δ subunit of RNA polymerase from Bacillus subtilis. Experimental data (presented by black vertical lines) were collected on 600MHz Bruker Avance III NMR spectrometer. Other colors present computational data by using three different force-fields in combination with TIP3P and TIP4P-D water models.

residual dipolar coupling (RDCs), paramagnetic relaxation enhancement (PREs) and small angular X-ray scattering (SAXS) data and also check the dynamic performance with respect to the NMR relaxation data.^{P13} Fig. 9 shows that the majority of biomolecular FFs parameters result in unrealistic predictions. However, this data allows us to detect those that are the most reliable for the given class of IDPs.

From NMR experimental data - chemical shifts (CS) of proteins are mostly used for the verification of structural data. These procedures use predicted CS data for the given protein structures by using the empirical databases and the related prediction procedures. There are however cases, such as post-translational modifications of proteins (e.g. phosphorylation), when such procedures are currently not available because of an insufficient number of experimental NMR data. In such cases, one can use QM predictions of CS. Previously QM predictions of NMR CSs were shown for small organic molecules or selected parts of the globular proteins. The applicant presented the procedure by which such an approach can be applied even for the IDP region hTH1 in the explicit water solvent.[P14]

Unstructured domains are challenging for structural characterization and for free-energy calculations as they should be described by ensembles of structures rather than by a single structure. The group of prof. Zidek has recently developed nonuniform sampling NMR experiments (NUS) which are suitable to address intrinsically disordered proteins.[39,40] This methodology was successfully applied for the δ-subunit of RNA polymerase.[41] Different computational approaches will be tested to generate structural ensembles of the hTH1 unstructured region and the δ-subunit of RNA polymerase. Ensemble averaged data will be compared with experimentally measured NMR properties: chemical shifts, NOEs, paramagnetic relaxation enhancement and relaxation data.

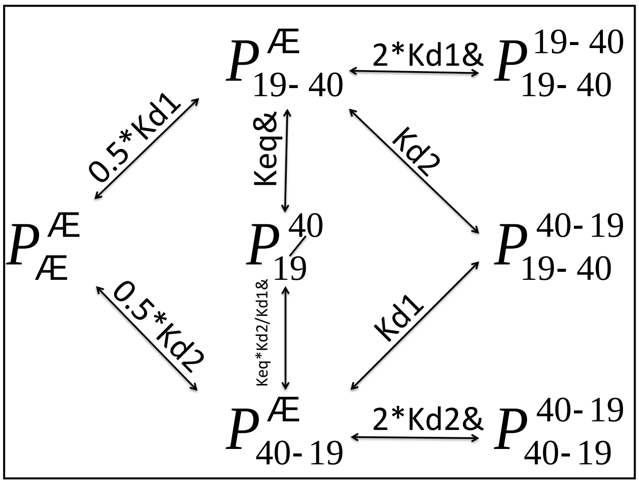


Figure 8: Binding scheme of dp_hTH1_50 phosphorylated on positions 19 and 40 with the 14-3-3ζ dimer, represented by a letter P.

Subscript and superscript next to P indicate which phosphorylation site (non, 19 or 40) is bound to the 14 3 3ζ. The whole scheme can be described by only three parameters Kd1, Kd2 and Keq when assuming independent binding of individual phosphoserines to different binding sites with 14 3 3ζ. This assumption does not hold for the cooperative binding and different values of equilibrium constants that have to be assigned to the individual equilibriums.

6.3 NMR investigation of IDP/IDRs binding properties

The NMR methodology is well known as one of the structural biology methods but it is often forgotten that it can also be used very well for following different kinds of changes within the protein. In the case of the studied IDPs/IDRs we applied NMR to follow the phosphorylation kinetic profiles of RD-hTH1[P12] and MAP2c.[P15] Subsequently NMR was applied for the determination of the interaction patches of those phosphorylated proteins towards their known binding partner 14-3-3ζ. In the case of multiple binding patches to the 14-3-3 proteins having two different binding sites determining the complete binding scheme is far from trivial (Fig. 8). The applicant used ^{31}P 1D NMR spectroscopy to determine a whole variety of existing binding complexes from their disordered fragments (region 1-50 of hTH1) containing single or double phosphorylation sites.[P11] We identified an interaction between singly S40-phosphorylated hTH1-50 and 14-3-3ζ that exhibits one order of magnitude weaker affinity than the singly S19-phosphorylated hTH1-50. Detailed analysis of the NMR data, giving information about individual phosphorylation sites, revealed that binding of 14-3-3ζ to the doubly phosphorylated peptide occurs in a complex fashion and three major distinct forms of the 14-3-3ζ complex with doubly phosphorylated hTH1-50 peptide were identified.[P11] **To the best of our knowledge, this is the first study to show the existence of more than the single complex form of 14-3-3 with bound phosphorylated peptide.**

Paper 11

Hritz J.; Byeon I-J.; Krzysiak T.; Martinez A.; Sklenář V.; Gronenborn A.M. Dissection of binding between a phosphorylated tyrosine hydroxylase peptide and 14-3-3ζ: a complex story elucidated by NMR. *Biophys. J.* **2014**, 107, 2185-2194

Article

Dissection of Binding between a Phosphorylated Tyrosine Hydroxylase Peptide and 14-3-3ζ: A Complex Story Elucidated by NMR

Jozef Hritz,[1,2] In-Ja L. Byeon,[1] Troy Krzysiak,[1] Aurora Martinez,[3] Vladimir Sklenar,[2] and Angela M. Gronenborn[1,*]
[1]Department of Structural Biology, University of Pittsburgh, School of Medicine, Pittsburgh, Pennsylvania; [2]Department of Structural Biology, Central European Institute of Technology, Masaryk University, Brno, Czech Republic; and [3]Department of Biomedicine, University of Bergen, Bergen, Norway

ABSTRACT Human tyrosine hydroxylase activity is regulated by phosphorylation of its N-terminus and by an interaction with the modulator 14-3-3 proteins. We investigated the binding of singly or doubly phosphorylated and thiophosphorylated peptides, comprising the first 50 amino acids of human tyrosine hydroxylase, isoform 1 (hTH1), that contain the critical interaction domain, to 14-3-3ζ, by ^{31}P NMR. Single phosphorylation at S19 generates a high affinity 14-3-3ζ binding epitope, whereas singly S40-phosphorylated peptide interacts with 14-3-3ζ one order-of-magnitude weaker than the S19-phosphorylated peptide. Analysis of the binding data revealed that the 14-3-3ζ dimer and the S19- and S40-doubly phosphorylated peptide interact in multiple ways, with three major complexes formed: 1), a single peptide bound to a 14-3-3ζ dimer via the S19 phosphate with the S40 phosphate occupying the other binding site; 2), a single peptide bound to a 14-3-3ζ dimer via the S19 phosphorous with the S40 free in solution; or 3), a 14-3-3ζ dimer with two peptides bound via the S19 phosphorous to each binding site. Our system and data provide information as to the possible mechanisms by which 14-3-3 can engage binding partners that possess two phosphorylation sites on flexible tails. Whether these will be realized in any particular interacting pair will naturally depend on the details of each system.

INTRODUCTION

Members of the 14-3-3 protein family are important modulators of several key signaling pathways that regulate critical biological activities, including progression through the cell cycle, growth, proliferation, and apoptosis (1–3). Extensive proteomics studies have revealed that the 14-3-3 family targets >200 distinct partners, ~0.6% of the entire cellular proteome (4,5).

In humans, seven different 14-3-3 isoforms (β, γ, ε, η, σ, τ, and ζ) are found, each encoded by a distinct gene, that function as homo- or heterodimers of ~30-kDa proteins (6,7). Crystal structures of all seven homodimeric mammalian proteins are available. The fold of the monomeric unit contains nine α-helices, arranged in an antiparallel fashion. Within the dimeric structure, two amphipathic grooves, one per monomer, constitute the binding pockets in which the conserved KRRY residues-Lys^{49}, Arg^{56}, Arg^{127}, and Tyr^{128} in 14-3-3ζ-are positioned for the recognition of a phosphorylated serine or threonine target (8). Three different phosphoserine (pS)- or phosphothreonine-containing sequence motifs can be recognized by 14-3-3 proteins: RSXp(S/T)XP (motif I), RXXXp(S/T)XP (motif II), and a carboxyl-terminal p(S/T)XCOOH motif (motif III), where X is a nonproline amino acid (9–11). Some 14-3-3 family members have also been reported to bind nonphosphorylated peptides as well as motifs that do not strictly display the above canonical sequences (6,11).

Tyrosine hydroxylase (TH) and tryptophan hydroxylase were the first proteins that were identified as binding partners for 14-3-3 proteins (12). TH catalyzes the tetrahydrobiopterin-dependent hydroxylation of L-tyrosine to L-3,4-dihydroxyphenylalanine, which is the rate-limiting step in the biosynthesis of dopamine and other catecholamines (13). In humans, several isoforms of TH are expressed, with isoform 1 (hTH1; 497 residues, NP-000351.2) being the best characterized (14–16). hTH1 can be phosphorylated at T8 and/or three serine residues (S19, S31, and S40) by several protein kinases (17) on its highly dynamic 50-residue N-terminal region that precedes the regulatory ACT domain, characteristic of the aromatic amino-acid hydroxylase family. 14-3-3ζ and other isoforms have been reported to bind to S19 phosphorylated TH (18–20). Although not strictly exhibiting the canonical binding motifs, recent crystallographic analysis of the complex between hTH1 S19-phosphorylated N-terminal peptide (1–43 residues) and 14-3-3γ revealed that the phosphate interacts with the KRRY motif in the 14-3-3γ binding grove, and the region surrounding S19 engages in similar interactions as seen for peptides containing canonical motifs I and II (21). In addition, phosphorylation of hTH1 on S40 stimulates TH activity (17), and induces binding to the yeast 14-3-3 proteins BMH1 and BMH2 and to sheep brain 14-3-3 proteins, but not to 14-3-3ζ (18,19). On the other hand, 14-3-3ζ has been reported

Submitted June 30, 2014, and accepted for publication August 28, 2014.

*Correspondence: amg100@pitt.edu

Editor: David Eliezer.

0006-3495/14/11/2185/10 $2.00

http://dx.doi.org/10.1016/j.bpj.2014.08.039

to interact with doubly S19- and S40-phosphorylated hTH1 (18). In a recent study (22) we have shown that the doubly phosphorylated hTH1 binds 14-3-3γ with similar stoichiometry and affinity as the singly phosphorylated enzyme on Ser[19]. However, 14-3-3γ interfered with the phosphorylation of TH1-pS19 on Ser[40], supporting that Ser[40] becomes inaccessible in the hTH1:14-3-3γ complex, although Ser[40] does not significantly contribute to the binding of 14-3-3γ. Therefore the question of whether S40 phosphorylated hTH1 can be bound by 14-3-3 proteins has not been settled unequivocally.

The 14-3-3 proteins are thought to function by inducing a conformational change in the target proteins, thereby modulating the target proteins' activities. In this way, 14-3-3 family members impart additional control onto key cellular enzymes, in addition to common enzymatic control mechanisms such as feedback loops or product inhibition. The contemporary prevalent model for the binding of doubly phosphorylated peptides by 14-3-3 suggests that one phosphorylated residue of the target protein engages 14-3-3 first, linking the two proteins together, followed by binding of the second phosphorylated residue, to effect the conformational change (23). The notion that binding of 14-3-3 to two phosphorylation sites is required to induce the activity altering conformational change, has led some authors to consider 14-3-3 as a binary computer (4,24).

We explored this model of 14-3-3 modulation using ^{31}P NMR in a simplified peptide system. We investigated the binding of hTH1 peptides (region 1–50) with a single phosphorylation at S19 (pS19) or S40 (pS40) to 14-3-3ζ and determined individual binding affinities. We found that 14-3-3ζ can interact with the hTH1 peptide through either phosphorylated-S19 or -S40, although the binding affinity of 14-3-3ζ for the pS40 peptide is significantly lower than that for the pS19 peptide. Furthermore, binding of 14-3-3ζ to the doubly phosphorylated hTH1 peptide exhibits a complex binding profile involving three major complexes. Our results suggest that when considering 14-3-3 interactions with multiply phosphorylated targets in general, and hTH1 in particular, avidity considerations should be taken into account.

METHODS

Cloning, expression, and purification of hTH1 peptides and 14-3-3ζ

The coding region for residues 1–50 of human tyrosine hydroxylase (hTH1–50) (Fig. 1 *A*) was inserted into a modified pET32a vector, in frame with N-terminal thioredoxin and 6×His tags, followed by a tobacco etch virus (TEV) cleavage site. A codon-optimized gene for 14-3-3ζ (GenScript, Piscataway Township, NJ) was inserted into pET15b with Cys[25] and Cys[189] codons converted to Ala using the QuikChange Site-Directed Mutagenesis Kit (Stratagene, La Jolla, CA). The DNA sequence for all constructs was confirmed by sequencing (Genewiz, South Plainfield, NJ).

Proteins were produced in *Escherichia coli* BL21(DE3) cells that were grown at 37°C in LB media to an OD_{600} of 0.8 and induced with 0.5 mM IPTG for 18 h at 18°C. Cells were harvested, resuspended in 50 mM TBS (50 mM Tris, 150 mM NaCl, pH 8.0) buffer, and lysed with a microfluidizer (Microfluidics, Westwood, MA). The cell lysate was clarified by centrifugation (21,000*g* for 60 min at 4°C), and the supernatant was loaded onto a HisTrap column (GE Healthcare, Piscataway, NJ). Proteins were eluted using a linear imidazole gradient (0.05–1 M in a total volume of 15 mL).

Fractions containing tagged hTH1 peptide were treated with TEV (1 mg/20 mg of peptide) at 4°C overnight, and applied to a Vydac reverse-phase C4 column (Grace Davison Discovery Sciences, Deerfield, IL). hTH1–50 was eluted from the reverse phase column using a linear gradient of 10–40% acetonitrile in 50 mL. Peptide-containing samples were further purified by gel filtration through a Superdex75 26/60 column (GE Healthcare) equilibrated in 20 mM Tris buffer, pH 7.5. The Trp-containing three amino-acid cloning artifact at the N-terminus of the hTH1–50 peptide (Fig. 1 *A*) allowed for detection during purification and concentration determination at 280 nm.

Fractions containing 14-3-3ζ were dialyzed into TBS buffer and treated with TEV overnight. Undigested protein, 6×His tag cleavage products, and TEV protease were removed by passing the sample over the HisTrap column for a second time and the flow-through containing untagged 14-3-3ζ was collected. The sample was further purified over anion exchange (HiTrap Q HP; GE Healthcare) followed by gel filtration on a Superdex75 26/60 column (GE Healthcare), equilibrated in 20 mM sodium phosphate buffer, pH 7.0.

The Cys[25]Ala/Cys[189]Ala 14-3-3ζ variant exhibited essentially the same stability as wild-type 14-3-3ζ protein, as evidenced by identical differential scanning calorimetry unfolding profiles.

Phosphorylation of hTH1–50

The serine residue at position 40 of the hTH1–50 peptide was phosphorylated using the catalytic subunit of cAMP-dependent protein kinase A (PKA; New England BioLabs, Ipswich, MA), whereas Ser[19] was phosphorylated using p38 regulated/activated protein kinase (PRAK; obtained from the protein phosphorylation unit, University of Dundee, Dundee, Scotland). For phosphorylation at S19, the reaction mixture, containing hTH1–50 S40A peptide in 50 mM Tris buffer, pH 7.5, 1 mM ATP, 10 mM $MgCl_2$, and PRAK kinase (50 μg for 10 mg of hTH1–50 S40A), was incubated for three days at 30°C. For phosphorylation at S40, hTH1–50 was incubated with PKA (0.5 μg for 15 mg hTH1–50) in 50 mM Tris buffer, pH 7.5,10 mM $MgCl_2$, 200 μM ATP at 30°C for 2 h. Reaction products were purified by gel filtration using a Superdex30 16/60 column (GE Healthcare) followed by anion exchange (HiTrap Q HP; GE Healthcare).

Doubly phosphorylated peptide was prepared using S40 phosphorylated hTH1–50 peptide (pS40-hTH1–50) and PRAK kinase, as described above. Approximately 70% of pS40-hTH1–50 was converted to the doubly phosphorylated product (pS19pS40-hTH1–50), which was purified as described above. For thiophosphorylation, adenosine 5′-[γ-thio]triphosphate (ATPγS) was added to the PKA reaction mixture, instead of ATP. In all cases, the correct masses and purity of the final products were verified by matrix-assisted laser desorption/ionization-time of flight mass spectrometry.

^{31}P NMR titrations

All ^{31}P NMR spectra were recorded at 37°C on an AVANCE 600 MHz NMR spectrometer equipped with a broadband frequency probe (Bruker, Billerica, MA). Samples containing 14-3-3ζ and singly or doubly phosphorylated hTH1–50 peptides were prepared in 20 mM sodium phosphate, pH 7.0, at a constant peptide concentration (0.83 or 0.71 mM) with molar ratios of 14-3-3ζ to peptide of 0:1–8:1.

Modeling of binding scenarios

Binding schemes for the different complexes between 14-3-3ζ and singly or doubly phosphorylated peptides are presented in Fig. 1 *C* and Fig. 2 *B*,

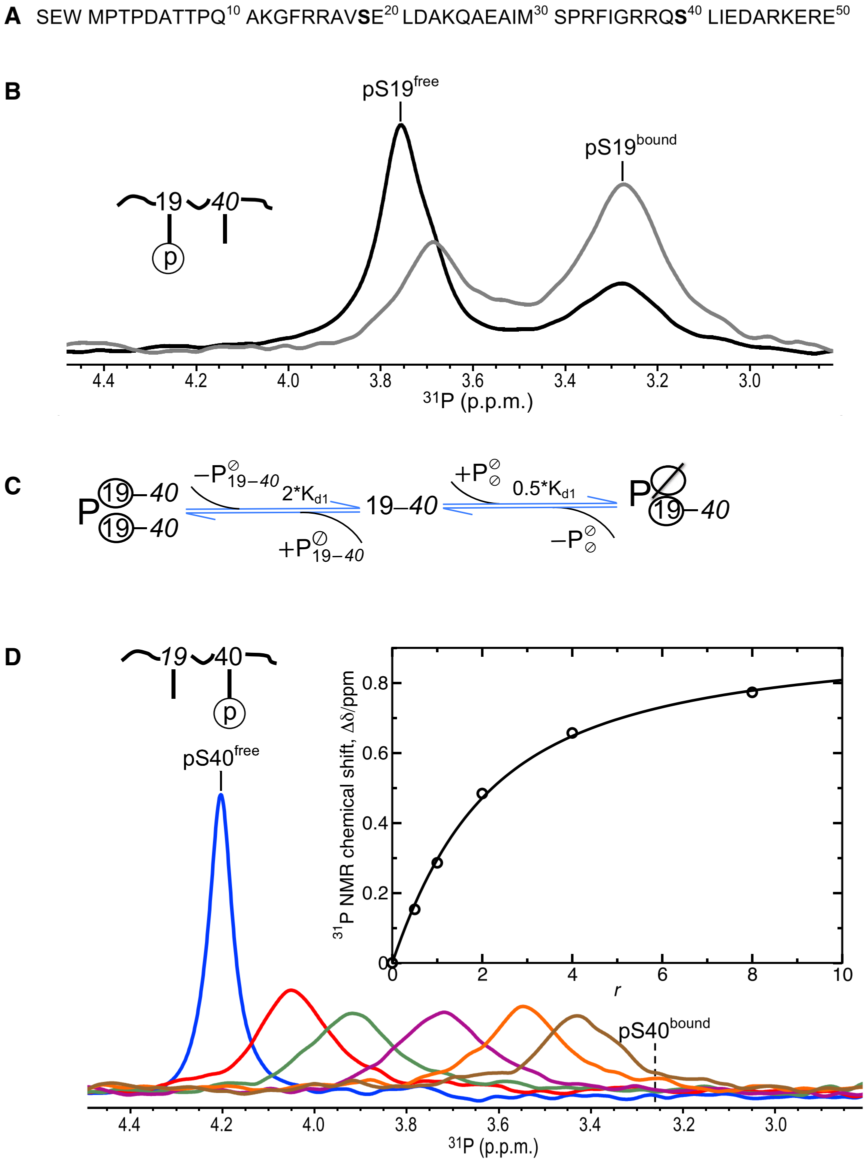


FIGURE 1 Binding of singly phosphorylated hTH1–50 peptides to 14-3-3ζ. (*A*) Amino-acid sequence of the hTH1–50 peptide with S19 and S40 highlighted (*boldface*). (*B*) ^{31}P spectra of pS19-hTH1–50 at 0.83 mM peptide concentration and molar ratios of 14-3-3ζ to peptide (*r*) 0.5:1 (*black*) and 1:1 (*gray*). (*C*) Schematic depiction of pS19-hTH1–50 (denoted by 19–*40*) binding to the 14-3-3ζ homodimer (free protein dimer is depicted by $P^{\varnothing}_{\varnothing}$, where Ø indicates an empty phosphoserine binding site). (*Circled*) Bound pS19 residues. (*D*) ^{31}P spectra of pS40-hTH1–50 at 0.71 mM peptide concentration and molar ratios (*r*) of 14-3-3ζ to peptide of 0:1 (*blue*), 0.5:1 (*red*), 1:1 (*dark green*), 2:1 (*purple*), 4:1 (*orange*), and 8:1 (*brown*). (*Inset*) Chemical shift changes upon 14-3-3ζ binding (Δ*δ*) are plotted as a function of molar ratio (*r*). The curve obtained by fitting Δ*δ* to the quadratic equation shown in Eq. 2 is shown. To see this figure in color, go online.

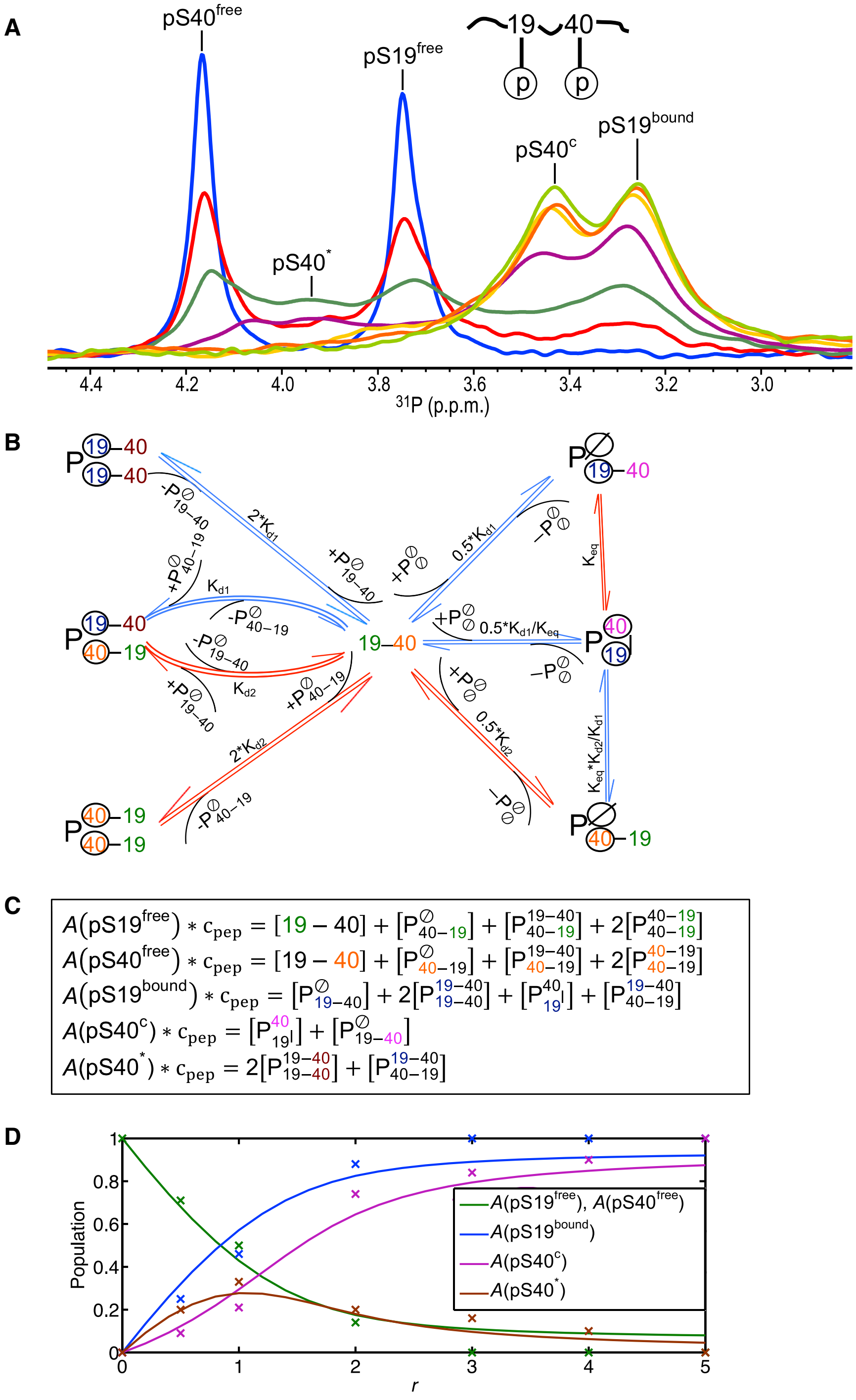
respectively. Equations for individual binding equilibria and equations describing mass conservation are provided in Fig. S1 and Fig. S2 in the Supporting Material. The equations were solved numerically using the software MATLAB (The MathWorks, Natick, MA) and results yield populations of the individual species as a function of the 14-3-3ζ to peptide molar ratios, *r*. Three different sets of solutions were obtained, only one of which was physically meaningful given that populations cannot be negative for any *r* value ≥0.

RESULTS

Binding of singly phosphorylated hTH1 peptides to 14-3-3ζ

Binding between 14-3-3ζ and the phosphorylated N-terminus of hTH1 was probed by ^{31}P NMR. hTH1 peptides, containing the first 50 residues (hTH1–50), were prepared and each serine in the peptide was phosphorylated individually, generating pS19-hTH1–50 and pS40-hTH1–50. PKA selectively and specifically phosphorylated only Ser^{40}, but PRAK phosphorylated both Ser^{19} and Ser^{40} (J. Hritz and A.M. Gronenborn, unpublished results). Therefore, we introduced the S40A change, permitting the preparation of singly phosphorylated pS19-hTH1–50 peptide. Titration of 14-3-3ζ into a solution containing pS19-hTH1–50 revealed the system to be in slow exchange on the NMR chemical shift timescale. Thus, resonances for both free and the 14-3-3ζ-bound peptide can be observed separately (Fig. 1 *B*), and the relative fractions of the free and bound peptide can be determined from resonance intensities (25,26). A dissociation constant (K_{d1}) of 0.15 ± 0.03 mM for the 14-3-3ζ/pS19-hTH1–50 complex was derived, using

$$K_{d1} = \frac{[\mathrm{pS19^{free}}] * [14-3-3\zeta^{\mathrm{free}}]}{[\mathrm{pS19^{bound}}]} = \frac{c_{\mathrm{pep}} * A(\mathrm{pS19^{free}}) * \{r - A(\mathrm{pS19^{bound}})\}}{A(\mathrm{pS19^{bound}})}, \qquad (1)$$



$$A(\mathrm{pS19}^{\mathrm{free}}) * c_{\mathrm{pep}} = [19-40] + [P^{\varnothing}_{40-19}] + [P^{19-40}_{40-19}] + 2[P^{40-19}_{40-19}]$$
$$A(\mathrm{pS40}^{\mathrm{free}}) * c_{\mathrm{pep}} = [19-40] + [P^{\varnothing}_{40-19}] + [P^{19-40}_{40-19}] + 2[P^{40-19}_{40-19}]$$
$$A(\mathrm{pS19}^{\mathrm{bound}}) * c_{\mathrm{pep}} = [P^{\varnothing}_{19-40}] + 2[P^{19-40}_{19-40}] + [P^{40}_{19|}] + [P^{19-40}_{40-19}]$$
$$A(\mathrm{pS40}^{\mathrm{c}}) * c_{\mathrm{pep}} = [P^{40}_{19|}] + [P^{\varnothing}_{19-40}]$$
$$A(\mathrm{pS40}^{*}) * c_{\mathrm{pep}} = 2[P^{19-40}_{19-40}] + [P^{19-40}_{40-19}]$$

FIGURE 2 Binding of doubly phosphorylated peptide to 14-3-3ζ. (*A*) ^{31}P spectra of pS19pS40-hTH1–50 at 0.83 mM and the molar ratio of 14-3-3ζ to the peptide (*r*) was 0:1 (*blue*), 0.5:1 (*red*), 1:1 (*dark green*), 2:1 (*purple*), 3:1 (*yellow*), 4:1 (*orange*), or 5:1 (*light green*). (*B*) Binding scheme of pS19pS40-hTH1–50 (denoted by 19–40) with respect to the 14-3-3ζ homodimer ($P^{\varnothing}_{\varnothing}$). Dissociation constants K_{d1} and K_{d2} correspond to the values determined for the singly phosphorylated peptides. (*Blue arrows*) Equilibria for which slow exchange on the ^{31}P chemical shift scale is observed; (*red arrows*) equilibria for which fast exchange on the ^{31}P chemical shift scale is observed. (*Circled*) Bound phosphorous groups. Phosphorous groups contributing intensity to the same peak in the NMR spectra are shown in the same color. (*C*) Equations governing the relative peak intensities (denoted as "*A*"), based on the scheme depicted in panel *B*; each phosphopeptide contributes to one or more of the five peaks. The specific phosphoserine that contributes to intensity of a particular peak is colored in the same color as in panel *B*. (*D*) Populations of phosphoserine species that contribute to the experimentally observed peak intensities (*x*) for different 14-3-3ζ to peptide molar ratios, *r*. (*Continuous lines*) For comparison, the theoretically predicted populations, based on the binding model in panel *B* using $K_{d1} = 0.15$ mM, $K_{d2} = 1.0$ mM, and $K_{eq} = 1.5$, are plotted. To see this figure in color, go online.

where c_{pep}, $A(\mathrm{pS19}^{\mathrm{free}})$, and $A(\mathrm{pS19}^{\mathrm{bound}})$ comprise the total peptide concentration, and the free and bound fractions of the pS19-hTH1–50 peptide, respectively, with *r* the molar ratio of 14-3-3ζ to peptide. For both singly phosphorylated peptides, dissociation constants are reported for the 14-3-3ζ monomer, assuming that binding to a single site in the 14-3-3ζ homodimer is not influenced by binding to the second site.

To analyze our binding data, we generated a scheme for the interaction between pS19-hTH1–50 and the 14-3-3ζ dimer (Fig. 1 *C*). A pS19-hTH1–50 peptide (represented by 19–40 in Fig. 1 *C*) can bind to the free 14-3-3ζ dimer (depicted by $P^{\varnothing}_{\varnothing}$ in Fig. 1 *C*) at two different sites, with the dissociation constant being one-half of the observed K_{d1}; the binding of a second pS19-hTH1–50 peptide to the 14-3-3ζ dimer in which one site is already occupied (depicted by $P_{19-40}^{\oslash}$ in Fig. 1 *C*) can only occur to the unoccupied site, and dissociation from a doubly occupied 14-3-3ζ dimer (depicted by P_{19-40}^{19-40} in Fig. 1 *C*) can occur from either site, resulting in $2{*}K_{d1}$. The equations describing these two equilibria, taking mass conservation of the peptide and 14-3-3ζ into account, are presented in Fig. S1 in the Supporting Material. Based on this model, in a 1:1 mixture of pS19-hTH1–50 and 14-3-3ζ (monomer concentration) with a K_{d1} of 0.15 mM, 34.4% of the peptide would be in the free state, 22.6% of peptide in the $P_{19-40}^{\oslash}$ bound form, and 43% in the P_{19-40}^{19-40} bound form.

In contrast to the result with the pS19 peptide, titration of 14-3-3ζ into a solution containing the pS40-hTH1–50 peptide revealed fast exchange on the NMR chemical shift scale, resulting in a single resonance at the weighted averaged chemical shifts of both the free and bound resonances (Fig. 1 *D*). By fitting the chemical shift change (Δδ) to Eq. 2, a K_{d2} value of 1.0 ± 0.1 mM and a maximum chemical shift difference between free and bound peptide ($\Delta\delta_{\max}$) of 0.94 p.p.m. were determined for the binding of 14-3-3ζ to pS40-hTH1–50:

$$\Delta\delta = \frac{\Delta\delta_{\max}}{2c_{\mathrm{pep}}}\left(K_d + c_{\mathrm{pep}} + rc_{\mathrm{pep}} - \sqrt{\left(K_d + c_{\mathrm{pep}} + rc_{\mathrm{pep}}\right)^2 - 4c_{\mathrm{pep}}^2 r}\right). \qquad (2)$$

These values allowed us to calculate the ^{31}P bound chemical shift, δ(pS40$^{\mathrm{bound}}$), which was 3.26 p.p.m. (Fig. 1 *D*).

Interestingly, the bound ^{31}P chemical shifts for both singly phosphorylated hTH1–50 peptides are essentially identical (3.26 p.p.m., Fig. 1, *B* and *D*), irrespective of the different neighboring amino acids around the phosphorylated Ser in the hTH1–50 peptide. This suggests that the binding of the phosphorylated residues to the identical binding pockets of 14-3-3ζ that are lined by residues Lys^{49}, Arg^{56}, Arg^{127}, and Tyr^{128} is the major determinant of the chemical shift, irrespective of the local peptide sequence. We speculate that the side chain of Tyr^{128} most likely causes the upfield shift of the pS19 and pS40 ^{31}P resonances.

Binding of doubly phosphorylated hTH1 peptide to 14-3-3ζ

Binding between doubly phosphorylated proteins and 14-3-3 has previously been characterized using various methods, including isothermal titration calorimetry and fluorescence polarization (19,27,28). Although these methods provide overall binding properties, such as K_d or stoichiometry, important details for a system involving a homodimer and two binding motifs on the target are lacking. For example, answers to the question of how individual phosphorylated amino acids interact with 14-3-3, and what kind of complexes are possible, are still incomplete. ^{31}P NMR is ideally suited to provide such information, since individual ^{31}P signals can be monitored throughout the course of titration experiments between various interacting partners. We therefore prepared doubly phosphorylated hTH1–50 peptide (pS19pS40-hTH1–50) and monitored 14-3-3ζ binding to this peptide by ^{31}P NMR. Our data reveal that a number of different bound species are present during the course of the titration (Fig. 2 *A*). The spectrum of free pS19pS40-hTH1–50 (*blue spectrum* in Fig. 2 *A*; $r = 0$) contains two well-separated resonances, at 3.76 p.p.m. and at 4.18 p.p.m. for the phosphorous atoms on pSer^{19} (pS19$^{\mathrm{free}}$) and pSer^{40} (pS40$^{\mathrm{free}}$), respectively. At the end of the titration (*light green trace* in Fig. 2 *A*; $r = 5$), the pS19pS40-hTH1–50 peptide is saturated with 14-3-3ζ, and resonances at 3.26 p.p.m. and 3.43 p.p.m. for the phosphorous atoms on pSer^{19} (pS19$^{\mathrm{bound}}$) and pSer^{40} (pS40$^{\mathrm{c}}$), respectively, are observed. The bound resonances are broader than those of the free peptide, reflecting the slower rotational correlation time of the peptide/14-3-3ζ complex (molecular mass ~62-67 kDa).

The chemical shift of 3.43 p.p.m. is different from the completely saturated position (3.26 p.p.m.) obtained with a peptide that contains only phosphorylated Ser^{40} (pS40-hTH1–50, Fig. 1 *D*), indicating that this bound resonance position contains a contribution from the free pS40 resonance (due to fast exchange on the chemical shift scale). This signal, therefore, is designated pS40$^{\mathrm{c}}$, instead of pS40$^{\mathrm{bound}}$. During the titration, the intensities of both the pS19$^{\mathrm{free}}$ and pS40$^{\mathrm{free}}$ resonances decrease to the same degree with increasing 14-3-3ζ concentration, but only the bound pS19 ^{31}P resonance at 3.26 p.p.m. increases in intensity at the same rate (Fig. 2 *A*, e.g., *red spectrum*). The bound pS40 ^{31}P resonance at 3.43 p.p.m. is only visible for excess 14-3-3ζ over peptide (e.g., *purple*, *yellow*, *orange*, and *light-green spectra*). These data suggest that for excess peptide and little 14-3-3ζ ($r < 1$), the phosphorous group on Ser^{19} (strong binder) is the one predominantly bound by 14-3-3ζ, and two pS19pS40-hTH1–50 peptides are likely bound by one 14-3-3ζ dimer. For conditions where 14-3-3ζ is in excess ($r > 1$), both phosphorous groups on a single peptide can be bound by one 14-3-3ζ dimer.

The complete binding scenario, considering six possible complexes for the interaction between doubly phosphorylated peptide (free peptide is depicted as 19–40) and 14-3-3ζ (free 14-3-3ζ dimer is depicted as $P_{\oslash}^{\oslash}$), is presented in Fig. 2 *B*. Based on this interaction scenario and taking into account that resonances can be in slow or fast exchange on the chemical shift scale (marked by *blue* and *red arrows*, respectively), five different resonances are expected to be observable in the ^{31}P NMR spectra (see below). Phosphorous groups that exhibit identical resonance frequencies, thus contributing intensity to the same peak, are labeled in the same colors in Fig. 2, *B* and *C*. The derived binding scheme is based on two assumptions:

1. The bound ^{31}P chemical shifts of pS19 and pS40 are identical when saturated with 14-3-3ζ, irrespective of whether the second binding site on the 14-3-3ζ dimer is occupied or not; and
2. The individual phosphorous resonances in pS19pS40-hTH1–50 exhibit the same NMR exchange regime as those for the singly phosphorylated peptides (Fig. 1, *B* and *D*, i.e., the pS19 resonance is in slow exchange and that of pS40 is in fast exchange on the ^{31}P NMR shift scale).

Note that the slow or fast exchange regimes apply not only to the interaction between free peptide (19–40) and the one site occupied 14-3-3ζ dimer complex ($P^{\oslash}_{19-40}$ or $P^{\oslash}_{40-19}$) but also to various additional 14-3-3ζ complexes (Fig. 2 *B*). For example, the pS40 phosphorous resonance of $P^{40|}_{19|}$, the complex in which both sites on the protein dimer engage the two phosphate groups on 19–40, is in fast exchange with that of $P^{\oslash}_{19-40}$, the complex in which one site on the dimer engages the pS19 phosphate group of the peptide, since very weak binding of the phosphate group of pS40 is involved. On the other hand, $P^{40|}_{19|}$ and $P^{\oslash}_{40-19}$ are in slow exchange, since the exchange involves tight binding of the phosphate group of pS19. In agreement with the above binding scheme, five distinct peaks are observed in the ^{31}P spectra during the titration (Fig. 2 *A*).

Multiple complexes contribute to the intensity of each phosphorous resonance. All peptides with the pS19 phosphate group free of 14-3-3ζ protein (19–40, $P^{\oslash}_{40-19}$, P^{40-19}_{19-40}, P^{40-19}_{40-19}) contribute to the peak labeled "pS19$^{\text{free}}$" (19 colored *green* in Fig. 2, *B* and *C*). Free peptide (19–40) as well as any peptides bound to 14-3-3ζ via the pS40 phosphate group ($P^{\oslash}_{40-19}$, P^{40-19}_{19-40}, P^{40-19}_{40-19}) (40 colored *orange* in Fig. 2, *B* and *C*) will contribute to the peak labeled "pS40$^{\text{free}}$" in Fig. 2 *A*. Given that the same peptide can contribute intensity to two different resonances (pS19$^{\text{free}}$ and pS40$^{\text{free}}$), the total intensity of the two resonances is expected to be similar, as is indeed observed in the ^{31}P spectra (Fig. 2 *A*).

Any complexes of 14-3-3ζ that contain peptide bound via the phosphate group on pS19 ($P^{\oslash}_{19-40}$, P^{19-40}_{19-40}, $P^{40|}_{19|}$, and P^{40-19}_{19-40}; 19 colored *blue* in Fig. 2, *B* and *C*) add intensity to the peak labeled "pS19$^{\text{bound}}$." The pS40$^{\text{c}}$ peak arises from the pS40 phosphate groups in the $P^{40|}_{19|}$ and $P^{\oslash}_{19-40}$ complexes (40 colored *purple* in Fig. 2, *B* and *C*), which are in fast exchange. The contribution from the free pS40 phosphate group in $P^{\oslash}_{19-40}$ causes the "pS40$^{\text{c}}$" resonance to slightly shift toward the free position (4.18 p.p.m.) and, therefore, it is observed at 3.43 p.p.m., slightly downfield from the bound position (3.26 p.p.m.) of the singly phosphorylated pS40 peptide (Fig. 1 *D*). In contrast, the position of the pS19$^{\text{bound}}$ peak is at 3.27 p.p.m., identical to the bound position of the singly phosphorylated pS19 peptide (Fig. 1 *B*). The reason for this is that the phosphate groups that contribute to the pS19$^{\text{bound}}$ peak are in slow exchange, thus no weighted chemical shift averaging comes into play.

Finally, the free pS40 phosphate groups in the P^{19-40}_{19-40} and P^{19-40}_{40-19} complexes (40 colored *brown* in Fig. 2, *B* and *C*), denoted pS40*, reside on the peptide that is bound to 14-3-3ζ via pS19, but they are not buried in the 14-3-3ζ binding site and are free in solution, surrounded by solvent. Note that pS40* is observed as a distinct peak at 3.93 p.p.m., in slow exchange with the pS40 phosphate resonance (4.18 p.p.m.) of the free peptide (Fig. 2 *A*), because the peptides in these complexes are bound to 14-3-3ζ via the pS19 phosphate group.

The relationships listed in Fig. 2 *C* imply that the total normalized area (denoted as "*A*") of all peaks in Fig. 2 *A* corresponds to 2, because the peptide contains two phosphate groups:

$$\begin{aligned}&\Big(A\big(pS19^{\text{free}}\big) + A\big(pS40^{\text{free}}\big) + A\big(pS19^{\text{bound}}\big) + A\big(pS40^{c}\big)\\&\quad + A\big(pS40^{*}\big)\Big) * c_{\text{pep}} = 2\Big([19-40] + \big[P^{\oslash}_{40-19}\big]\\&\quad + \big[P^{\oslash}_{19-40}\big] + \big[P^{40|}_{19|}\big] + 2\big[P^{19-40}_{19-40}\big] + 2\big[P^{19-40}_{40-19}\big]\\&\quad + 2\big[P^{40-19}_{40-19}\big]\Big) = 2 * c_{\text{pep}}.\end{aligned} \tag{3}$$

Assignment of phosphorous resonances by selective thiophosphorylation

The ^{31}P resonance assignment of the pS19 and pS40 in the pS19pS40-hTH1–50 peptide (Fig. 2 *A*) was initially obtained using the assignment of singly phosphorylated hTH1–50 peptides (Fig. 1, *B* and *D*). We confirmed the assignment using a hTH1–50 peptide that was phosphorylated on Ser^{19} and thiophosphorylated on Ser^{40} (pS19tpS40-hTH1–50). Thiophosphorylation causes a large downfield shift of the ^{31}P resonance from ~4.2 p.p.m. (*blue spectra* in Fig. 1 *D* and Fig. 2 *A*) to 43.95 p.p.m. (*blue spectra* in Fig. 3, *A* and *B*). Thus, in the spectrum of the pS19tpS40-hTH1–50 peptide (Fig. 3 *B*), both resonances are clearly separated.

A thiophosphate group is assumed to functionally mimic a phosphate group, given its close chemical structure. To confirm that no gross changes to 14-3-3ζ binding were introduced into the system by thiophosphorylation, we carried out titration experiments with tpS40-hTH1–50 and 14-3-3ζ under the same conditions that were used for pS40-hTH1–50. We determined a K_d value of 3.8 mM for tpS40-hTH1–50 (Fig. 3 *A*), compared to 1.0 mM for the pS40-hTH1–50 (Fig. 1 *D*). This demonstrates that binding is slightly reduced by the replacement of an oxygen atom with a sulfur atom. Interestingly, binding of 14-3-3ζ to tpS40-hTH1–50 causes a downfield ^{31}P chemical shift (Fig. 3 *A*), whereas binding to pS40-hTH1–50 induces an upfield shift (see above, and Fig. 1 *D*), illustrating that chemical shifts are extremely sensitive to any changes and hard to predict.

Whereas 14-3-3ζ binding to the pS19pS40-hTH1–50 peptide yields three distinct pS40 phosphorous signals that

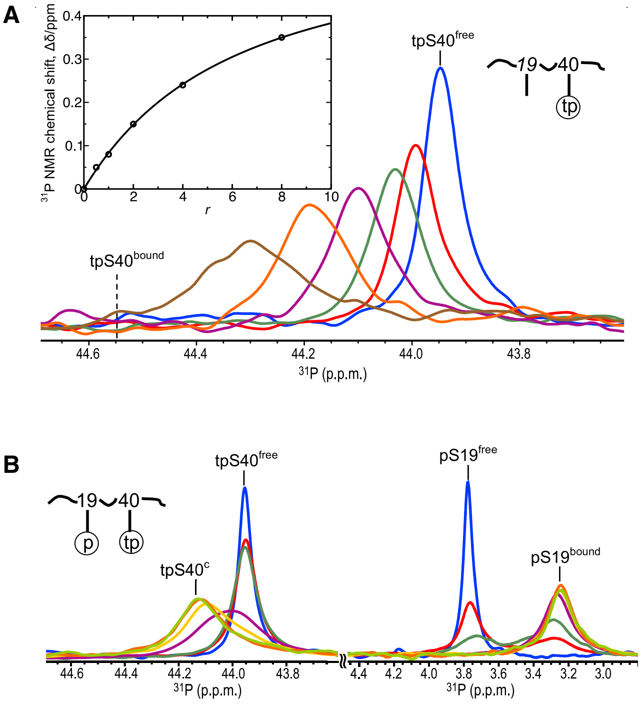


FIGURE 3 Binding of thiophosphorylated hTH1–50 peptides to 14-3-3ζ. (*A*) ^{31}P spectra of tpS40-hTH1–50 peptide (0.71 mM) for varying molar ratios of 14-3-3ζ to the peptide (*r*); 0:1 (*blue*), 0.5:1 (*red*), 1:1 (*dark green*), 2:1 (*purple*), 4:1 (*orange*), and 8:1 (*brown*). (*Inset*) The chemical shift changes upon 14-3-3ζ binding (Δδ) are plotted as a function of molar ratio (*r*). (*Continuous line*) Predicted Δδ from fitting the experimental Δδ to the quadratic equation shown in Eq. 2. (*B*) ^{31}P spectra of pS19tpS40-hTH1–50 (0.83 mM) for varying molar ratios of 14-3-3ζ to the peptide (*r*): 0:1 (*blue*), 0.5:1 (*red*), 1:1 (*dark green*), 2:1 (*purple*), 3:1 (*yellow*), 4:1 (*orange*), and 5:1 (*light green*). To see this figure in color, go online.

exhibit slow exchange, pS40^{free}, pS40*, and pS40^{c} (Fig. 2 *A*), binding to the pS19tpS40-hTH1–50 peptide results in only two signals that are in fast-intermediate exchange (Fig. 3 *B*), caused in part by the smaller chemical shift difference between free (pS40^{free}) and bound (pS40^{c}) resonances (0.60 p.p.m) compared to that for pS19pS40-hTH1–50 (0.94 p.p.m.). However, even with this difference, it seems reasonable to assume that the binding scheme in Fig. 2 *B* can be applied to pS19tpS40-hTH1–50, with $K_{d2} = 3.8$ mM instead of 1 mM and $K_{eq} = 0.26$ (see below) rather than 1.5. Using the well-separated pS19^{bound} ^{31}P resonance in the thiophosphorylated peptide (Fig. 3 *B*) proved beneficial for deconvoluting the overlapping pS19 resonances in the 14-3-3ζ-bound pS19pS40-hTH1–50 spectra (see Fig. S3).

Analysis of the binding equilibrium between doubly phosphorylated hTH1 peptide to the 14-3-3ζ homodimer

Fig. 2 *B* illustrates a comprehensive equilibrium binding scenario for the interaction between the doubly phosphorylated hTH1 peptide and the 14-3-3ζ homodimer. Six possible complexes are considered. Because insufficient experimental data for the accurate determination of all individual dissociation constants are available, it is assumed that each subunit monomer in the 14-3-3ζ homodimer binds equivalent phosphate groups on pS19pS40-hTH1–50 with identical affinity, irrespective of the state of the other 14-3-3ζ monomer, i.e., no cooperativity is present. Therefore, three independent equilibrium constants, K_{d1}, K_{d2}, and K_{eq} (Fig. 2 *B*), come into play. K_{d1} (0.15 ± 0.03 mM) and K_{d2} (1.0 ± 0.1 mM) are the experimentally determined dissociation constants for the interaction between singly phosphorylated hTH1–50 peptides, pS19-hTH1–50 and pS40-hTH1–50, respectively, and the 14-3-3ζ dimer (using the monomer at the concentration unit); K_{eq} is the equilibrium constant defined by the concentration ratio of the ($P^{40}_{19|}$) complex to the ($P^{\oslash}_{19-40}$) complex. This is treated as the only variable parameter in this simplistic equilibrium binding model.

We used three approaches for determining K_{eq}, as follows:

In the first approach, we derived a set of six equations for the individual, independent equilibria, along with two equations that are necessary to ensure mass conservation of the peptide and 14-3-3ζ (see Fig. S2 *A*). These equations were solved numerically in the software MATLAB. The resulting solutions provide the molar concentration of peptide complexes as a function of the 14-3-3ζ to doubly phosphorylated peptide ratio, *r* (see Fig. S2 *B*). The equations in Fig. 2 *C* link the theoretically predicted molar concentrations of individual complexes (see Fig. S2 *B*) with the intensities of five experimentally observable resonances. By fitting the theoretical profiles (*continuous lines* in Fig. 2 *D*) with the experimentally observed intensities (individual data points depicted by x in Fig. 2 *D*), for several 14-3-3ζ to peptide molar ratios, *r*, a K_{eq} value of 1.5 was determined.

In the second approach, the value of K_{eq} could, in principle, also be determined from the observed difference in the "pS40^{c}" resonance position from that of "pS40^{bound}," if the resonance position of pS40 in the $P^{\oslash}_{19-40}$ complex is known. Unfortunately, this is not known experimentally, but, assuming that it is identical to that of (pS40*), then the following equation holds:

$$\delta(pS40^{c}) = \frac{\left[P^{\oslash}_{19-40}\right]\delta(pS40^{*}) + \left[P^{40}_{19|}\right]\delta(pS40^{\mathrm{bound}})}{\left[P^{\oslash}_{19-40}\right] + \left[P^{40}_{19|}\right]}. \quad (4)$$

Considering that

$$K_{eq} = \frac{\left[P^{40}_{19|}\right]}{\left[P^{\oslash}_{19-40}\right]},$$

then

$$K_{eq} = \frac{\delta(pS40^{c}) - \delta(pS40^{*})}{\delta(pS40^{\mathrm{bound}}) - \delta(pS40^{c})}. \quad (5)$$

If the values of δ(pS40*), 3.93 p.p.m.; δ(pS40^{c}), 3.43 p.p.m.; and δ(pS40^{bound}), 3.26 p.p.m. are applied to Eq. 5, a K_{eq} value of 3.0 is found.

In the third approach, the value of K_{eq} can be estimated, based on K_{d2} and the local concentration of the pS40 phosphate group in the $P^{\oslash}_{19-40}$ complex. In an extended chain model of the doubly phosphorylated peptide, the distance between the Ser^{19} and Ser^{40} phosphate groups is estimated to be ~6 nm. The local concentration of the Ser^{40} phosphate that resides within a sphere of 6 nm radius, if the Ser^{19} phosphate from the same peptide is fixed to the first binding site on the 14-3-3ζ dimer, would become c_{loc} ~ 1/(6.022*10^{23}*4*10^{3}*(6.10^{-9})^{3})M, i.e., 1.9 mM. Thus, an estimated value of K_{eq} is

$$\left[P^{40}_{19}|\right]/\left[P^{\oslash}_{19-40}\right] = c_{\text{loc}}/K_{d2} = 1.9.$$

The latter two estimated K_{eq} values of 3.0 and 1.9 are very similar to the value determined from fitting (K_{eq} =1.5), considering that K_{eq} values of 3.0, 1.9, and 1.5 result in relative populations of 75, 66, and 60%, respectively, of the $P^{40}_{19}|$ complex.

Using the simplified binding model (Fig. 2 *B*), with $K_{d1} =$ 0.15 mM, $K_{d2} = 1.0$ mM, and $K_{eq} = 1.5$, populations of $P^{\oslash}_{40-19}$, P^{19-40}_{40-19}, and P^{40-19}_{40-19} complexes were found to be significantly lower than other complexes (<5.5%; see Fig. S2 *B*). For example, at a 1:1 molar ratio of doubly phosphorylated peptide to 14-3-3ζ ($r = 1$), 37.0% of the peptide is unbound; 17.7% of the peptide is in the $P^{40}_{19}|$ complex; 11.8% is in $P^{\oslash}_{19-40}$ complex; and 24.1% (2 × 12.05%) is in the P^{19-40}_{19-40} complex. Together, all other species account for 9.4% of the total population. At $r = 5$, the theoretical binding model predicts 52.5% of the peptide is in the $P^{40}_{19}|$ complex and 35% is in the $P^{\oslash}_{19-40}$ complex. Note that the ^{31}P resonances of the pS40 phosphate groups in these two complexes are in fast exchange (Fig. 2 *B*), resulting in a single pS40^{c} peak (Fig. 2 *A*).

DISCUSSION

In the prevailing model of 14-3-3 binding to doubly phosphorylated partners, the partners contain a high affinity, sometimes called a "gatekeeper" residue, which initially interacts with 14-3-3, and a secondary, weaker epitope, which only binds when the gatekeeper site is already bound to 14-3-3 (23). hTH1 that is singly phosphorylated on Ser^{40} is thought to be unable to bind 14-3-3ζ, whereas the pS19 phosphate is the high affinity epitope (18,19). Our NMR results with the N-terminal 50 residue peptide, pS40-hTH1–50, indicate, however, that binding of pS40 to 14-3-3ζ is indeed possible, as evidenced by the change from the free peptide phosphorous resonance to the bound one. This interaction is clearly weaker ($K_d = 1.0 \pm 0.1$ mM) than that between 14-3-3ζ and pS19-hTH1–50, containing the purported gatekeeper phosphate, which binds a factor-of-10 tighter ($K_d = 0.15 \pm 0.03$ mM). Therefore, even under conditions where 14-3-3ζ is present in excess over the pS19pS40-hTH1–50 peptide, two complexes, $P^{40}_{19}|$ and $P^{\oslash}_{19-40}$, form in a 60:40% ($K_{eq} = 1.5$) ratio, irrespective of the 14-3-3ζ concentration. Our results show that it is possible that multiple complexes for doubly phosphorylated binding partners of 14-3-3ζ may exist, even if in the phosphorylated full-length hTH1 additional mechanisms may come into play and alter the relative binding affinities (18).

Indeed, for the full-length pS40-hTH1, the binding to 14-3-3γ was almost undetectable by native mass spectroscopy and surface plasmon resonance, while for the doubly phosphorylated enzyme Ser^{40} is involved in binding to 14-3-3γ (22). In addition, binding of one 14-3-3γ dimer to one of the monomers in tetrameric hTH1—which is organized as a dimer of dimers—appears to interfere with the binding of an additional 14-3-3γ dimer to the adjacent monomer in hTH1, but not to the dimer at the opposite end, as seen in electron micrographs of the complex (22). Thus, in the complex involving the full-length enzyme, steric factors appear to be responsible for moving the equilibrium toward a scenario where the pS19 site binds to a single monomer in the 14-3-3 dimer. Whether this is the case for all 14-3-3 interactions with phosphorylated regulatory regions is unclear. In general, the presence of additional interactions and their contributions to the overall regulatory mechanisms have to be ascertained and confirmed for each individual system.

At this point, it may be of interest to analyze the underlying thermodynamics that govern the observed interaction between 14-3-3ζ and pS19pS40-hTH1–50 (or other similar peptides, denoted as 1–2, with two phosphorous sites), with the $P^{2}_{1}|$ bound form as the dominant complex when 14-3-3ζ is in excess and P^{1-2}_{1-2} dominant when peptide is in excess. For simplicity, we assume that the binding of site 1 (primary site) is independent of that of site 2 (secondary site) and that peptide binding to one binding site on the 14-3-3ζ dimer leads to a global entropy decrease ($\Delta S_G < 0$), whereas engagement of the second site on the same peptide only results in a local entropy decrease ($\Delta S_L < 0$). The enthalpy changes corresponding to the binding of the primary and secondary sites to 14-3-3 protein are denoted as ΔH_1 and ΔH_2. Given these assumptions, the condition where the amount of P^{1-2}_{1-2} is larger than $P^{2}_{1}|$ (when peptide is in excess over 14-3-3ζ; $r < 1$), i.e.,

$$\Delta G\left(P^{1-2}_{1-2}\right) - \Delta G\left(P^{2}_{1}|\right)<0,$$

results in

$$\Delta H_1 - \Delta H_2<T(\Delta S_G - \Delta S_L). \tag{6}$$

The opposite inequality

$$(\Delta H_1 - \Delta H_2>T(\Delta S_G - \Delta S_L))$$

is expected for the case when the amount of $P^{2}_{1}|$ is larger than P^{1-2}_{1-2} (when $r > 1$). In the case that the two phosphorous groups exhibit comparable binding affinities (e.g., such is the case for the PKCε-V3 peptide that contains tandem repeat sequences (27)),

$$\Delta G\left(P_{1}^{2|}\right) - \Delta G\left(P_{1-2}^{1-2}\right) = (\Delta H_2 - \Delta H_1) + T(\Delta S_G - \Delta S_L).$$

Because $\Delta H_1 - \Delta H_2 \sim 0$ and $T(\Delta S_G - \Delta S_L) < 0$, the free energy difference

$$\left(\Delta G\left(P_{1}^{2|}\right) - \Delta G\left(P_{1-2}^{1-2}\right)\right)$$

becomes negative, making the amount of $P_{1}^{2|}$ complex larger than the amount of P_{1-2}^{1-2} complex, irrespective of the r value.

Considering the large number of 14-3-3 binding partners and the above-described different binding modes, both cases are thermodynamically possible and may contribute to the regulation imparted by 14-3-3 on its targets.

CONCLUSIONS

A detailed analysis of the interaction between singly or doubly phosphorylated hTH1–50 peptides and 14-3-3ζ is presented. Using phosphorous NMR, dissociation constants for the binding of singly phosphorylated peptides, pS19-hTH1–50 and pS40-hTH1–50, to 14-3-3ζ were determined. They differ by a factor of 10, with the Ser^{19} phosphorylated peptide exhibiting a higher affinity ($K_d = 0.15 \pm 0.03$ vs. 1.0 ± 0.1 mM). Furthermore, for the doubly phosphorylated peptide, pS19pS40-hTH1–50, a mixture of distinct complexes was observed, which permitted us to derive a comprehensive equilibrium binding scheme. All six possible different complexes are taken into account, and the system is described by three equilibrium constants: $K_{d1} = 0.15$ mM, $K_{d2} = 1.0$ mM, and $K_{eq} = 1.5$. For excess pS19pS40-hTH1–50 peptide, the major complex is the 14-3-3ζ dimer with two peptides bound via the phosphate group on Ser^{19} (P_{19-40}^{19-40}), whereas, for excess 14-3-3ζ, two major complexes are seen, each one with a doubly phosphorylated peptide bound to one 14-3-3ζ dimer, with the Ser^{19} phosphate group bound to one 14-3-3ζ subunit, and the Ser^{40} phosphate group either bound to the second subunit $\left(P_{19}^{40|}\right)$ or free in solution $\left(P_{19-40}^{\oslash}\right)$ in a 60:40 ratio.

For other systems, the underlying thermodynamics will contribute to whether a doubly phosphorylated partner will bind with a 1:1 stoichiometry to the 14-3-3ζ dimer or if a mixture of complexes is formed. It is likely that most 14-3-3ζ binding partners will possess phosphorylated residues that exhibit different affinities for the binding sites on 14-3-3ζ. The presented methodology derived by monitoring individual ^{31}P resonances has general applicability for any complex system involving 14-3-3 and doubly phosphorylated ligands, and it will be interesting to extend this approach to full-length hTH1 or other target proteins.

SUPPORTING MATERIAL

Three figures are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(14)01012-1.

AUTHOR CONTRIBUTIONS

J.H., A.M., and A.M.G. conceived this study; J.H. and I.-J.L.B. acquired experimental data; J.H., T.K., I.-J.L.B., and A.M.G. performed data analysis and interpretation; J.H., I.-J.L.B., and A.M.G. drafted the article; all authors contributed to the final article.

We thank Jinwoo Ahn and Leonardus Koharudin for advice on protein purification, Doug Bevan and Phil Greer for computer technical support, and Michael J. Delk for NMR instrumental support. We also thank Teresa Brosenitsch for critical reading of the manuscript. We are grateful to Dr. Damodaran Krishnan, Department of Chemistry, University of Pittsburgh, for access to the Bruker 600 MHz NMR spectrometer with a BBF probe.

J.H. acknowledges financial support by an International Outgoing Fellowship of the European Community program "Support for Training and Career Development of Researchers (Marie Curie)", under contract No. PIOF-GA-2009-235902. This work was supported in part by startup funds from the University of Pittsburgh School of Medicine.

REFERENCES

1. Morrison, D. K. 2009. The 14-3-3 proteins: integrators of diverse signaling cues that impact cell fate and cancer development. *Trends Cell Biol.* 19:16–23.
2. Gardino, A. K., and M. B. Yaffe. 2011. 14-3-3 proteins as signaling integration points for cell cycle control and apoptosis. *Semin. Cell Dev. Biol.* 22:688–695.
3. Freeman, A. K., and D. K. Morrison. 2011. 14-3-3 proteins: diverse functions in cell proliferation and cancer progression. *Semin. Cell Dev. Biol.* 22:681–687.
4. Johnson, C., S. Crowther, ..., C. MacKintosh. 2010. Bioinformatic and experimental survey of 14-3-3-binding sites. *Biochem. J.* 427:69–78.
5. Jin, J., F. D. Smith, ..., T. Pawson. 2004. Proteomic, functional, and domain-based analysis of in vivo 14-3-3 binding proteins involved in cytoskeletal regulation and cellular organization. *Curr. Biol.* 14:1436–1450.
6. Obsil, T., and V. Obsilova. 2011. Structural basis of 14-3-3 protein functions. *Semin. Cell Dev. Biol.* 22:663–672.
7. Yang, X., W. H. Lee, ..., J. M. Elkins. 2006. Structural basis for protein-protein interactions in the 14-3-3 protein family. *Proc. Natl. Acad. Sci. USA.* 103:17237–17242.
8. Gardino, A. K. 2006. Structural determinants of 14-3-3 binding specificities and regulation of subcellular localization of 14-3-3-ligand complexes: a comparison of the x-ray crystal structures of all human 14-3-3 isoforms. *Semin. Cancer Biol.* 16:173–182.
9. Yaffe, M. B., K. Rittinger, ..., L. C. Cantley. 1997. The structural basis for 14-3-3: phosphopeptide binding specificity. *Cell.* 91:961–971.
10. Ganguly, S., J. L. Weller, ..., D. C. Klein. 2005. Melatonin synthesis: 14-3-3-dependent activation and inhibition of arylalkylamine *n*-acetyltransferase mediated by phosphoserine-205. *Proc. Natl. Acad. Sci. USA.* 102:1222–1227.
11. Aitken, A. 2006. 14-3-3 proteins: a historical overview. *Semin. Cancer Biol.* 16:162–172.
12. Ichimura, T., T. Isobe, ..., H. Fujisawa. 1987. Brain 14-3-3 protein is an activator protein that activates tryptophan 5-monooxygenase and tyrosine 3-monooxygenase in the presence of Ca^{2+} calmodulin-dependent protein kinase II. *FEBS Lett.* 219:79–82.

13. Nagatsu, T., M. Levitt, and S. Udenfriend. 1964. Tyrosine hydroxylase: the initial step in norepinephrine biosynthesis. *J. Biol. Chem.* 239:2910–2917.

14. Grima, B., A. Lamouroux, …, J. Mallet. 1987. A single human gene encoding multiple tyrosine hydroxylases with different predicted functional characteristics. *Nature.* 326:707–711.

15. Kaneda, N., K. Kobayashi, …, T. Nagatsu. 1987. Isolation of a novel cDNA clone for human tyrosine hydroxylase: alternative RNA splicing produces four kinds of mRNA from a single gene. *Biochem. Biophys. Res. Commun.* 146:971–975.

16. Fitzpatrick, P. F. 1999. Tetrahydropterin-dependent amino acid hydroxylases. *Annu. Rev. Biochem.* 68:355–381.

17. Daubner, S. C., T. Le, and S. Wang. 2011. Tyrosine hydroxylase and regulation of dopamine synthesis. *Arch. Biochem. Biophys.* 508:1–12.

18. Kleppe, R., K. Toska, and J. Haavik. 2001. Interaction of phosphorylated tyrosine hydroxylase with 14-3-3 proteins: evidence for a phosphoserine 40-dependent association. *J. Neurochem.* 77:1097–1107.

19. Obsilova, V., E. Nedbalkova, …, T. Obsil. 2008. The 14-3-3 protein affects the conformation of the regulatory domain of human tyrosine hydroxylase. *Biochemistry.* 47:1768–1777.

20. Itagaki, C., T. Isobe, …, T. Ichimura. 1999. Stimulus-coupled interaction of tyrosine hydroxylase with 14-3-3 proteins. *Biochemistry.* 38:15673–15680.

21. Skjevik, A. A., M. Mileni, …, A. Martinez. 2013. The N-terminal sequence of tyrosine hydroxylase is a conformationally versatile motif that binds 14-3-3 proteins and membranes. *J. Mol. Biol.* 426:150–168.

22. Kleppe, R., S. Rosati, …, A. Martinez. 2014. Phosphorylation dependence and stoichiometry of the complex formed by tyrosine hydroxylase and 14-3-3γ. *Mol. Cell. Proteomics.* 13:2017–2030.

23. Yaffe, M. B. 2002. How do 14-3-3 proteins work? Gatekeeper phosphorylation and the molecular anvil hypothesis. *FEBS Lett.* 513:53–57.

24. Kleppe, R., S. Ghorbani, …, J. Haavik. 2014. Modeling cellular signal communication mediated by phosphorylation dependent interaction with 14-3-3 proteins. *FEBS Lett.* 588:92–98.

25. Lian, L. Y., and G. C. K. Roberts. 1993. Effects of chemical exchange on NMR spectra. *In* NMR of Macromolecules. A Practical Approach. R. C. K. Robert, editor. IRL Press, Oxford, United Kingdom, pp. 153–182.

26. Bain, A. D. 2003. Chemical exchange in NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* 43:63–103.

27. Kostelecky, B., A. T. Saurin, …, N. Q. McDonald. 2009. Recognition of an intra-chain tandem 14-3-3 binding site within PKCε. *EMBO Rep.* 10:983–989.

28. Molzan, M., and C. Ottmann. 2012. Synergistic binding of the phosphorylated S233-and S259-binding sites of C-RAF to one 14-3-3 ζ-dimer. *J. Mol. Biol.* 423:486–495.

Paper 12

Louša, P.; Nedozrálová, H.; Župa, E.; Nováček, J.; Hritz, J.*: Phosphorylation of the regulatory domain of human tyrosine hydroxylase 1 monitored using non-uniformly sampled NMR. *Biophysical Chemistry* **2017**, 223, 25-29

Biophysical Chemistry 223 (2017) 25–29

Contents lists available at ScienceDirect

Biophysical Chemistry

journal homepage: http://www.elsevier.com/locate/biophyschem

Phosphorylation of the regulatory domain of human tyrosine hydroxylase 1 monitored using non-uniformly sampled NMR

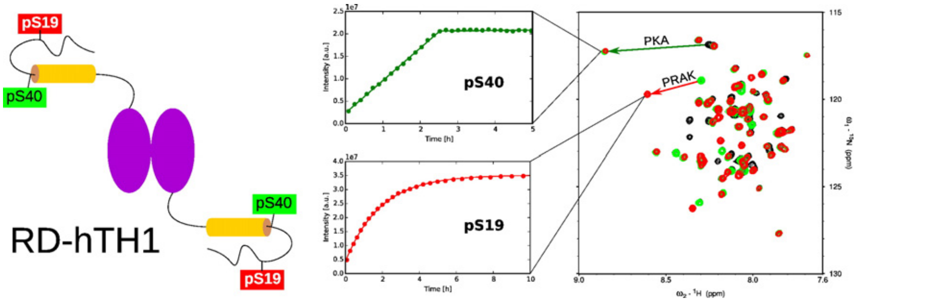
Petr Louša, Hana Nedozrálová, Erik Župa, Jiří Nováček, Jozef Hritz *

CEITEC MU, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic

HIGHLIGHTS

- Disordered part of regulatory domain of human tyrosine hydroxylase 1 was assigned.
- Transient alpha-helices are present next to phosphorylation sites S40 and S19.
- The secondary structure does not change after phosphorylation.
- The phosphorylation kinetic rates were measured efficiently using time resolved NMR.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 30 December 2016
Accepted 24 January 2017
Available online 27 January 2017

Keywords:
Human tyrosine hydroxylase
IDP
NMR
Phosphorylation
Kinetics
Time-resolved NMR
Non-uniform sampling
SSP

ABSTRACT

Human tyrosine hydroxylase 1 (hTH1) activity is regulated by phosphorylation of its regulatory domain (RD-hTH1) and by an interaction with the 14-3-3 protein. The RD-hTH1 is composed of a structured region (66-169) preceded by an intrinsically disordered protein region (IDP, hTH1_65) containing two phosphorylation sites (S19 and S40) which are highly relevant for its increase in activity. The NMR signals of the IDP region in the non-phosphorylated, singly phosphorylated (pS40) and doubly phosphorylated states (pS19_pS40) were assigned by non-uniformly sampled spectra with increased dimensionality (5D). The structural changes induced by phosphorylation were analyzed by means of secondary structure propensities. The phosphorylation kinetics of the S40 and S19 by kinases PKA and PRAK respectively were monitored by non-uniformly sampled time-resolved NMR spectroscopy followed by their quantitative analysis.

1. Introduction

Tyrosine hydroxylase (TH) is an enzyme which converts L-tyrosine to L-DOPA. This reaction is the rate limiting step in the biosynthetic pathway producing important catecholamine neurotransmitters: dopamine, noradrenaline and adrenaline [1,2]. The human enzyme isoform 1 is a tetramer each consisting of three domains: an N-terminal regulatory domain (RD-hTH1, 1-169 aa), a catalytic domain (170-450 aa), and a short C-terminal tetramerization domain (451-497 aa) [3]. The first 65 residues of the regulatory domain form an intrinsically disordered protein region (IDP, hTH1_65) important for regulation. The activity of hTH1 is controlled by the phosphorylation of its IDP region (S19, S31, S40) and by the interaction with 14-3-3 protein [4]. Phosphorylation sites S19 and S40 are the most relevant phosphorylation sites regarding 14-3-3 binding [5,6].

Recently, the NMR structure of the ordered region (65-159) of the dimeric regulatory domain of rat tyrosine hydroxylase (the sequence

* Corresponding author.
E-mail address: jozef.hritz@ceitec.muni.cz (J. Hritz).

http://dx.doi.org/10.1016/j.bpc.2017.01.003
0301-4622/© 2017 Elsevier B.V. All rights reserved.

identity between rat and human RD-hTH1 is 81.8%) was determined by Zhang et al. [7]. Authors claimed that the problems with unstable sample and low signal dispersion of the IDP region prevented structural characterization of the full RD. We overcame these problems by modifying the preparation of full length RD-hTH1 (1-169) sample in non- and phosphorylated states and by applying non-uniform sampling (NUS) NMR approaches allowing much faster data collection in comparison with the uniformly sampled experiments.

In the past, the rate of the phosphorylation of hTH1 was monitored semi-quantitatively by Toska et al. [8] using radioactively labeled ATP. Such methodology has several drawbacks, especially the necessity of working with radioactive material, laborious sample preparation and low temporal resolution. The alternative methodology for a monitoring of protein phosphorylation is NMR spectroscopy where individual NMR spectra are collected sequentially during the course of a phosphorylation reaction [9]. The time resolution of such an approach is on the order of the measurement time of one particular NMR spectra. Another possibility for monitoring is to measure one long 2D spectrum and afterwards analyze lineshapes of signals modulated by the reaction kinetics [10]. In this way, the time resolution can be reduced at cost of significant increase in difficulty of analysis. The recent advances in NMR allow us to overcome these limitations by applying non-uniform sampled time-resolved NMR spectroscopy to monitor the reaction course with time resolution down to seconds [11]. This approach was already successfully utilized for the monitoring of the phosphorylation of cytoplasmic domain of human B cell receptor protein CD79b [11].

In this study, we present the resonance assignment of the IDP region (hTH1_65) of RD-hTH1 in the non-, singly- and doubly-phosphorylated states using non-uniformly sampled NMR experiments with increased dimensionality. Next, we analyze the structural changes induced by the phosphorylation of S40 and S19 by PKA and PRAK kinase respectively and the kinetics of these processes.

2. Experimental

2.1. Protein expression and purification

The regulatory domain (residues 1-169) of human tyrosine hydroxylase 1 (RD-hTH1) in a pET15b plasmid containing a TEV-cleavable His-tag was expressed in *E. coli* BL21(DE3)RIL cells. For preparation of ^{15}N-labeled and ^{13}C,^{15}N-labeled samples, the cells were cultured in M9 medium with $^{15}NH_4Cl$, ampicillin (100 mg/ml), chloramphenicol (35 mg/ml) and with addition of $^{13}C_6$-glucose for the double labeled sample. Cells grew at 37 °C until OD_{600} ~ 0.8 then were induced with 0.5 mM IPTG and further cultured at 18 °C for ~18 h. Cells were harvested and homogenized in 50 mM Tris pH = 8, 150 mM NaCl, 3 mM NaN_3. Cell lysate was centrifuged for 1 h at 21,040*g*. Supernatant was then applied on a Ni^{2+} affinity column (HisTrap HP, GE Healthcare) equilibrated in 50 mM Tris pH = 8, 500 mM NaCl, 3 mM NaN_3. Sample was eluted by gradient of elution buffer (equilibration buffer + 1 M imidazole) at its ~70% concentration. Sample was then gel filtrated on a Superdex 75 column (HiLoad 16/600 Superdex 75 pg, GE Healthcare) equilibrated in 50 mM Tris pH = 8, 100 mM NaCl, 3 mM NaN_3. The eluted sample was treated with TEV protease (protein:protease ratio 20:1) at 4 °C overnight and then dialysed into phosphate buffer (20 mM sodium phosphate buffer pH = 6, 3 mM NaN_3). The His-tag cleaved protein was loaded on a cation exchange column (Resource S, GE Healthcare) equilibrated in phosphate buffer (pH = 6.0), and sample was eluted by a gradient of elution buffer (phosphate buffer, pH = 6.0 + 1 M NaCl) at conductivity 23–45 mS·cm^{-1}. Fractions containing our protein sample were again dialysed into phosphate buffer (pH = 6.0) and then concentrated for final gel filtration on a Superdex 75 column equilibrated in phosphate buffer.

2.2. NMR samples and phosphorylation

The NMR backbone assignment of the IDP region (1–65 aa) was performed on [^{15}N, ^{13}C] labeled samples of non- and doubly phosphorylated RD-hTH1 at 1.0 mM concentration in 20 mM sodium phosphate buffer pH = 6 with 8% D_2O and 3 mM NaN_3. The phosphorylation kinetic studies were performed with [^{15}N] samples at 0.3 mM concentration in phosphorylation buffer containing 50 mM sodium phosphate buffer pH = 6, 10 mM ATP, 10 mM $MgCl_2$ with 8% D_2O and 3 mM NaN_3. For phosphorylation of S40, PKA (the catalytic subunit of cAMP-dependent protein kinase, New England BioLabs Inc.) was used in 0.5 μg/ml (13.2 nM) concentration. Afterwards, S19 phosphorylation using PRAK (p38 regulated/activated protein kinase, obtained from University Dundee, Scotland) was performed. The concentration of PRAK was 0.11 mg/ml (2.02 μM). Both phosphorylation reactions were monitored over the course of 40 h.

2.3. NMR experiments

The assignment of non-phosphorylated RD-hTH1 was performed using a Bruker 850 MHz US^2 spectrometer equipped with cryogenic triple-resonance probe head (5 mm CPTCI 1H/19F-13C/15N/D). The kinetic measurements as well as assignment of phosphorylated RD-hTH1 were performed using a Bruker 600 MHz spectrometer equipped with a cryogenic triple-resonance probe head (5 mm CPQCI 1H-31P/13C/15N/D). Both probe heads are equipped with z-axis gradient coils. All measurements were done at a temperature of 293.2 K.

For the assignment of non-phosphorylated and doubly S19_S40-phosphorylated RD-hTH1, 3D HNCO [12], 5D HN(CA)CONH and 5D HabCabCONH [13,14] were measured. All experiments were carried out with non-uniform sampling of the indirectly detected domains. The time schedule was generated using Poisson disk sampling on a grid, introducing distance constraints between points. The density of points was set according to a Gaussian distribution ($\sigma = 0.5$).

The phosphorylation was monitored using 2D HSQC experiments. In both cases, the maximal evolution times were set to 102 and 64 ms, respectively. The time resolution was achieved using non-uniform sampling in the indirect domain. The sampling schedule comprised 16,000 points in total for both phosphorylations. The size of the schedules exceeded the size of regular Nyquist grid 125 times. The total measuring time was 40 h.

2.4. Data processing

The non-uniform 3D HNCO spectra were processed using a Multidimensional Fourier Transform [15], while 5D HN(CA)CONH and 5D HabCabCONH spectra were processed using a Sparse Multidimensional Fourier Transform (SMFT) algorithm [16]. The direct dimension was square cosine weighted and zero-filled to 4096 complex points, followed by a standard FFT. The 5D spectra were processed by the SMFT algorithm using the program *reduced*, using fixed frequencies of ^{13}C′, ^{15}N and ^{1}H^{N} identified in the 3D HNCO spectrum, providing sets of 2D slices. The assignment and visualization of the NMR spectra was performed in the software Sparky 3.115 [18].

The secondary structure propensities were calculated by the program SSP [19] using chemical shifts of ^{1}H$^{\alpha}$, ^{13}C$^{\alpha}$ and ^{13}C$^{\beta}$. The data from RefDB [17] were used for random coil referencing. These values were used for calculation of the phosphorylated state as well.

The time-resolved HSQC spectra were processed using a coprocessed Multidimensional Decomposition (co-MDD) [11]. The initial window size was set to 64 points. In the case of PRAK phosphorylation, the window size was incremented by a factor of 1.05 to reduce the fitting errors in subsequent analyses. The processing yielded 250 individual frames for PKA and 52 frames for PRAK phosphorylation.

3. Results and discussion

The ^{15}N- and ^{13}C,^{15}N-labeled samples of RD-hTH1 (region 1-169) protein were expressed and purified as described in the Methods section. Final purity of prepared non-, singly and doubly phosphorylated variants was verified by MALDI-TOF-MS spectroscopy (Fig. S1 in Suppl. mat.). We want to emphasize that in order to study the IDP region of RD-hTH1, very high purity of samples is needed because of proteolytic degradation.

3.1. NMR assignment

The assignment of the disordered part (IDP, first 65 residues) within the RD-hTH1 was performed by employing non-uniform high-dimensional 5D spectra HabCabCONH and HN(CA)CONH. NUS 5D NMR spectroscopy allows measurement of well resolved spectra of the regions with fast tumbling (i.e. slow relaxation) in a very efficient way. The HN(CA)CONH spectrum provided sequential information leading to linkage of several long fragments. This assignment was then verified by amino acid classification based on chemical shifts of $^1H^\alpha$, $^1H^\beta$, $^{13}C^\alpha$, $^{13}C^\beta$ derived from cross-sections of HabCabCONH spectra. Using this approach, we were able to assign all amide resonances in region 1-65 and most resonances of H^α, H^β, C^α, C^β, and carbonyl atoms (including prolines) except for those directly preceding prolines, i.e. M1, T3, T8, S31, and V60 and except for H^β, C^β atoms of L21 and I42 (Tables S1, S2, S3 in Suppl. mat.). In comparison to previously published NMR assignment of RD of rat TH [7], hereby presented assignment for RD of human TH is more complete, especially in the region around second phosphorylation site (S40).

The described assignment procedure was applied to the non- and double-phosphorylated variants of RD-hTH1. The single phosphorylated variant (pS40) was assigned based on the very close similarities in region 1-30 with respect to the non-phosphorylated variant, and region 30-65 was rather similar to the doubly phosphorylated variant. Fig. 1A presents the superposed assigned HSQC spectra of the IDP region of RD-hTH1 in the non-, singly- (pS40) and doubly- (pS19_pS40) phosphorylated states.

Naturally, the largest changes in the chemical shifts (Fig. 1B) are observed for the phosphorylated residue itself and its neighbors in the primary sequence. The effect on longer distances seems to be more

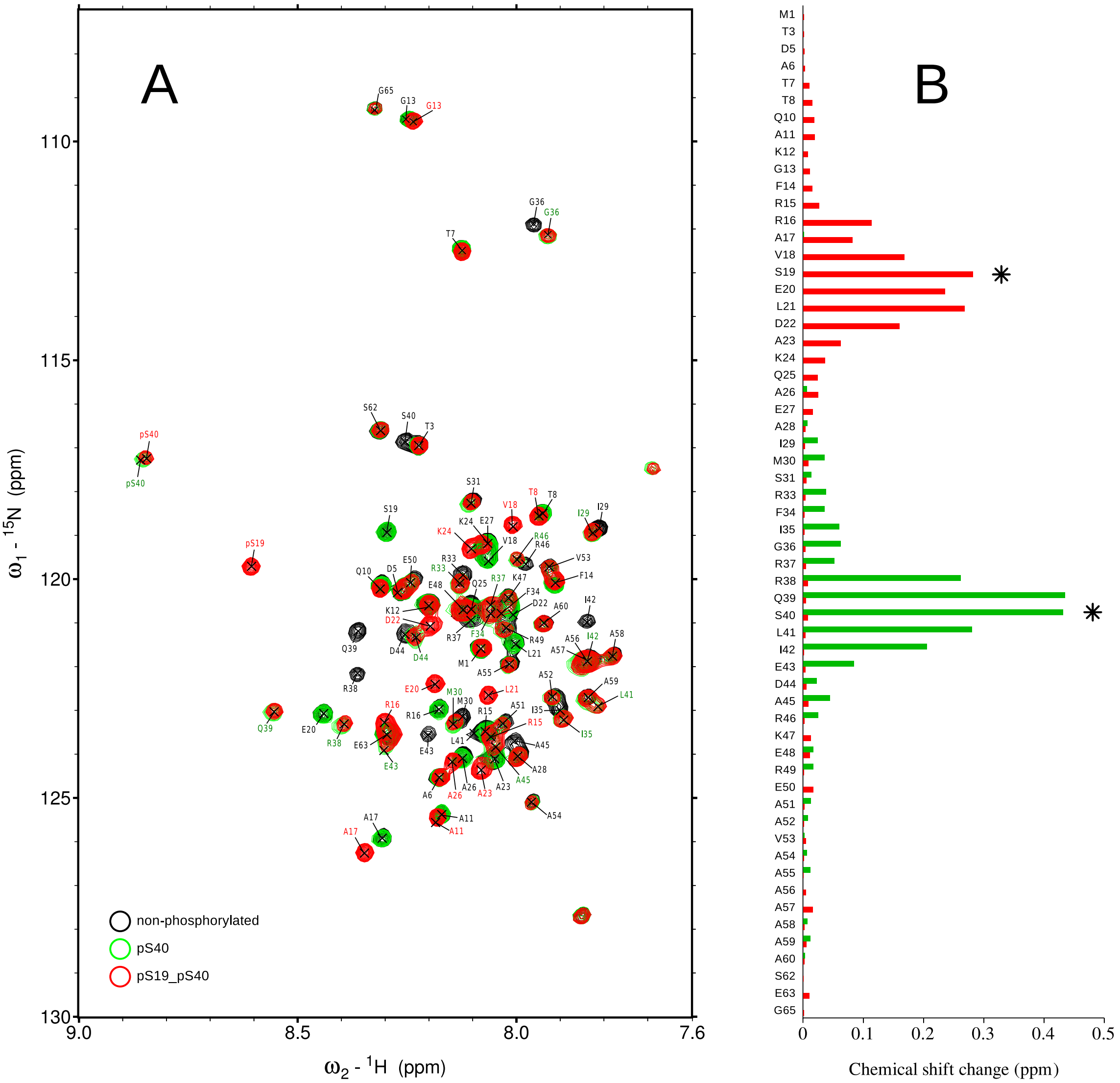


Fig. 1. Superposition of HSQC spectra of differently phosphorylated RD-hTH1 (IDP region). The spectra of non-phosphorylated (black), single (pS40) phosphorylated (green) and doubly (pS19_pS40) phosphorylated spectrum (red) are shown in Panel A. The assignment labels are black for peak positions in non-phosphorylated RD-hTH1, green for signals that changed significantly during S40 phosphorylation (by PKA) and red for signals that changed during the subsequent S19 phosphorylation by PRAK. Changes of chemical shifts of RD-hTH1 observed in HSQC spectra during phosphorylation of S40 (green) and of S19 (red) are shown in Panel B. The asterisks denote the position of phosphorylation sites.

Table 1
Differences of chemical shifts of nuclei used in SSP calculation for residues close to phosphorylation sites.

	A17	V18	S19	E20	L21	R38	Q39	S40	L41	I42
$\Delta\delta C^{\alpha}$ (ppm)	0.23	0.23	−0.08	−0.51	−0.17	−0.40	0.61	−0.04	−0.26	−0.04
$\Delta\delta C^{\beta}$ (ppm)	−0.07	−0.19	1.78	0.25	0.14	0.03	−0.19	1.73	0.09	–
$\Delta\delta H^{\alpha}$ (ppm)	−0.03	−0.05	−0.01	0.00	−0.01	0.03	−0.09	−0.03	0.00	−0.02

pronounced around the pS40 site with change observed even at M30. This can be due to the transient secondary structure elements described in the following section.

3.2. Secondary structural changes induced by phosphorylation

The secondary structure propensities (SSP) for non- and doubly phosphorylated RD-hTH1 were determined by the approach described in Methods. The phosphoserines were excluded from this analysis due to large change in chemical shift of their beta carbon (Table 1). The SSP values close to zero mean that the particular amino acids are in a random-coil conformation; the larger positive and negative values indicate the alpha-helical or beta-sheet conformations, respectively [19]. Fig. 2A indicates slight alpha-helical propensities in region 17-23 and more pronounced in the region 35-55 for both non- and doubly phosphorylated variants of RD-hTH1. The intensities of signals in 3D HNCO spectra in Fig. 2B further support the existence of alpha-helical structure in the region 35-55. The intensity is related to relaxation properties of residues. The residues with slower tumbling, i.e. inside structured regions, relax faster which manifests as lowered intensity of the resulting signal. The highest intensity is found for the flexible N-terminal residues as expected, while the lowest intensity corresponds to the alpha-helical region suggested by SSP.

Both secondary structure propensities (Fig. 2A) as well as the peak intensities (Fig. 2B) over residues within the IDP region are very similar for non- and doubly phosphorylated RD-hTH1. There is slight increase in alpha-helical propensity for region 48-58 and slight decrease of alpha-helical propensity in region 18-22 around the second phosphorylation site (pS19). This observation indicates negligible conformational changes within the RD-hTH1 due to the phosphorylation of S19 and S40. This is in agreement with the proposed molecular mechanism of hTH1 activation suggesting that the phosphorylation at S40 and S19 induce conformational changes of the whole RD with respect to the catalytic domain rather than within the RD itself [20].

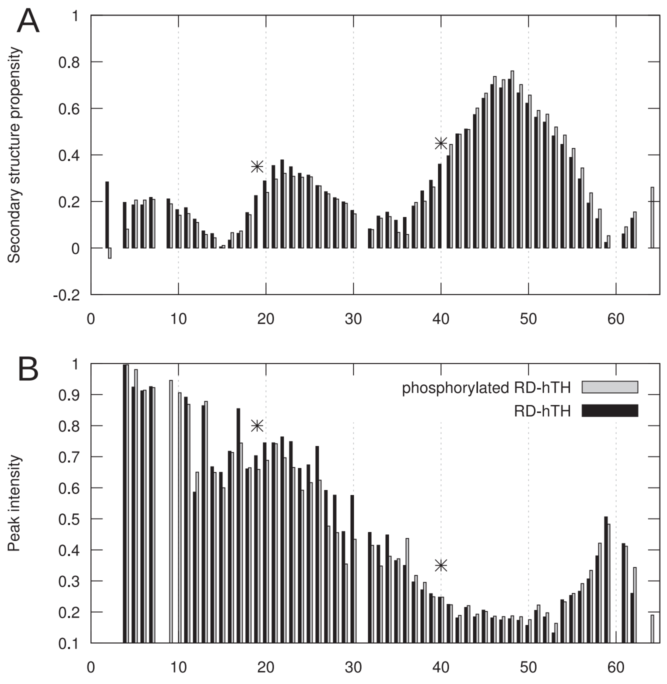


Fig. 2. Secondary structure propensities (A) and intensities (B) of signals in 3D HNCO spectra of non-phosphorylated (black bars) and doubly-phosphorylated (pS19_pS40, grey bars) RD-hTH1 (IDP region). The asterisks denote phosphorylation sites.

From a methodological point of view, we found it quite surprising that in contrast to amide groups, the chemical shifts of nuclei that are usually used for the SSP analysis ($^1H^{\alpha}$, $^{13}C^{\alpha}$ and $^{13}C^{\beta}$) are relatively insensitive to the phosphorylation of serines (Table 1). The standard deviation of chemical shifts for SSP analysis is on the order of 1 ppm, which is much more than the differences in Table 1. The only significant difference is found only for the phosphorylated serine itself and its beta carbon. This insensitivity of chemical shifts of regions around phosphoserines in IDP regions allows us to perform SSP analysis using non-phosphorylated reference values also for other phosphorylated proteins without the need for intrinsic chemical shift referencing [21].

3.3. Phosphorylation kinetics

In Fig. 3 we present the time progress of phosphorylation for involved serine and surrounding well resolved residues. We plot together decrease in intensity of the original peak and increase in intensity of newly formed peak after phosphorylation. First, S40 was phosphorylated using PKA. The progress of phosphorylation was monitored by changes in the intensity of the HSQC spectrum, which provided the information to derive a 0^{th} order rate constant of $k = 0.128 \pm 0.003\ mM \cdot h^{-1}$ (Fig. 3A). After 10 h when S40 was fully phosphorylated, PRAK was added, and phosphorylation of S19 was observed. The intensity changes can be described by 1^{st} order kinetics. Fitting provides a rate constant of $k = 0.556 \pm 0.008\ h^{-1}$ (Fig. 3B). After the normalization of rate constants to micromolar kinase concentrations, we obtained these values: $9.7\ mM \cdot h^{-1}/\mu M$ for PKA and $0.27\ h^{-1}/\mu M$ for PRAK.

4. Conclusions

Advanced techniques of non-uniform sampled NMR experiments were applied to gain information about the intrinsically disordered region of the regulatory domain of human tyrosine hydroxylase 1 (RD-hTH1) and its phosphorylated variants. The assignment of non-phosphorylated and doubly phosphorylated (pS19_pS40) samples was performed using non-uniformly sampled 5D NMR spectroscopy. These two sets of chemical shifts were also sufficient to perform assignment of singly phosphorylated (pS40) RD-hTH1. Although the phosphorylation has quite significant impact on amidic chemical shifts of phosphoserine and its neighboring residues, it has negligible effect on aliphatic chemical shifts with the exception of C^{β} of the phosphoserine itself. Therefore, the aliphatic chemical shifts could be employed to monitor alpha-helical propensity in the region between residues 35 and 55 without the need of phosphoprotein reference. Phosphorylation has no significant impact on the secondary structure propensities of RD-hTH1.

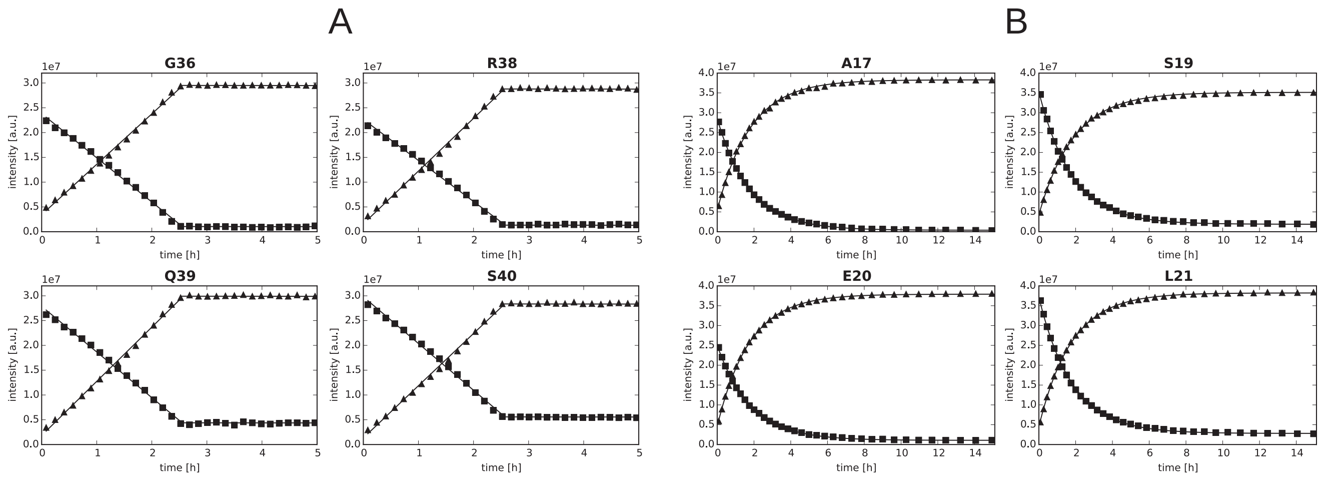


Fig. 3. Progress of phosphorylation by PKA kinase is shown in Panel A. The intensities of signals belonging to non-phosphorylated protein (squares) and S40-phosphorylated (triangles) are plotted against reaction time. Other residues close to S40 were not used due to severe overlap, leading to errors in analysis. Panel B: Progress of phosphorylation by PRAK. Intensities belonging to singly (pS40) (squares) and to doubly (pS19_pS40) phosphorylated sample (triangles) are plotted against reaction time. Other residues were strongly influenced by signal overlap and therefore not used in subsequent analysis.

The kinetics of phosphorylation of S40 by PKA and of S19 by PRAK were determined using non-uniformly sampled time-resolved NMR spectroscopy. The phosphorylation of S40 was determined as a 0^{th} order reaction with normalized rate constant value 9.7 mM·h^{-1}/μM, while phosphorylation of S19 was observed as 1^{st} order reaction with value 0.27 h^{-1}/μM. The described time-resolved NMR spectroscopy approach has general applicability over large time scales of various post-translational processes and can be easily employed in other protein systems.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.bpc.2017.01.003.

Acknowledgment

We thank Dr. Tanvir Shaikh for critical reading of the manuscript. The project is financed from the SoMoPro II programme. The research leading to this invention has acquired a financial grant from the People Programme (Marie Curie action) of the Seventh Framework Programme of EU according to the REA Grant Agreement No. 291782. The research is further co-financed by the South-Moravian Region. The article/paper reflects only the author's views and the Union is not liable for any use that may be made of the information contained therein. In addition, this work was also supported by the Czech Science Foundation (15-34684L). The CIISB research infrastructure project LM2015043 funded by MEYS CR is gratefully acknowledged for the partial financial support of the measurements at the Josef Dadok National NMR Centre and Proteomics Core Facilities, CEITEC – Masaryk University.

References

[1] T. Nagatsu, M. Levitt, S. Udenfriend, Tyrosine hydroxylase: the initial step in norephinephrine biosynthesis, J. Biol. Chem. 239 (1964) 2910–2917.
[2] P.B. Molinoff, J. Axelrod, Biochemistry of catecholamines, Annu. Rev. Biochem. 40 (1971) 465–500.
[3] S.C. Daubner, D.L. Lohse, P.F. Fitzpatrick, Expression and characterization of catalytic and regulatory domains of rat tyrosine hydroxylase, Protein Sci. 2 (1993) 1452–1460.
[4] P.F. Fitzpatrick, Tetrahydropterin-dependent amino acid hydroxylases, Annu. Rev. Biochem. 68 (1999) 355–381.
[5] J. Hritz, I.-J. Byeon, T. Krzysiak, A. Martinez, V. Sklenář, A.M. Gronenborn, Dissection of binding between a phosphorylated tyrosine hydroxylase peptide and 14-3-3zeta: a complex story elucidated by NMR, Biophys. J. 107 (2014) 2185–2194.
[6] R. Kleppe, K. Toska, J. Haavik, Interaction of phosphorylated tyrosine hydroxylase with 14-3-3 proteins: evidence for a phosphoserine 40-dependent association, J. Neurochem. 77 (2001) 1097–1107.
[7] S. Zhang, T. Huang, U. Ilangovan, A.P. Hinck, P.F. Fitzpatrick, The solution structure of the regulatory domain of tyrosine hydroxylase, J. Mol. Biol. 426 (2014) 1483–1497.
[8] K. Toska, R. Kleppe, C.G. Armstrong, N.A. Morrice, P. Cohen, J. Haavik, Regulation of tyrosine hydroxylase by stress-activated protein kinases, J. Neurochem. 83 (2002) 775–783.
[9] F.X. Theillet, H.M. Rose, S. Liokatis, A. Binolfi, R. Thongwichian, M. Stuiver, P. Selenko, Site-specific NMR mapping and time-resolved monitoring of serine and threonine phosphorylation in reconstituted kinase reactions and mammalian cell extracts, Nat. Protoc. 8 (2013) 1416–1432.
[10] I. Landrieu, L. Lacosse, A. Leroy, J.-M. Wieruszeski, X. Trivelli, A. Sillen, N. Sibille, H. Schwalbe, K. Saxena, T. Langer, G. Lippens, NMR analysis of a tau phosphorylation pattern, J. Am. Chem. Soc. 128 (2006) 3575–3583.
[11] M. Mayzel, J. Rosenlöw, L. Isaksson, V.Y. Orekhov, Time-resolved multidimensional NMR with non-uniform sampling, J. Biomol. NMR 58 (2014) 129–139.
[12] L.E. Kay, M. Ikura, R. Tschudin, A. Bax, Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins, J. Magn. Reson. 89 (1990) 496–514.
[13] K. Kazimierczuk, A. Zawadzka-Kazimierczuk, W. Kozminski, Non-uniform frequency domain for optimal exploitation of non-uniform sampling, J. Magn. Reson. 205 (2010) 286–292.
[14] V. Motackova, J. Novacek, A. Zawadzka-Kazimierczuk, K. Kazimierczuk, L. Zidek, H. Sanderova, L. Krasny, W. Kozminski, V. Sklenar, Strategy for complete NMR assignment of disordered proteins with highly repetitive sequences based on resolution-enhanced 5D experiments, J. Biomol. NMR 48 (2010) 169–177.
[15] J. Stanek, W. Kozminski, Iterative algorithm of discrete Fourier transform for processing randomly sampled NMR data sets, J. Biomol. NMR 47 (2010) 65–77.
[16] K. Kazimierczuk, A. Zawadzka, W. Kozminski, Narrow peaks and high dimensionalities: exploiting the advantages of random sampling, J. Magn. Reson. 197 (2009) 219–228.
[17] H. Zhang, S. Neal, D.S. Wishart, RefDB: a database of uniformly referenced protein chemical shifts, J. Biomol. NMR 25 (2003) 173–195.
[18] T.D. Goddard and D. G. Kneller, SPARKY 3, University of California, San Francisco, USA.
[19] J.A. Marsh, V.K. Singh, Z. Jia, J.D. Forman-Kay, Sensitivity of secondary structure propensities to sequence differences between α- and γ-synuclein: implications for fibrillation, Protein Sci. 15 (2006) 2795–2804.
[20] S.C. Daubner, T. Lee, S. Wang, Tyrosine hydroxylase and regulation of dopamine synthesis, Arch. Biochem. Biophys. 508 (2011) 1–12.
[21] K. Modig, V.W. Jürgensen, K. Lindorff-Larsen, W. Fieber, H. Bohr, F.M. Poulsen, Detection of initiation sites in protein folding of the four helix bundle ACBP by chemical shift analysis, FEBS Lett. 581 (25) (2007) 4965–4971.

Paper 13

Zapletal, V.; Mládek, A.; Melková, K.; Louša, P.; Nomilner, E.; Jaseňáková, Z.; Kubáň, V.; Makovická, M.; Laníková, A.; Žídek L.; Hritz, J.* Choice of force field for proteins containing structured and intrinsically disordered regions. *Biophys. J.* **2020**, 118, 1621–1633

Biophysical*Journal*
Article

Choice of Force Field for Proteins Containing Structured and Intrinsically Disordered Regions

Vojtěch Zapletal,[1,2] Arnošt Mládek,[2] Kateřina Melková,[1,2] Petr Louša,[2] Erik Nomilner,[1] Zuzana Jaseňáková,[1,2] Vojtěch Kubáň,[1,2] Markéta Makovická,[1] Alice Laníková,[1] Lukáš Žídek,[1,2] and Jozef Hritz[2,*]

[1]National Centre for Biomolecular Research, Faculty of Science and [2]Central European Institute of Technology, Masaryk University, Brno, Czech Republic

ABSTRACT Biomolecular force fields optimized for globular proteins fail to properly reproduce properties of intrinsically disordered proteins. In particular, parameters of the water model need to be modified to improve applicability of the force fields to both ordered and disordered proteins. Here, we compared performance of force fields recommended for intrinsically disordered proteins in molecular dynamics simulations of three proteins differing in the content of ordered and disordered regions (two proteins consisting of a well-structured domain and of a disordered region with and without a transient helical motif and one disordered protein containing a region of increased helical propensity). The obtained molecular dynamics trajectories were used to predict measurable parameters, including radii of gyration of the proteins and chemical shifts, residual dipolar couplings, paramagnetic relaxation enhancement, and NMR relaxation data of their individual residues. The predicted quantities were compared with experimental data obtained within this study or published previously. The results showed that the NMR relaxation parameters, rarely used for benchmarking, are particularly sensitive to the choice of force-field parameters, especially those defining the water model. Interestingly, the TIP3P water model, leading to an artificial structural collapse, also resulted in unrealistic relaxation properties. The TIP4P-D water model, combined with three biomolecular force-field parameters for the protein part, significantly improved reliability of the simulations. Additional analysis revealed only one particular force field capable of retaining the transient helical motif observed in NMR experiments. The benchmarking protocol used in our study, being more sensitive to imperfections than the commonly used tests, is well suited to evaluate the performance of newly developed force fields.

SIGNIFICANCE We compared the performance of several force fields in molecular dynamics simulations of three proteins differing in the content of ordered and disordered regions. From the obtained trajectories, we predicted a set of measurable quantities and compared them with their experimental values. Among the predicted parameters, NMR relaxation data were particularly sensitive to the choice of force field parameters, especially those defining the water model. The presented benchmarking protocol will help to select force fields that reliably simulate properties of physiologically important intrinsically disordered proteins.

INTRODUCTION

The fact that many proteins of biological relevance contain considerably large intrinsically disordered regions (IDRs), contradicting the classical structure-function paradigm, has been accepted by the structural biology community during the past two decades (1–3). Intrinsically disordered proteins (IDPs) represent a relatively diverse class of molecules differing in their biophysical properties. One polypeptide chain often contains fully structured domains together with IDRs. Moreover, IDRs are not random polymer chains but exhibit various degree of partial ordering. They typically contain a short segment with an increased propensity to form secondary transient structures, described in the literature as prestructured motifs (4), preformed structural elements (5), or molecular recognition features (6). Such hybrid systems present methodological challenges because an IDR tethered to a well-ordered domain is a molecule consisting of regions with highly diverse dynamics. As a result, it is difficult for experimental and computational methods to accurately capture both types of behavior in these systems.

Submitted April 17, 2019, and accepted for publication February 5, 2020.

*Correspondence: jozef.hritz@ceitec.muni.cz

Editor: Rohit Pappu.

https://doi.org/10.1016/j.bpj.2020.02.019

NMR represents a method of choice for studies of proteins with IDRs at atomic resolution, as disordered systems are difficult to investigate using single-crystal x-ray diffraction or single-particle reconstruction of cryo-electron microscopic images. Currently available NMR methods provide sufficient resolution to overcome the narrow distribution of chemical shifts of IDPs (7,8). However, it should be noted that the resolution improvement of IDP-targeted NMR experiments relies on the slow relaxation of IDRs. Therefore, the sensitivity of such experiments is often too low for the rapidly relaxing signals of amino acids in the well-ordered regions of hybrid proteins.

Molecular dynamics (MD) simulations could, in principle, serve as an ideal tool to study behavior of hybrid proteins at an atomic level. Moreover, most of the NMR parameters can be reliably predicted from structural models, which allows for direct comparison of MD results with experimental data. In practice, several problems complicate MD simulations of IDPs. The energy landscapes of IDRs are expected to be weakly funneled such that the search for a specific functionally competent conformation could be extremely inefficient (9). In this study, we examined the applicability of MD simulations to hybrid proteins and assessed their reliability by predicting several measurable parameters from the obtained trajectories and comparing them with experimental data. Moreover, we provide an example of prediction of measurable parameters as a guide to select optimal setup for future experiments, such as positions with paramagnetic relaxation enhancement (PRE) labels. We tested the currently available force fields Amber99SB-ILDN (A99), CHARMM22* (C22*), and CHARMM36m (C36m) in combination with explicit solvent models TIP3P, TIPS3P, and TIP4P-D. Chemical shift, residual dipolar coupling (RDC), PRE, relaxation rate, and small-angle x-ray scattering (SAXS) experimental data were used for validation. It should be pointed out that our goal was not to describe the properties of the studied proteins as faithfully as possible but to look for features that would distinguish the performance of various force fields in MD simulations of a microsecond time range. The hybrid proteins investigated in this study included 1) δ subunit of RNA polymerase from *Bacillus subtilis* (δRNAP), 2) regulatory domain of human tyrosine hydroxylase (RD-hTH), and 3) a fragment consisting of residues 159–254 of rat microtubule-associated protein 2c (MAP2c$^{159-254}$). δRNAP makes RNA polymerase sensitive to the concentration of initiating nucleoside triphosphates, which is important for rapid changes in gene expression (10). It consists of two domains of a similar size. The N-terminal half of δRNAP folds into a well-ordered, mostly α-helical domain, whereas the C-terminal half is disordered and highly negatively charged, with the exception of a lysine-rich motif ^{96}KAKKKKAKK^{104} and C-terminal Lys173 (11). Experimental data showed that the lysine stretch makes transient electrostatic contacts with various residues in the acidic C-terminal region (1). No sign of formation of transient α-helical structures was observed in the C-terminal domain, which prefers extended backbone conformations, presumably because of the electrostatic repulsion of the acidic side chains. RD-hTH catalyzes hydroxylation of L-tyrosine to L-3,4-dihydroxyphenylalanine (L-DOPA) and is a key and rate-limiting enzyme in biosynthesis of important catecholamine neurotransmitters (12,13). Its N-terminal region (~40% of its sequence) is disordered but contains a segment with ~80% propensity to form four turns of α-helix (14). MAP2c$^{159-254}$ corresponds to the central region of MAP2c, where proteins regulating the microtubule-stabilizing activity of MAP2c bind in a phosphorylation-dependent manner (15,16). MAP2c$^{159-254}$ is mostly disordered but exhibits an ~20% propensity to form four turns of an α-helix presumably important for intermolecular interactions (16,17).

MATERIALS AND METHODS

NMR spectroscopy

NMR assignments of δRNAP and of the disordered part of RD-hTH were published previously (14,18). MAP2c$^{159-254}$ was assigned as described for full-length MAP2c (19). PRE data of δRNAP were published previously (11). RDCs were calculated as a difference between splitting observed in in-phase/anti-phase (IPAP) spectra (20) obtained for proteins in stretched 5% polyacrylamide gel and in isotropic medium. The RD-hTH data were acquired at 20°C on a 600 and 850 MHz Bruker Avance III spectrometer (Bruker, Billerica, MA), and the MAP2c$^{159-254}$ data were acquired at 27°C on a 600 MHz Bruker Avance III spectrometer. The backbone amide ^{15}N relaxation data were published previously for δRNAP (21) and measured using standard pulse sequences (22) for 1.0 mM [^{15}N]-RD-hTH and 0.34 mM [^{15}N]-MAP2c$^{159-254}$ at 27°C on 850 and 950 MHz Bruker Avance III spectrometers, respectively. The interscan delays were set to 1.5, 2, 6, and 25 s for R_1, R_2, and steady-state heteronuclear Overhauser effect (ssNOE) measurements without and with ssNOE transfer, respectively. The relaxation delays for the R_1 experiment were 11.2, 16.8, 28.0, 44.8, 61.6, 95.2, 196, and 308 ms, and the delays for the R_2 experiment were 0, 14.4, 28.8, 43.2, 57.6, 86.4, and 129.6 ms. R_1 and R_2 data were fitted using a two-parameter exponential. The errors were obtained using the bootstrap procedure (23). The ssNOE values were calculated as a ratio between signal intensities obtained from spectra with or without ^{1}H saturation. The errors were derived from background noise levels in each individual spectrum.

SAXS

The SAXS data sets were collected using a BioSAXS-1000 (Rigaku, Tokyo, Japan) instrument with an x-ray beam wavelength of 1.54 Å at 27°C. The distance between the sample and the detector (PILATUS 100K; Dectris, Baden-Daettwil, Switzerland) was 0.48 m, covering a scattering vector ($q = 4\pi \sin(\theta)/\lambda$) range from 0.009 to 0.65 Å^{-1}. For solvent and sample, one two-dimensional image was collected with 1 h exposure time per image. Radial averaging of two-dimensional scattering images and the solvent subtractions were performed using SAXSLab3.0.0r1 (Rigaku). All data sets were truncated to a maximal scattering vector of 0.3 Å^{-1} for further analysis. Radii of gyration were determined using PRIMUS from ATSAS v2.7.2 (24) with Guinier analysis (implemented in the PRIMUS Guinier Wizard) in the range 2–52 for δRNAP and in the range 9–30 for MAP2c$^{159-254}$; the globular particle type was used for both proteins. The molecular form factor analysis (25) was performed online at http://sosnick.uchicago.edu/SAXSonIDPs.

Computational details

The MD simulations were performed using Amber99SB-ILDN (26), CHARMM22* (27), and CHARMM36m (28) force-field parameters for the protein atoms and the TIP3P, TIPS3P (29,30), or TIP4P-D (31) water models. The proteins were solvated using a rhombic dodecahedral box of waters with a minimal distance between the box walls and solute of 2 nm. The charge of the system was neutralized by adding Cl^- and Na^+ ions, and the concentration of salt was adjusted to 100 mM. All simulations were performed under periodic boundary conditions. Before the MD runs, in vacuo and solvent energy minimizations with the steepest descent algorithm were carried out. The lengths of bonds with hydrogen atoms were constrained using the LINCS algorithm (32). An integration time step of 2 fs was used. A cutoff of 1.0 nm was applied for the Lennard-Jones interactions and short-range electrostatic interactions. Long-range electrostatic interactions were calculated by particle mesh Ewald summation with a grid spacing of 0.12 nm and a fourth-order interpolation (33). The four-step 8-ns-long equilibration protocol consisted of the following parts: 1) 2-ns relaxation of water molecules at 300 K during the NVT equilibration with restrained (1000 kJ mol^{-1} nm^{-2}) solute coordinates, 2) 2-ns NVT (300 K) run with restrained (1000 kJ mol^{-1} nm^{-2}) backbone atom coordinates, 3) 2-ns NpT (300 K, 1 atm) run with restrained (1000 kJ mol^{-1} nm^{-2}) backbone atom coordinates, and 4) 2 ns of unrestrained NpT simulation (300 K, 1 atm). The length of the follow-up production NpT simulations (300 K, 1 atm) was 200 ns. The simulations using the TIP4P-D water model were further prolonged to 1 μs, and additional independent simulations (500 ns for δRNAP and MAP2c$^{159–254}$, 400 ns for RD-hTH) starting from different initial conditions were run for all listed force fields and protein systems. The temperature and pressure were maintained using the Berendsen coupling scheme (34) during the equilibration steps; the production NpT simulations were performed using the velocity rescaling thermostat with a stochastic term (35) and the Parrinello-Rahman barostat algorithms (36). Atomic coordinates were recorded every 1 ps.

Predictions of NMR parameters

Chemical shifts were calculated using the prediction algorithm SPARTA+ (37) for each structure, and averaged secondary chemical shifts (SCSs) were calculated by subtracting the random-coil values (38). RDC calculations using a local alignment window were performed. For calculations using a local alignment window, the RDC, calculated using the program PALES (39), for the central amino acid of the local 15-amino-acid segment was calculated for each conformer (40). The resulting RDC profile along the primary sequence was calculated by averaging each value over the whole trajectory and multiplying by the corresponding scaled absolute value of the generic baseline to account for long-range effects (41). The scale was chosen so that the lowest root mean-square deviations (RMSDs) from the experimental values were obtained in the disordered regions. PRE was calculated for the spin label used in the experiments, i.e., for thiol-reactive methanethiosulfonate (MTSL) attached to a side chain of cysteine introduced by site-directed mutagenesis. Sterically allowed MTSL side-chain conformations were sampled using previously published rotameric distributions (42) and built explicitly for each spin-label site of each individual structure backbone. 600 side-chain conformers were calculated, and the sterically allowed conformers were retained. Relaxation effects were averaged over these conformers as described by Salmon et al. (41).

The strategy described previously (43) was used to calculate relaxation rates. The autocorrelation function $C_i(\tau)$ was calculated from subtrajectories of each simulation, probing two timescales ($\tau_{\max} \leq 5$ or 50 ns). Each trajectory was thus divided into the blocks of 10 and 100 ns, respectively, and averaged. For each averaged block, $C_i(\tau)$ was described as a sum of $N = 512$ exponentials $e^{-\tau/\tau_{c,i}}$ whose amplitudes A_i were obtained using a Tikhonov regularization procedure (44). For each averaged block, the spectral densities are then defined as

$$J_i(\omega) = \sum_{i=1}^{N} \frac{A_i \tau_{c,i}}{1 + \omega^2 \tau_{c,i}^2} \tag{1}$$

and used to predict spin relaxation rates.

RESULTS AND DISCUSSION

Impact of selected water model

Our first goal was to examine the performance of various models of water in simulations of three hybrid proteins used as test molecules in this study. It is well described in the literature (31) that the water models typically used in MD simulations (e.g., TIP3P) significantly underestimate London dispersion interactions. To prevent this problem, TIPS3P and TIP4P-D water models were also tested, and the reliability of the behavior of IDPs or the IDRs was analyzed. Modified TIP3P containing nonzero van der Waals parameters was introduced originally in 1998 (30) and shown to provide more realistic results for IDPs when combined with C36m (28). Simulations using this model typically result in extended states when other models of water tend to produce ensembles that are structurally too compact relative to experiments (31).

Our preliminary 200-ns MD runs, using the TIP3P water model in combination with A99 (26), confirmed the artificial behavior of TIP3P.

The inferior performance of TIP3P was manifested most clearly by the calculated radius of gyration (R_g) of MAP2c$^{159–254}$. All simulations started with conformations having R_g close to the experimental value of 2.5 nm obtained from SAXS. During the initial 100 ns of simulations with TIP3P, R_g dropped to ~1.5 nm regardless of the protein force field used (Fig. 1 *b*). This result confirms that attractive interactions leading to formation of compact structures are unnaturally enhanced when TIP3P is used, as was also reported previously in (31,45). We also ran 200-ns simulations with the TIPS3P model (30) in combination with the C22* (26) and C36m (28) force fields. Remarkably, TIPS3P did not prevent the artificial compaction of MAP2c$^{159–254}$ (Fig. 1 *b*; a comparison of trajectories calculated using C36m with TIP3P and TIPS3P is presented in Fig. S1). A similar trend was observed also for RD-hTH (Fig. 1 *c*), although a direct comparison with the experiment was not possible because of the RD-hTH dimerization in real samples.

Finally, we performed MD simulations with the TIP4P-D water model combined with A99, C22*, and C36m. In agreement with the literature (31) and in a sharp contrast with the simulations run with TIP3P and TIPS3P, R_g of MAP2c$^{159–254}$ did not drop below the value calculated from the SAXS data (Fig. 1 *e*). In the case of RD-hTH, TIP4P-D also prevented unexpected compaction but only in combination with C36m (see the discussion of the impact of the protein force field parameters in the following section).

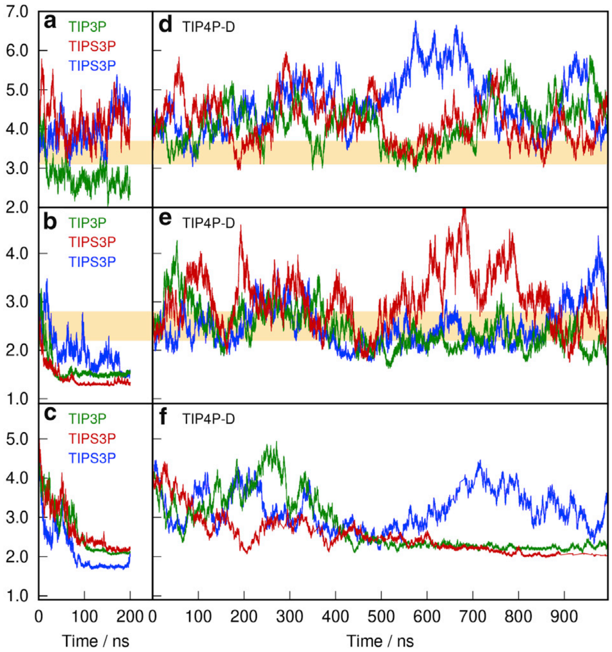


FIGURE 1 Simulated R_g of δRNAP (*a* and *d*), MAP2c$^{159–254}$ (*b* and *e*), and RD-hTH (*c* and *f*) obtained using the TIP3P (TIPS3P in the case of C22* and C36m) water model (*a–c*) and TIP4P-D water model (*d–f*). Experimental data are shown in gold, and values calculated using A99, C22*, and C36m are shown in green, red, and blue, respectively. To see this figure in color, go online.

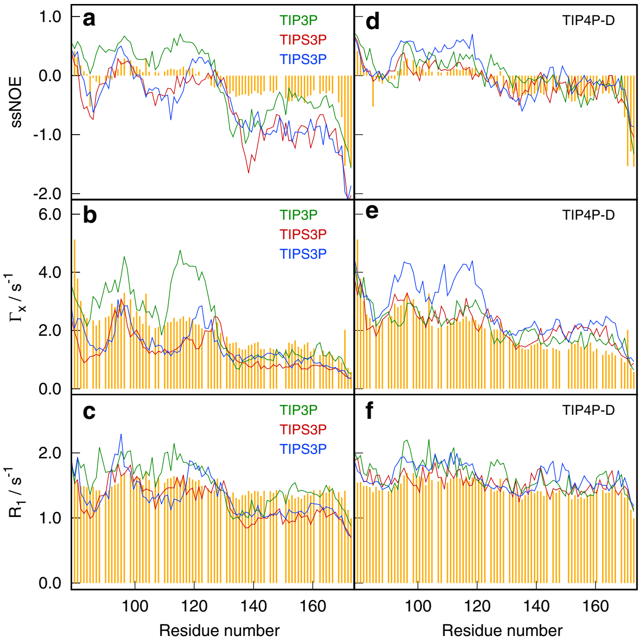


FIGURE 2 ssNOE (*a* and *d*) and relaxation rates Γ_x (*b* and *e*) and R_1 (*c* and *f*) of the C-terminal IDR of δRNAP obtained with the 5-ns sliding window using the TIP3P (TIPS3P in the case of C22* and C36m) water model (*a–c*) and TIP4P-D water model (*d–f*). Experimental data are shown in gold, and values calculated using A99, C22*, and C36m are shown in green, red, and blue, respectively. To see this figure in color, go online.

Interestingly, the water model influenced simulated R_g of δRNAP less than R_g of MAP2c$^{159–254}$ and RD-hTH. We observed the artificial collapsed structure only rarely in simulations of δRNAP with TIP3P or TIPS3P (see A99 data in Fig. 1 *a*). The most likely explanation is that the C-terminal IDR of δRNAP is very strongly negatively charged, preventing its collapsed structure regardless of applied water model (Fig. S2). To examine the influence of water models in the simulations of δRNAP more closely, we also calculated NMR relaxation parameters from the MD trajectories. As discussed below in more details, the prediction of relaxation parameters is a challenging task. Therefore, we hoped that a comparison of calculated and experimental relaxation parameters might reveal more subtle effects. Indeed, we observed much lower predicted values of ssNOE for residues 134–173 of δRNAP (Fig. 2 *a*). Such values indicate an artificially high disorder of the highly acidic region further than 30 residues from the positively charged lysine stretch (residues 96–104). Importantly, this effect was observed not only for the original TIP3P model but also for the modified version TIPS3P (*red* and *blue traces* in Fig. 2 *a*). Limitations of TIPS3P were already noticed by Huang et al., who reported that this model with C36m provided correct R_g for the arginine-serine but underestimated the R_g of the *Thermotoga maritima* cold-shock protein and of the N-terminal domain of HIV-1 integrase (28). A further modification of TIPS3P improved the R_g prediction for the cold-shock proteins but worsened the agreement for the other two proteins (28).

In conclusion, significant differences between water models were already observed during the first 200 ns of the simulations. Considering the results for TIP3P, TIPS3P, and TIP4P-D, we continued our study only with the TIP4P-D model, extending the MD simulations to the microsecond range.

Impact of force fields on global shape of proteins

We first examined the global shape of studied proteins in terms of R_g and SAXS curves. The R_g-values calculated from 1-μs (Fig. 1) and 0.5-μs (Figs. S3 and S4) A99, C22*, and C36m trajectories of δRNAP (simulated using TIP4P-D) oscillated between 3 and 6 nm. The experimental R_g-value, obtained by the Guinier analysis of the SAXS curve, was (3.45 ± 0.3) nm. The R_g alone did not reveal any systematic difference among the force fields, except for an extending event observed around 550 ns in the C36m simulation.

The R_g-values calculated from MD trajectories of MAP2c$^{159–254}$ oscillated between 2 and 4.5 nm, close to the experimental value of (2.5 ± 0.3) nm (Fig. 1 *e*).

Figs. S5 and S6 show the comparisons between the predicted and measured SAXS profiles for δRNAP and MAP2c$^{159–254}$ and for individual force fields. The lowest *Q*-factors, indicating the best fit (Table 1), were obtained

TABLE 1 Comparison of Calculated RMSD from Experimental Values and Normalized scores

Metric	δRNAP			RD-hTH			MAP2c$^{159–254}$		
	A99	C22*	C36m	A99	C22*	C36m	A99	C22*	C36m
RMSD:									
$\Delta\delta C^{\alpha}$/ppm	0.92	0.88	0.82	0.92	0.84	0.64	0.74	0.44	0.55
$\Delta\delta C^{\beta}$/ppm	0.86	0.78	0.79	0.63	0.52	0.43	0.45	0.45	0.41
$\Delta\delta$C(O)/ppm	0.74	0.67	0.62	0.96	0.90	0.56	0.84	0.55	0.65
$\Delta\delta$N/ppm	1.71	1.92	1.67	3.55	3.40	2.88	1.54	1.90	1.07
Q for $D(\mathrm{NH^N})$	0.30	0.33	0.33	1.21	1.12	0.89	0.76	0.69	0.74
Local PRE[a] by MTSL at L110C	0.25	0.41	0.26	n.d.[b]	n.d.	n.d.	n.d.	n.d.	n.d.
Local PRE by MTSL at L132C	0.07	0.17	0.16	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
Local PRE by MTSL at L151C	0.12	0.13	0.14	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
Local PRE by MTSL at L168C	0.06	0.06	0.07	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
PRE (all data)	0.08	0.10	0.09	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
ssNOE	0.15	0.17	0.20	0.11	0.38	0.25	0.15	0.20	0.16
R_2 or Γ_x/s^{-1}	1.50	1.56	2.46	2.62	2.32	0.93	1.13	1.32	1.09
R_1/s^{-1}	0.12	0.15	0.14	0.19	0.24	0.13	0.21	0.19	0.23
Q for SAXS (0.1 nm^{-1} < q < 1 nm^{-1})	0.057	0.040	0.084	n.d.	n.d.	n.d.	0.070	0.048	0.075
Q for SAXS (all data)	0.060	0.045	0.085	n.d.	n.d.	n.d.	0.084	0.066	0.083
Score:									
s_{CS}	1.11	1.08	1.00	1.46	1.32	1.00	1.42	1.22	1.11
s_{RDC}	1.00	1.09	1.08	1.33	1.20	1.00	1.06	1.00	1.36
s_{PRE} (all data)	1.00	1.25	1.11	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
s_{relax}	1.00	1.13	1.37	1.74	2.57	1.44	1.04	1.16	1.07
s_{NMR}	1.00	1.13	1.44	1.54	1.89	1.22	1.06	1.12	1.07
s_{SAXS} (all data)	1.32	1.00	1.87	n.d.	n.d.	n.d.	1.28	1.00	1.26
s_{all}	1.08	1.08	1.21	1.55	1.78	1.17	1.24	1.15	1.11
R_g/nm	4.13	4.03	4.66	n.d.	n.d.	n.d.	2.60	2.94	2.78
R_g penalty	0.13	0.10	0.28	n.d.	n.d.	n.d.	0	0.06	0
s_{combined}	1.21	1.18	1.49	1.55	1.78	1.17	1.24	1.21	1.11

[a] Calculated for PRE of residues 84–104 because of the indicated spin label.

[b] Input experimental data not determined.

for the C22* force field. Inspection of individual SAXS profiles revealed that C36m and A99 overestimated SAXS intensities for MAP2c$^{159–254}$ and underestimated SAXS intensities for δRNAP in the medium- and low-q region (0.1 nm^{-1} < q < 1 nm^{-1}; see Table 1), whereas C22* predicted the medium- and low-q SAXS intensities well.

We also fitted the experimental and predicted SAXS profiles to molecular form factors (MFFs) developed by Riback et al. (25). Fitting the experimental MAP2c$^{159–254}$ data to an MFF provided $R_g = (2.87 \pm 0.3)$ nm and a value of the Flory exponent typical for a fully unfolded polypeptide ($\nu = 0.59 \pm 0.02$). Very similar values were obtained by fitting the profile simulated by C22*, $R_g = (2.952 \pm 0.003)$ nm and $\nu = 0.603 \pm 0.001$. As expected, the MFF did not fit well the SAXS profile of δRNAP, containing a large well-ordered domain.

The effect of a force field on the global conformations of RD-hTH was examined as well (Fig. 1 *f*). Compact structures with $R_g \approx 2$ nm were formed after 500 ns in simulations with A99 and C22* force fields but not with C36m. Therefore, it seems that CH36m in combination with TIP4P-D most efficiently prevents the collapse of structures of RD-hTH and MAP2c, representing proteins not exhibiting extraordinary electrostatic repulsion.

Impact of force fields on long-range contacts

After evaluating the effect of different force-field parameters on the global shapes of the studied proteins, we compared the ability of the force fields to properly reproduce contacts between residues further apart in the sequence. We started by inspecting δRNAP as a test system for which long-range contacts are observed experimentally, yet no transient helicity is observed in the C-terminal IDR.

Residue pairwise distance maps represent an efficient way to evaluate contacts formed during individual MD runs. Rectangular boxes in the maps presented in Fig. 3 highlight distances 1) between the lysine-rich stretch ^{96}KAKKKKAKK^{104} and highly acidic residues in the C-terminal region and 2) between the ordered N-terminal and disordered C-terminal domains. Differences between distances represented by colors document that the relative average distances varied depending on the force field used. For example, the lysine tract interacted with the residues in the vicinity of Glu120 more strongly in the A99 simulation than in the runs with the C22* and C36m force fields (*blue rectangles* in Fig. 3). The average distances between the N-terminal domain and vicinity of Glu120 (*red rectangles* in Fig. 3) also differed. We want to emphasize that the reliability of distance maps is influenced not only by the capabilities of individual force fields applied for

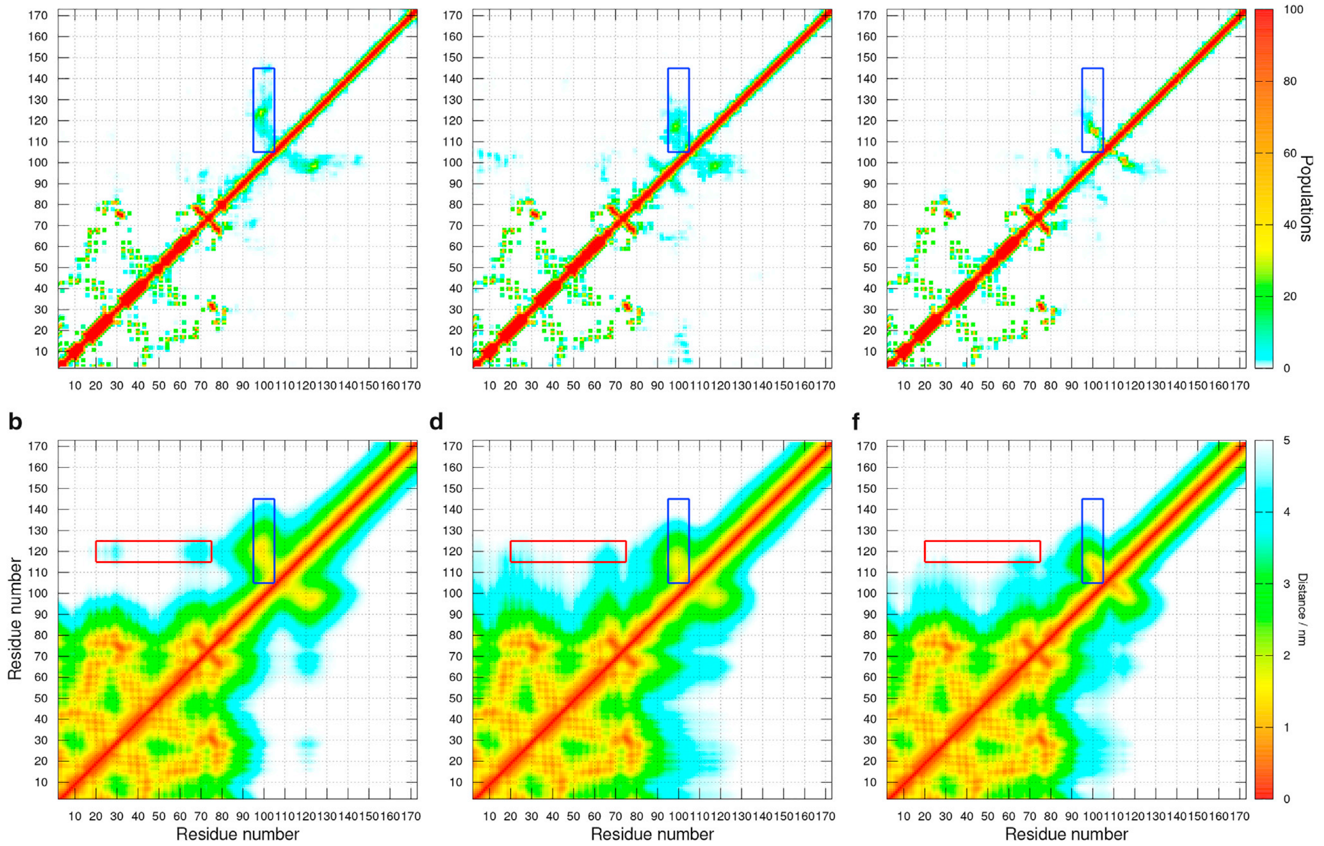


FIGURE 3 Maps describing the distances between residue pairs of δRNAP simulated using A99 (*a* and *b*), C22* (*c* and *d*), and C36m (*e* and *f*) force fields in combination with TIP4P-D. The scales shown at the right indicate color coding of mean inter-residue C^{α}-C^{α} distances (*lower row*) and of populations of events when the distance between any pair of residue atoms was shorter than 0.4 nm (*upper row*). To see this figure in color, go online.

the δRNAP in combination with the TIP4P-D water model but also by limited sampling within the trajectories of the cumulative length of 1.5 μs.

To directly compare the contacts observed during the MD calculations with the experimental data, we simulated PRE of individual residues of δRNAP for a series of spin-label positions examined in a previous experimental study (11). The results showed that A99 realistically reproduced experimentally observed contacts of the lysine stretch ^{96}KAKKKKAKK^{104} with labels placed at L110C and L132C in the highly acidic C-terminal sequence (Fig. 4, *a* and *b*). The experimentally detected contact with L151C was also predicted. In agreement with the distance map, C22* overestimated contacts of the lysines with the closest label at L110C and underestimated contacts with more distant labels, including L132C. C36m did not overestimate contacts with the label at L110C but underestimated contacts with the label at L132C similarly to C22* and A99 (see RMSD for PRE in the lysine stretch for individual spin labels, listed in Table 1).

In conclusion, correct prediction of electrostatic contacts between residues apart in the sequence is a challenging task. In the case of δRNAP, A99 with TIP4P-D performed best, being able to predict reliably distances between amino acids separated by up to 50 residues in the sequence.

Impact of protein force-field parameters on local conformations

In the next step, we analyzed the accuracy of description of local backbone conformations of δRNAP in the simulations. For this purpose, we compared experimental values of several NMR parameters reflecting the local backbone conformation with the values calculated from snapshots of the MD simulations. First, we checked values of RDC. In principle, RDCs depend both on local conformation and on the overall shape of the molecule, determining the distribution of orientations of the molecule in a partially aligned environment (20). However, it is very demanding to achieve a good sampling of conformations and orientations to faithfully reproduce experimental data without any prior information. Therefore, we used the knowledge of long-range electrostatic contacts in the δRNAP molecule, obtained experimentally as PRE, and applied the local averaging window (40) to predict RDC values. The predicted RDC values are plotted in Fig. 5 *a*. The prediction varied especially in the vicinity of the lysine stretch, in

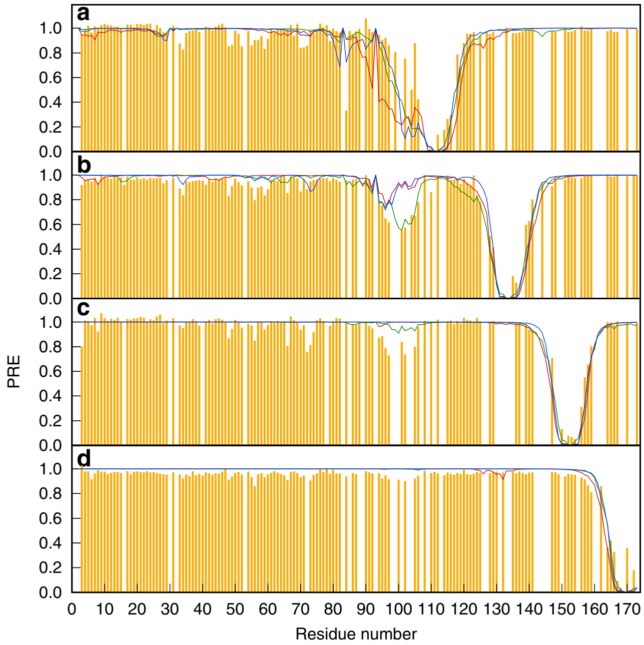


FIGURE 4 Simulated PRE of δRNAP with the spin label at L110C (*a*), L132C (*b*), L151C (*c*), and E168C (*d*). Experimental data are shown in gold, and values calculated from MD simulations using the TIP4P-D water model combined with the A99, C22*, and C36m force fields are shown in green, red, and blue, respectively. To see this figure in color, go online.

which A99 and C36m achieved better agreement with the experiment than C22*.

The second examined NMR parameter reflecting local conformation was the chemical shift. In comparison with RDC, the chemical shifts are measured with higher precision, and their prediction does not require a prior knowledge of molecular orientation. SCSs (deviations of predicted chemical shifts from their random-coil values (38)) are compared with the experimental data in Fig. 5, *b–e*. Values provided by all force fields agree well with the experimental chemical shifts, except for some mismatch of data predicted by C22* for the lysine stretch.

In summary, all force fields predicted local conformation of the extended disordered region of δRNA reasonably well. The slightly worse prediction in the lysine-rich motif by C22* most likely reflects less accurate description of electrostatic contacts by the force field.

Simulation of transient helical regions

In the next step, we tested how different force fields describe transient α-helical elements in (partially) disordered proteins. A propensity to adopt α-helical secondary structure represents another level of complexity of IDP conformations not present in the mostly extended C-terminal domain of δRNAP. To explore its effect on the simulations, we inspected MD trajectories of RD-hTH and MAP2c$^{159-254}$. These proteins were chosen so that they differ in α-helical propensity.

Experimental values of chemical shifts (17,19) indicate that RD-hTH and MAP2c$^{159-254}$ form α-helices in the regions 40–53 and 200–216 with ~80 and 20% propensity, respectively (cf. Fig. 5). We performed a set of simulations with different force fields, starting from structures containing ideal α-helices in the experimentally identified α-helical regions, and observed the stability of the helices (Figs. S10–S12). During MD simulations using A99 and C22* force fields, the initially present α-helix unfolded in less than 80 ns (Fig. S12). Huang et al. (28) reported that C36m optimized with TIPS3P 1) correctly simulated transient α-helices and 2) provided correct R_g for the arginine-serine peptide but not for other two tested IDPs. Therefore, we were curious whether C36m would keep its ability to maintain the transient α-helices with TIP4P-D. For RD-hTH, C36m with TIP4P-D maintained the α-helical conformation in the whole 1-μs trajectory and for 220 ns in an independent 400-ns run (Fig. S13). In the case of MAP2c$^{159-254}$, the transient α-helix unfolded after ~35, 85, and 830 ns in three independent runs starting from conformations including the helix at the beginning (Fig. S9). The lower stability of the MAP2c$^{159-254}$ helix in simulations is in agreement with its low (20%) population observed experimentally. Formation of the well-defined helix was not observed in the remaining 465, 415, and 170 ns of the trajectories or during two 0.5-μs runs started from conformations without the helix. However, temporary formation of the helix (for ~50 ns) was observed in simulations of MAP2c$^{159-254}$ using C22* (Fig. S12 *a*) and A99 (Fig. S12 *d*). It should be emphasized that the calculated trajectories do not fully sample the equilibrium canonical ensembles. Therefore, we do not expect to observe quantitative agreement between the experimental and simulated populations of transient α helices.

The ability of the force fields to reliably describe transient α-helices was directly reflected by the predicted SCS values (Fig. 5). Outside of the helical region, a good agreement of the predicted and experimental chemical shifts was obtained for all force fields tested. Deviations from the experimental values were observed for the A99 and C22* simulations in the region where the originally present α-helix unfolded. For the C36m simulations, SCSs typical for α-helices were obtained in the regions where the α-helix was modeled (Fig. 5, *g–i* and *l–n*). Quantitatively, the values of predicted versus SCSs corresponded to 87% populations of the helix in the simulations vs. ~80% population in the real sample of RD-hTH. In the case of MAP2c$^{159-254}$, data from all trajectories were combined in a ratio corresponding to the experimentally estimated 20% population of the helix.

In conclusion, the ability of C36m and TIP4P-D to keep the transient α-helices during the simulation and to prevent the artificial collapse of IDR structures suggests that this combination is most robust for our studied proteins. This is noteworthy considering that C36m was not optimized in combination with TIP4P-D, but with TIPS3P and

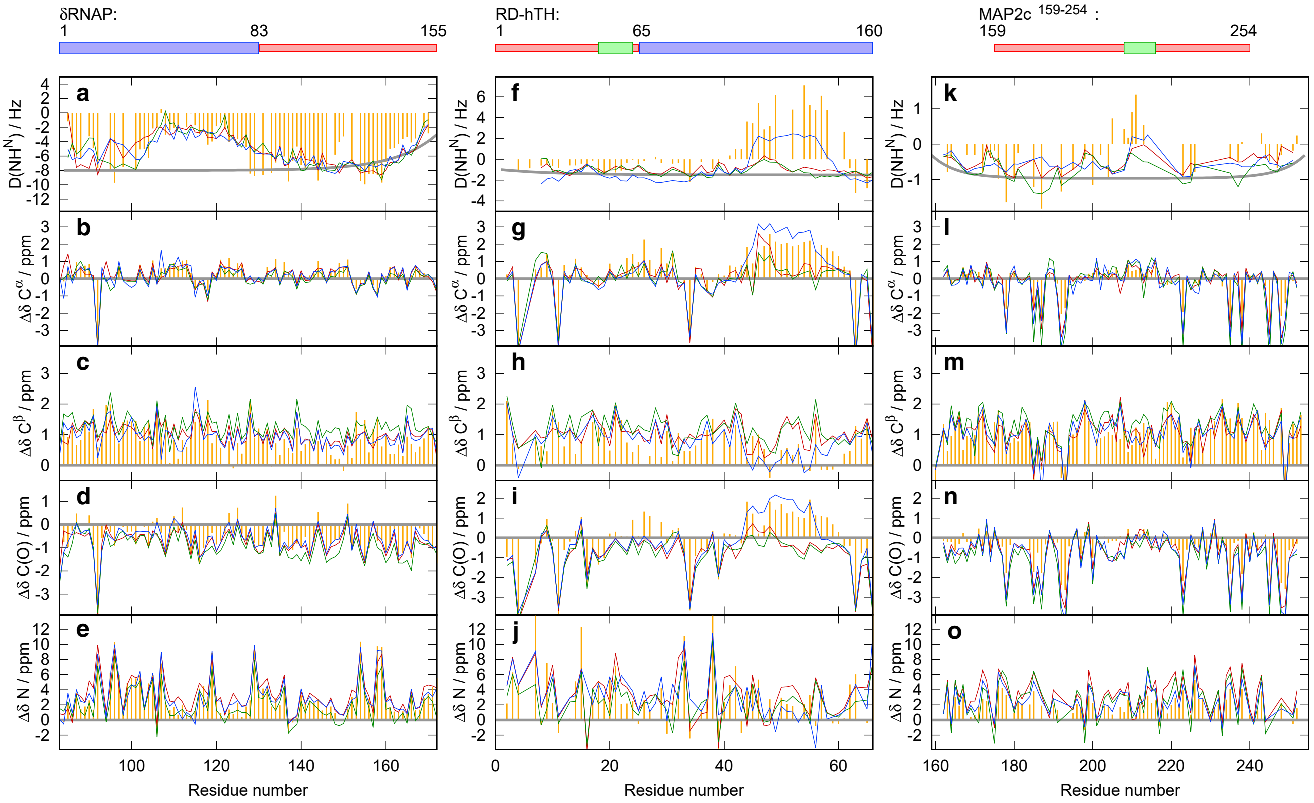


FIGURE 5 Values of RDC (*a*, *f*, and *k*) and SCS (*b–e*, *g–j*, and *l–o*) in C-terminal IDR (residues 85–173) of δRNAP (*a–e*), N-terminal IDR (residues 1–65) of RD-hTH (*f–j*), and MAP2c$^{159–254}$ (*k–o*) simulated with the TIP4P-D water model. Experimental data are shown in gold, and values calculated using A99, C22*, and C36m are shown in green, red, and blue, respectively. The random-coil limits are shown in gray. The MD simulations started from structures with α-helices modeled for residues 40–53 of RD-hTH and 200–216 of MAP2c$^{159–254}$. For the sake of clarity, the statistical errors are not displayed here but separately in Figs. S7–S9. To see this figure in color, go online.

its variants (28). We believe that the C36m and TIP4P-D combination is a promising candidate for a benchmarking on a wider range of proteins, as was done, e.g., by Robustelli et al. (45).

It should be noted that the tests discussed so far utilize a limited set of the most readily available experimental values. We already discussed that NMR relaxation data distinguished the performance of TIP3P and TIP4P-D water models better than R_g in the A99 and C36m simulations of δRNAP (details are presented in the section Impact of Selected Water Model). In the next section, we examine whether extending the benchmarking to NMR relaxation helps to discriminate the ability of force fields to describe dynamic properties of hybrid proteins.

Simulation of NMR relaxation data

To test the examined force fields within the MD framework more thoroughly, we compared their abilities to reproduce NMR relaxation rates. The NMR signal relaxes because of the local magnetic fields that fluctuate as a result of stochastic molecular motions. The stochastic reorientation of vectors describing the interactions contributing to relaxation is described by the correlation function. In real samples, large numbers of molecules with an almost isotropic distribution of orientations are measured. Consequently, the correlation functions have a simple analytical form (series of exponential functions). In simulations, the sufficient sampling of the orientations is difficult to achieve (in principle, data of large sets of independent trajectories should be averaged), which deteriorates the calculated correlation function regardless of the force field employed. In our analyses of limited numbers of trajectories, the ensemble averaging was approximated by calculating averages of correlation functions of different regions of the trajectories. The simulations should be also sufficiently long because the correlation function reflects only the effects of motions on the timescale covered by the simulation. The obtained predictions must be interpreted carefully to not confuse the artifacts of the force fields with the effects of the sampling scheme used.

The δRNAP molecule is particularly well suited to test the ability of force fields to predict relaxation rates because its domains greatly differ in their stochastic motions. Dynamics of the well-ordered N-terminal domain is dominated by overall tumbling and probes the ability of the force fields to reproduce hydrodynamic properties. In contrast, internal motions contribute most significantly to the relaxation rates

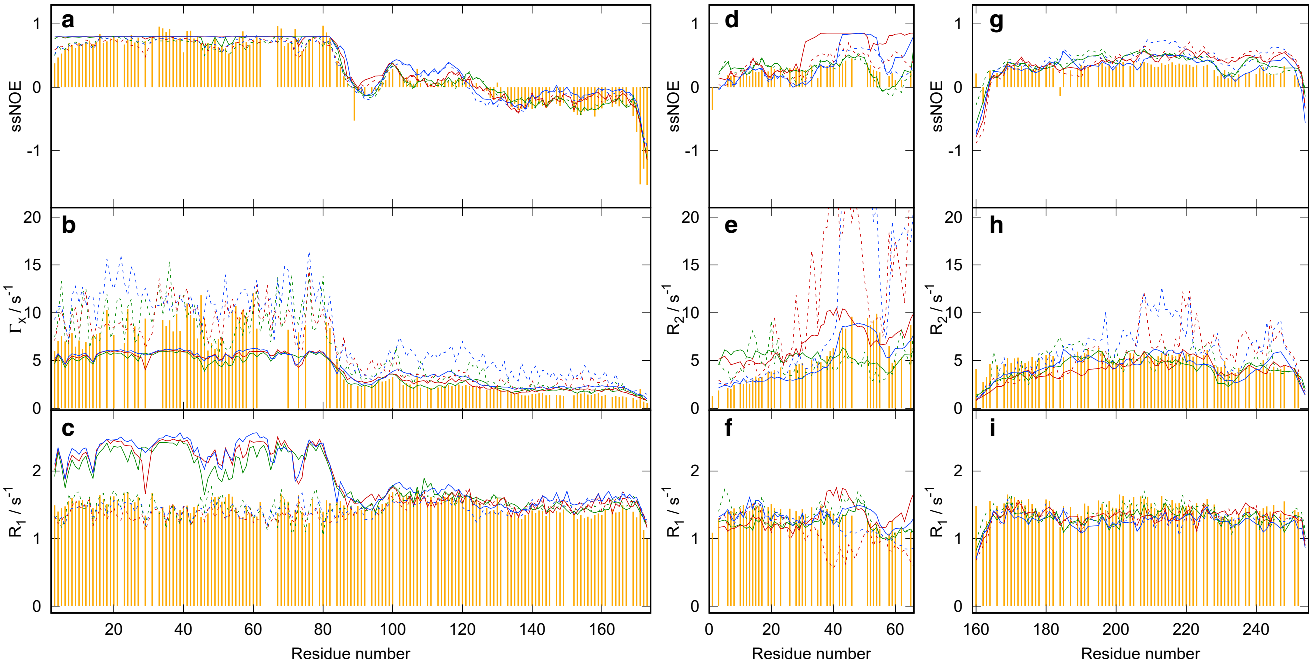


FIGURE 6 ssNOE (*a*, *d*, and *g*) and relaxation rates Γ_x (*b*), R_2 (*e* and *h*), and R_1 (*c*, *f*, and *i*) of δRNAP (*a–c*), N-terminal IDR (residues 1–65) of RD-hTH (*d–f*), and MAP2c$^{159–254}$ (*g–i*). Experimental data are shown in gold, and values calculated using A99, C22*, and C36m are shown in green, red, and blue, respectively. Solid and dashed lines indicate data obtained from correlation functions calculated using 5- and 50-ns windows, respectively. For the sake of clarity, the statistical errors are not displayed here but separately in Figs. S13–S15. To see this figure in color, go online.

of residues in the disordered C-terminal domain. Analysis of the simulated trajectories showed that correlation functions calculated for 5-ns sliding time windows (*solid lines* in Fig. 6, *a–c*) describe relaxation in the C-terminal IDR sufficiently well but fail to match the experimental data in the well-ordered region, where slower motions, most notably the overall tumbling, dominate the dynamics. Smooth profiles of the calculated NMR relaxation parameters, resembling the experimental profiles in the IDR, document that the use of a short sliding window allowed us to average a sufficient number of correlation functions and to capture most important modes of motion.

To obtain relaxation rates close to the experimental values also in the well-ordered N-terminal region, the time window was extended to 50 ns, and averages of 12 correlation functions were calculated (*dashed lines* in Fig. 6, *a–c*). Prediction of the cross-correlated Γ_x rate, which is most sensitive to slow motions, was most informative (Γ_x was used to monitor the slow motions instead of the more frequently measured R_2 rates because Γ_x could be obtained with a higher accuracy than R_2 for δRNAP, which has an extremely poor dispersion of chemical shifts in its C-terminal IDR (21)). Predicted and experimental values were comparable for tested force fields, albeit scattered for individual residues because of the contribution of slow motions that were not sufficiently averaged in the small set of the 30 independent correlation functions. The match of the experimental and (average) simulated Γ_x-values is in the line with the fact that TIP4P-D reproduces the water diffusion coefficient more reliably than the TIP3P model (31).

The general agreement was also good in the disordered region for all three force field, but significant differences were observed when different regions of the δRNAP sequence were compared. The trend of ssNOE values, most sensitive to fast motions, was best reproduced by A99 (*green line* in Fig. 6 *a*). Also, the predictions by CHARMM force fields reflected their abilities to predict long-range contacts. Somewhat higher ssNOE values (and elevated Γ_x) in the vicinity of residues 100 and 125 indicated that C36m slightly overestimated partial ordering in the most rigid regions of the C-terminal domain, where long-range contacts were predicted (see the section Impact of the Force Fields on Long-Range Contacts). This is in agreement with the difficulty of predicting contacts between residues far in the sequence, as discussed above. C22* overestimated ssNOE around residues 115 and 90, in agreement with its tendency to prefer contacts of the lysine stretch with residues closer in the sequence.

In conclusion, we calculated NMR relaxation rates from MD trajectories using two time windows (50 and 5 ns) to cover different timescales of motions in the ordered and disordered regions, respectively. The results independently confirmed the observed moderate differences in predicting long-range contacts and showed that all force fields describe well the hydrodynamic properties for the TIP4P-D water model. For δRNAP with a well-ordered domain and highly charged disordered domains not forming transient helical structures, A99 performed best. The C36m force field described long-range electrostatic contacts slightly worse, but its accuracy was acceptable. C22* somewhat overestimated electrostatic contacts between residues close in the sequence, which resulted in noticeable, but not dramatic, deviations of simulated NMR parameters from the experiment.

NMR relaxation data were also predicted for RD-hTH and MAP2c$^{159–254}$. In the case of RD-hTH, the comparison was possible only in the N-terminal IDR because the protein dimerizes in real samples. As a consequence, relaxation of the well-ordered portion of RD-hTH is incomparable (faster) with that simulated for the monomer. Moreover, broadening of the peaks of the well-ordered region did not allow us to obtain sufficiently sensitive NMR spectra under the conditions used. Comparison of the simulated and experimental relaxation rates in the N-terminal IDR of RD-hTH (Fig. 6, *d–f*) and MAP2c$^{159–254}$ (Fig. 6, *g–i*) led to the same general conclusions as for δRNAP. However, the simulation of RD-hTH allowed us to address a particular feature not manifested by δRNAP, namely the effect of formation of the transient α-helix on the calculated relaxation rates. In the experimental data, the propensity to form an α-helix is reflected by elevated R_2 values. Comparison of the relaxation rates calculated from the C36m trajectories (in which the helix was present during most of the simulation time) with those obtained from the A99 and C22* runs (in which the helix quickly unfolded) revealed that the presence of the α-helix in the simulated structures is needed to reproduce the values of R_2 (Fig. 6). The 5-ns sliding window was sufficient to match the experimental R_2 profile in the C36m simulations, indicating that the dynamics of the IDR of RD-hTH is dominated by relatively short correlation times. Less frequent conformational changes occurring in the α-helical region were not sampled sufficiently and resulted in high standard deviations of R_2. Outside of the transient α-helix, the data obtained with the 50-ns window matched the experimental profile well.

In the case of MAP2c$^{159–254}$, which has a much lower α-helical propensity, the increase of R_2 is hardly visible in the experimental data. Similarly to RD-hTH, relaxation data predicted using the 5-ns window matched the experimental values reasonably well.

Quantitative comparison of the force-field reliability

To express the discussed differences of the force field performance quantitatively, we applied the metrics developed by Robustelli et al. (45) and calculated normalized force-field scores based on the RMSDs of the experimentally obtained parameters from the corresponding values predicted from the simulations (Table 1). The calculated combined force-field scores (s_{combined}) show that none of the tested force fields provided superior prediction of all parameters for all proteins.

In the case of δRNAP with a highly charged IDR forming no transient α-helices, relative accuracy of predicting experimental data varied for different parameters. Predicted chemical shifts, reflecting the local backbone conformation, was similar for all force fields (best for C36m). A99 best reproduced NMR data sensitive to long-range intramolecular interactions (RDC, PRE, and relaxation data, described by s_{NMR}). The chemical shift RMSD values are within the reported standard deviation of the SPARTA+ predictor (2.45, 1.09, 0.94, and 1.14 ppm for backbone N, C(O), C$^{\alpha}$, and C$^{\beta}$ nuclei (37)). The quality factor Q of RDC is comparable with typical RMSDs of data predicted from x-ray structures with 2-Å resolution (46). PRE and relaxation data deviated from the experimental ranges by less than 10%, with the exception of Γ_x, which is particularly difficult to predict, as discussed above. C22* provided the best prediction of parameters describing the overall shape of δRNAP (R_g and SAXS profiles), with the Q-factor of SAXS data equal to 5%. If the individual scores are averaged with the same weights, the overall scores s_{all} are better for C22* and A99 than for C36m.

In the case of RD-hTH, with the experimentally determined 80% propensity to form an α-helix in its IDR, C36m predicted all experimental data much better than A99 or C22*. The superior performance of C36m, reflected by low s_{all}, can be clearly attributed to its ability to maintain the experimentally observed transient α-helix. The quantitative parameters showed that C36m predicted the experimental data well (compared with the reference values discussed above), with the exception of RDC.

In the case of MAP2c$^{159–254}$, which lacks a well-ordered domain and exhibits only ~20% propensity to form an α-helix, all force fields predicted the experimental data with similar accuracy, reflected by small differences between their scores. It documents that the ability of C36m to maintain transient α-helices loses its significance if the populations of the helical structures are low. The performance of all force fields was good, based on comparison with the reference values discussed in the details of the δRNAP simulations.

In conclusion, the quantitative comparison confirmed the superior performance of C36m in the case when a transient α-helix was present. A99 predicted most accurately parameters influenced by long-range electrostatic interactions ($s_{\text{NMR}} = 1$ for δRNAP).

Prediction of suitable spin-label positions

Calculation of already measured experimental parameters is important for benchmarking of the simulations as illustrated above. However, prediction of so far unknown measurable values is also very useful because it can facilitate experimental design. Selection of residues for placement of paramagnetic labels can serve as an example. Preparation of paramagnetically labeled samples is a time-consuming procedure, including site-directed mutagenesis, expression, purification, and paramagnetic spin labeling of the protein before the NMR PRE measurements. A choice of the label position not providing information about long-range contacts thus represents a considerable waste of time and sources. Reliable MD simulations can be used for in silico prediction how much structural information a label in a certain position can provide and how much such labeling perturbs the native structural ensemble. An example of such prediction for all solvent accessible positions within the RD-hTH is presented in Fig. 7.

Based on the simulations using C36m and TIP4P-D, we calculated PRE profiles for all possible positions. To localize solvent accessible positions reporting on long-range contacts with the disordered region, we defined the following score. Integrals of the areas between PRE profiles and a threshold of 0.8 were summed in the disordered region (residues 1–70), except for ±10 residues in the vicinity of the spin label. In addition, the score was set to zero for residues with solvent accessibility lower than 0.6 according to Fraczkiewicz and Braun (47) because the label should be freely accessible on the surface of protein and should not interfere with any protein conformation. The analysis presented in Fig. 7 *a* indicates four to five areas well suited for attachment of spin labels for future PRE experiments. Predicted PRE profiles for labels placed in the centers of the suggested areas are presented in Fig. 7, *b–e*.

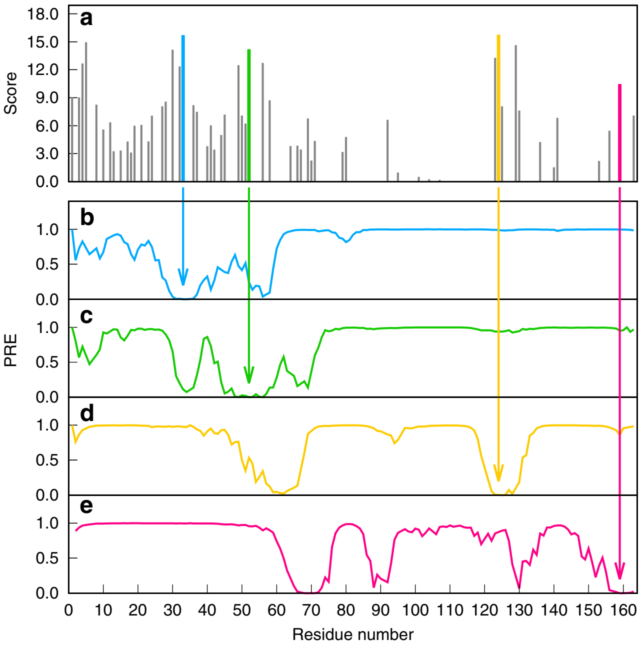


FIGURE 7 Predicted preferential positions of spin labels in RD-hTH (*a*) and simulated PRE profiles for four selected spin-label positions (*b–e*). To see this figure in color, go online.

CONCLUSIONS

The goal of our study was to assess the applicability of currently available MD approaches to hybrid proteins consisting of ordered and disordered regions. We tested the performance of the A99, C22*, and C36m force fields in combination with the TIP3P and TIP4P-D water models. The performance was examined for mostly disordered MAP2c$^{159–254}$ containing a low population of prestructured α-helix, for δRNAP consisting of comparably large well-ordered and disordered domains without transient α-helices, and for RD-hTH consisting of ordered and disordered domains with a highly populated α-helical prestructured motif. Considering the functional importance of transient helical elements, we paid particular attention to the ability to preserve the α-helix during the simulation. The generated structural ensembles were used for predicting a variety of NMR and SAXS parameters and subsequently compared with the experimental data.

The TIP4P-D water model performed substantially better than TIP3P or TIPS3P. The differences between force fields were less distinct. The optimal (most universal) combination was CHARMM36m with the TIP4P-D water model, which most efficiently prevented artificial collapse of disordered regions and retained transient α-helical structure elements within the disordered regions. The study also showed that the performance of different force fields and models can vary depending on the actual physical properties of investigated IDRs.

Considering the fast pace of force-field development (45,48), the major value of this study is not identification of the best force field available at the time of testing but presentation of a generally applicable benchmarking approach, including NMR relaxation rates. An important feature of the approach is the combination of checked parameters that report on abilities of the force fields to reproduce a broad range of physical properties of the studied molecules.

SUPPORTING MATERIAL

Supporting Material can be found online at https://doi.org/10.1016/j.bpj.2020.02.019.

AUTHOR CONTRIBUTIONS

J.H. and L.Ž. designed the research. V.Z., A.M., P.L., A.L., E.N., and V.K. carried out calculations, performed the experiment, and analyzed the data. Z.J., M.M., and K.M. prepared the samples and performed the experiment. V.Z., A.M., P.L., J.H., and L.Ž. wrote the article, and all authors reviewed the manuscript.

ACKNOWLEDGMENTS

This research was funded by the Ministry of Education, Youth, and Sport of the Czech Republic (MEYS CR), grant numbers LTC17078

(Inter-Excellence Inter-Cost), LTAUSA18168 (Inter-Excellence Inter-Action), and LQ1601 (National Sustainability Programme II Project CEITEC 2020). Computational resources were provided by CESNET (LM2015042) and the CERIT Scientific Cloud (LM2015085) under the program "Projects of Large Research, Development, and Innovations Infrastructures" funded by MEYS CR and by IT4Innovations National Supercomputing Center (LM2015070) under the program "Large Infrastructures for Research, Experimental Development and Innovations" funded by MEYS CR. The Czech Infrastructure for Integrative Structural Biology (CIISB) research infrastructure project LM2018127 funded by MEYS CR is gratefully acknowledged for the partial financial support of the measurements at the Josef Dadok National NMR Centre and at X-Ray Diffraction and Bio-SAXS Core Facilities, CEITEC-Masaryk University.

REFERENCES

1. Uversky, V. N. 2002. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* 11:739–756.
2. Dunker, A. K., C. J. Brown, …, Z. Obradović. 2002. Intrinsic disorder and protein function. *Biochemistry.* 41:6573–6582.
3. Tompa, P. 2011. Unstructural biology coming of age. *Curr. Opin. Struct. Biol.* 21:419–425.
4. Chi, S.-W., D.-H. Kim, …, K. H. Han. 2007. Pre-structured motifs in the natively unstructured preS1 surface antigen of hepatitis B virus. *Protein Sci.* 16:2108–2117.
5. Fuxreiter, M., I. Simon, …, P. Tompa. 2004. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.* 338:1015–1026.
6. Vacic, V., C. J. Oldfield, …, A. K. Dunker. 2007. Characterization of molecular recognition features, MoRFs, and their binding partners. *J. Proteome Res.* 6:2351–2366.
7. Nováček, J., L. Židek, and V. Sklenář. 2014. Toward optimal-resolution NMR of intrinsically disordered proteins. *J. Magn. Reson.* 241:41–52.
8. Nowakowski, M., S. Saxena, …, W. Koźmiński. 2015. Applications of high dimensionality experiments to biomolecular NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* 90–91:49–73.
9. Papoian, G. A. 2008. Proteins with weakly funneled energy landscapes challenge the classical structure-function paradigm. *Proc. Natl. Acad. Sci. USA.* 105:14237–14238.
10. Rabatinová, A., H. Šanderová, …, L. Krásný. 2013. The δ subunit of RNA polymerase is required for rapid changes in gene expression and competitive fitness of the cell. *J. Bacteriol.* 195:2603–2611.
11. Papoušková, V., P. Kadeřávek, …, L. Žídek. 2013. Structural study of the partially disordered full-length δ subunit of RNA polymerase from Bacillus subtilis. *ChemBioChem.* 14:1772–1779.
12. Nagatsu, T., M. Levitt, and S. Udenfriend. 1964. Tyrosine hydroxylase. The initial step in norepinephrine biosynthesis. *J. Biol. Chem.* 239:2910–2917.
13. Molinoff, P. B., and J. Axelrod. 1971. Biochemistry of catecholamines. *Annu. Rev. Biochem.* 40:465–500.
14. Louša, P., H. Nedozrálová, …, J. Hritz. 2017. Phosphorylation of the regulatory domain of human tyrosine hydroxylase 1 monitored using non-uniformly sampled NMR. *Biophys. Chem.* 223:25–29.
15. Jansen, S., K. Melková, …, L. Žídek. 2017. Quantitative mapping of microtubule-associated protein 2c (MAP2c) phosphorylation and regulatory protein 14-3-3ζ-binding sites reveals key differences between MAP2c and its homolog Tau. *J. Biol. Chem.* 292:6715–6727.
16. Melková, K., V. Zapletal, …, L. Žídek. 2018. Functionally specific binding regions of microtubule-associated protein 2c exhibit distinct conformations and dynamics. *J. Biol. Chem.* 293:13297–13309.
17. Melková, K., V. Zapletal, …, L. Žídek. 2019. Structure and functions of microtubule associated proteins Tau and MAP2c: similarities and differences. *Biomolecules.* 9:E105.
18. Motáčková, V., J. Nováček, …, V. Sklenář. 2010. Strategy for complete NMR assignment of disordered proteins with highly repetitive sequences based on resolution-enhanced 5D experiments. *J. Biomol. NMR.* 48:169–177.
19. Nováček, J., L. Janda, …, V. Sklenář. 2013. Efficient protocol for backbone and side-chain assignments of large, intrinsically disordered proteins: transient secondary structure analysis of 49.2 kDa microtubule associated protein 2c. *J. Biomol. NMR.* 56:291–301.
20. Ottiger, M., F. Delaglio, and A. Bax. 1998. Measurement of J and dipolar couplings from simplified two-dimensional NMR spectra. *J. Magn. Reson.* 131:373–378.
21. Srb, P., J. Nováček, …, L. Žídek. 2017. Triple resonance ^{15}N NMR relaxation experiments for studies of intrinsically disordered proteins. *J. Biomol. NMR.* 69:133–146.
22. Korzhnev, D. M., M. Billeter, …, V. Y. Orekhov. 2001. NMR studies of Brownian tumbling and internal motions in proteins. *Prog. Nucl. Magn. Reson. Spectrosc.* 38:197–266.
23. Efron, B. 1979. Bootstrap methods: another look at the jackknife. *Ann. Stat.* 7:1–26.
24. Petoukhov, M. V., D. Franke, …, D. I. Svergun. 2012. New developments in the *ATSAS* program package for small-angle scattering data analysis. *J. Appl. Cryst.* 45:342–350.
25. Riback, J. A., M. A. Bowman, …, T. R. Sosnick. 2017. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science.* 358:238–241.
26. Lindorff-Larsen, K., S. Piana, …, D. E. Shaw. 2010. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins.* 78:1950–1958.
27. Piana, S., K. Lindorff-Larsen, and D. E. Shaw. 2011. How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* 100:L47–L49.
28. Huang, J., S. Rauscher, …, A. D. MacKerell, Jr. 2017. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods.* 14:71–73.
29. Jorgensen, W. L. 1981. Quantum and statistical mechanical studies of liquids. 10. Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water. *J. Am. Chem. Soc.* 103:335–340.
30. MacKerell, A. D., D. Bashford, …, M. Karplus. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* 102:3586–3616.
31. Piana, S., A. G. Donchev, …, D. E. Shaw. 2015. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B.* 119:5113–5123.
32. Hess, B., H. Bekker, …, J. G. E. M. Fraaije. 1997. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* 18:1463–1472.
33. Essmann, U., L. Perera, …, L. G. Pedersen. 1995. A smooth particle mesh Ewald method. *J. Chem. Phys.* 103:8577–8593.
34. Berendsen, H. J. C., J. P. M. Postma, …, J. R. Haak. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–3690.
35. Bussi, G., D. Donadio, and M. Parrinello. 2007. Canonical sampling through velocity rescaling. *J. Chem. Phys.* 126:014101.
36. Parrinello, M., and A. Rahman. 1981. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* 52:7182–7190.
37. Shen, Y., and A. Bax. 2010. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR.* 48:13–22.
38. Nielsen, J. T., and F. A. A. Mulder. 2018. POTENCI: prediction of temperature, neighbor and pH-corrected chemical shifts for intrinsically disordered proteins. *J. Biomol. NMR.* 70:141–165.
39. Zweckstetter, M. 2008. NMR: prediction of molecular alignment from structure using the PALES software. *Nat. Protoc.* 3:679–690.
40. Nodet, G., L. Salmon, …, M. Blackledge. 2009. Quantitative description of backbone conformational sampling of unfolded proteins at

amino acid resolution from NMR residual dipolar couplings. *J. Am. Chem. Soc.* 131:17908–17918.

41. Salmon, L., G. Nodet, …, M. Blackledge. 2010. NMR characterization of long-range order in intrinsically disordered proteins. *J. Am. Chem. Soc.* 132:8407–8418.
42. Sezer, D., J. H. Freed, and B. Roux. 2008. Simulating electron spin resonance spectra of nitroxide spin labels from molecular dynamics and stochastic trajectories. *J. Chem. Phys.* 128:165106.
43. Salvi, N., A. Abyzov, and M. Blackledge. 2016. Multi-timescale dynamics in intrinsically disordered proteins from NMR relaxation and molecular simulation. *J. Phys. Chem. Lett.* 7:2483–2489.
44. Urbańczyk, M., D. Bernin, …, K. Kazimierczuk. 2013. Iterative thresholding algorithm for multiexponential decay applied to PGSE NMR data. *Anal. Chem.* 85:1828–1833.
45. Robustelli, P., S. Piana, and D. E. Shaw. 2018. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. USA.* 115:E4758–E4766.
46. Bax, A. 2003. Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Sci.* 12:1–16.
47. Fraczkiewicz, R., and W. Braun. 1998. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comput. Chem.* 19:319–333.
48. Song, D., R. Luo, and H.-F. Chen. 2017. The IDP-specific force field ff14IDPSFF improves the conformer sampling of intrinsically disordered proteins. *J. Chem. Inf. Model.* 57:1166–1178.

Paper 14

Pavlíková Přecechtělová, J.; Mládek, A.; Zapletal, V.; Hritz, J. Quantum Chemical Calculations of NMR Chemical Shifts in Phosphorylated Intrinsically Disordered Proteins, JCTC **2019**, 15, 5642-5658

JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2019, 15, 5642−5658

pubs.acs.org/JCTC

Quantum Chemical Calculations of NMR Chemical Shifts in Phosphorylated Intrinsically Disordered Proteins

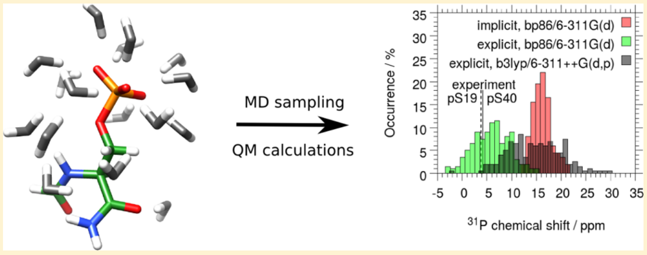
Jana Pavlíková Přecechtělová,*[†,‡] Arnošt Mládek,[†] Vojtěch Zapletal,[†,§] and Jozef Hritz[†]

[†]CEITEC and [§]NCBR, Faculty of Science, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic

[‡]Faculty of Pharmacy in Hradec Králové, Charles University, Akademika Heyrovského 1203, 500 05 Hradec Králové, Czech Republic

Supporting Information

ABSTRACT: Quantum mechanics (QM) calculations are applied to examine ^{1}H, ^{13}C, ^{15}N, and ^{31}P chemical shifts of two phosphorylation sites in an intrinsically disordered protein region. The QM calculations employ a combination of (1) structural ensembles generated by molecular dynamics, (2) a fragmentation technique based on the adjustable density matrix assembler, and (3) density functional methods. The combined computational approach is used to obtain chemical shifts (i) in the S19 and S40 residues of the non-phosphorylated and (ii) in the pS19 and pS40 residues of the doubly phosphorylated human tyrosine hydroxylase 1 as the system of interest. We study the effects of conformational averaging and explicit solvent sampling as well as the effects of phosphorylation on the computed chemical shifts. Good to great quantitative agreement with the experiment is achieved for all nuclei, provided that the systematic error cancellation is optimized by the choice of a suitable NMR standard. The effect of the standard reference on the computed ^{15}N and ^{31}P chemical shifts is demonstrated by employing three different referencing methods. Error bars associated with the statistical averaging of the computed ^{31}P chemical shifts are larger than the difference between the ^{31}P chemical shift of pS19 and pS40. The sequence trend of ^{31}P shifts therefore could not be reliably reproduced. On the contrary, the calculations correctly predict the change of the ^{13}C chemical shift for CB induced by the phosphorylation of the serine residues. The present work demonstrates that QM calculations coupled with molecular dynamics simulations and fragmentation techniques can be used as an alternative to empirical prediction tools in the structure characterization of intrinsically disordered proteins.



INTRODUCTION

Motivation. Intrinsically disordered proteins (IDPs) or regions[1] within the so-called hybrid proteins[2] regulate a vast of molecular processes in neurobiochemistry and their misfunction leads to neurodegenerative diseases, for example, Parkinson's and Alzheimer disease.[3] IDPs are characterized by polypeptide chains that do not fold into a stable and well-defined tertiary structure. More than 30% of the proteins in the eukaryotic genomes contain long (>30 residue) intrinsically disordered regions[4] and they are predicted to be common in functional proteins.[5,6]

A typical example of a hybrid protein comprising the structured as well as disordered regions is the human tyrosine hydroxylase 1 (hTH1).[7] It controls the rate-limiting step in the synthesis of dopamine and adrenaline. The hTH1 contains the regulatory domain (RD) composed from ~70 amino acid IDP region and the structured part (~110 amino acids). Its function is regulated by the phosphorylation of RD-hTH1 within its IDP region and the subsequent binding to the 14-3-3 proteins.[8−10]

The structural characterization of IDPs is extremely challenging due to their flexibility that renders the applicability of standard techniques such as X-ray crystallography and cryo-electron microscopy limited.[11] An efficient technique suitable for the structural characterization of this challenging class of proteins is nuclear magnetic resonance (NMR) spectroscopy.[12] The structural interpretation of measured NMR data combined with computational tools provides an ensemble of diverse structures that characterize the flexible nature of IDPs.[13] The obtained structural ensembles are verified through a comparison of experimental NMR properties including chemical shifts (CSs) with software-aided predictions.[14] Unfortunately, the currently available software tools,[15−20] applicable to thousands of large biomolecular structures, are not tailored to predict CSs for phosphorylated IDPs. The absence of a suitable prediction tool is caused by the lack of relevant experimental CS data that could be used as training sets in the design of prediction algorithms. Instead, CSs obtained from quantum mechanical (QM) calculations[21−24] can potentially be used to aid the structural characterization of phosphorylated IDPs.

In recent years, automatic fragmentation techniques have emerged[25−30] that allow for feasible QM calculations of NMR

Received: March 9, 2018
Published: September 5, 2019

DOI: 10.1021/acs.jctc.8b00257

CSs[31−36] in large biomolecular systems such as proteins, protein−protein complexes, and nucleic acids. Fragment-based approaches perform very well, yet their accuracy still does not surpass that of empirical methods, especially for sensitive nuclei such as H^N and ^{15}N.[32] Fragmentation techniques thus become a valuable prediction alternative in cases when empirical parameterization is prevented by the lack of relevant experimental NMR data. One such case is the prediction of ^{31}P NMR CSs for phosphorylated proteins. Previous validation studies[25,31,32,34−36] that employed fragment-based methods, for example, the adaptive density matrix assembler (ADMA),[25] focused exclusively on the computation of conventional protein nuclei, that is, ^1H, ^{13}C, ^{15}N in nonphosphorylated well-structured proteins. It thus remains unknown how the ADMA approach performs for CSs in proteins that are intrinsically disordered and—in addition—modified by a phosphorylation. Of particular interest is what results can be achieved if the ADMA methodology is combined with density functional theory (DFT) and explicit solvent sampling by molecular dynamics (MD). In this respect, three principal questions arise: (1) do the MD/ADMA/DFT calculations of ^1H, ^{13}C, and ^{15}N CSs for the phosphorylated IDPs perform equally well as in a previous study of *N*-methyl acetamide (NMA) and an example of a structured protein?[32] (2) What is the accuracy of the MD/ADMA/DFT scheme for ^{31}P CSs? (3) How does the phosphorylation affect the computed CSs of other nuclei? The span of phosphorus CSs observed in proteins is only a few ppm.[37−39] This raises a great challenge for QM calculations that have to be very accurate to be able to reproduce CS variations due to structural changes. Previous studies for proteins[32,40−43] and nucleic acids[44−47] indicate that QM CS calculations have to be combined with MD simulations to reflect the ensemble nature of the protein conformation and solvent distribution. This becomes especially crucial for flexible IDPs. Yet, a study combining MD and a fragment-based approach has so far been performed only for NMA, a textbook minimal-size model[32,43] mimicking the protein peptide bond, and a small well-structured protein.[32] The example applications demonstrate that the accuracy of the combined MD/QM approach is affected by the quality of MD geometries. Because the simulation of IDPs requires the use of specifically designed water models[48] and/or even force fields designed for that purpose,[49] the MD-rooted errors in IDP geometries can differ from those of well-structured proteins.

Computational methodologies are typically validated against experimental data. The quantitative agreement is measured by the root-mean-square-deviation (rmsd) while the qualitative agreement is reflected by the correlation coefficient (*R*). A previous study has shown that the dynamical averaging yields a reasonable correlation of the computed and experimental ^{15}N CSs while the rmsd's remain high. It was proposed that the problem could be alleviated by the use of a suitable secondary standard.[50−52] This referencing choice typically helps to suppress systematic errors stemming from the computational approach as well as from the fact that the computed and experimental chemical shieldings are often obtained in different phases (gas phase vs solution).[53] Despite this experience, computational studies of ^{15}N CSs in proteins have so far employed secondary standards rather rarely.[51]

^{31}P CS calculations also suffer from referencing issues. The quest for a suitable ^{31}P NMR reference compound is a long-standing problem.[53] The experimental ^{31}P NMR spectroscopy experiments typically employs the 85% water solution of H_3PO_4 as the standard reference. In calculations, however, the referencing to the gas-phase computed value of the ^{31}P chemical shielding in H_3PO_4 has been long perceived as problematic for two reasons.[54] First, the proper species in the concentrated phosphoric acid solution are not known experimentally. Second, water as a polar solvent can be involved in associations with the H_3PO_4 molecules. Therefore, ^{31}P CS studies typically avoid the use of the calculated H_3PO_4 as the NMR standard. Instead, a secondary referencing scheme employing PH_3 has been proposed[52] and extensively applied to reference the computed ^{31}P CSs in inorganic[55,56] and organophosphorus[53,57] compounds as well as in biomolecules.[44,45,58,59] Nevertheless, a recent study carried out by Fukal et al.[46] shows that referencing to the computed H_3PO_4 in fact leads to a better agreement of the calculated ^{31}P CSs with experiment. The study also proposes an alternative relative referencing scheme, which obviates the use of an external reference.

Objectives. In the present work, we extend on the previous experience with QM calculations employing ADMA[25,31,32] to obtain NMR CSs in proteins. We aim to address three areas of ADMA calculations that have not been studied in previous works. First, the performance of the MD/ADMA/DFT computational scheme is examined for ^1H, ^{13}C, and ^{15}N CSs in an example IDP region. We analyze the performance in terms of absolute agreement with experiment, CS variations within the structural ensemble, as well as in terms of statistical errors. Conclusions based on the results obtained here are compared to those previously drawn for NMA and an example of a structured protein.[32] Second, we examine the effect of the conformational averaging and solvent on protein ^{31}P CSs for the first time. Finally, the effect of phosphorylation on the computed ^1H, ^{13}C, and ^{15}N CSs is inspected. In order to understand the quantitative differences between the computed ^{15}N and ^{31}P CSs and experiment, we additionally aim to analyze three referencing schemes for each nuclei and to identify the extent of systematic error cancellations affecting the computed CSs.

In order to implement the defined objectives, we perform MD simulations for the nonphosphorylated and doubly phosphorylated hTH1 and apply ADMA to construct fragments of two phosphorylation sites of hTH1, the serine residues at positions 19 and 40. The two residues will be in the following referred to as S19/S40 in the nonphosphorylated hTH1 and pS19/pS40 in the doubly phosphorylated hTH1. Density functional calculations are performed for the two residues of interest.

■ METHODS

MD Simulations. The structure of the doubly phosphorylated hTH1 used as a starting structure for the MD simulations is provided in the Supporting Information. The starting structure of the nonphosphorylated hTH1 was produced by replacing (two) phosphate groups by hydrogens. MD simulations of the nonphosphorylated and doubly phosphorylated hTH1 were carried out using the protein Amber99SB-ILDN force field,[60,61] and parameters for the phosphorylated serine residues (charge −2) were taken from the work of Homeyer et al.[62] The systems were solvated using a rhombic dodecahedral box of TIP4P-D waters[48] with a minimum distance between the box walls and solute of 4 nm. The TIP4P-D water model was designed to produce more reliable

DOI: 10.1021/acs.jctc.8b00257
J. Chem. Theory Comput. 2019, 15, 5642−5658

conformational ensembles of IDPs.[48] The charge of the system was neutralized by adding Na^+ and Cl^- ions; the concentration of the salt was adjusted to the physiological concentration of 150 mM. All simulations were performed under periodic boundary conditions. Prior to the MD runs, energy minimizations with a steepest descent algorithm were performed. The lengths of hydrogen atom covalent bonds were constrained using the LINCS algorithm.[63] An integration time step of 2 fs was used. A cutoff of 1.0 nm was applied for the Lennard-Jones interactions and short-range electrostatic interactions. Long-range electrostatic interactions were calculated by particle-mesh Ewald summation with a grid spacing of 0.12 nm and a fourth order interpolation.[64] The 8 ns long equilibration protocol consisted of 4 steps during which backbone restraints were progressively released, starting at 1000 kJ mol^{-1} nm^{-2}. Two follow-up production *NpT* simulations (300 K, 1 atm) were then carried out. The resulting MD trajectories were 10 ns (simulation A) and 100 ns (simulation B) long. While both the temperature and pressure were maintained using a Berendsen coupling scheme,[65] the production simulations were performed using Nose−Hoover thermostat[66] and Parrinello−Rahman barostat algorithms.[67] Atomic coordinates were recorded every 1 ps.

DFT Calculations of NMR CSs. MD snapshots have been extracted in 100 ps steps from both MD trajectories. A total of 100 snapshots from the simulation A (A/100 ensemble) have been obtained and further used. Calculations building on simulation B employed the first 100 (B/100 ensemble) or 500 frames (B/500 ensemble). For each snapshot, the ADMA fragmentation[25] procedure has been performed using a Python suite of codes. ADMA divides the protein into small fragments in such a way that one fragment is generated for alanine, glycine, and proline residues, while two fragments are generated for the remaining amino acids (one for the backbone part and one for the side chain part of the residue in question). Fragments are then embedded into the surroundings of neighboring macromolecule moieties, water molecules, and ions to build the so-called parent molecules. The detailed description of the procedure for the construction of parent molecules can be found elsewhere.[25,32] The surroundings within the radius *r* from atoms of the original fragment is taken into account, where we use *r* = 5 Å (for A/100 ensemble) and *r* = 3.5 Å (for B/100 and B/500 ensembles). The size of the parent molecules is 120−216 atoms for the former or 58−132 atoms for the latter, respectively. The geometries of parent molecules constructed from the MD snapshots are directly used as input coordinates for CS calculations, that is, no geometry optimization is performed. The CS calculations employed the Gaussian16[68] implementation of the GIAO formalism within the coupled perturbed DFT method.[69−73] The bp86[74,75] and b3lyp[76−78] density functionals as implemented in Gaussian have been used. The former has previously provided favorable agreement between computed and experimental ^{31}P CSs in nucleic acids[44,45] while the latter has been recommended by preceding validation studies of CSs in structured proteins.[31,32] For calculations with the A/100 ensemble, the DFT functionals were combined with the 6-311G(d)[79,80] and 6-31G(d)[81,82] basis sets to keep the QM methodology consistent with works[31,32] demonstrating the results of the MD/ADMA/DFT framework in ordered protein structures. Using the same methodology facilitates a comparison between the results achieved here and those obtained previously.[31,32] Additionally, b3lyp/6-311++G(d,p) and b3lyp/pcs4/6-311++G(d,p) levels of theory were applied to B/100 and B/500 ensembles to account for the effects of larger basis sets and larger statistical data sets. The pcs4/6-311++G(d,p) notation refers to the pcs4 basis set placed only on atoms of the original fragment in question (N, H^N, CA, HA, C′, O for the backbone fragment and CB, HB1, HB2, O, P, O1P, O2P, O3P for the side chain fragment, see Figure 1) and 6-311++G(d,p) basis set placed on atoms that were included in the parent molecule as surroundings. In order to account for the solvent effects, two sets of calculations have been carried out for A/100. In the first one, water molecules and ions were removed from the parent molecules and replaced by an implicit solvent. The conductor-like implicit solvent model developed within the framework of the polarizable continuum model[83,84] based on the self-consistent reaction field has been applied. Results obtained with this setup are referred to as "implicit" throughout the text. The second set of calculations has been performed with both implicit and explicit solvent as well as with ions and will be from now on referred to as "explicit". From the NMR calculation of each parent molecule, we only extract the chemical shielding of atoms belonging to the original fragment (i.e., atoms constituting the surroundings are disregarded). The chemical shielding σ is converted to the CS δ using various referencing schemes. ^1H and ^{13}C CSs are referenced to tetramethylsilane (TMS) computed at the same level of theory as the atom of interest (X)

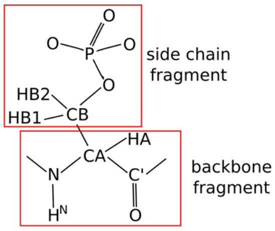


Figure 1. Labeling of atoms in the phosphorylated serine.

$$\delta_X^{calc} = \sigma_{TMS}^{calc} - \sigma_X \tag{1}$$

Unless otherwise stated, ^{15}N CS calculations employ a secondary standard referencing scheme based on methylamine[51]

$$\delta_X^{calc} = \sigma_{CH_3NH_2(gas)}^{calc} - \sigma_X^{calc} + (\sigma_{NH_3(liq)}^{exp} - \sigma_{CH_3NH_2(gas)}^{exp}) \tag{2}$$

where $\sigma_{CH_3NH_2(gas)}^{calc}$ is the chemical shielding of methylamine computed in the gas phase, $\sigma_{NH_3(liq)}^{exp} = 244.6$ ppm is the absolute ^{15}N chemical shielding of liquid ammonia at 25 °C,[85] and $\sigma_{CH_3NH_2(gas)}^{exp} = 249.5$ ppm is the experimental ^{15}N chemical shielding of methylamine.[86] The referencing for ^{31}P employs the absolute experimental chemical shielding of the 85% H_3PO_4 (328.4 ppm)[87] as the standard reference

$$\delta_X^{calc} = \sigma_{85\%\ H_3PO_4(liq)}^{exp} - \sigma_X^{calc} \tag{3}$$

Alternative referencing schemes have also been used (see Results and Discussion) for ^{15}N as well as ^{31}P CSs to inspect how the choice of the NMR standard influences the agreement of the computed and experimental data.

Because we here focus on the two phosphorylation sites of hTH1, the CS calculations are performed for four fragments of the nonphosphorylated and doubly phosphorylated hTH1. The four fragments correspond to the backbone and side-chain part of S19 and S40 or pS19 and pS40, respectively. The total number of geometries subject to CS calculations is (100 snapshots) × (4 fragments) or (500 snapshots) × (4 fragments) per each of the two systems, respectively.

For every atom of a given fragment, the CS is calculated as the statistical average, $\overline{x} = \sum_{i=1}^{N} x_i/N$, where x_i is the value of the CS for the atom in question in the i-th snapshot and $N = 100$ is the number of CS values that correspond to the individual snapshots of the MD trajectory. The statistical distribution of the calculated CSs within the chosen set of MD snapshots is evaluated by the standard deviation of the sample mean $s_{\overline{x}}$ defined as

$$s_{\overline{x}} = \sqrt{\frac{1}{N(N-1)} \sum_{i=1}^{N} (x_i - \overline{x})^2} \tag{4}$$

In order to estimate the error caused by using a sample, which is much smaller than the true population of the IDP in question, we report the 95% confidence interval (CI) expressed as

$$\mathrm{CI} = \overline{x} \pm z s_{\overline{x}} \tag{5}$$

where $z = 1.96$ is the z-value from the standard normal z-table. The term $z s_{\overline{x}}$ is called the maximum error of estimate (MEE). The averaged CSs are compared with the experimental data taken from the Supporting Information of Louša et al.[9] (^{1}H, ^{13}C, and ^{15}N CSs) and from Hritz et al.[8] (^{31}P CSs). CSs are obtained for HA, HB1, HB2, H^N, CA, CB, C′, N, and P atoms (Figure 1).

RESULTS AND DISCUSSION

Effects of Conformational Changes and Solvent on the CSs. In order to get an idea how changes in the protein conformation and the constant rearrangement of solvent molecules during the simulation time affect the computed CSs, we will first inspect the ranges of CSs obtained for the MD ensembles. The conformational changes are easy to track through the time dependence of the backbone torsion angles. Figure 2a shows the time evolution of the torsion angles ϕ and ψ for the residues pS19 and pS40 in the doubly phosphorylated hTH1. It reveals instant, fast fluctuations of the torsion angles in the equilibrated state while no major conformational switches are observed. The CSs of atoms in the backbone fluctuate similarly (for an example of ^{15}N CSs, see Figure 2b), though not necessarily in direct correlation with the torsion angle changes. This is obvious as many other effects (besides the backbone conformation) influence the CSs, for example, side-chain conformation, interactions of the protein with the explicit solvent, interactions with other protein moieties, and so forth. Particularly sensitive to the influence of the explicit solvation are the polar protons, H^N, as they are directly involved in the hydrogen bonding with the solvent molecules. Figure 2c shows the time-dependence of the H^N CS and does not reveal any patterns that would indicate unusual solvent arrangements. The observed variations in CSs reflect most importantly the combined effect of conformational changes and changes in the positions and orientation of water molecules. The impact of this combined effect is large, not only for the polar protons but for other atoms as well (see the discussion below) and leads to CSs spanning several to tens of ppm depending on the type of the nucleus in question.

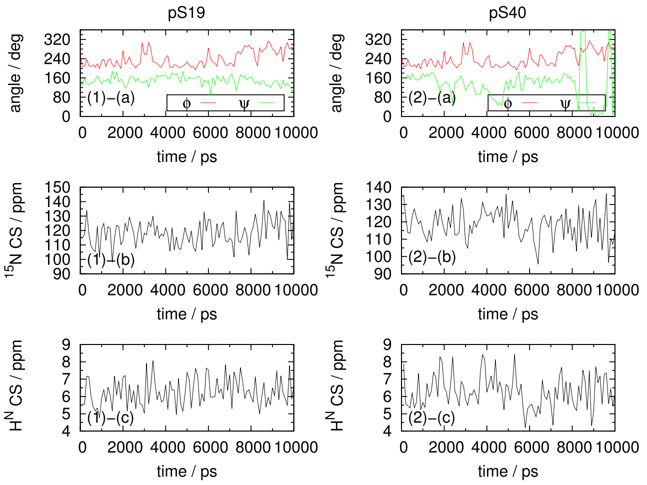


Figure 2. Time evolution of (a) the backbone torsion angles ϕ(C′−N−CA−C′), ψ(N−CA−C′−N), (b) H^N CSs, and (c) ^{15}N CSs in the (1) pS19 and (2) pS40 residues of the doubly-phosphorylated hTH1 (trajectory A).

In general, the CSs of the explicitly solvated pS19 and pS40 fragments as extracted from the MD-generated ensemble of hTH1 structures are in line with previous findings obtained for the NMA,[32] a minimal-size model of the protein peptide bond. The nonpolar protons of pS19 and pS40 (i.e., HA, HB1, and HB2) span a range of ∼3−4 ppm (Figure 3) when calculated with bp86/6-311G(d) in the explicit solvent and >6 ppm when

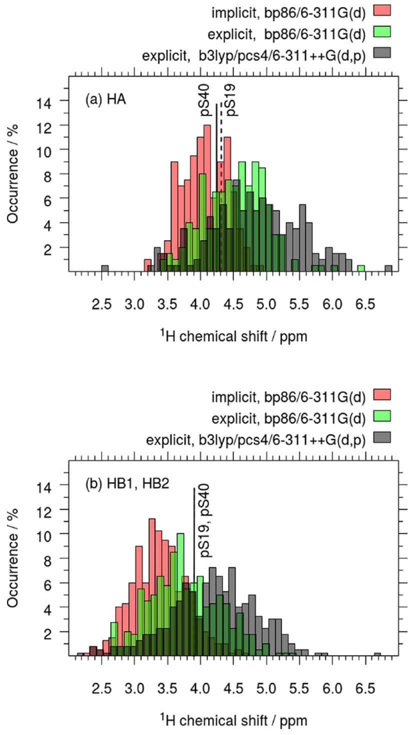


Figure 3. ^{1}H CSs of (a) HA and (b) HB1, HB2 atoms as obtained from calculations in the implicit and explicit solvent. The histograms include CSs of both pS19 and pS40. The dashed and solid vertical lines mark the experimental values of the ^{1}H CSs in pS19 and pS40, respectively. The experimental HB1 and HB2 CSs are identical and differ negligibly between pS19 and pS40 (see Table S1 of the Supporting Information).

calculated with b3lyp/pcs4/6-311++G(d,p). We recall that the two sets of calculations build on different ensembles of parent molecules constructed with $r = 5$ Å and $r = 3.5$ Å cutoff for the surroundings, respectively. It is thus both, the basis set size and the solvent shell size, that are responsible for the distinctions in the variations of proton CSs. HA CSs are larger by about ∼0.5 ppm than the CSs of HB1, HB2. The distributions of nonpolar proton CSs obtained with b3lyp/6-311G(d) in implicit and explicit solvents overlap to a large extent. Yet, the addition of explicit water molecules to the implicit solvent model causes a notable shift of the CS histogram by up to 0.5 ppm to larger CSs and thus also to an improved agreement with experimental values (see Figure 3 and Table S1).

For the polar protons, H^N, the histograms of computed CSs are shown in Figure 4. It was observed already previously[32] that calculations of H^N CSs in the implicit solvent provide rather narrow distributions while the combination of the implicit and explicit solvents leads to the distribution broadening. We can see this behavior for IDPs here as well (Figure 4). The H^N CSs computed with bp86/6-311G(d) span a range of up to 3 ppm in the implicit solvent with the average value being around 5 ppm (see Table 1). Upon adding the explicit solvent on top of the implicit solvent model, the span of H^N CSs increases to ∼4.5 ppm. Consequently, the statistical errors increase as well. Note that MEEs are larger for the polar protons (up to 0.19 ppm, Table 1) than for the nonpolar ones (up to 0.11 ppm; Table S1). The inclusion of the explicit water molecules shifts the CS histogram toward a better agreement with the experiment. Further improvement is achieved by extending the basis set. At the b3lyp/pcs4/6-311++G(d,p) level, the CS histogram spans more than 6 ppm and leads to overall best statistical averages in terms of agreement with experiment. The computed average CSs amount to 7.90 (pS19) and 7.33 ppm (pS40). This is underestimated by 0.7 ppm and ∼1.5 ppm compared to the experimental values of 8.6 ppm (pS19) and 8.8 ppm (pS40), respectively (Table 1). It was proposed before[32] that the remaining error in the computed H^N CSs results from too long hydrogen bonds caused by the AMBER force field parameterization, especially by the approximations used in the parameterization of the TIP3P water model. This hypothesis was confirmed by combined ab initio MD/DFT calculations of NMR CSs in NMA.[43] Here, we employ the TIP4P-D model for the solvent, which is more suitable for IDPs than TIP3P; however, it suffers from similar problems with respect to NMR CS calculations using MD geometries. CS results obtained with B/100 and B/500 ensembles demonstrate that the augmentation of the statistical data set significantly reduces the MEE. For H^N CSs, the MEEs approximately halved after increasing the number of MD snapshots from 100 to 500.

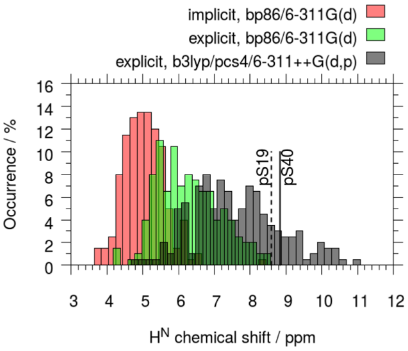


Figure 4. ^1H CSs of H^N atoms as obtained from calculations in the implicit and explicit solvent. The histograms include CSs of both pS19 and pS40, respectively. The dashed and solid vertical lines mark the experimental values of the H^N CSs in pS19 and pS40, respectively.

The ^{13}C CSs for both CA and CB obtained here with bp86/6-311G(d) and the A/100 ensemble span a range of ∼15−20 ppm (Figure 5). The typical MEEs then amount to 0.6 ppm (Table 2). On the contrary, the CSs computed using the B/100 ensemble display larger variations of ∼25 ppm. The increased fluctuations are likely caused by the fact that only the first solvation shell was explicitly included in the QM calculations. Because the two carbon atoms do not directly interact with the solvent molecules, the addition of the explicit solvent does not significantly change the shape and location of the ^{13}C CS distributions on the CS scale. Yet, the histograms obtained with explicit solvation seem to be closer to the approximate Gaussian shape, and there is an obvious difference in the percentage occurrences of the individual CS values when comparing the two solvent model setups, respectively. The average CS of CA is little affected by the explicit solvation (Table 2(a)). For CB (Table 2(b)), explicit solvent decreases the CS by 2.0 (pS19) and 1.2 ppm (pS40) compared to calculations in the implicit solvent. Consequently, the difference between the calculated and experimental CSs decreases from 3.9 ppm (pS19) and 2.7 ppm (pS40) to 1.9 and 1.5 ppm, respectively, upon explicit solvation. These findings are in line with the downfield shift of methyl carbons observed in NMR calculations of *N*-methyl-acetamide.[32] Interestingly, calculations employing B/100 and B/500 ensembles provide much worse estimates of CA CSs than the calculations building on molecular structures constructed from A/100 snapshots. This is likely caused by the differences in the sampling of torsion angles within the A and B trajectory, respectively. The histogram of b3lyp/pcs4/6-311++G(d,p)-

Table 1. ^1H CSs (in ppm) of H^N Atoms Calculated by Different QM Setups[a]

method	ensemble	solvent	δ(pS19)/ppm	δ(pS40)/ppm
b3lyp/pcs4/6-311++G(d,p)	B/100	explicit	7.90 ± 0.25	7.33 ± 0.23
b3lyp/6-311++G(d,p)	B/100	explicit	7.43 ± 0.28	7.12 ± 0.32
b3lyp/6-311++G(d,p)	B/500	explicit	7.42 ± 0.13	7.23 ± 0.16
bp86/6-311G(d)	A/100	implicit	5.06 ± 0.11	5.02 ± 0.12
bp86/6-311G(d)	A/100	explicit	6.26 ± 0.16	6.28 ± 0.19
b3lyp/6-311G(d)	A/100	explicit	6.14 ± 0.15	6.19 ± 0.18
b3lyp/6-31G(d)	A/100	explicit	6.08 ± 0.14	6.13 ± 0.17
experiment[b]		aq solution	8.599	8.841

[a] The CSs are reported as 95% CIs (eq 5). [b] See the Supporting Information of ref 9.

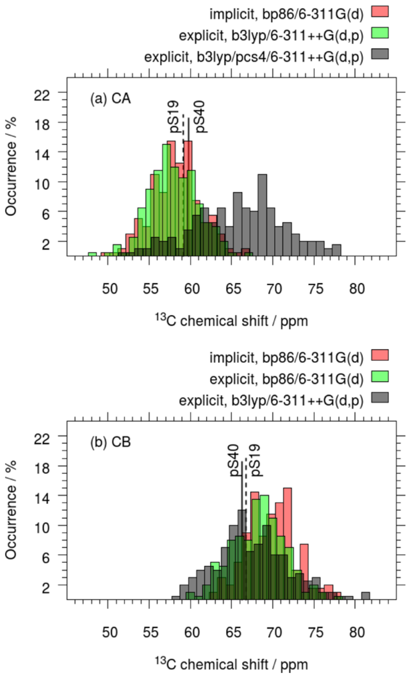


Figure 5. ^{13}C CSs of (a) CA atoms and (b) CB atoms as obtained from calculations in the implicit and explicit solvent. The histograms include CSs of both pS19 and pS40, respectively. The dashed and solid vertical lines mark the experimental values of the CA/CB CSs in pS19 and pS40, respectively.

computed CSs is located significantly further away from the experimental values. Therefore, the results obtained at the lower level of theory (bp86/6-311G(d)) using the large (r = 5 Å) radius for the explicit water surroundings seem to be more reliable. The effect of replacing the bp86 functional by b3lyp on CA and CB CSs is about the same as the effect of replacing the 6-311++G(d,p) basis set on fragment atoms by pcs4 (1−2 ppm). We will show below that the influence of the functional and basis set on carbonyl carbons is significantly larger.

The span of carbonyl carbon CSs is comparable to that of CA and CB and amounts to ∼30 ppm in both implicit and explicit solvents (Figure 6) and the MEEs are up to 0.6 ppm (Table 3). The distributions of CSs in the explicit solvent is shifted downfield compared to the results in the implicit solvent. Consequently, the average CS values increase by ∼1 and ∼2 ppm for pS19 and pS40 (Table 3), respectively. Frank et al.[31] recommended the b3lyp functional for the calculation of carbonyl carbon CSs. We therefore also provide the results obtained at the b3lyp/6-311G(d) level of theory for comparison. Note that the change of the functional from bp86 to b3lyp causes a larger CS change than the addition of the explicit solvent (6.6 ppm vs 0.9 ppm for pS19). The results in Table 3 also confirm previous suggestions[32] that a basis set larger than 6-31G(d) is needed to obtain an accurate prediction for carbonyl carbon CSs. For an example of the pS19 residue, the b3lyp/6-31G(d) level of theory underestimates the carbonyl carbon CSs by 13 ppm. On the contrary, the b3lyp/6-311++G(d,p) level of theory overestimates the computed CSs by ∼3 ppm only and gives the

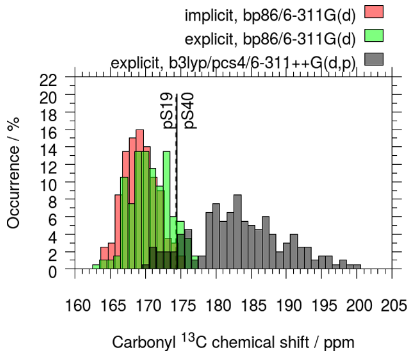


Figure 6. ^{13}C CSs of carbonyl carbons as obtained from calculations in the implicit and explicit solvent. The histograms include CSs of both pS19 and pS40, respectively. The dashed and solid vertical lines mark the experimental values of the carbonyl ^{13}C CSs in pS19 and pS40, respectively.

Table 2. ^{13}C CSs (in ppm) of CA and CB Atoms Calculated by Different QM Setups[a]

method	ensemble	solvent	δ(pS19)/ppm	δ(pS40)/ppm
		(a) CA CSs		
b3lyp/pcs4/6-311++G(d,p)	B/100	explicit	67.3 ± 1.0	64.5 ± 1.1
b3lyp/6-311++G(d,p)	B/100	explicit	64.8 ± 1.6	62.4 ± 1.6
b3lyp/6-311++G(d,p)	B/500	explicit	64.7 ± 0.6	64.3 ± 0.7
bp86/6-311G(d)	A/100	implicit	59.1 ± 0.6	57.4 ± 0.6
bp86/6-311G(d)	A/100	explicit	58.7 ± 0.6	57.2 ± 0.6
b3lyp/6-311G(d)	A/100	explicit	57.6 ± 0.5	56.1 ± 0.5
b3lyp/6-31G(d)	A/100	explicit	54.3 ± 0.5	52.8 ± 0.5
experiment[b]		aq solution	59.1	59.8
		(b) CB CSs		
b3lyp/pcs4/6-311++G(d,p)	B/100	explicit	67.9 ± 0.9	66.7 ± 1.5
b3lyp/6-311++G(d,p)	B/100	explicit	65.1 ± 1.2	63.5 ± 2.3
b3lyp/6-311++G(d,p)	B/500	explicit	66.2 ± 0.6	65.7 ± 0.7
bp86/6-311G(d)	A/100	implicit	70.7 ± 0.6	69.0 ± 0.6
bp86/6-311G(d)	A/100	explicit	68.7 ± 0.6	67.8 ± 0.6
b3lyp/6-311G(d)	A/100	explicit	66.6 ± 0.6	65.8 ± 0.6
b3lyp/6-31G(d)	A/100	explicit	62.4 ± 0.6	61.5 ± 0.5
experiment[b]		aq solution	66.8	66.3

[a] The CSs are reported as 95% CIs (eq 5). [b] See the Supporting Information of ref 9.

Table 3. ^{13}C CSs (in ppm) of C′ Atoms Calculated by Different QM Setups[a]

method	ensemble	solvent	δ(pS19)/ppm	δ(pS40)/ppm
b3lyp/pcs4/6-311++G(d,p)	B/100	explicit	182.9 ± 1.3	183.8 ± 1.3
b3lyp/6-311++G(d,p)	B/100	explicit	176.0 ± 1.9	177.6 ± 1.8
b3lyp/6-311++G(d,p)	B/500	explicit	177.6 ± 0.9	177.5 ± 0.8
bp86/6-311G(d)	A/100	implicit	169.6 ± 0.5	168.9 ± 0.4
bp86/6-311G(d)	A/100	explicit	170.5 ± 0.6	170.9 ± 0.5
b3lyp/6-311G(d)	A/100	explicit	177.1 ± 0.6	177.2 ± 0.5
b3lyp/6-31G(d)	A/100	explicit	161.3 ± 0.5	160.9 ± 0.5
experiment[b]		aq solution	174.3	174.4

[a]The CSs are reported as 95% CIs (eq 5). [b]See the Supporting Information of ref 9.

overall best agreement with the experiment. Introducing pcs4 on backbone or side chain atoms of the phosphorylated serine increases the computed CS by as much as 7−8 ppm. Carbonyl carbons are thus sensitive to the choice of the DFT functional (hybrid vs non-hybrid) as well as the basis set.

The variations of the CSs within the MD ensemble of pS19 and pS40 fragments are particularly large for the nitrogen CSs. The range of ^{15}N CSs spans over 40−50 ppm (Figure 7), regardless of the level of theory and solvent model used. Consequently, the average ^{15}N CSs have the largest MEEs out of all nuclei (up to 2.4 ppm), see Table 4. The inclusion of explicit water molecules causes a notable shift of the ^{15}N CS distributions toward larger CS values. Consequently, the average ^{15}N CSs of 111.8 ppm (pS19) and 112.0 ppm (pS40) obtained with bp86/6-311G(d) in the implicit solvent increase to 118.4 and 118.2 ppm, respectively, upon explicit solvation. The agreement with the experimental values 119.7 (pS19) and 117.2 ppm (pS40) thus significantly improves. Note that results obtained with the b3lyp/6-311G(d) level of theory provides ^{15}N CSs smaller by ~3 ppm than the bp86-computed counterparts, and the replacement of bp86 by b3lyp increases the difference between the computed and experimental data. The augmentation of the basis set for fragment atoms from 6-311++G(d,p) to pcs4 leads to an overestimation of the computed CSs by up to 5 ppm, which again worsens the agreement with experiment. Comparison of the basis set, solvent, and DFT functional effect reveals that the effects are of about the same order, which is in line with previous findings.[51] Cai et al.[51] pointed out that the difficulty of the theoretical ^{15}N CS prediction lies in its complexity. The ^{15}N CS is affected by a multitude of effects (conformation, H-bonding, neighboring residue types, solvent, electrostatic interactions), none of which appears dominant.

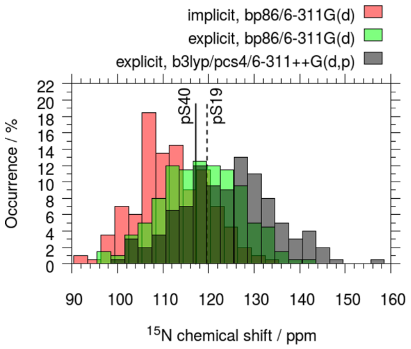


Figure 7. ^{15}N CSs (in ppm) of backbone N atoms as obtained from calculations in the implicit and explicit solvent. The histograms include CSs of both pS19 and pS40, respectively. The dashed and solid vertical lines mark the experimental values of the ^{15}N CSs in pS19 and pS40, respectively.

It is striking that the b3lyp/6-31G(d) (Table 4) functional/basis set combination is the most unfavorable model chemistry tested here, judged by the agreement with the experiment. Previous study[31] of CSs in proteins recommended b3lyp/6-31G(d) as a method of choice that benefits from error cancellations and gives rather an accurate prediction of the ^{15}N CSs in NMA and leads to a reasonable correlation of the computed and experimental ^{15}N CSs in the HA2 domain. A closer analysis reveals that the seeming discrepancy between the conclusions made here and the conclusions of Exner et al.[32] can be explained by the different choice of the referencing method as will be discussed in further details in a separate section below.

The computed ^{31}P CSs in pS19 and pS40 change within 20−25 ppm range in the explicit solvent. The extent of the variations is sizable considering the fact that the span of experimental ^{31}P CSs in O-phosphorylated amino acids,[37] peptides,[38] and proteins[39] is typically very small (a few ppm). A large part of the CS changes is therefore likely to stem from fluctuations in the H-bonding network rather than from the

Table 4. ^{15}N CSs (in ppm) of Backbone N Atoms Calculated by Different QM Setups[a]

method	ensemble	solvent	δ(pS19)/ppm	δ(pS40)/ppm
b3lyp/pcs4/6-311++G(d,p)	B/100	explicit	123.9 ± 2.1	124.7 ± 1.9
b3lyp/6-311++G(d,p)	B/100	explicit	119.0 ± 2.1	120.2 ± 2.4
b3lyp/6-311++G(d,p)	B/500	explicit	118.5 ± 1.2	120.0 ± 1.0
bp86/6-311G(d)	A/100	implicit	111.8 ± 1.5	112.0 ± 1.6
bp86/6-311G(d)	A/100	explicit	118.4 ± 1.8	118.2 ± 1.7
b3lyp/6-311G(d)	A/100	explicit	115.1 ± 1.7	115.2 ± 1.6
b3lyp/6-31G(d)	A/100	explicit	95.4 ± 1.3	95.6 ± 1.2
experiment[b]		aq solution	119.7	117.2

[a]The CSs are reported as 95% CIs (eq 5) and referenced to NH_3/CH_3NH_2 (eq 2). [b]See the Supporting Information of ref 9.

changes in the conformation of the phosphoserine residues. The plot of the ^{31}P CS histograms in the implicit and explicit solvents (Figure 8) corroborates the hypothesis as the 20 ppm CS span in the explicit solvent reduces to 10 ppm after removing the explicit solvent molecules from the calculations. A similar CS behavior has already been observed for the polar protons (see above). As a result of forming hydrogen bonds between phosphate oxygens and water hydrogens, the sharp and high histogram of CSs transforms into a broad one with overall lower percentage occurrences. Also, note that the use of the explicit solvent shifts the CS distributions more upfield and as a result, the computed CS ensemble better fits the experimental ^{31}P CSs of pS19 and pS40 (Figure 8). The average CSs in the implicit solvent are 16.2 ± 0.4 ppm (pS19) and 16.1 ± 0.4 ppm (pS40). H-bonding interactions between the solute and explicit solvent decrease the average CSs by ~10 ppm, which leads to δ(pS19) = 5.8 ± 0.7 ppm and δ(pS40) = 6.0 ± 0.8 ppm (Table 5), respectively. The quantitative agreement with the experimental values of 3.76 ppm (pS19) and 4.18 ppm (pS40), respectively, is therefore significantly improved. The best results, obtained with b3lyp/6-311G(d), differ from experiment by only 0.3 and 0.8 ppm, respectively. However, note that explicit solvation increases the MEEs from 0.4 ppm in the implicit solvent to 0.8 ppm in the explicit solvent. The difference between the experimental CSs of pS19 and pS40 amounts to 0.42 ppm and is thus about the same as the error of the explicit solvent calculations. This makes reproducing the sequence trends hardly possible, as will be further demonstrated in a separate section below.

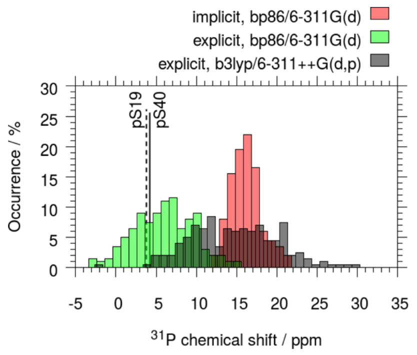


Figure 8. ^{31}P CSs as obtained from calculations in the implicit and explicit solvents. The CSs were referenced to 85% H_3PO_4 (eq 3). The histograms include CSs of both pS19 and pS40, respectively. The dashed and solid vertical lines mark the experimental values of the ^{31}P CSs in pS19 and pS40, respectively.

^{31}P CS calculations that employ pcs4 differ by as much as 18 ppm in the ensemble average from the calculations that apply 6-311++G(d,p) to all atoms. The basis set effect is thus almost twice as large as the effect of explicit solvation. The agreement with experimental data significantly deteriorates upon the replacement of 6-311++G(d,p) on atoms of the original fragments, which indicates an accumulation of systematic errors. The compensation of systematic errors is analyzed in more detail below.

Previous MD/DFT studies[25,31,32,88] of CSs in proteins opted to apply only double-zeta quality basis sets to make the computations tractable even for an increasing number of MD snapshots and their fragments. Because this option might be unavoidable also in future calculations of ^{31}P CSs in proteins, we test the performance of the 6-31G(d) basis set (along with the b3lyp functional) here as well. For the ^{31}P CSs referenced using 85% H_3PO_4, the results computed with b3lyp/6-31G(d) differ from experiment by >70 ppm (Table 5) for both pS19 and pS40. However, we will demonstrate below that the agreement with the experimental data can be dramatically influenced by the choice of the reference compound and/or the referencing approach. The section Choice of the Referencing Method for ^{31}P CSs discusses the ^{31}P CS referencing in detail and explains the reasons for the observed deviations of the computed and experimental ^{31}P NMR CS data.

Choice of the Referencing Method for ^{15}N CSs. For ^{15}N, three methods have become common in the literature to reference the computed CSs: (method 1) referencing to the absolute ^{15}N chemical shielding of liquid ammonia employing the calculated ^{15}N chemical shielding of CH_3NH_2[51] (see eq 2) used as a secondary standard (the multistandard approach),[50] (method 2) referencing to the absolute ^{15}N chemical shielding of liquid ammonia at 25 °C[43,85,88]

$$\delta_X^{calc} = \sigma_{NH_3(liq)}^{exp} - \sigma_X^{calc} \tag{6}$$

and (method 3) referencing to the ^{15}N chemical shielding of NH_3 calculated at the same level of theory as the molecule of interest[25,31,32,34]

$$\delta_X^{calc} = \sigma_{NH_3(gas)}^{calc} - \sigma_X^{calc} \tag{7}$$

The three methods/references will be in the following referred to as calc-CH_3NH_2, liq-NH_3, and calc-NH_3, respectively. In order to inspect the effect of referencing, we calculated the ^{15}N CSs using the three referencing methods listed above and compare the outcomes for three combinations of the DFT functional and basis set. Figure 9 demonstrates that CS distribution obtained with bp86/6-311G(d)/calc-CH_3NH_2

Table 5. ^{31}P **CSs (in ppm) of P Atoms in Phosphate Groups**[a]

method	ensemble	solvent	δ(pS19)/ppm	δ(pS40)/ppm
b3lyp/pcs4/6-311++G(d,p)	B/100	explicit	14.5 ± 1.0	15.0 ± 1.1
b3lyp/6-311++G(d,p)	B/100	explicit	−1.9 ± 1.5	−2.9 ± 1.7
b3lyp/6-311++G(d,p)	B/500	explicit	−3.0 ± 0.9	−2.2 ± 0.7
bp86/6-311G(d)	A/100	implicit	16.2 ± 0.4	16.1 ± 0.4
bp86/6-311G(d)	A/100	explicit	5.8 ± 0.7	6.0 ± 0.8
b3lyp/6-311G(d)	A/100	explicit	3.5 ± 0.6	3.4 ± 0.7
b3lyp/6-31G(d)	A/100	explicit	−67.2 ± 1.0	−68.7 ± 1.0
experiment[b]		aq solution	3.76	4.18

[a] The CSs are reported as 95% CIs (eq 5) and referenced to 85% H_3PO_4 (eq 3). [b] ref 8.

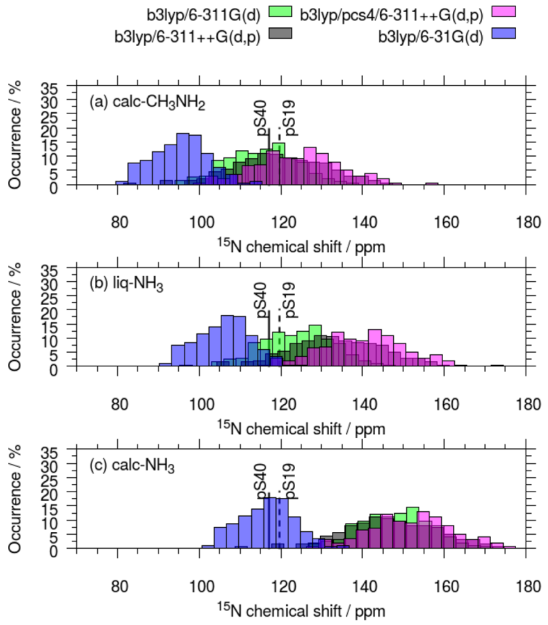


Figure 9. ^{15}N CSs obtained for the MD ensemble of the doubly phosphorylated hTH1. The CSs are referenced using three referencing schemes: (a) referencing to liquid ammonia using the secondary standard CH_3NH_2 computed at the same level of theory, (b) referencing to liquid ammonia, and (c) referencing to NH_3 computed at the same level of theory. The histograms include CSs of both pS19 and pS40, respectively. The dashed and solid vertical lines mark the experimental values of the ^{15}N CSs in pS19 and pS40, respectively.

covers about the same range as the distribution obtained with b3lyp/6-31G(d)/calc-NH_3. For the calculations carried out with the larger basis set, the best agreement between the statistically averaged CSs and the experimental values of 119.7 ppm (pS19) and 117.2 ppm (pS40) is achieved when bp86/6-311G(d)/calc-CH_3NH_2 is used. This yields δ_{iso}(pS19) = 118.4 ppm and δ_{iso}(pS40) = 118.2 ppm (Table 6), respectively. Similarly, a favorable agreement with the experiment is achieved with b3lyp/6-311++G(d,p).

In the particular case of using methylamine as a secondary standard in ^{15}N CS calculations, the referencing follows eq 2. Assuming that the ^{15}N chemical shielding only depends on the local electron density, deviations of the computed ^{15}N chemical shielding should be the same for all nitrogens with similar chemical surroundings.[51] If this is indeed the case, subtracting $\sigma_{CH_3NH_2(gas)}^{calc}$ and σ^{calc} is supposed to mitigate systematic errors due to the finite basis set and approximations made in the DFT exchange−correlation functional. The subsequent addition of the experimental CS of methylamine with respect to liquid ammonia, typically used[89] in experimental protein NMR to reference ^{15}N CSs, makes the computed results well suited for a comparison with experimental data. This holds provided that the solution NMR conditions are modeled by the explicit solvent in calculations of σ^{calc}. Note that although eq 2 is designed to suppress systematic errors of the calculation, only a part of the error is eliminated. The value of $\sigma_{CH_3NH_2(gas)}^{calc}$ is obtained here from a simple static DFT calculation, whereas the calculation of σ_X^{calc} employs DFT as well as MD. As a result, only the systematic errors associated with the DFT calculations can cancel out. Errors introduced by the approximations inherent to the MD cannot be mitigated unless the calculation of the standard includes the MD averaging as well. Because this is not the case here, the errors caused by the MD force field deficiencies still affect the data computed for hTH1.

Referencing the computed ^{15}N CSs to the calculated value of the ^{15}N chemical shielding in NH_3 (eq 7) can be interpreted as calculating the CS using

$$\delta_X^{calc} = \sigma_{NH_3(gas)}^{calc} - \sigma_X^{calc} + (\sigma_{NH_3(liq)}^{exp} - \sigma_{NH_3(gas)}^{exp}) \qquad (8)$$

where the term in the round brackets is neglected (set to zero). In fact, abandoning the gas-to-liquid shift of NH_3 is a very rough approximation because $\sigma_{NH_3(liq)}^{exp}$ = 244.6 ppm[85] while $\sigma_{NH_3(gas)}^{exp}$[86] and therefore $\sigma_{NH_3(liq)}^{exp} - \sigma_{NH_3(gas)}^{exp}$ = −19.9 ppm.

Table 6. ^{15}N CSs (in ppm) of Backbone N Atoms as Obtained Using Different References for the Conversion from the Chemical Shielding to the CS[a]

method	ensemble	calc-CH_3NH_2[b]	liq-NH_3[c]	calc-NH_3[d]
		^{15}N CSs/ppm		
		pS19		
b3lyp/pcs4/6-311++G(d,p)	B/100	123.9 ± 2.1	139.8 ± 2.1	151.7 ± 2.1
b3lyp/6-311++G(d,p)	B/100	119.0 ± 2.1	132.8 ± 2.1	146.3 ± 2.1
bp86/6-311G(d)	A/100	118.4 ± 1.8	126.9 ± 1.8	151.1 ± 1.8
b3lyp/6-311G(d)	A/100	115.1 ± 1.7	124.1 ± 1.7	148.0 ± 1.7
b3lyp/6-31G(d)	A/100	95.4 ± 1.3	106.2 ± 1.3	116.6 ± 1.3
experiment[e]		119.7		
		pS40		
b3lyp/pcs4/6-311++G(d,p)	B/100	124.7 ± 1.9	140.6 ± 1.9	152.5 ± 1.9
b3lyp/6-311++G(d,p)	B/100	120.2 ± 2.4	134.0 ± 2.4	147.5 ± 2.4
bp86/6-311G(d)	A/100	118.2 ± 1.7	126.7 ± 1.7	150.9 ± 1.7
b3lyp/6-311G(d)	A/100	115.2 ± 1.6	124.1 ± 1.6	148.0 ± 1.6
b3lyp/6-31G(d)	A/100	95.6 ± 1.2	106.4 ± 1.2	116.8 ± 1.2
experiment[e]		117.2		

[a] The CSs are reported as 95% CIs (eq 5). [b] CSs are referenced to liquid ammonia using the secondary standard CH_3NH_2 computed at the same level of theory (see eq 2). [c] Liquid ammonia is used as a reference (see eq 6). [d] NH_3 computed at the same level of theory is used as a reference (see eq 7). [e] See the Supporting Information of ref 9.

DOI: 10.1021/acs.jctc.8b00257

Referencing according to eq 7 thus introduces a systematic overestimation by ~20 ppm. Overall, the agreement between the calculated ^{15}N CSs referenced using eq 7 and experimental data is affected in three ways. First, the computed and experimental data are referenced to different standards (gas-phase vs liquid NH_3). Second, the chemical environment of ^{15}N in NH_3 resembles the chemical environment of ^{15}N in the protein amide group less than, for example, CH_3NH_2 or other potential secondary standards. As a result, the compensation of systematic errors of the DFT calculation is likely less effective. Third, as in the case of referencing according to eq 2, the computed results suffer from the MD force field inaccuracies. The approach of eq 7 therefore implies extra sources of errors when compared to the approach of eq 2 and is thus expected to provide inferior results in terms of agreement with experimental data. While this is indeed the case for ^{15}N CSs obtained here with the larger 6-311G(d) basis set, the b3lyp/6-31G(d) calculations perform better with the referencing to the computed gas-phase value of NH_3 (Table 6). It can be assumed that with this type of referencing, the errors introduced by the use of the small 6-31G(d) basis set is accidentally compensated by the errors stemming from neglecting the gas-to-liquid shift of the ^{15}N chemical shielding in NH_3 and therefore from the use of a different standard in calculations and experiment, respectively. An analysis of the computed ^{15}N chemical shielding in NH_3, CH_3NH_2, pS19, and pS40 (Table S2) reveals that this is indeed the case.

Truncation of the basis set leads to an increase of the ^{15}N chemical shielding in pS19 and pS40 by ~18 ppm, which is accompanied by the decrease of the ^{15}N chemical shielding by ~2 ppm in CH_3NH_2 and by as much as ~13.5 ppm in NH_3, respectively. Consequently, the computed ^{15}N CSs in pS19 and pS40 decrease by 20 ppm when referenced to calc-CH_3NH_2 and by more than 30 ppm when referenced to calc-NH_3. This leads to b3lyp/6-311G(d)/calc-CH_3NH_2 results being too underestimated while b3lyp/6-31G(d)/calc-$_3NH_3$ results agree well with experiment (Table 6).

The systematic errors associated with the calc-NH_3 referencing were already noticed by Frank et al.[31] and Exner et al.[32] It was suggested that the problem could potentially be solved in future studies by using a secondary standard. The results presented here strongly support the proposed solution. The accuracy of the computed ^{15}N CSs referenced to the calculated chemical shielding of NH_3 can be significantly improved by adding the gas-to-liquid shift of NH_3. Yet, the compensation of errors originating from the QM methodology is likely to work better between the protein NH groups and CH_3NH_2, especially if large basis sets are employed. The present work thus highly supports the previous suggestion[32] that very accurate ^{15}N CSs can be achieved if the dynamical averaging of conformational changes and solvent rearrangements is applied along with large basis sets and standards similar to the groups typical for proteins.

Choice of the Referencing Method for ^{31}P CSs. For the sake of computing ^{31}P CSs, we test three referencing methods here in analogy with the referencing methods discussed above for ^{15}N CSs. (Method 1) references the ^{31}P CSs to 85% H_3PO_4 using the secondary standard PH_3 as proposed by van Wüllen[52]

$$\delta_X^{\text{calc}} = \sigma_{PH_3(\text{gas})}^{\text{calc}} - \sigma_X^{\text{calc}} + (\sigma_{H_3PO_4(85\%\ \text{solution})}^{\text{exp}} - \sigma_{PH_3(\text{gas})}^{\text{exp}}) \tag{9}$$

where $\sigma_{PH_3(\text{gas})}^{\text{calc}}$ is the chemical shielding of ^{31}P in PH_3 calculated at the same level of theory as the parent molecules constructed from the protein structure, $\sigma_{H_3PO_4(85\%\ \text{solution})}^{\text{exp}}$ is the absolute experimental chemical shielding of the 85% H_3PO_4 (328.4 ppm),[87] and $\sigma_{PH_3(\text{gas})}^{\text{exp}}$ is the absolute experimental chemical shielding of PH_3 (594.5 ppm).[87] (Method 2) references to the experimental absolute ^{31}P chemical shielding of 85% H_3PO_4 (eq 3). To the best of our knowledge, the approach of eq 3 has not been commonly applied. Yet, we propose to employ it here as an equivalent of the widely accepted use of liquid NH_3 in ^{15}N CS referencing. Finally, (method 3) references the ^{31}P CSs to the chemical shielding of ^{31}P in H_3PO_4 calculated at the same level of theory

$$\delta_X^{\text{calc}} = \sigma_{H_3PO_4(\text{gas})}^{\text{calc}} - \sigma_X^{\text{calc}} \tag{10}$$

The three referencing methods will be in the following referred to as calc-PH_3, liq-H_3PO_4, and calc-PH_3, respectively.

In Figure 10, we compare the distributions of the ^{31}P CSs calculated with the three referencing schemes. By far, the best agreement with the experiment is achieved using the liq-H_3PO_4 referencing and the b3lyp/6-311G(d) level of theory, for which we obtain δ_{iso}(pS19) = 3.5 ± 0.6 ppm and δ_{iso}(pS40) = 3.4 ± 0.7 ppm (Table 7) while the experimental values are δ_{iso}(pS19) = 3.76 ppm and δ_{iso}(pS40) = 4.18 ppm. Note that with the liq-H_3PO_4 referencing, the cancellation of errors stemming from the QM methodology cannot occur as eq 3 subtracts a calculated chemical shielding from an experimental one. This inconsistency obviously leads to an error cancellation as well, which is reflected in the very favorable agreement with

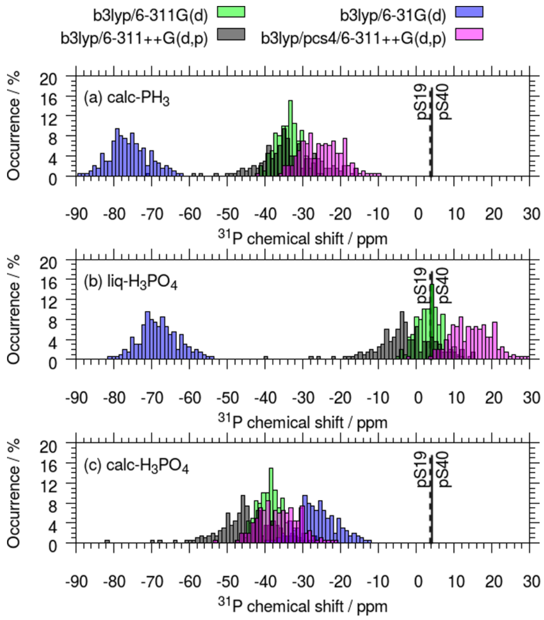


Figure 10. ^{31}P CSs obtained using different references for the conversion from the chemical shielding to the CS. CSs are given in ppm. (a) CSs are referenced to 85% H_3PO_4 using the secondary standard PH_3 computed at the same level of theory, (b) liquid 85% H_3PO_4 is used as a reference, and (c) H_3PO_4 computed at the same level of theory is used as a reference. The histograms include CSs of both pS19 and pS40, respectively. The dashed and solid vertical lines mark the experimental values of the ^{31}P CSs in pS19 and pS40, respectively.

Table 7. ^{31}P CSs (in ppm) of P Atoms in Phosphate Groups as Obtained Using Different References for the Conversion from the Chemical Shielding to the CS[a]

		^{31}P CSs/ppm		
method	ensemble	calc-PH_3[b]	liq-H_3PO_4[c]	calc-H_3PO_4[d]
		pS19		
b3lyp/pcs4/6-311++G(d,p)	B/100	−25.2 ± 1.0	14.5 ± 1.0	−36.5 ± 1.0
b3lyp/6-311++G(d,p)	B/100	−33.0 ± 1.5	−1.9 ± 1.5	−43.9 ± 1.5
bp86/6-311G(d)	A/100	−26.7 ± 0.7	5.8 ± 0.7	−39.0 ± 0.7
b3lyp/6-311G(d)	A/100	−33.8 ± 0.6	3.5 ± 0.6	−39.0 ± 0.6
b3lyp/6-31G(d)	A/100	−75.2 ± 1.0	−67.2 ± 1.0	−25.5 ± 1.0
experiment[e]		3.76		
		pS40		
b3lyp/pcs4/6-311++G(d,p)	B/100	−24.7 ± 1.1	15.0 ± 1.1	−36.0 ± 1.1
b3lyp/6-311++G(d,p)	B/100	−34.0 ± 1.7	−2.9 ± 1.7	−44.9 ± 1.7
bp86/6-311G(d)	A/100	−26.5 ± 0.8	6.0 ± 0.8	−38.8 ± 0.8
b3lyp/6-311G(d)	A/100	−33.9 ± 0.7	3.4 ± 0.7	−39.1 ± 0.7
b3lyp/6-31G(d)	A/100	−76.7 ± 1.0	−68.7 ± 1.0	−27.0 ± 1.0
experiment[e]		4.18		

[a]The CSs are reported as 95% CIs (eq 5). [b]CSs are referenced to 85% H_3PO_4 using the secondary standard PH_3 computed at the same level of theory (see eq 9). [c]85% solution of H_3PO_4 is used as a reference (see eq 3). [d]H_3PO_4 computed at the same level of theory is used as a reference (see eq 10). [e]ref 8.

experimental data. Yet, the errors have different origins, one being, for example, the neglect of the ∼37 ppm difference between the calculated (557.2 ppm, see the b3lyp/6-311G(d) result in Table S3) and experimental (594.5 ppm) chemical shielding of PH_3 (eq 9).

For the referencing with calc-PH_3 (Figure 10a) and calc-H_3PO_4 (Figure 10c), respectively, a very different dependence of the calculated CSs on the QM methodology is observed. This is a consequence of the fact that the chemical shielding of PH_3 and H_3PO_4 changes differently when one of the three methods is replaced by the other (Table S2). The cancellation of the QM-related errors does not work the same. For instance, when going from bp86/6-311G(d) to b3lyp/6-311G(d), the chemical shielding of PH_3 decreases by 4.8 ppm while the chemical shielding of H_3PO_4 increases by 2.4 ppm. Simultaneously, the chemical shielding of pS19/pS40 increases by 2.4/2.6 ppm. The chemical shieldings of PH_3 and the protein phosphates, respectively, thus change in opposite directions. This leads to the 7 ppm difference between bp86/6-311G(d) and b3lyp/6-311G(d) results already mentioned above. In contrast, the chemical shieldings of H_3PO_4 and the protein phosphates change in the same direction by an equal amount. The results obtained with bp86/6-311G(d) and b3lyp/6-311G(d) are equal as a result when the referencing to H_3PO_4 is employed. The described changes of the chemical shieldings and CSs demonstrate the effect of the DFT approximation.

The influence of the basis set on the ^{31}P NMR parameters can be traced by comparing the b3lyp results in the last two columns of Table S2. The replacement of 6-311G(d) by 6-31G(d) increases the chemical shielding of PH_3 by 29.3 ppm and the chemical shielding of H_3PO_4 by as much as 84.2 ppm. This is to be compared with the increase of the pS19 and pS40 chemical shielding by 70.7 and 72.1 ppm, respectively. PH_3 thus eliminates only less than a half of the error caused by the truncation of the basis set, thereby leaving >40 ppm difference. On the contrary, H_3PO_4 overcompensates the error merely by 10 ppm. The described effects overall lead to the trend unveiled by Figure 10. Under the H_3PO_4 referencing, the ^{31}P CSs obtained with the smaller 6-31G(d) basis set display a superior agreement with the experimental data than the ^{31}P CSs obtained with 6-311G(d). We have witnessed a similar behavior for ^{15}N CSs referenced to NH_3 computed at the same level of theory as the atoms of interest (see the paragraph Choice of the Referencing Method for ^{15}N CSs).

The ^{31}P CSs results just described can be easily understood. The chemical environment of phosphorus in the phosphoserine phosphate closely resembles that in H_3PO_4. Consequently, the systematic error in the chemical shielding is also similar, which facilitates the cancellation of systematic errors due to the finite basis set and approximations to the exchange−correlation energy of DFT. The compensation of errors between PH_3 and the protein phosphates is less effective. In general, the results presented here further support the findings of Fukal et al.,[46] who performed a benchmarking study of the ^{31}P chemical shielding in PH_3 and H_3PO_4. The dependence of $\delta_{iso}(^{31}P)$ on the basis set was found not very systematic, which led to notable variations of CSs in the phosphates of interest.

Effects of Systematic Errors on the Computed CSs. The agreement between the CSs computed here and the experimental data is affected by several systematic errors. The main source of errors are (1) the basis set dependence of the calculated chemical shielding, (2) the sample size, that is, the number of MD frames used for the statistical averaging, and (3) the quality of molecular geometries as provided by the MD force field. In the case of calculations employing the B/100 ensemble of MD frames, the error associated with the truncation of the explicit solvation to the first solvation shell adds to the list of systematics errors.

The basis set related systematic errors can be compensated by the choice of a suitable NMR reference, as we have thoroughly demonstrated above. The error compensation works only partially and especially in the case of ^{31}P its extent is hard to estimate in advance. In agreement with our findings, Fukal et al.[46] previously pointed out that the chemical shielding of ^{31}P in the reference compounds (PH_3, H_3PO_4) show a different dependence on the basis set than the molecule of interest. This is the consequence of the differences in the underlying electronic structures. As a result, ^{31}P CSs referenced with the calculated chemical shielding of PH_3 and H_3PO_4,

DOI: 10.1021/acs.jctc.8b00257

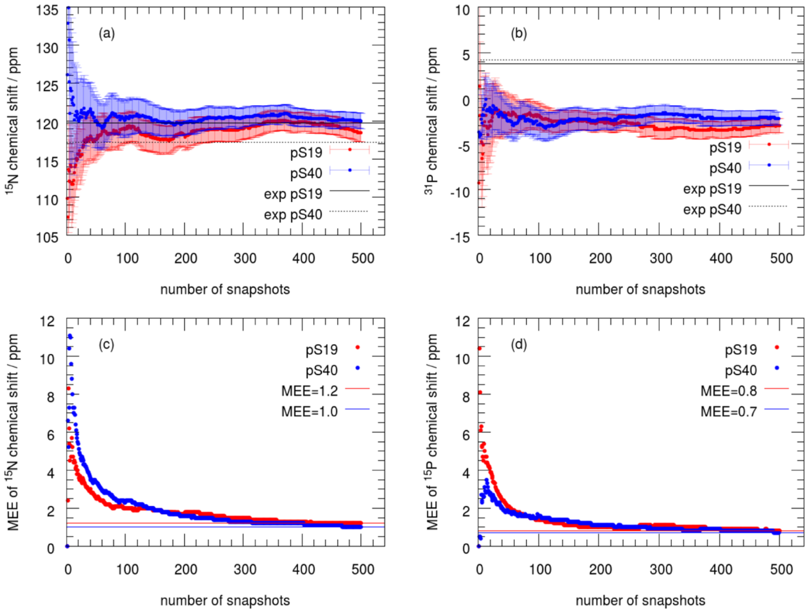


Figure 11. Dependence of the average ^{15}N (a) and ^{31}P (b) CS and the associated MEE (c,d) on the number of MD frames used in the statistical averaging. The dashed and solid lines in pictures (a,b) indicate the experimental values of the corresponding CSs in pS19 and pS40, respectively.

respectively, display numerical instabilities. That is why we propose to use the liq-H_3PO_4 referencing as a more transparent alternative to the established referencing schemes.

In order to inspect the influence of the statistical data set size, the ensemble of MD frames subject to MD/ADMA/DFT calculations was enlarged from 100 to 500 MD frames. Using the B/500 ensemble, CSs were calculated with b3lyp/6-311++G(d,p). To make the calculations computationally tractable, only water molecules of the first solvation shell were included in the molecular clusters. This corresponds to applying the r = 3.5 Å cutoff in the construction of parent molecules. Although samples >500 MD frames would lead to even beer converged ensemble averages (Figure 11a,b), the MEEs are converged sufficiently attfter 500 frames. The MEE for the ^{15}N and ^{31}P CS does not exceed 1.2 and 0.8 ppm, respectively. The additional 400 MD frames reduce the MEE approximately by a factor of 2 (Figure 11c,d). The width of the CI for ^{15}N and ^{31}P CSs reduces by ∼2−3 and ∼1−2 ppm, respectively. Consequently, the effect of the sample size on ^{15}N CSs is comparable to the effect of the basis set or the DFT functional (Table 4). On the contrary, the influence of the sample size on ^{31}P CSs is much smaller than the basis set effect (Table 5).

Because molecular geometries are overtaken directly from the MD simulation without any further geometry optimization, they suffer from incorrect bond lengths and angles caused by the approximations inherent to the force field of the classical MD simulation. While we are fully aware of the problem, there are several reasons for which we opt to dismiss the geometry optimization step here. First, the computed CSs are to be compared with data from NMR spectroscopy experiments. We therefore strive to calculate the properties of a thermodynamic ensemble of structures that is substantially different from an ensemble of energy-minimized structures resulting from a geometry optimization. Second, while a partial geometry optimization could serve here as a compromise solution, its successful execution is often not very straightforward. The parent molecules built from protein fragments tend to be sizable and have numerous degrees of freedom both in the protein of interest and in the H-bonding network. In such a case, the choice of geometrical parameters to be frozen/optimized during the partial optimization is not obvious. The goal of the partial optimization is to preserve the overall geometry of the MD snapshot and improve the bond lengths and valence angles at the same time. Several different combinations of frozen and relaxed parameters are possible, each introducing a substantial bias on the resulting geometries. Consequently, the NMR parameters computed for the partially optimized molecular geometries can significantly differ depending on the choice of frozen parameters. In addition, the number of fixed torsion angles often becomes large, which causes the system to be overconstrained, and the optimization fails to converge as a result. Further potential complications include the disruption of a realistic orientation of water molecules with respect to the solute. This could have severe consequences, given the demonstrated importance of the explicit solvent for the quantitative agreement with experiment. Equally importantly, the geometry optimization raises the computational time/memory/storage requirements of the MD/DFT calculations. The computational costs thus quickly become too high and make the MD/DFT calculations intractable, especially if a large number of fragments need to be calculated. For these reasons, the state-of-the-art MD/DFT studies of NMR CSs in biomolecules opt to dismiss the geometry optimization step, especially if realistic protein structures rather than model systems are studied. In the recent work by Fukal et al.,[46] geometry optimization was carried out for snapshots obtained from an MD simulation of diethylphosphate. Depending on the referencing scheme used for ^{31}P CSs, the statistically averaged CSs computed with b3lyp/IGLO-III increased by 10−20 ppm. However, we have to point out that the optimization employed only implicit solvation.

While a geometry reoptimization is beyond the scope of the present work, two alternative optimization approaches can potentially be applied in future to avoid the difficulties described above. First, the use of a normal mode geometry

DOI: 10.1021/acs.jctc.8b00257

optimization[90] can be considered to preserve the overall geometry of MD snapshots. Another option would be to apply a penalty function to restrain the key internal coordinates.[91] The two approaches can potentially be applied provided that the number of MD frames stays within reasonable limits.

CS Differences between pS19 and pS40. It was proposed[46] that errors related to the ^{31}P CS referencing and the quality of MD geometries can potentially be eliminated or reduced by calculating the relative $\Delta\delta(^{31}P)$ CS, defined as $\Delta\delta(^{31}P) = \delta(^{31}P,\text{phosphate 1}) - \delta(^{31}P,\text{phosphate 2})$, which is equivalent to the difference between the chemical shieldings $\Delta\delta(^{31}P) \equiv \delta(^{31}P,\text{phosphate 1}) - \delta(^{31}P,\text{phosphate 2})$. Because the pS19 and pS40 satisfy the condition of chemical equivalence, they are well suited for the relative $\Delta\delta$ calculation that avoids the use of an external reference. Table 8 shows the MD-averaged ^{31}P CSs of pS19 and pS40. The CS difference $\delta(\text{pS19}) - \delta(\text{pS40})$ of −0.2 ppm does closely approach the experimental CS difference of −0.42 ppm. However, MEEs obtained for the ^{31}P CSs of both residues are as large as 0.8 ppm. The average CS of pS19 lies within the error bar of pS40 and vice versa, and the two ^{31}P CSs are therefore identical numbers in practice. The size of the error is too large also for the chemical shielding. The difference between the chemical shielding of pS19 (322.6 ppm) and pS40 (322.4 ppm) is 0.2 ppm while the MEE amounts to 0.8 ppm. The average chemical shielding for one residue thus falls within the error bar of the other residue. Calculating the relative $\Delta\delta$ is therefore not relevant here. To be able to reproduce the very small size of the ^{31}P CS difference ($\Delta\delta_{exp}(\text{pS19} - \text{pS40}) = -0.42$ ppm), the MEEs would have to reduce significantly. This could be potentially achieved[46] by the geometry optimization of MD snapshots as discussed above.

Table 8. Comparison of the ^{31}P CSs and Chemical Shieldings of pS19 and pS40, Obtained at the bp86/6-311G(d) Level

	$\delta_{\text{liq-H}_3\text{PO}_4}$/ppm[a]	σ/ppm	δ_{exp}/ppm[b]
pS19	5.8 ± 0.8	322.6 ± 0.8	3.76
pS40	6.0 ± 0.8	322.4 ± 0.8	4.18
$\Delta_{\text{pS19−pS40}}$[c]	−0.2	0.2	−0.42

[a] CSs referenced to the 85% H_3PO_4 (eq 3). [b] Experimental CSs taken from ref 8. [c] $\Delta_{\text{pS19−pS40}}$ is the difference between the ^{31}P CS or chemical shielding of pS19 and pS40, respectively, that is, $\Delta_{\text{pS19−pS40}} = \delta(\text{pS19}) - \delta(\text{pS40}) = \sigma(\text{pS19}) - \sigma(\text{pS40})$.

Effect of Phosphorylation. It is known that the phosphorylation of amino acids influences the conformation of peptides.[92−94] The recent NMR study of the RD of hTH1 demonstrated that phosphorylation slightly decreases the alpha-helical propensity around the S19 phosphorylation site. Figure 12 shows the backbone torsion angles as obtained from the MD simulation of the nonphosphorylated and doubly phosphorylated hTH1. The ϕ torsion angle in S19 spans a broad interval of −20° to 180°, whereas the same angle in S40 covers a smaller interval of (50°, 180°). In contrast, the range of ψ is narrower in pS19 (80−180°) than in pS40 (−10° to 180° degrees). Differences between the CSs of the non-phosphorylated and doubly phosphorylated hTH1 thus reflect not only the direct electronic effect of the phosphate group but also the effect of the induced conformational change. In the experiment, the most significant CS change (2.7 and 2.8 ppm for pS19 and pS40, respectively) upon phosphorylation is observed for the CB carbon (Table 9). The MD/DFT calculations correctly predict the downfield shift induced by phosphorylation as well and estimate it to 3.6 ppm (pS19) and 2.3 ppm (pS40), respectively. The CB CS can thus safely be used as a probe of effects coupled with the phosphorylation. The same observation was recently reported in Louša et al.[9] based on the experimental NMR data. The CB CS change should therefore be incorporated into the databases underlying the CS prediction programs.

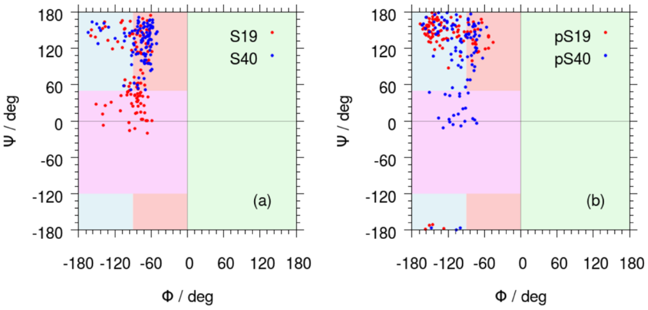


Figure 12. Ramachandran plot of torsion angles in (a) residues S19 and S40 of the nonphosphorylated hTH1 and in (b) residues pS19 and pS40 of the doubly phosphorylated hTH1.

Table 9. CSs in the S19, S40, and pS19, pS40 Residues of the Nonphosphorylated and Doubly Phosphorylated hTH1, Respectively[a]

	non-phosphorylated		phosphorylated	
atom	δ_{calc}/ppm[b]	δ_{exp}/ppm[c]	δ_{calc}/ppm[b]	δ_{exp}/ppm[c]
	S19		pS19	
P	n.a.	n.a.	5.8 ± 0.7	3.76
H^N	6.77 ± 0.18	8.429	6.26 ± 0.16	8.599
N	120.5 ± 1.3	119.0	118.4 ± 1.8	119.7
C′	168.3 ± 0.6	174.9	170.5 ± 0.6	174.3
Ca	58.5 ± 0.7	58.27	58.7 ± 0.6	59.13
Ha	4.63 ± 0.08	4.454	4.52 ± 0.10	4.316
Cb	65.1 ± 0.6	64.02	68.7 ± 0.6	66.75
Hb1	4.38 ± 0.07	3.848	3.81 ± 0.10	3.905
Hb2	4.05 ± 0.09	3.917	3.56 ± 0.09	3.905
	S40		pS40	
P	n.a.	n.a.	6.0 ± 0.8	4.18
H^N	6.75 ± 0.17	8.392	6.28 ± 0.19	8.841
N	120.6 ± 1.6	117.1	118.2 ± 1.7	117.2
C′	169.9 ± 0.5	174.7	170.9 ± 0.6	174.4
Ca	58.6 ± 0.7	58.88	57.2 ± 0.6	59.78
Ha	4.61 ± 0.08	4.394	4.5 ± 0.09	4.238
Cb	65.6 ± 0.6	63.63	67.9 ± 0.6	66.30
Hb1	4.31 ± 0.06	3.873	3.85 ± 0.10	3.904
Hb2	4.11 ± 0.08	3.873	3.85 ± 0.11	3.904

[a] The CSs are reported as 95% CIs (eq 5). ^{15}N CSs were referenced to NH_3/CH_3NH_2 (eq 2) and ^{31}P CSs were referenced to 85% H_3PO_4 (eq 3). [b] Computed CSs obtained with bp86/6-311G(d). [c] Experimental ^{1}H, ^{13}C, and ^{15}N CSs are taken from the Supporting Information of ref 9, ^{31}P CSs are taken from ref 8.

■ CONCLUSIONS

We have demonstrated that the accuracy of the MD/ADMA/DFT approach for the calculation of CSs in phosphorylation sites of IDPs is affected by many factors, including the basis set, molecular geometries, statistical sample size, CS referencing, and the size of explicitly treated surroundings. Calculations of CSs for atoms affected by hydrogen-bonding with other protein moieties and solvent are particularly difficult to compute. This finding is in line with conclusions previously made for the NMA model[32,43] of the protein peptide bond.

The MD ensemble of pS19 and pS40 geometries within the hTH1 peptide (region 1−50) augmented by the explicit solvent and protein surroundings displays large CS variations. The nonpolar (HA, HB1, HB2) and polar (H^N) proton CSs span a range of ∼3−4 ppm and up to 6 ppm, respectively, depending on the conformation of the fragment in question and the arrangement of the surrounding explicit water molecules. The ^{13}C CSs vary within up to 20 ppm for the aliphatic (CA, CB) and up to 30 ppm for the carbonyl (C′) carbons. Particularly large ensemble variations have been observed for ^{15}N CSs that range within as much as 50 ppm. Finally, the conformational and solvent shell changes lead to ^{31}P CSs fluctuating within 25 ppm. The range of ^{31}P CSs obtained computationally thus substantially exceeds the several-ppm-wide range found in the experiment.

The CS variations in the MD structural ensemble are reflected by MEEs, which were systematically higher for average CSs computed with the B/100 ensemble than for the A/100 ensemble. Nevertheless, MEEs improved with averaging over the extended B/500 set of MD snapshots. The resulting MEE is about the same for the nonpolar and polar proton CSs (up to 0.20 ppm). MEE for ^{13}C CSs amounts to 0.7−0.9 ppm, whereas it is as large as 1.2 ppm for ^{15}N CSs. This indicates that reproducing the sequence trends will be more difficult for ^{15}N CSs than for ^{13}C CSs, which agrees with previous findings of other authors.[32,35] The average ^{31}P CSs are calculated with an MEE of 0.9 ppm, which is close to the 0.42 ppm difference between the experimental ^{31}P CSs of pS19 and pS40. This makes the discrimination between the two residues of hTH1 a difficult task within the computational setup employed here.

Provided a suitable reference is chosen, the use of an explicit solvent in ensemble calculations uniformly improves the quantitative agreement with experiment for all nuclei considered in this work. The H-bonding interactions significantly broaden the CS distributions of polar protons and phosphoruses. Consequently, the CS ranges increase by 1.5 and 10 ppm for H^N and P atoms, respectively, compared to implicit solvent-only calculations.

Regardless of the level of theory used, the results for nonpolar proton CSs deviate from the experimental values by ∼0.5 ppm or less. If a large basis set is applied, rather good ensemble averages are obtained also for the polar protons. The b3lyp/pcs4/6-311++G(d,p) level of theory underestimates the CSs for polar protons by 1−2 ppm. The best predictions of CSs for the aliphatic carbons (CA, CB) were achieved with the modest b3lyp/6-311G(d) theoretical description. b3lyp/6-311++G(d,p) overestimates the carbonyl carbon CSs by 3−4 ppm and the replacement of the 6-311++G(d,p) basis set on atoms of the original fragment deteriorates the results. The absolute agreement with the experiment for ^{15}N and ^{31}P CSs depends on the efficiency of the error cancellation, which is strongly affected by the choice of the referencing method. The best agreement of the computed ^{15}N CSs with the experiment has been found using a secondary reference scheme employing liquid ammonia and CH_3NH_2 as the primary and secondary standards, respectively. Within this referencing scheme, the ^{15}N CSs computed with b3lyp/6-311++G(d,p) differ from the experiment by no more than ∼3 ppm. For ^{31}P, referencing to the 85% solution of H_3PO_4 yields computed CSs that differ from experiment by 6−7 ppm (b3lyp/6-311++G(d,p)) or less than 1 ppm (b3lyp/6-311G(d)).

^{15}N CSs computed with bp86/6-311G(d) are overestimated by ∼30 ppm when referenced to NH_3 calculated at the same level theory. 20 ppm out of the 30 ppm overestimation comes from the neglect of the gas-to-liquid shift of NH_3. ^{31}P CSs are underestimated by 30 ppm or more when referenced to either PH_3 secondary reference or H_3PO_4 computed at the same level of theory. This systematic error, partly caused by inaccurate MD molecular geometries, is eliminated here by changing the reference to the absolute ^{31}P chemical shielding of 85% H_3PO_4. In this way, a favorable quantitative agreement with experiment is achieved by neglecting the ∼37 ppm difference between the calculated and experimental ^{31}P chemical shielding of PH_3.

Calculations employing b3lyp/6-311G(d) and 85% H_3PO_4 reference provide the ^{31}P CSs of 3.5 ± 0.6 and 3.4 ± 0.7 ppm in pS19 and pS40, respectively. Because of the small difference between the two MD-averaged values and large standard deviations of the mean, the MD/ADMA/DFT calculations were not able to discriminate the ^{31}P CSs of the two residues in question.

Upon the phosphorylation of hTH1, the computed CB CS increases by 3.6 ppm in the pS19 and by 2.3 ppm in the pS40 residue, respectively. This is in good agreement with the experimental 2.7 and 2.8 ppm downfield shift of pS19 and pS40, respectively. It can be thus concluded that unlike the standard CS prediction protocols designed for well-structured proteins, the MD/ADMA/DFT calculations are capable of correctly predicting the phosphorylation-induced CS changes.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.8b00257.

> ^1H CSs of HA, HB1, and HB2; calculated ^{15}N chemical shielding in the reference compounds as well in the pS19 and pS40 residues; calculated ^{31}P chemical shielding in the reference compounds as well in the pS19 and pS40 residues; time dependence of the ^{31}P chemical shielding in pS19 and pS40; time dependence of the difference between the ^{31}P chemical shielding of pS19 and pS40; and structures of the doubly phosphorylated and nonphosphorylated hTH1 used as starting coordinates for the MD simulation (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: precechj@faf.cuni.cz. Phone: +420 549 49 3847.

ORCID

Jana Pavlíková Přecechtělová: 0000-0003-0844-2666
Jozef Hritz: 0000-0002-4512-9241

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Thomas E. Exner (University of Tübingen, Germany) for providing the Python suite of codes for the ADMA workflow procedure and the related technical assistance. This research was financed by Czech Science Foundation (no. GF15-34684L) and by the Ministry of Education, Youth and Sports of the Czech Republic (MEYS CR) under the National Sustainability Programme II, project CEITEC 2020 (LQ1601). Computational resources were provided by the projects CESNET (LM2015042) and CERIT Scientific Cloud (LM2015085) under the program "Projects of Large Research, Development, and Innovations Infrastructures" funded by MEYS CR, and by the IT4Innovations National Supercomputing Center (LM2015070) under the program "Large Infrastructures for Research, Experimental Development and Innovations" funded by MEYS CR.

■ REFERENCES

(1) Oldfield, C. J.; Dunker, A. K. Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. *Annu. Rev. Biochem.* **2014**, *83*, 553−584.

(2) Dunker, A. K.; Babu, M. M.; Barbar, E.; Blackledge, M.; Bondos, S. E.; Dosztányi, Z.; Dyson, H. J.; Forman-Kay, J.; Fuxreiter, M.; Gsponer, J.; Han, K.-H.; Jones, D. T.; Longhi, S.; Metallo, S. J.; Nishikawa, K.; Nussinov, R.; Obradovic, Z.; Pappu, R. V.; Rost, B.; Selenko, P.; Subramaniam, V.; Sussman, J. L.; Tompa, P.; Uversky, V. N. What's in a Name? Why These Proteins are Intrinsically Disordered. *Intrinsically Disord. Proteins* **2013**, *1*, No. e24157.

(3) Dyson, H. J.; Wright, P. E. Intrinsically Unstructured Proteins and Their Functions. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197−208.

(4) Ward, J. J.; Sodhi, J. S.; McGuffin, L. J.; Buxton, B. F.; Jones, D. T. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol.* **2004**, *337*, 635−645.

(5) Uversky, V. N. Natively Unfolded Proteins: A Point Where Biology Waits for Physics. *Protein Sci.* **2002**, *11*, 739−756.

(6) Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradović, Z. Intrinsic Disorder and Protein Function. *Biochemistry* **2002**, *41*, 6573−6582.

(7) Steinacker, P.; Aitken, A.; Otto, M. 14-3-3 Proteins in Neurodegeneration. *Semin. Cell Dev. Biol.* **2011**, *22*, 696−704.

(8) Hritz, J.; Byeon, I.-J. L.; Krzysiak, T.; Martinez, A.; Sklenar, V.; Gronenborn, A. M. Dissection of Binding between a Phosphorylated Tyrosine Hydroxylase Peptide and 14-3-3ζ: A Complex Story Elucidated by NMR. *Biophys. J.* **2014**, *107*, 2185−2194.

(9) Louša, P.; Nedozrálová, H.; Župa, E.; Nováček, J.; Hritz, J. Phosphorylation of the Regulatory Domain of Human Tyrosine Hydroxylase 1 Monitored Using Non-Uniformly Sampled NMR. *Biophys. Chem.* **2017**, *223*, 25−29.

(10) Shimada, T.; Fournier, A. E.; Yamagata, K. Neuroprotective Function of 14-3-3 Proteins in Neurodegeneration. *BioMed Res. Int.* **2013**, *2013*, 564534.

(11) Oroguchi, T.; Ikeguchi, M.; Sato, M. Towards the Structural Characterization of Intrinsically Disordered Proteins by SAXS and MD Simulation. *J. Phys.: Conf. Ser.* **2011**, *272*, 012005.

(12) Kosol, S.; Contreras-Martos, S.; Cedeño, C.; Tompa, P. Structural Characterization of Intrinsically Disordered Proteins by NMR Spectroscopy. *Molecules* **2013**, *18*, 10802−10828.

(13) Ytreberg, F. M.; Borcherds, W.; Wu, H.; Daughdrill, G. W. Using chemical shifts to generate structural ensembles for intrinsically disordered proteins with converged distributions of secondary structure. *Intrinsically Disord. Proteins* **2015**, *3*, No. e984565.

(14) Nielsen, J. T.; Eghbalnia, H. R.; Nielsen, N. C. Chemical Shift Prediction for Protein Structure Calculation and Quality Assessment Using an Optimally Parameterized Force Field. *Prog. Nucl. Magn. Reson. Spectrosc.* **2012**, *60*, 1−28.

(15) Shen, Y.; Bax, A. SPARTA+: A Modest Improvement in Empirical NMR Chemical Shift Prediction by Means of an Artificial Neural Network. *J. Biomol. NMR* **2010**, *48*, 13−22.

(16) Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S. SHIFTX2: Significantly Improved Protein Chemical Shift Prediction. *J. Biomol. NMR* **2011**, *50*, 43−57.

(17) Meiler, J. PROSHIFT: Protein Chemical Shift Prediction Using Artificial Neural Networks. *J. Biomol. NMR* **2003**, *26*, 25−37.

(18) Vila, J. A.; Arnautova, Y. A.; Martin, O. A.; Scheraga, H. A. Quantum-Mechanics-Derived ^{13}Cα Chemical Shift Server (CheShift) for Protein Structure Validation. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 16972−16977.

(19) Xu, X.-P.; Case, D. A. Automated Prediction of ^{15}N, ^{13}C$^{\alpha}$, ^{13}C$^{\beta}$ and ^{13}C′ Chemical Shifts in Proteins Using a Density Functional Database. *J. Biomol. NMR* **2001**, *21*, 321−333.

(20) Larsen, A. S.; Bratholm, L. A.; Christensen, A. S.; Channir, M.; Jensen, J. H. ProCS15: a DFT-Based Chemical Shift Predictor for Backbone and Cβ atoms in Proteins. *PeerJ* **2015**, *3*, No. e1344.

(21) Sumowski, C. V.; Hanni, M.; Schweizer, S.; Ochsenfeld, C. Sensitivity of ab Initio vs Empirical Methods in Computing Structural Effects on NMR Chemical Shifts for the Example of Peptides. *J. Chem. Theory Comput.* **2014**, *10*, 122−133.

(22) Mulder, F. A. A.; Filatov, M. NMR Chemical Shift Data and Ab Initio Shielding Calculations: Emerging Tools for Protein Structure Determination. *Chem. Soc. Rev.* **2010**, *39*, 578−590.

(23) Sitkoff, D.; Case, D. A. Density Functional Calculations of Proton Chemical Shifts in Model Peptides. *J. Am. Chem. Soc.* **1997**, *119*, 12262−12273.

(24) Case, D. A. Chemical Shifts in Biomolecules. *Curr. Opin. Struct. Biol.* **2013**, *23*, 172−176.

(25) Frank, A.; Onila, I.; Möller, H. M.; Exner, T. E. Toward the Quantum Chemical Calculation of Nuclear Magnetic Resonance Chemical Shifts of Proteins. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 2189−2202.

(26) He, X.; Wang, B.; Merz, K. M. Protein NMR Chemical Shift Calculations Based on the Automated Fragmentation QM/MM Approach. *J. Phys. Chem. B* **2009**, *113*, 10380−10388.

(27) He, X.; Zhu, T.; Wang, X.; Liu, J.; Zhang, J. Z. H. Fragment Quantum Mechanical Calculation of Proteins and Its Applications. *Acc. Chem. Res.* **2014**, *47*, 2748−2757.

(28) Zhu, T.; Zhang, J. Z. H.; He, X. In *Advance in Structural Bioinformatics*; Wei, D., Xu, Q., Zhao, T., Dai, H., Eds.; Springer Netherlands: Dordrecht, 2015; pp 49−70.

(29) Jose, K. V. J.; Raghavachari, K. Fragment-Based Approach for the Evaluation of NMR Chemical Shifts for Large Biomolecules Incorporating the Effects of the Solvent Environment. *J. Chem. Theory Comput.* **2017**, *13*, 1147−1158.

(30) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chem. Rev.* **2012**, *112*, 632−672.

(31) Frank, A.; Möller, H. M.; Exner, T. E. Toward the Quantum Chemical Calculation of NMR Chemical Shifts of Proteins. 2. Level of Theory, Basis Set, and Solvents Model Dependence. *J. Chem. Theory Comput.* **2012**, *8*, 1480−1492.

(32) Exner, T. E.; Frank, A.; Onila, I.; Möller, H. M. Toward the Quantum Chemical Calculation of NMR Chemical Shifts of Proteins. 3. Conformational Sampling and Explicit Solvents Model. *J. Chem. Theory Comput.* **2012**, *8*, 4818−4827.

(33) Victora, A.; Möller, H. M.; Exner, T. E. Accurate Ab Initio Prediction of NMR Chemical Shifts of Nucleic Acids and Nucleic Acids/Protein Complexes. *Nucleic Acids Res.* **2014**, *42*, No. e173.

(34) Zhu, T.; He, X.; Zhang, J. Z. H. Fragment Density Functional Theory Calculation of NMR Chemical Shifts for Proteins with Implicit Solvation. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7837−7845.

(35) Zhu, T.; Zhang, J. Z. H.; He, X. Automated Fragmentation QM/MM Calculation of Amide Proton Chemical Shifts in Proteins with Explicit Solvent Model. *J. Chem. Theory Comput.* **2013**, *9*, 2104−2114.

(36) Swails, J.; Zhu, T.; He, X.; Case, D. A. AFNMR: Automated Fragmentation Quantum Mechanical Calculation of NMR Chemical Shifts for Biomolecules. *J. Biomol. NMR* **2015**, *63*, 125−139.

(37) Potrzebowski, M. J.; Assfeld, X.; Ganicz, K.; Olejniczak, S.; Cartier, A.; Gardiennet, C.; Tekely, P. An Experimental and Theoretical Study of the ^{13}C and ^{31}P Chemical Shielding Tensors

in O-Phosphorylated Amino Acids. *J. Am. Chem. Soc.* **2003**, *125*, 4223−4232.

(38) Hoffmann, R.; Reichert, I.; Wachs, W. O.; Zeppezauer, M.; Kalbitzer, H. R. ^{1}H and ^{31}P NMR Spectroscopy of Phosphorylated Model Peptides. *Int. J. Pept. Protein Res.* **1994**, *44*, 193−198.

(39) Matheis, G.; Whitaker, J. R. ^{3}1P NMR Chemical Shifts of Phosphate Covalently Bound to Proteins. *Int. J. Biochem.* **1984**, *16*, 867−873.

(40) Scheurer, C.; Skrynnikov, N. R.; Lienin, S. F.; Straus, S. K.; Brüschweiler, R.; Ernst, R. R. Effects of Dynamics and Environment on N-15 Chemical Shielding Anisotropy in Proteins. A Combination of Density Functional Theory, Molecular Dynamics Simulation, and NMR Relaxation. *J. Am. Chem. Soc.* **1999**, *121*, 4242−4251.

(41) Vícha, J.; Babinský, M.; Demo, G.; Otrusinová, O.; Jansen, S.; Pekárová, B.; Žídek, L.; Munzarová, M. L. The Influence of Mg2+ Coordination on ^{13}C and ^{15}N Chemical Shifts in CKI1RD Protein Domain from Experiment and Molecular Dynamics/Density Functional Theory Calculations. *Proteins: Struct., Funct., Bioinf.* **2016**, *84*, 686−699.

(42) Dračínský, M.; Bouř, P. Computational Analysis of Solvent Effects in NMR Spectroscopy. *J. Chem. Theory Comput.* **2010**, *6*, 288−299.

(43) Dračínský, M.; Möller, H. M.; Exner, T. E. Conformational Sampling by Ab Initio Molecular Dynamics Simulations Improves NMR Chemical Shift Predictions. *J. Chem. Theory Comput.* **2013**, *9*, 3806−3815.

(44) Přecechtělová, J.; Novák, P.; Munzarová, M. L.; Kaupp, M.; Sklenář, V. Phosphorus Chemical Shifts in a Nucleic Acid Backbone from Combined Molecular Dynamics and Density Functional Calculations. *J. Am. Chem. Soc.* **2010**, *132*, 17139−17148.

(45) Přecechtělová, J.; Munzarová, M. L.; Vaara, J.; Novotný, J.; Dračínský, M.; Sklenář, V. Toward Reproducing Sequence Trends in Phosphorus Chemical Shifts for Nucleic Acids by MD/DFT Calculations. *J. Chem. Theory Comput.* **2013**, *9*, 1641−1656.

(46) Fukal, J.; Páv, O.; Buděšínský, M.; Šebera, J.; Sychrovský, V. The Benchmark of ^{31}P NMR Parameters in Phosphate: a Case Study on Structurally Constrained and Flexible Phosphate. *Phys. Chem. Chem. Phys.* **2017**, *19*, 31830−31841.

(47) Benda, L.; Schneider, B.; Sychrovský, V. Calculating the Response of NMR Shielding Tensor σ(31P) and 2J(31P,13C) Coupling Constants in Nucleic Acid Phosphate to Coordination of the Mg2+ Cation. *J. Phys. Chem. A* **2011**, *115*, 2385−2395.

(48) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **2015**, *119*, 5113−5123.

(49) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D., Jr CHARMM36m: an Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2017**, *14*, 71−73.

(50) Sarotti, A. M.; Pellegrinet, S. C. A Multi-Standard Approach for GIAO ^{13}C NMR Calculations. *J. Org. Chem.* **2009**, *74*, 7254−7260.

(51) Cai, L.; Fushman, D.; Kosov, D. S. Density Functional Calculations of ^{15}N Chemical Shifts in Solvated Dipeptides. *J. Biomol. NMR* **2008**, *41*, 77−88.

(52) van Wüllen, C. A Comparison of Density Functional Methods for the Calculation of Phosphorus-31 NMR Chemical Shifts. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2137−2144.

(53) Latypov, S. K.; Polyancev, F. M.; Yakhvarov, D. G.; Sinyashin, O. G. Quantum Chemical Calculations of ^{31}P NMR Chemical Shifts: Scopes and Limitations. *Phys. Chem. Chem. Phys.* **2015**, *17*, 6976−6987.

(54) Chesnut, D. B. Theoretical Study of ^{31}P NMR Chemical Shielding Models for Concentrated Phosphoric Acid Solution. *J. Phys. Chem. A* **2005**, *109*, 11962−11966.

(55) Patchkovskii, S.; Ziegler, T. Phosphorus NMR Chemical Shifts with Self-Interaction Free, Gradient-Corrected DFT. *J. Phys. Chem. A* **2002**, *106*, 1088−1099.

(56) Pascual-Borràs, M.; López, X.; Poblet, J. M. Accurate Calculation of ^{31}P NMR Chemical Shifts in Polyoxometalates. *Phys. Chem. Chem. Phys.* **2015**, *17*, 8723−8731.

(57) Chernyshev, K. A.; Larina, L. I.; Chirkina, E. A.; Krivdin, L. B. The Effects of Intramolecular and Intermolecular Coordination on ^{31}P Nuclear Shielding: Phosphorylated Azoles. *Magn. Reson. Chem.* **2012**, *50*, 120−127.

(58) Přecechtělová, J.; Munzarová, M. L.; Novák, P.; Sklenář, V. Relationships between P-31 Chemical Shift Tensors and Conformation of Nucleic Acid Backbone: A DFT Study. *J. Phys. Chem. B* **2007**, *111*, 2658−2667.

(59) Přecechtělová, J.; Padrta, P.; Munzarová, M. L.; Sklenář, V. P-31 Chemical Shift Tensors for Canonical and Non-Canonical Conformations of Nucleic Acids: A DFT Study and NMR Implications. *J. Phys. Chem. B* **2008**, *112*, 3470−3478.

(60) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950−1958.

(61) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* **2006**, *65*, 712−725.

(62) Homeyer, N.; Horn, A. H. C.; Lanig, H.; Sticht, H. AMBER Force-Field Parameters for Phosphorylated Amino Acids in Different Protonation States: Phosphoserine, Phosphothreonine, Phosphotyrosine, and Phosphohistidine. *J. Mol. Model.* **2006**, *12*, 281−289.

(63) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.

(64) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.

(65) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(66) Evans, D. J.; Holian, B. L. The Nose-Hoover Thermostat. *J. Chem. Phys.* **1985**, *83*, 4069−4074.

(67) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.

(68) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16*, Revision A.03; Gaussian, Inc.: Wallingford, CT, 2016.

(69) London, F. Théorie quantique des courants interatomiques dans les combinaisons aromatiques. *J. Phys. Radium* **1937**, *8*, 397−409.

(70) McWeeny, R. Perturbation Theory for the Fock-Dirac Density Matrix. *Phys. Rev.* **1962**, *126*, 1028−1034.

(71) Ditchfield, R. Self-Consistent Perturbation Theory of Diamagnetism. *Mol. Phys.* **1974**, *27*, 789−807.

(72) Wolinski, K.; Hinton, J. F.; Pulay, P. Efficient Implementation of the Gauge-Independent Atomic Orbital method for NMR Chemical Shift Calculations. *J. Am. Chem. Soc.* **1990**, *112*, 8251−8260.

(73) Cheeseman, J. R.; Trucks, G. W.; Keith, T. A.; Frisch, M. J. A Comparison of Models for Calculating Nuclear Magnetic Resonance Shielding Tensors. *J. Chem. Phys.* **1996**, *104*, 5497−5509.

(74) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098−3100.

(75) Perdew, J. P. Density-Functional Approximation for the Correlation Energy of the Inhomogeneous Electron Gas. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1986**, *33*, 8822−8824.

(76) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(77) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785−789.

(78) Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. Results Obtained with the Correlation Energy Density Functionals of Becke and Lee, Yang and Parr. *Chem. Phys. Lett.* **1989**, *157*, 200−206.

(79) McLean, A. D.; Chandler, G. S. Contracted Gaussian Basis Sets for Molecular Calculations. I. Second Row Atoms, Z=11−18. *J. Chem. Phys.* **1980**, *72*, 5639−5648.

(80) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-Consistent Molecular Orbital Methods. XX. A Basis Set for Correlated Wave Functions. *J. Chem. Phys.* **1980**, *72*, 650−654.

(81) Hariharan, P. C.; Pople, J. A. The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies. *Theor. Chim. Acta* **1973**, *28*, 213−222.

(82) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. Self-Consistent Molecular Orbital Methods. XXIII. A Polarization-Type Basis Set for Second-Row Elements. *J. Chem. Phys.* **1982**, *77*, 3654−3665.

(83) Barone, V.; Cossi, M. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J. Phys. Chem. A* **1998**, *102*, 1995−2001.

(84) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. Energies, Structures, and Electronic Properties of Molecules in Solution with the C-PCM Solvation Model. *J. Comput. Chem.* **2003**, *24*, 669−681.

(85) Jameson, C. J.; Jameson, A. K.; Oppusunggu, D.; Wille, S.; Burrell, P. M.; Mason, J. ^{15}N Nuclear Magnetic Shielding Scale from Gas Phase Studies. *J. Chem. Phys.* **1981**, *74*, 81−88.

(86) Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*, 2nd ed.; Wiley: New York, 1961; p 347.

(87) Jameson, C. J.; De Dios, A.; Keith Jameson, A. Absolute Shielding Scale for ^{31}P from Gas-Phase NMR studies. *Chem. Phys. Lett.* **1990**, *167*, 575−582.

(88) Vícha, J.; Babinský, M.; Demo, G.; Otrusinová, O.; Jansen, S.; Pekárová, B.; Žídek, L.; Munzarová, M. L. The Influence of Mg^{2+} Coordination on ^{13}C and ^{15}N Chemical Shifts in CKI1RD Protein Domain from Experiment and Molecular Dynamics/Density Functional Theory Calculations. *Proteins: Struct., Funct., Bioinf.* **2016**, *84*, 686−699.

(89) Wishart, D. S.; Bigam, C. G.; Yao, J. ^{1}H, ^{13}C and ^{15}N Chemical Shift Referencing in Biomolecular NMR. *J. Biomol. NMR* **1995**, *6*, 135−140.

(90) Bouř, P.; Keiderling, T. A. Partial Optimization of Molecular Geometry in Normal Coordinates and Use as a Tool for Simulation of Vibrational Spectra. *J. Chem. Phys.* **2002**, *117*, 4126−4132.

(91) Kruse, H.; Šponer, J. Towards Biochemically Relevant QM Computations on Nucleic Acids: Controlled Electronic Structure Geometry Optimization of Nucleic Acid Structural Motifs Using Penalty Restraint Functions. *Phys. Chem. Chem. Phys.* **2015**, *17*, 1399−1410.

(92) Broncel, M.; Wagner, S. C.; Paul, K.; Hackenberger, C. P. R.; Koksch, B. Towards Understanding Secondary Structure Transitions: Phosphorylation and Metal Coordination in Model Peptides. *Org. Biomol. Chem.* **2010**, *8*, 2575−2579.

(93) Tholey, A.; Lindemann, A.; Kinzel, V.; Reed, J. Direct Effects of Phosphorylation on the Preferred Backbone Conformation of Peptides: A Nuclear Magnetic Resonance Study. *Biophys. J.* **1999**, *76*, 76−87.

(94) Fujitani, N.; Kanagawa, M.; Aizawa, T.; Ohkubo, T.; Kaya, S.; Demura, M.; Kawano, K.; Nishimura, S.-i.; Taniguchi, K.; Nitta, K. Structure Determination and Conformational Change Induced by Tyrosine Phosphorylation of the N-Terminal Domain of the α-Chain of Pig Gastric H+/K+-ATPase. *Biochem. Biophys. Res. Commun.* **2003**, *300*, 223−229.

Paper 15

Jansen, S.;Melková, K.; Trošanová, Z.; Hanáková, K.; Zachrdla, M.; Nováček, J.; Župa, E.; Zdráhal, Z.; Hritz, J.*; Žídek, L.*: Quantitative Mapping of MAP2c Phosphorylation and 14-3-3ζ Binding Sites Reveals Key Differences Between MAP2c and Tau. *J. Biol. Chem.* **2017**, 292, 6715-6727

Quantitative mapping of microtubule-associated protein 2c (MAP2c) phosphorylation and regulatory protein 14-3-3ζ-binding sites reveals key differences between MAP2c and its homolog Tau

Received for publication, December 7, 2016, and in revised form, March 1, 2017 Published, Papers in Press, March 3, 2017, DOI 10.1074/jbc.M116.771097

Séverine Jansen[‡§], **Kateřina Melková**[‡§], **Zuzana Trošanová**[‡§], **Kateřina Hanáková**[§], **Milan Zachrdla**[‡§], **Jiří Nováček**[§], **Erik Župa**[‡§], **Zbyněk Zdráhal**[§], **Jozef Hritz**[‡§1], **and Lukáš Žídek**[‡§2]

From the [‡]*National Centre for Biomolecular Research, Faculty of Science, and the* [§]*Central European Institute of Technology, Masaryk University, Kamenice 5, CZ-62500 Brno, Czech Republic*

Edited by Norma Allewell

Microtubule-associated protein 2c (MAP2c) is involved in neuronal development and is less characterized than its homolog Tau, which has various roles in neurodegeneration. Using NMR methods providing single-residue resolution and quantitative comparison, we investigated molecular interactions important for the regulatory roles of MAP2c in microtubule dynamics. We found that MAP2c and Tau significantly differ in the position and kinetics of sites that are phosphorylated by cAMP-dependent protein kinase (PKA), even in highly homologous regions. We determined the binding sites of unphosphorylated and phosphorylated MAP2c responsible for interactions with the regulatory protein 14-3-3ζ. Differences in phosphorylation and in charge distribution between MAP2c and Tau suggested that both MAP2c and Tau respond to the same signal (phosphorylation by PKA) but have different downstream effects, indicating a signaling branch point for controlling microtubule stability. Although the interactions of phosphorylated Tau with 14-3-3ζ are supposed to be a major factor in microtubule destabilization, the binding of 14-3-3ζ to MAP2c enhanced by PKA-mediated phosphorylation is likely to influence microtubule-MAP2c binding much less, in agreement with the results of our tubulin co-sedimentation measurements. The specific location of the major MAP2c phosphorylation site in a region homologous to the muscarinic receptor-binding site of Tau suggests that MAP2c also may regulate processes other than microtubule dynamics.

Cytoskeletal microtubule-associated proteins (MAPs)[3] are proteins of critical importance for regulating the stability and dynamics of microtubules (1). MAP2 and Tau represent MAP subfamilies expressed in neurons; MAP2 is localized in dendrites, whereas Tau is found mainly in axons (2). Tau and MAP2 belong to the class of intrinsically disordered proteins (IDPs), which lack a unique structure and which exist in multiple, quickly interconverting conformations (3–7). NMR is the method of choice for structural investigation of this type of protein. Tau and MAP2 differ in their N-terminal projection domains, which contain acidic and proline-rich subdomains, whereas the C-terminal parts, containing the microtubule-binding domain (MTBD) and the C-terminal region, are homologous (8). Tau is expressed in several splice variants. The human brain isoforms differ in the number of microtubule-binding regions (MTBRs) in the C-terminal portion and in the presence of two inserts near the N terminus. For the sake of simplicity, only the 441-residue variant lacking exons 6, 8, and 10 (9) is discussed in this paper. This variant has been studied in detail and was shown to form paired helical filaments and neurofibillary tangles in brains of patients suffering from Alzheimer's disease (10). The MAP2 family is composed of two high-molecular-weight proteins, MAP2a and MAP2b, each consisting of 1830 amino acids, and two low-molecular-weight proteins, MAP2c and MAP2d, consisting of 467 and 498 amino acids, respectively. The MAP2 isoforms differ mainly in their projection domains (8), with MAP2c being the shortest functional isoform. MAP2c mainly is expressed perinatally (2). Postnatally, its expression is restricted to regions exhibiting postnatal plasticity, such as the olfactory bulb (11), suggesting a role in neuronal development.

The phosphorylation of MAPs, which regulates their binding to microtubules (12–17) and consequently the microtubule dynamics (18), is implicated in neuronal development and plasticity (19). *In vivo* MAPs are substrates of various protein kinases, such as cAMP-dependent protein kinase A (PKA), protein kinase C (PKC), cdc2 kinase, and the like (16, 17, 20, 21). Phosphorylation of Tau weakens its interaction with microtubules and increases its affinity to regulatory 14-3-3 proteins (vide infra).

This work was supported by Grants GA15-14974S (to S. J., K. M., K. H., M. Z., and L. Z.) and GF15-34684L (Z. T., E. Z., and J. H.) from the Czech Science Foundation. The authors declare that they have no conflicts of interest with the contents of this article.

The chemical shifts were deposited in the Biological Magnetic Resonance Bank under BMRB accession no. 26960.

[1] To whom correspondence may be addressed. Tel.: 420-54949-3847; E-mail: jozef.hritz@ceitec.muni.cz.

[2] To whom correspondence may be addressed: National Centre for Biomolecular Research, Masaryk University, Kamenice 5, CZ-62500 Brno, Czech Republic. Tel.: 420-54949-8393; Fax: 420-54949-2556; E-mail: lzidek@chemi.muni.cz.

[3] The abbreviations used are: MAP, microtubule-associated protein; IDP, intrinsically disordered protein; MTBD, microtubule-binding domain; MTBR, microtubule-binding region; SSP, secondary structure propensity; MST, microscale thermophoresis; TCEP, tris(2-carboxyethyl)phosphine.

The 14-3-3 family includes highly conserved ubiquitous proteins (22) mostly expressed in the brain, and in particular in regions exhibiting neuroplasticity, and having important roles in neuronal development (23). The seven 14-3-3 isoforms present in mammals are β, γ, ϵ, ζ, τ, σ, and η. The 14-3-3ζ isoform was reported to be associated with microtubules in the brain (24). *In vitro*, 14-3-3ζ interacts with unphosphorylated Tau; however the phosphorylation of Tau increases its affinity for 14-3-3ζ (24–27).

Our goal was to compare the phosphorylation of Tau and MAP2c by PKA and the interactions of Tau and MAP2c with 14-3-3ζ. To the best of our knowledge, no data have been published on the interaction of 14-3-3ζ with MAP2c until now. A prerequisite for such studies is the knowledge of the phosphorylated residues of MAP2c. MAP2c phosphorylation by PKA has been studied previously, but the reported data are partially contradictory (28, 29). Therefore, we re-addressed this issue and determined the PKA phosphorylation sites of MAP2c by NMR and mass spectrometry (MS). We then characterized the interaction of MAP2c with 14-3-3ζ at a molecular level, and we studied the effect of 14-3-3ζ on tubulin polymerization induced by MAP2c. We found that the same regions of unphosphorylated MAP2c and Tau interact with 14-3-3ζ and that the binding affinity of MAP2c is greatly increased by phosphorylation, as described previously for Tau (25, 39). However, NMR analysis of the phosphorylation kinetics revealed that, despite their highly homologous sequences, PKA phosphorylates MAP2c and Tau in a different manner. Consequently, the high-affinity binding sites of phosphorylated MAP2c and Tau differ substantially.

Results

NMR assignment of the phosphorylated MAP2c

Chemical shifts of unphosphorylated MAP2c were assigned in our previous study (30). The same assignment strategy for PKA-phosphorylated MAP2c was used in this study. The resonance frequencies of the phosphorylated MAP2c were assigned using 5D CACONCACO, 3D (H)CANCO, and 5D HC(CC-TOCSY)CACON NMR experiments. The spectra were measured on a 1.1 mM [^{13}C,^{15}N]MAP2c sample phosphorylated by PKA for 24 h. The assigned backbone chemical shifts were similar to those obtained for unphosphorylated MAP2c other than the phosphorylated residues and their neighbors. The major changes in the CACONCACO spectrum were observed near Ser-435, where six neighboring residues showed a significant change in backbone chemical shifts, up to 1 ppm for ^{15}N and 0.3 ppm for protons.

Identification of phosphorylation sites

The phosphorylated MAP2c residues were first detected by NMR spectroscopy. A ^{1}H,^{15}N HSQC spectrum of [^{15}N,^{13}C] MAP2c was measured after 24 h of phosphorylation. Four new intense peaks and several weaker signals appeared at proton frequencies downfield from 8.6 ppm, corresponding to a chemical shift of the NH groups influenced by phosphorylation (Fig. 1). The 5D CACONCACO spectra of phosphorylated MAP2c provided an unambiguous assignment of the major new peaks in the ^{1}H,^{15}N HSQC spectrum to pSer-435, pSer-184, pThr-220, and to Arg-221 following pThr-220. The 3D HNCACB spectrum allowed us to classify the types of amino acids preceding the four additional phosphoserines/phosphothreonines detected in the ^{1}H,^{15}N HSQC spectrum.

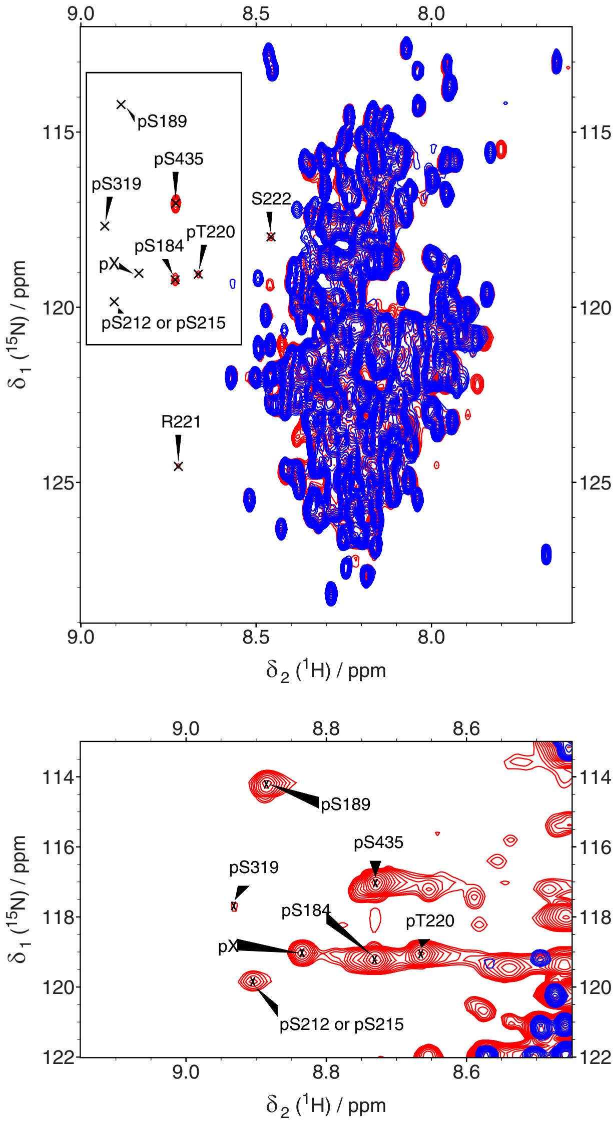


Figure 1. Overlaid ^{1}H,^{15}N HSQC spectra of unphosphorylated [^{15}N,^{13}C] MAP2c (*blue*), and [^{15}N,^{13}C]MAP2c phosphorylated by PKA for 24 h (*red*). The well-resolved peaks of phosphorylated residues (pSer-184, pSer-189, pThr-220, and pSer-435) at proton frequencies downfield from 8.6 ppm and of their neighbors (Arg-221 and Ser-222) are labeled. *Bottom*, close-up of the region of phosphorylated residues plotted with a lowered signal threshold level to show minor peaks. *pX*, this peak could not be assigned.

The phosphorylation sites were also identified by MS (Table 1). After incubation with PKA for 24 h, MAP2c was digested with trypsin. The phosphorylated peptides were separated from the unphosphorylated ones by TiO$_2$ fractionation, and the phosphopeptides were subjected to LC-MS/MS analysis. In agreement with the NMR data, the most intense peaks corresponded to peptides containing pSer-184 and pSer-435. Phosphopeptides containing pThr-220 were not detected in our tryptic digests (possibly because of the high frequency of tryptic cleavage sites in the vicinity of Thr-220 resulting in peptides too short for analysis under the conditions used in our study) but were unambiguously identified previously (29). The peak areas of the obtained phosphopeptides relative to the sum of the areas of peaks containing pSer-435 are presented in Fig. 2. The phosphopeptides identified by MS are presented in Table 1.

Verification of NMR assignment of phosphorylated residues by site-directed mutagenesis

Because the sensitivity of the 5D CACONCACO spectra was not sufficient to assign minor phosphorylation sites, we designed several mutants (S184D, S189D/S435D, S199D/S435D, T220E, S319D/S435D, T320E/S435D, S342D/S435D, S350D/S435D, S367D/S435D, S382D/S435D, and S435D) based on our MS data and on phosphorylation sites reported in the literature (28). The mutants were uniformly ^{15}N-labeled and phosphorylated, and the ^{1}H,^{15}N HSQC spectra were recorded. In the spectra of the S184D, T220E, S435D, S189D/S435D, and S319D/S435D mutants, the peaks of the mutated phosphorylated residues disappeared, confirming our previous assignment and allowing us to assign two minor peaks to pSer-189 and pSer-319. No significant changes were observed in the region of the phosphorylated serines and threonines in the HSQC spectra of the other mutants, suggesting that their degree of phosphorylation is low.

Table 1

Phosphopeptides identified by MS in percentage compared with the most intense peak, Ser-435

Only the phosphopeptides that have an intensity higher than 1% of the intensity of the most intense peak are presented here. In the absence of PKA, none of the phosphopeptides identified has an intensity higher than 1% of the most intense peak. The phosphorylated residues are indicated in bold.

Residue	Peptide	% Intensity
Ser-184	^{183}S**S**LPRPSSILPPR^{195}	65
Ser-199	^{196}RGV**S**GDREENSFSLNSSISSAR^{217}	5
Ser-212	^{197}GVSGDREENSFSLNS**S**ISSAR^{217}	1
Ser-214	^{197}GVSGDREENSFSLNSSI**S**SAR^{217}	2
Ser-231	^{228}AGK**S**GTSTPTTPGSTAITPGTPPSYSSR^{255}	2
Ser-319	^{315}SKIG**S**TDNIK^{324}	7
Ser-367	^{363}VKIE**S**VKLDFK^{373}	2
Ser-435	^{433}RL**S**NVSSSGSINLLESPQLATLAEDVT AALAK^{464}	100
Ser-435, Ser-438	^{433}RL**S**NV**S**SSGSINLLESPQL^{451}	3
Ser-435, Ser-442	^{433}RL**S**NVSSSG**S**IN^{444}	1

Sequence-based prediction of phosphorylation sites

In principle, the target sites for specific kinases are encoded in the amino acid sequence and by their accessibility to the kinase. However, currently available Web server predictors of phosphorylation sites have been optimized for structured proteins, whereas their reliability for IDPs is not well documented. Fig. 3 compares the PKA phosphorylation sites of MAP2c determined by NMR in this study with three popular Web server predictors. One can see that the PKA phosphorylation of Ser-435 was predicted by all three predictors. Sites Ser-184, Ser-189, and Thr-220 were predicted by Scansite3 (31) and GPS3 (32) but not by Kinasephos2 (33). This limited comparison indicates that Scansite3 and GPS3 can identify the most important PKA phosphorylation sites within the intrinsically disordered protein MAP2c. On the other hand, all three predictors generated significant number of false-positive predictions with respect to the NMR data (Fig. 3). Interestingly, the same type of prediction for Tau protein shows better agreement between GPS3 and Kinasephos2 in contrast to Scansite3 (Fig. 3). New experimental results are therefore needed to improve the reliability of prediction of phosphorylated sites within IDPs.

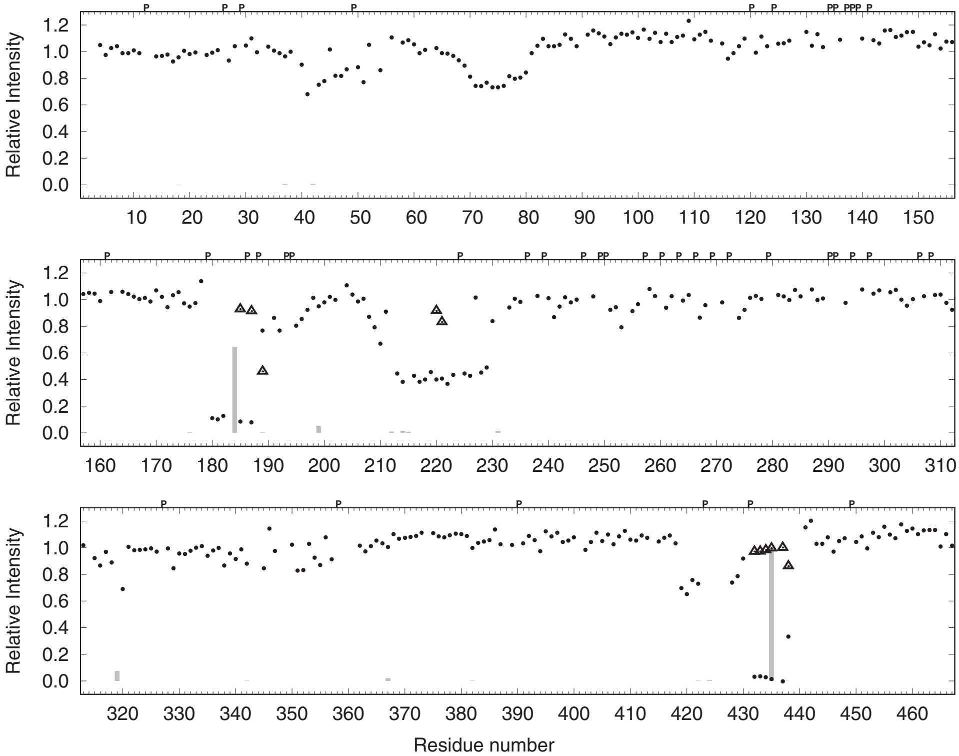


Figure 2. Phosphorylation of MAP2c. The *dots* represent the ratio U_u/U_p, where U_u and U_p are peak intensities of unphosphorylated residues in the unphosphorylated and phosphorylated MAP2c samples, respectively. *Triangles* represent the ratio $P_p/(U_p + P_p)$, where P_p are peak intensities of phosphorylated residues in the phosphorylated MAP2c sample. The *bars* represent the sum of areas of peptide peaks containing the given phosphorylated residue in mass spectra (relative to the sum of areas of peptides peaks containing Ser-435). *Letter P* (above the plot), indicates proline residues.

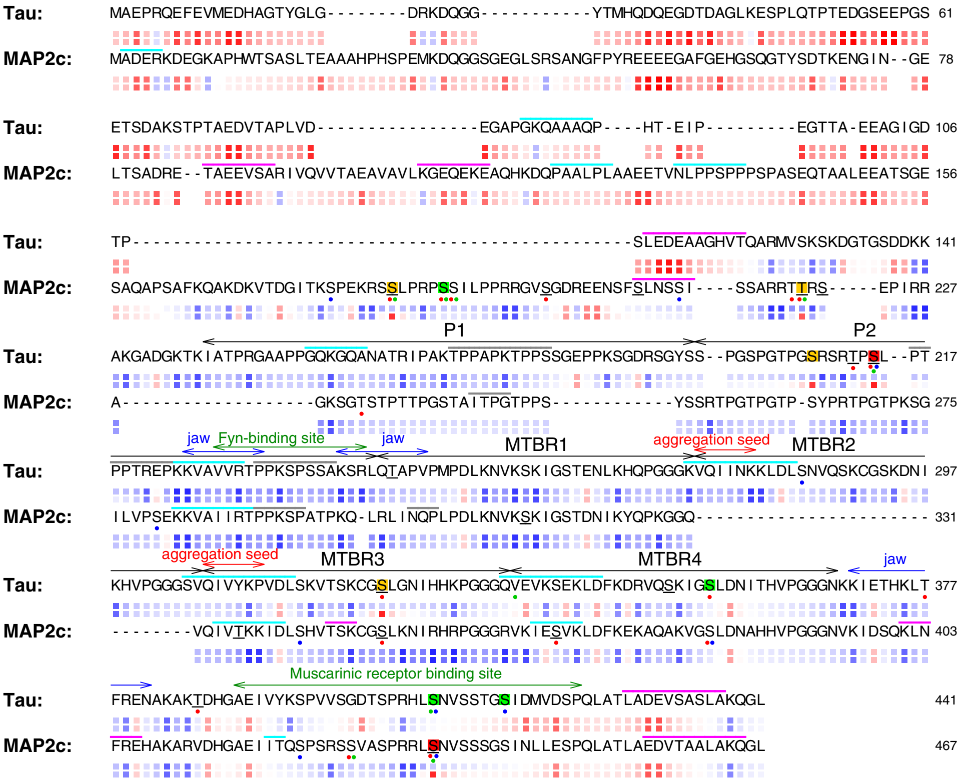

Tau: MAEPRQEFEVMEDHAGTYGLG--------DRKDQGG-----------YTMHQDQEGDTDAGLKESPLQTPTEDGSEEPGS 61
MAP2c: MADERKDEGKAPHWTSASLTEAAAHPHSPEMKDQGGSGEGLSRSANGFPYREEEEGAFGEHGSQGTYSDTKENGIN--GE 78
Tau: ETSDAKSTPTAEDVTAPLVD----------------EGAPGKQAAAQP----HT-EIP---------EGTTA-EEAGIGD 106
MAP2c: LTSADRE--TAEEVSARIVQVVTAEAVAVLKGEQEKEAQHKDQPAALPLAAEETVNLPPSPPPSPASEQTAALEEATSGE 156
Tau: TP--SLEDEAAGHVTQARMVSKSKDGTGSDDKK 141
MAP2c: SAQAPSAFKQAKDKVTDGITKSPEKRSSLPRPSSILPPRRGVSGDREENSFSLNSSI----SSARRTTRS-----EPIRR 227
P1
P2
Tau: AKGADGKTKIATPRGAAPPGQKGQANATRIPAKTPPAPKTPPSSGEPPKSGDRSGYSS--PGSPGTPGSRSRTPSL--PT 217
MAP2c: A--------------------GKSGTSTPTTPGSTAITPGTPPS-----------YSSRTPGTPGTP-SYPRTPGTPKSG 275
jaw
Fyn-binding site
jaw
MTBR1
aggregation seed
MTBR2
Tau: PPTREPKKVAVVRTPPKSPSSAKSRLQTAPVPMPDLKNVKSKIGSTENLKHQPGGGKVQIINKKLDLSNVQSKCGSKDNI 297
MAP2c: ILVPSEKKVAIIRTPPKSPATPKQ-LRLINQPLPDLKNVKSKIGSTDNIKYQPKGGQ---------------------- 331
aggregation seed
MTBR3
MTBR4
jaw
Tau: KHVPGGGSVQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQVEVKSEKLDFKDRVQSKIGSLDNITHVPGGGNKKIETHKLT 377
MAP2c: --------VQIVTKKIDLSHVTSKCGSLKNIRHRPGGGRVKIESVKLDFKEKAQAKVGSLDNAHHVPGGGNVKIDSQKLN 403
Muscarinic receptor binding site
Tau: FRENAKAKTDHGAEIVYKSPVVSGDTSPRHLSNVSSTGSIDMVDSPQLATLADEVSASLAKQGL 441
MAP2c: FREHAKARVDHGAEIITQSPSRSSVASPRRLSNVSSSGSINLLESPQLATLAEDVTAALAKQGL 467

Figure 3. Comparison of sequences, phosphorylation sites, and electrostatic properties of Tau and MAP2c. Experimentally determined phosphorylation sites are indicated by background colors: *red*, class I; *orange*, class II; *green*, class III (according to the classification used by Laudrieu *et al.* (43)). The predicted phosphorylation sites are indicated by *dots* below the sequences: *red*, prediction by Scansite3; *green*, prediction by GPS3; *blue*, prediction by Kinasephos2. Residues predicted by 1433pred (48) to interact with 14-3-3ζ are *underlined*. Regions with a propensity to form helical, β-strand, and polyproline II structures according to Mukrasch *et al.* (38) and Nováček *et al.* (30) are indicated by *magenta*, *cyan*, and *gray bars* above the sequences, respectively. The relative electrostatic potential is shown as *color-coded boxes* below the sequences for unphosphorylated (*upper row*) and phosphorylated (*lower row*) form of each protein. The potential is approximated by $\Sigma_j CQ_j/(d_0 + d_1|n_i - n_j|)$, where Q_i and n_j are the charge and sequential number of the i-th residue, C is a constant including the electric permittivity, and d_k are distance constants. The ratio d_1/d_0 was set to 2.0, and the colors were chosen so that *red* and *blue* correspond to the highest negative and positive potential, respectively, which makes the color code independent of C/d_0. Gaps in the sequences were inserted manually to optimize the alignment of regions with similar trends to form transient secondary structures. The annotation of functionally important regions of Tau, defined by the *double-headed arrows* above the sequences, was taken from the literature (50, 38, 9).

Quantitative determination of phosphorylation

Strong signals indicating a high degree of phosphorylation at Ser-184 and Ser-435 were observed using both NMR and MS methods. NMR was used to determine the degree of phosphorylation quantitatively. A comparison of the peak intensities in the 3D HNCO spectra of phosphorylated and unphosphorylated MAP2c revealed three regions showing a significant decrease in the intensity of the peaks, corresponding to unphosphorylated residues in the phosphorylated sample (Fig. 2). The intensity levels decreased to ~40% in the vicinity of Thr-220, to less than 10% in the vicinity of Ser-184 and below the detection limit in the vicinity of Ser-435. Such a comparison provided an overall quantitative picture of the phosphorylation and allowed us to estimate the degree of phosphorylation of individual amino acids (with the exception of residues close in sequence to Ser-184, Thr-220, and Ser-435, where the effect of possible minor phosphorylation was obscured by the decrease in the signal intensity due to a major phosphorylation site). In summary, both NMR and MS identified Ser-184 and Ser-435 as major PKA phosphorylation sites of MAP2c, and significant phosphorylation was observed also at Thr-220 (Fig. 2).

Kinetics of phosphorylation

The kinetics of phosphorylation was followed by real-time NMR spectroscopy from 2 min to 30 h after the addition of PKA. The signal of the first phosphorylated residue, pSer-435, already appeared in the first spectrum recorded after the addition of PKA and reached saturation after 1.5 h of incubation (Fig. 4). The kinetic curve was exponential even for short reaction times, indicating that PKA was not saturated by MAP2c. The other residues were phosphorylated much more slowly, and their buildup curves exhibited lag phases documenting competition of the phosphorylation sites for PKA. After the lag phase, the ratio of phosphorylation rates of Ser-184, Ser-189, and Thr-220 was almost constant (~4:1:2).

To compare the phosphorylation of individual sites quantitatively, apparent rate constants (k_{obs}) were estimated as described under "Experimental procedures." The evaluation of k_{obs} was facilitated by the kinetic separation of Ser-435, allowing us to assume that the other sites were unphosphorylated during phosphorylation of Ser-435 by the aforementioned kinetic partitioning of Ser-184, Ser-189, and Thr-220. The estimated k_{obs} values were 9700 ± 1500 $\text{M}^{-1}\text{s}^{-1}$ for Ser-435, 200 ± 14 $\text{M}^{-1}\text{s}^{-1}$ for Ser-184, 50 ± 6 $\text{M}^{-1}\text{s}^{-1}$ for Ser-189, and 95 ± 21 $\text{M}^{-1}\text{s}^{-1}$ for Thr-220. In summary,

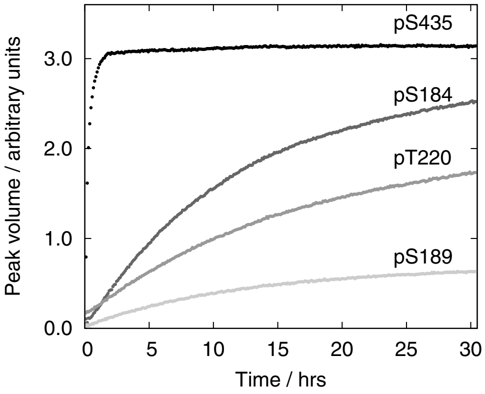


Figure 4. Volumes of pSer-435 (*black*), pSer-184 (*dark gray*), pThr-220 (*medium gray*), and pSer-189 (*light gray*) peaks in the SOFAST-HMQC spectra recorded during MAP2c phosphorylation by PKA.

the real-time NMR measurements revealed a dramatic difference between phosphorylation of Ser-435 and other residues.

Effect of phosphorylation on the secondary structure propensities of MAP2c

To determine the effect of phosphorylation on the secondary structure, secondary structure propensities (SSP) of unphosphorylated and phosphorylated MAP2c were calculated. To identify SSP changes close to the phosphorylation sites, the SSP was calculated from the chemical shifts of $^{13}C^{\alpha}$, $^{13}C^{\beta}$, and $^{1}H^{\alpha}$, recently found to be insensitive to the presence of the phosphate group attached to neighboring residues (34). The most significant changes were observed in two regions close to the major phosphorylation sites. In the vicinity of pSer-184 and especially pThr-220, the SSP value indicates that phosphorylation stabilizes the already existing secondary structures (Fig. 5). The opposite effect was observed in a region preceding pSer-435, where phosphorylation alters the SSP of residues Arg-425–Pro-431 from the propensity to form an extended secondary structure in the unphosphorylated state to a helical propensity in the phosphorylated state (Fig. 5). The less pronounced changes observed for residues distant from the phosphorylation sites (including the N-terminal region) indicate that phosphorylation by PKA also influences the intramolecular interactions of MAP2c.

Apparent dissociation constants of MAP2c-14-3-3ζ complexes

The affinities of 14-3-3ζ to MAP2c in phosphorylated and non-phosphorylated forms were compared by microscale thermophoresis (MST). To achieve a sufficient sensitivity, MAP2c was labeled fluorescently. MAP2c contains a single cysteine residue (Cys-348), but it is located in a potential binding region (MTBR3). Therefore, an E52C/C348S MAP2c construct was prepared and labeled specifically by the fluorescent dye Alexa Fluor 647 at the introduced Cys-52.

Titration of E52C/C348S MAP2c with 14-3-3ζ revealed that both unphosphorylated and phosphorylated MAP2c bind 14-3-3ζ, but phosphorylation greatly enhances the affinity. The titration curve of phosphorylated E52C/C348S MAP2c exhibited a well-defined inflection point in the low micromolar range of 14-3-3ζ concentrations and another less-defined inflection point at a concentration 2 orders of magnitude higher (Fig. 6). Fitting the first sigmoidal region separately (for 14-3-3ζ monomer concentrations lower than 100 μM) resulted in an apparent dissociation constant (expressed for the 14-3-3ζ monomer concentration as described under "Experimental procedures") of 0.57 ± 0.37 μM and stoichiometry of $n = 1.53 \pm 0.14$ (14-3-3ζ monomer/MAP2c molecule). The fractional stoichiometry may indicate the contribution of several binding modes, but the size of the experimental errors made a precise determination of the stoichiometry difficult. Fitting the second region gives $K'_D \approx 280 \pm 120$ μM, but full saturation could not be achieved.

The binding curve of unphosphorylated E52C/C348S MAP2c showed an interaction at least 1 order of magnitude weaker than observed for the phosphorylated form. Fitting the data provided $K'_D = 92 \pm 12$ μM, but the shape of the titration curve indicates that this number is probably affected by additional binding events that occur at a higher 14-3-3ζ concentration and cannot be fully separated.

The obtained apparent dissociation constants are useful for a rough comparison of the binding affinities. However, they should not be overinterpreted because the MAP2c-14-3-3ζ interaction is more complex than the simple binding model used for the data fitting (35).

Identification of 14-3-3ζ-binding sites in MAP2c

NMR spectroscopy was used to identify individual residues of MAP2c that interact with 14-3-3ζ. Unphosphorylated [^{13}C,^{15}N]MAP2c and [^{13}C,^{15}N]MAP2c phosphorylated for 24 h by PKA were mixed with 14-3-3ζ in ratios of 1:0.125, 1:0.25, 1:0.5, 1:1, and 1:2 (MAP2c:14-3-3ζ monomer). A 3D HNCO spectrum was recorded before titration and after each addition of 14-3-3ζ. When titrating unphosphorylated MAP2c with 14-3-3ζ (Fig. 7), we observed a gradual diminution of the intensity of the peaks of the residues in the MTBD (Thr-296–Ser-380) and in the C-terminal region with a strong helical propensity (Ile-443–Leu-467), indicating that these regions bind 14-3-3ζ and become less flexible. Note that the broad region of residues with reduced peak intensity due to the interaction with 14-3-3ζ is interrupted by several short stretches of amino acids already exhibiting significant line broadening in free MAP2c, and therefore it is little affected by 14-3-3ζ binding. This effect is particularly notable for the P*X*GG motifs immediately preceding the regions with high β-strand propensity in the MTBD. These regions initiate aggregation in Tau (36, 37). On the other hand, a decrease in the peak intensity was also observed for several residues in the N-terminal portion of MAP2c, indicating that interdomain interactions exist in MAP2c as in Tau (38).

The addition of 14-3-3ζ to phosphorylated MAP2c resulted, in addition to the changes described above, in a strong decrease of peak heights of amino acids in the vicinity of pSer-184 and pThr-220 in the proline-rich domain and of pSer-435 in the region corresponding to the muscarinic receptor-binding site of Tau (Fig. 3). The fact that the peak intensities were substantially reduced already in the presence of substoichiometric amounts of 14-3-3ζ suggests that the spectra of the 14-3-3ζ-bound MAP2c are also influenced by an intermediate chemical exchange. Therefore, we did not use our NMR data for quantitative determination of the binding affinity.

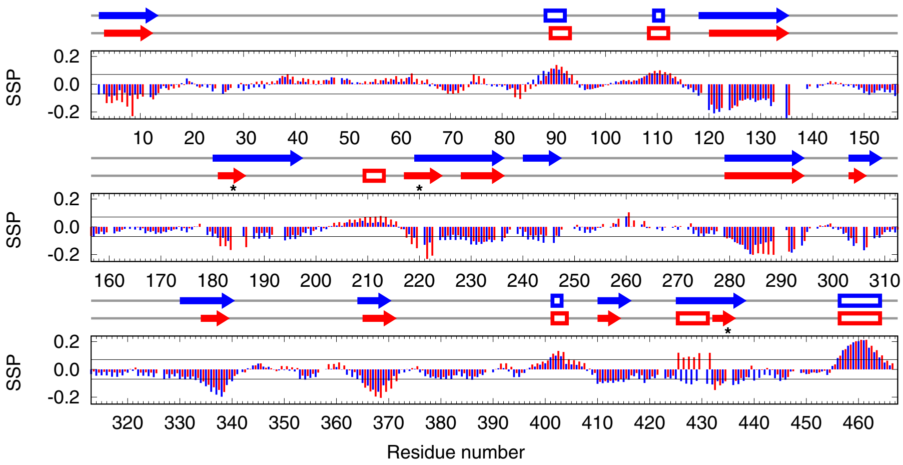


Figure 5. Effects of phosphorylation on the secondary structure propensity of unphosphorylated MAP2c (*blue*) and phosphorylated MAP2c (*red*). The *symbols* above the graphs indicate continuous regions of unphosphorylated (*blue*) and phosphorylated (*red*) with the propensity to form helical (*empty boxes*) or extended (*arrows*) structures (see "Experimental procedures" for details). *Asterisks* indicate the major PKA phosphorylation sites.

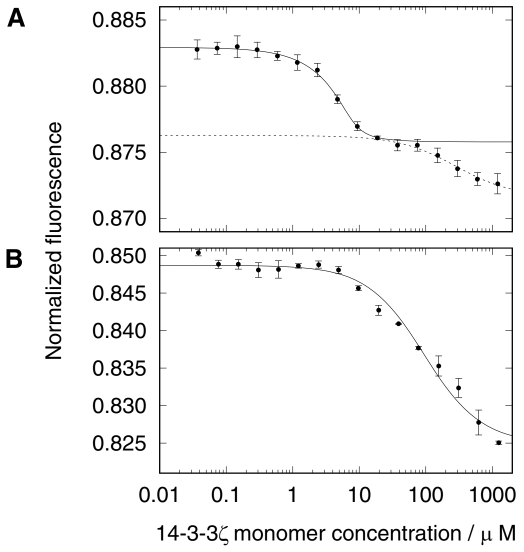


Figure 6. Microscale thermophoretic analysis of interaction between 5 μM E52C/C348S MAP2c and 14-3-3ζ in 50 mM Tris buffer. *A*, unphosphorylated MAP2c; *B*, phosphorylated MAP2c. The plots show interactions in the 14-3-3ζ monomer concentration range from 36.6 nM to 1.2 mM. The mean values ± S.D. for each concentration point were calculated from triplicate measurements.

Effect of 14-3-3ζ on MAP2c-induced tubulin polymerization

The effect of 14-3-3ζ on the polymerization of tubulin induced by unphosphorylated and phosphorylated MAP2c was measured by co-sedimentation assays (Fig. 8). To measure the direct effect of MAP2c and 14-3-3ζ on tubulin polymerization, tubulin was not stabilized by Taxol in the assay. 14-3-3ζ alone was not able to induce the polymerization of microtubules (data not shown). About 20% of the tubulin polymerized spontaneously. In the presence of 2 μM unphosphorylated and phosphorylated MAP2c, tubulin polymerization increased by 13 ± 2% and 12 ± 1%, respectively. The addition of up to 2 μM 14-3-3ζ (monomer concentration) had a negligible effect, but 20 μM 14-3-3ζ reduced the amount of tubulin in the pellet to a level comparable with the background of its spontaneous polymerization. The effect of MAP2c phosphorylation was negligible. Thus the data show that phosphorylation of MAP2c by PKA does not influence its affinity to tubulin under the conditions used and that 14-3-3ζ competes with tubulin for binding unphosphorylated and PKA-phosphorylated MAP2c with a comparable efficiency.

Discussion

The key role of phosphorylation in regulating the interactions of MAPs with microtubules has been well-known for decades (12–17). 14-3-3 proteins are proposed to be involved in the phosphorylation-dependent control of microtubule dynamics by competing for phosphorylated Tau with tubulin (24, 39, 40). PKA seems to be the key kinase in this mechanism. Phosphorylation of Tau by PKA or PKB, but not by GSK-3β, CDK2, or CK1, enhances the interaction with 14-3-3ζ (25), whereas Tau phosphorylated by PKA binds to tubulin with 7-fold lower affinity (41, 20).

In our study, we searched for possible differences between Tau and its homolog, MAP2c, that would implicate distinct regulatory roles for these proteins. As phosphorylation by PKA is sufficient to form the sites critical for 14-3-3ζ binding in Tau, we also used PKA in our studies of MAP2c.

Phosphorylation of MAP2c by PKA has been studied already in the past. However, early studies of MAP2c phosphorylation resulted in partially contradictory conclusions. Using two-dimensional electrophoresis combined with mass spectrometry and site-directed mutagenesis, Ozer and Halpain (28) identified Ser-319, Ser-350, and Ser-382 as early phosphorylation sites. Later, Ser-184, Thr-220, and Ser-435 were reported by Alexa *et al.* (29) to be the major phosphorylation sites in MAP2c, in agreement with the prediction based on the sequence (42). A single completely phosphorylated residue, Thr-220, was found in a 20-kDa peptide (Asn-205–Glu-366) of MAP2c phosphorylated by PKA (29). Although the Asn-205–Glu-366 peptide also contains Ser-319, Ser-350, and Ser-382, phosphorylation of these serines was not observed by Alexa *et al.* (29). Therefore, we revisited the topic of phosphorylation of MAP2c by PKA and, taking advantage of the recent methodological progress in MS and NMR spectroscopy, clarified which amino acids are predominantly phosphorylated under the given conditions. Our results confirm that Ser-184, Thr-220, and Ser-435 are

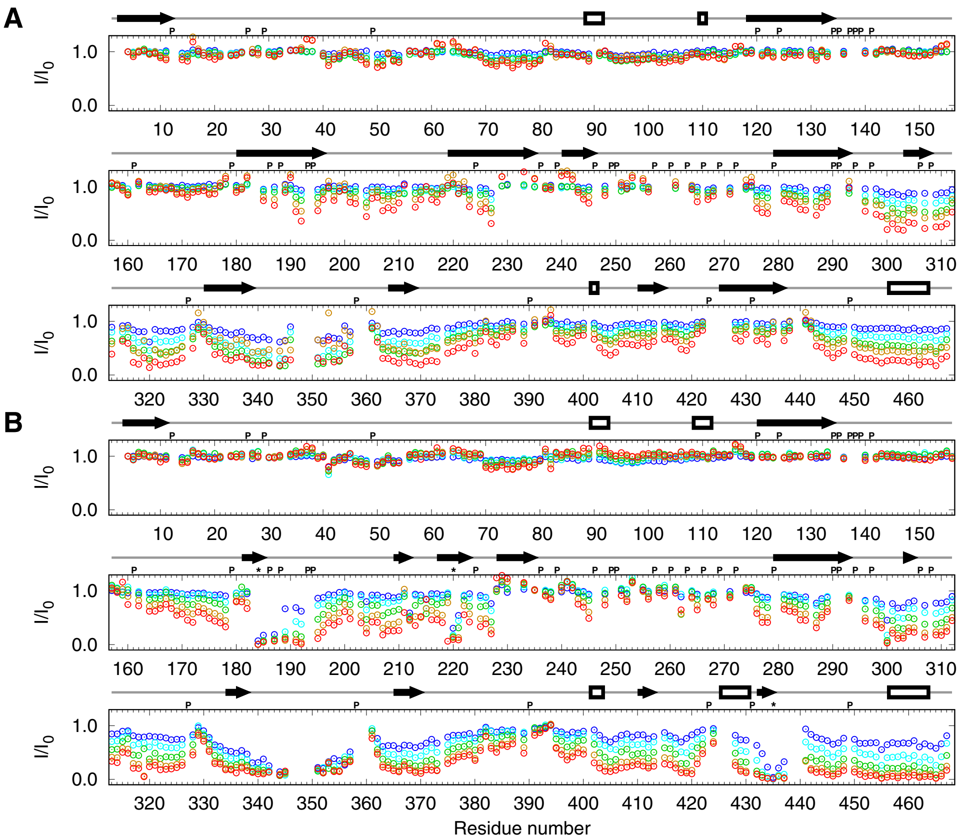


Figure 7. Intensities of the peaks in MAP2c HNCO spectra upon addition of 14-3-3ζ compared with intensities of the peaks in free MAP2c with the MAP2c:14-3-3ζ monomer ratios of 1:0.125 (*blue*), 1:0.25 (cyan), 1:0.5 (*green*), 1:1 (*orange*), and 1:2 (*red*). *A*, unphosphorylated MAP2c; *B*, phosphorylated MAP2c. The *empty boxes* and *arrows* above the graphs indicate propensity to form helical and extended structures, respectively, according to the data presented in Fig. 5. *Letter P* and *asterisks* (above the plot), indicate proline residues and major phosphorylation sites, respectively.

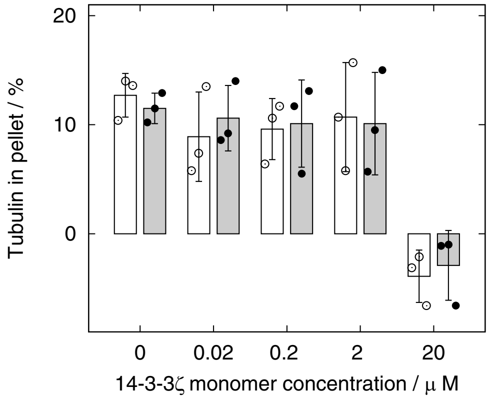


Figure 8. Results of the tubulin-MAP2c co-sedimentation assay. The relative amount of tubulin in the pellet (after substraction of the background of spontaneous tubulin polymerization) with increasing concentrations of 14-3-3ζ in the presence of unphosphorylated MAP2c and phosphorylated MAP2c is shown as *white* and *black boxes*, respectively. The *error bars* correspond to the standard deviations calculated from triplicate measurements. Only the effect of 20 μM 14-3-3ζ is statistically significant at $\alpha = 0.05$.

phosphorylated most efficiently. The Web server predictors also identified Ser-184, Thr-220, and Ser-435 as phosphorylation sites, but the number of false positive predictions documents that experimental determination of phosphorylation sites is still necessary (Fig. 3).

The locations of the phosphorylation sites in MAP2c, and especially the kinetics of phosphorylation, differ significantly from Tau. Our results can be compared quantitatively with the data published by Landrieu *et al.* (43), which were obtained under similar conditions.

The major phosphorylation site of Tau (class I according to the classification introduced by Landrieu *et al.* (43), corresponding to $k_{\mathrm{obs}} \approx 10^4\ \mathrm{M}^{-1}\mathrm{s}^{-1}$ and highlighted by the *red background* in Fig. 3) is ^{211}RTPpSLP^{216} in the P2 region of the proline-rich domain (Fig. 3), representing a typical 14-3-3ζ-binding motif, RS/T*X*pS*X*P (25). The P2 region is involved in interactions with tubulin (38, 44), and phosphorylation of Ser-214 reduces tubulin binding (44) and the ability of Tau to promote microtubule assembly (45). The corresponding sequence in MAP2c, ^{255}RTPGTP^{260}, does not contain a phosphorylatable serine or threonine in the position of pSer-214 as in Tau. This correlates with the fact that the phosphorylation by PKA inhibits the microtubule-stabilizing (growth promoting) activity of Tau (46) but not of MAP2 (47). However, the proline-rich region of MAP2c also contains a similar phosphorylated 14-3-3ζ-binding site (48) but in a different position (^{181}KRSpSLP^{186}), in a region exhibiting a certain sequence homology with the P1 region (see Fig. 3) of Tau. Ser-184 of MAP2c is phosphorylated more slowly than Ser-214 of Tau, exhibiting the second highest phosphorylation rate among the MAP2c phosphorylation sites (class II, with $k_{\mathrm{obs}} \approx 10^2\ \mathrm{M}^{-1}\mathrm{s}^{-1}$, *yellow background* in Fig. 3). Both Ser-214 of Tau and Ser-184 of MAP2c are accompanied by weaker phosphorylation sites (Ser-208 and Ser-189, respectively) in close proximity.

The proline-rich domain of MAP2c contains another class II phosphorylation site, Thr-220, in a region that has little homology with Tau (Fig. 3). The SSP calculated from the chemical shifts indicates that phosphorylation by PKA stabilizes the secondary structure in the vicinity of Thr-220 (Fig. 5). This site has been studied extensively by Alexa *et al.* (29), who found that pThr-220 is extraordinarily sensitive to phosphatases and that MAP2c phosphorylated at Ser-184 and Ser-435, but not at Thr-220, stabilizes microtubules significantly more strongly than the other combinations of phosphorylation.

The remaining phosphorylation sites are located in a C-terminal region with a high sequence homology between MAP2c and Tau. The site with the second highest phosphorylation rate (class II) in Tau is ^{321}KCGpS^{324} in MTBR3 (Fig. 3), representing the typical K*X*GS motif responsible for microtubule binding (16, 17). Ozer and Halpain (28) detected a rapid PKA phosphorylation of the corresponding residue Ser-350 (and Ser-319 and Ser-382 in other K*X*GS motifs) in MAP2c and found that the mutation of serines in the K*X*GS motifs of MAP2c has a great impact on interactions with microtubules. However, our quantitative data clearly show that the degree of phosphorylation at Ser-350 is low under our conditions. On the contrary, PKA in our study preferentially phosphorylated Ser-435 in the C-terminal region of MAP2c, corresponding to the muscarinic receptor-binding site of Tau (Fig. 3). The rate of Ser-435 phosphorylation was comparable with that of Ser-214 of Tau (class I). Despite the high sequence homology, the corresponding serine (Ser-409) is only a minor phosphorylation site (class III, indicated by the *green background* in Fig. 3) of Tau.

The listed striking differences between phosphorylation patterns and kinetics of MAP2c and Tau, determined by the same experimental approach (real-time NMR spectroscopy), are surprising considering the high sequence homology (especially in MTBDs) and have an important implication for 14-3-3ζ binding. Our interpretation is that these highly homologous regions have different local structures within MAP2c and Tau with a direct consequence: differing accessibility for PKA. This explanation is in agreement with the current view of IDPs as proteins that exhibit structural features significantly different from a random coil organization.

A recent study by Joo *et al.* (39) investigating the interaction of phosphorylated Tau with 14-3-3σ allowed us to make a direct comparison of the interactions of 14-3-3 with MAP2c and Tau. Unphosphorylated forms of MAP2c and Tau bind the 14-3-3 proteins, but phosphorylation by PKA significantly increases the binding affinities. The phosphorylation-independent 14-3-3ζ-binding sites of MAP2c are located in the MTBD and in the C-terminal domain. The same regions have been shown to bind 14-3-3ζ in unphosphorylated Tau (24, 25). Residue-specific data for unphosphorylated Tau are not reported by Joo *et al.* (39), but their analysis of the set of all residues of phosphorylated Tau influenced by 14-3-3σ binding leads to the same conclusion.

Phosphorylated forms of both MAP2c and Tau seem to interact with 14-3-3 proteins predominantly via two phosphoserines, one located in the proline-rich domain (25, 27, 39) and the other in the C-terminal portion (27, 39). As discussed above, the phosphorylation sites of MAP2c and Tau differ significantly even in regions of high sequential homology, which has a direct impact on phosphorylation-dependent 14-3-3 binding.

Charge distribution along the MAP2c and Tau sequences provides a physicochemical explanation of the observed differences in 14-3-3ζ binding. Positively charged regions of Tau are known to bind to acidic residues of tubulin and polyanions (49, 50), and MAP2c is likely to behave in a similar manner. We propose that the stretches of positively charged residues in MAP2c and Tau are also involved in the interaction with electronegative regions within the highly acidic 14-3-3ζ protein (pI ≈ 4.7). To explain the observed differences in the interactions of MAP2c and Tau, one should search for regions of the proteins that differ in charge distributions.

Fig. 3 shows that the electrostatic potential in the compared MAPs is similar in MTBR1 and MTBR4 but very different in MTBR3. A long stretch of positively charged residues is present in this region of MAP2c. As PKA did not phosphorylate Ser-350 under the conditions of our study, phosphorylation did not influence the charge distribution in that region of MAP2c. In contrast, the corresponding MTBR3 region in Tau is efficiently phosphorylated at Ser-324 (class II kinetics), which interrupts a stretch of positively charged residues with a negative patch. Therefore, the possible electrostatic interaction with acidic regions of 14-3-3 proteins is likely to be greatly suppressed by phosphorylation in the case of Tau but is little affected by phosphorylation in the case of MAP2c. This working hypothesis is consistent with our finding that 14-3-3ζ competes with MAP2c-induced tubulin polymerization regardless of MAP2c phosphorylation (Fig. 8), in contrast to the model of the Tau-regulated microtubule polymerization (51).

Another region with differing electrostatic potential between MAP2c and Tau is located 30 to 40 residues downstream of MTBR4. Phosphorylation of Ser-435, representing the major PKA target in MAP2c (but not in Tau), significantly reduces the positive charge in the mentioned region of MAP2c. This phosphorylation also induced the propensity to form a helical structure in the region Arg-425–Pro-431 (Fig. 5). The specific phosphorylation of MAP2c Ser-435 and its physicochemical consequences are particularly interesting because this region is responsible for the ability of Tau to act as a muscarinic agonist (52).

The observed phosphorylation patterns and interactions with 14-3-3 are related to the biological functions of Tau and MAP2c. Phosphorylation of Tau and MAP2c is controlled by various neurotransmitter receptors (53–55). PKA plays a key role in the signaling cascades triggered by receptors activating adenyl cyclase directly but is also involved, via Ca^{2+}-stimulated adenyl cyclases, in pathways employing Ca^{2+} as a second messenger (56). PKA phosphorylates Tau and MAP2c directly but also activates downstream kinases able to phosphorylate residues not targeted by PKA (57–60). The effects of direct phosphorylation by PKA addressed in this study differ not only between Tau and MAP2 isoforms but also between different activities of the MAPs. PKA moderately decreases the binding of Tau and MAP2 isoforms to microtubules (29, 41, 47), suppresses the microtubule-nucleating activity of both Tau and high-molecular-weight MAP2 (46, 47), and inhibits the microtubule-stabilizing activity of Tau (46), but it does not affect the microtubule-stabilizing activity of high-molecular-weight MAP2 (47). Differences in the phosphorylation of Tau and MAP2c also influence the interactions with 14-3-3 proteins, which are proposed to regulate interactions with microtubules and modulate phosphorylation of Tau by various kinases (for review, see Ref. 51). Moreover, binding to microtubules is in equilibrium with the interactions of MAPs with other components of the cytoskeleton. Phosphorylation in MTBRs promotes MAP2c localization to actin-rich regions of neurons (28), with a direct impact on neuronal development and plasticity (19, 61, 62). It is evident that the kinetics of phosphorylation is an important factor in determining the balance of the aforementioned effects and contributes to the specificity of the control of microtubule stability by Tau and MAP2c under different physiological conditions and in different regions of neurons (8). Some of the functional differences may be directly related to the observed distinct kinetics of phosphorylation: PKA phosphorylation of Ser-214 in Tau inhibits the microtubule-stabilizing activity of Tau but does not affect the microtubule-stabilizing activity of MAP2 (46, 47); and the kinetics of phosphorylation and dephosphorylation of Thr-220 in MAP2c should be comparable if Thr-220 dephosphorylation plays a regulatory role (29).

In conclusion, PKA phosphorylates Tau and MAP2c differently even in highly homologous regions, which is reflected by the interactions of different sites of Tau and MAP2c with 14-3-3 proteins. The involvement of two proteins responding differently to the same signal (phosphorylation by PKA) may represent a branching point in the signaling pathways controlling microtubule stability and other important events such as activation of cholinergic receptors.

Experimental procedures

Preparation of recombinant proteins

MAP2c expression and purification was performed as described previously (1, 30). The protein was then phosphorylated, or dialyzed, against NMR buffer consisting of 50 mM MOPS, pH 6.9, 150 mM NaCl, and 0.7 mM TCEP. MAP2c was expressed in M9 medium containing ^{15}N NH_4Cl and/or ^{13}C glucose for NMR measurement.

MAP2c was phosphorylated at 30 °C with 650 units of the catalytic subunit of PKA (New England Biolabs)/mg of MAP2c in a buffer containing 50 mM Tris-HCl, 10 mM $MgCl_2$, 0.1 mM EDTA, 2 mM DTT, pH 7.5, and 20 mM ATP. After 24 h of incubation, PKA was deactivated by heating to 95 °C for 20 min. For NMR measurement, the protein was dialyzed against the NMR buffer.

14-3-3ζ was purified as described previously (35). Two surface-exposed cysteines were mutated for alanine (C25A and C189A). This allowed us to work with highly concentrated samples over longer times without the risk of disulfide bond formation. We showed earlier that the mutation of these two exposed cysteines does not change the fold or stability of the protein (35). The purity of the final protein samples, including phosphorylated and labeled samples, was checked by MALDI-MS.

Mutagenesis

Single-point mutants of MAP2c (S435D, S184D, T220E, and C348S) were produced using the QuikChange Lightning site-directed mutagenesis kit (Agilent Technologies, Santa Clara, CA) following the manufacturer's protocol and using MAP2c in the pET3a vector as a template. To prepare the double mutants (S189D/S435D, S199D/S435D, S319D/S435D, T320E/S435D, S342D/S435D, S350D/S435D, S367D/S435D, and S382D/S435D), S435D MAP2c in pET3a was used as a template, and C348S MAP2c was used as template for the double mutant E52C/C348S. The results of the mutations were confirmed by sequencing. For NMR measurement, the MAP2c mutants were expressed in 1 liter of ^{15}N M9 medium. Phosphorylation was performed as for the wild-type MAP2c.

Mass spectrometry

MAP2c at 0.2 mM concentration was phosphorylated for 24 h with 650 units of PKA/mg of MAP2c. Phosphorylated MAP2c and 0.16 mM unphosphorylated MAP2c were processed by the filter-aided sample preparation (FASP) method (63, 64). Proteins were alkylated and digested by trypsin on the filter unit membrane, and the resulting peptides were eluted by ammonium bicarbonate. One-tenth of the peptide mixture was analyzed directly, and the rest of the sample was used for phosphopeptide enrichment. Both peptide mixtures were separately analyzed on a LC-MS/MS system (RSLCnano connected to Orbitrap Elite, Thermo Fisher Scientific).

The MS data were acquired using a data-dependent strategy by selecting up to the top six precursors based on precursor abundance in the survey scan (350–2000 m/z). High-resolution HCD or ETD MS/MS spectra were acquired in the Orbitrap analyzer. The analysis of the mass spectrometric RAW data files was carried out using Proteome Discoverer software (version 1.4, Thermo Fisher Scientific) with an in-house Mascot (version 2.4.1, Matrix Science) search engine utilization. Peptides with a false discovery rate lower than 1%, rank 1, and search engine rank 1 and with at least 6 amino acids were considered. Quantitative information assessment was performed using Skyline (Skyline Software Systems).

NMR spectroscopy

NMR experiments were performed using a 950-MHz Bruker Avance III spectrometer equipped with a ^{1}H-^{13}C/^{15}N/D TCI cryogenic probe head with z axis gradients and a 700-MHz Bruker Avance III spectrometer equipped with a ^{1}H/^{13}C/^{15}N TXO cryogenic probe head with z axis gradients. All experiments were performed at 27 °C with the temperature calibrated according to the chemical shift differences of pure methanol peaks. The indirect dimensions in 3D and 5D experiments were acquired in a non-uniformly sampled manner. On-grid Poisson disk sampling with a Gaussian probability distribution (65) was applied.

The ^{1}H-^{15}N HSQC (66, 67) spectrum of wild-type MAP2c was recorded with spectral widths set to 11,904 Hz in the direct dimension and to 2500 Hz in the indirect dimension. 2048 and 128 complex points were acquired in the direct and the indirect dimensions, respectively. The ^{1}H-^{15}N HSQC spectra of the MAP2c mutants were recorded with spectral widths set to

17,045 Hz in the direct dimension and to 1000 Hz in the indirect dimension. 2048 and 256 complex points were acquired in the direct and the indirect dimensions, respectively. 16 scans with recycle delay set to 1 s were recorded.

The ^{1}H-^{15}N SOFAST-HMQC (68) spectra were recorded with spectral widths set to 13,297 Hz in the direct dimension and to 2500 Hz in the indirect dimension. 2048 and 64 complex points were acquired in the direct and the indirect dimensions, respectively. 16 scans with the recycle delay set to 228 ms were recorded in each experiment.

The 3D (CACO)NCACO spectrum (30) was acquired with spectral widths set to 7042 (acquired dimension (aq)) × 2000 (^{15}N) × 4000 (^{13}C$^{\alpha}$) Hz and with maximal evolution times of 46 ms (^{15}N) and 26 ms (^{13}C$^{\alpha}$) in the indirectly detected dimensions. The overall number of 1024 complex points was acquired in the acquisition dimension, and 3000 hypercomplex points were randomly distributed over the indirectly detected dimensions. The 5D CACONCACO spectrum (30) was acquired with spectral widths set to 7042 (aq) × 4000 (^{13}C$^{\alpha}$) × 2000 (^{15}N) × 2000 (^{13}C′) × 4000 (^{13}C$^{\alpha}$) Hz. The maximal evolution times in the indirectly detected dimensions were set to 26 ms for the ^{13}C$^{\alpha}$ dimensions, 46 ms for the ^{15}N dimension, and 28 ms for the ^{13}C′ dimension. The overall number of 1024 complex points was acquired in the acquisition dimension, and 3000 hypercomplex points were distributed over the indirectly detected dimensions.

The 3D (H)CANCO spectrum (69) was acquired with spectral widths set to 7042 (aq) × 2000 (^{15}N) × 4000 (^{13}C$^{\alpha}$) Hz and with maximal evolution times of 32 ms (^{15}N) and 26 ms (^{13}C$^{\alpha}$) in the indirectly detected dimensions. The overall number of 1024 complex points was acquired in the acquisition dimension, and 1500 hypercomplex points were distributed over the indirectly detected dimensions.

The 5D HC(CC-TOCSY)CACON (30) experiment was measured with the spectral widths set to 7042 (aq) × 2500 (^{15}N) × 4000 (^{13}C$^{\alpha}$) × 12500 (^{13}C$^{\text{ali}}$) × 5000 (^{1}H$^{\text{ali}}$) Hz where "ali" indicates nuclei in the aliphatic side chain. The maximal evolution times in the indirectly detected dimensions were set to 46 ms for the ^{15}N dimension, 26 ms for the ^{13}C$^{\alpha}$ dimension, 8 ms for the ^{13}C$^{\text{ali}}$ dimension, and 10 ms for the ^{1}H$^{\text{ali}}$ dimension. The overall number of 1024 complex points was acquired in the acquisition dimension, and 2000 hypercomplex points were distributed over the indirectly detected dimensions. The 3D (HC(CC-TOCSY))CACON (30) experiment was acquired with spectral widths set to 7042 (aq) × 1956 (^{15}N) × 4000 (^{13}C$^{\alpha}$) Hz and maximal evolution times of 46 ms for ^{15}N and 26 ms for ^{13}C$^{\alpha}$ indirectly detected dimensions. The overall number of 1024 complex points was acquired in the acquisition dimension, and 2000 hypercomplex points were distributed over the indirectly detected dimensions.

The 3D HNCO (70) spectra were acquired with spectral widths set to 18939 (aq) × 2000 (^{15}N) × 2000 (^{13}C′) Hz and maximal evolution times of 120 ms for ^{15}N and 80 ms for ^{13}C′ indirectly detected dimensions. The overall number of 2048 complex points was acquired in the acquisition dimension, and 2000 hypercomplex points were distributed over the indirectly detected dimensions.

The 3D HNCACB experimental data (71) were acquired with spectral widths set to 18939 (aq) × 2500 (^{15}N) × 17921 (^{13}C′) Hz. The number of recorded complex points was 2048, 40, and 128 for the ^{1}H, ^{15}N, and ^{13}C dimensions, respectively.

Uniformly sampled data processing and direct dimension processing of non-uniformly sampled data were done using NMRPipe software (72). Multidimensional Fourier transform with iterative algorithm for artifact suppression (73) was employed to process indirect dimensions in three-dimensional experiments. Indirect dimensions in five-dimensional experiments were processed using the sparse multidimensional Fourier transform (74). Spectral analysis was done using the software Sparky 3.115 (T. D. Goddard and D. G. Kneller, University of California, San Francisco).

The transient secondary structure propensities of phosphorylated and unphosphorylated MAP2c were calculated using the program SSP (75) as described previously (30). To obtain values not biased by the direct chemical effect of the presence of the phosphate group, SSP values were calculated only from the chemical shifts of ^{13}C$^{\alpha}$, ^{13}C$^{\beta}$, and ^{1}H$^{\alpha}$, which are insensitive to the presence of the phosphate group attached to neighboring residues, with the exception of ^{13}C$^{\beta}$ of the phosphorylated side chain (34). Residues pSer-184, pThr-220, and pSer-435 were excluded from the analysis. Regions of propensity to form helical structures were defined as stretches of at least three residues with SSP higher than 0.07 (interrupted by no more than two residues with SSP lower than 0.07). Regions of propensity to form extended structures were defined as stretches of at least three residues with SSP lower than −0.07 (interrupted by no more than two residues with SSP higher than −0.07).

Phosphorylation kinetics

Real-time phosphorylation was followed by recording ^{1}H,^{15}N-SOFAST-HMQC spectra at 27 °C in 5-min intervals for 30 h after the addition of PKA. [^{15}N]MAP2c at a 0.65 mM concentration in 50 mM MOPS, pH 6.9, 150 mM NaCl, 0.7 mM TCEP, 10 mM MgCl$_2$, 20 mM ATP, and 0.1 mM EDTA was mixed with 12,500 units of PKA directly in the NMR tube. According to the manufacturer, the PKA concentration was ~0.1 μM. The first SOFAST-HMQC experiment was started 2 min after the addition of PKA. The relative concentrations of individual phosphorylation sites were assumed to be proportional to the corresponding peak volumes, and PKA was assumed to be saturated by ATP and magnesium. Phosphorylation rates were modeled as

$$-\frac{d[\mathrm{S}]_i}{dt} = \frac{d[\mathrm{P}]_i}{dt} = \frac{k_{\mathrm{cat},i}c_{\mathrm{PKA}}[\mathrm{S}]_i/K_{m,i}}{1 + \Sigma_i[\mathrm{S}]_i/K_{m,i}} \qquad \text{(Eq. 1)}$$

where t is time, $[\mathrm{S}]_i$ and $[\mathrm{P}]_i$ are concentrations of the i-th unphosphorylated and phosphorylated phosphorylation sites, respectively, $k_{\mathrm{cat},i}$ and $K_{m,i}$ are catalytic and Michaelis constants, respectively, and c_{PKA} is the total concentration of PKA. The apparent rate constants, $k_{\mathrm{obs},i}$, were defined as

$$k_{\mathrm{obs},i} = \frac{k_{\mathrm{cat},i}/K_{m,i}}{1 + \Sigma_i c_{\mathrm{MAP2c}}/K_{m,i}} \qquad \text{(Eq. 2)}$$

where c_{MAP2c} is the total concentration of MAP2c.

The k_{obs} value for Ser-435 was estimated by fitting the peak volumes to the integrated form of Equation 1, assuming that concentrations of other phosphorylation sites were equal to

c_{MAP2c}. The value of k_{cat} could not be reliably separated from k_{obs} for the obtained data. To estimate k_{obs} for Ser-184, Ser-189, and Thr-220, the ratios of peak volumes of pSer-184, pSer-189, and pThr-220 were assumed to be constant, and peak volumes, V, were fitted to the equation

$$V = V(0)\left(1 - \frac{a\exp(-bt) - b\exp(-at)}{a - b}\right) \qquad \text{(Eq. 3)}$$

where $a < b$, $a = k_{obs}c_{PKA}$ at [Ser-435] $= 0$, and b approximates the rate of the preceding phosphorylation of Ser-435.

Interaction with 14-3-3ζ by NMR

A sample containing 550 μl of 0.69 mM unphosphorylated [^{15}N,^{13}C]MAP2c was titrated with 17, 34, 69, 139, and 280 μl of 2.5 mM 14-3-3ζ (monomer concentration). After each addition, a 3D HNCO spectrum was recorded. 500 μl of 0.61 mM phosphorylated [^{15}N,^{13}C]MAP2c was titrated with 17, 34, 68, 136, and 272 μl of 2.2 mM 14-3-3ζ. For each titration point, 3D HNCO and 2D ^{1}H,^{15}N HSQC spectra were measured.

Microscale thermophoresis

Binding experiments between 14-3-3ζ and MAP2c in the unphosphorylated and phosphorylated forms were measured by MST at 20 °C. E52C/C348S MAP2c was specifically labeled at position Cys-52 with the fluorophore Alexa Fluor 647 C2 maleimide (AF647, Thermo Fisher Scientific). Titration experiments were performed in buffer containing 50 mM Tris, pH 7.5. In addition, 0.5 mg/ml BSA and 0.05% Tween 20 were added to prevent aggregation of the studied proteins on standard capillary walls. Binding studies were performed at 30% laser power with a Monolith NT.115 device (NanoTemper Technologies, Munich, Germany) in combination with three different MST power setups (at 40, 60, and 80%). The data were fitted to the following model,

$$\Phi = \Phi_0 + \Delta\frac{x + nc + K'_D - \sqrt{(x + nc + K'_D)^2 - 4ncx}}{2n} \qquad \text{(Eq. 4)}$$

where $\Phi = F/F_{max}$ is the normalized fluorescence, c is the total concentration of MAP2c, x is the total monomer concentration of 14-3-3ζ, and K'_D is the apparent dissociation constant expressed for the 14-3-3ζ monomer concentration.

Microtubular co-sedimentation assay

Co-sedimentation assays were adapted from Valencia *et al.* (76). Samples of 5 μM porcine brain tubulin in 30 μl of tubulin buffer (80 mM PIPES, pH 6.9, 2 mM $MgCl_2$, and 0.5 mM EGTA) were polymerized by the addition of 2 μM MAP2c and incubated for 15 min at 37 °C. 20 nM, 200 nM, 2 μM, and 20 μM concentrations of 14-3-3ζ (monomer concentrations) in tubulin buffer were added to the MAP2c-tubulin samples before the incubation. Microtubule bundles were pelleted by centrifugation (50,000 × *g*, 30 min, 37 °C). The pellet was washed by tubulin buffer and adjusted to the same volume as the supernatant with the buffer. Both supernatant and pellet were mixed with the SDS gel loading buffer. Equal volumes of supernatant and pellet were separated by SDS-PAGE. The Coomassie Brilliant Blue-stained protein bands were analyzed using the densitometer software QuantiScan 3.0.

Author contributions—S. J., L. Z., J. H., J. N., and K. M. conceived and designed the research. S. J., E. Z., K. M., and Z. T. prepared the protein samples. Z. T. and K. M., respectively, performed the MST and co-sedimentation assays. L. Z., M. Z., S. J., and K. M. performed the NMR experiments, and K. H. and Z. Z. performed the MS experiments. S. J., L. Z., and J. H. wrote the paper. All authors reviewed the results and approved the final version of the manuscript.

Acknowledgments—CIISB research infrastructure project LM2015043, funded by Ministry of Education, Youth and Sports of the Czech Republic (MEYS CR), is gratefully acknowledged for partial financial support of the measurements at the Josef Dadok National NMR Centre, Biomolecular Interactions and Crystallization and Proteomics Core Facilities, Central European Institute of Technology (CEITEC), Masaryk University. We thank Tanvir Shaikh for language corrections and Arnošt Mládek and Irmgard Fischer for technical help.

References

1. Gamblin, T. C., Nachmanoff, K., Halpain, S., and Williams, R. C. (1996) Recombinant microtubule-associated protein 2c reduces the dynamic instability of individual microtubules. *Biochemistry* **35,** 12576–12586
2. Jalava, N. S., Lopez-Picon, F. R., Kukko-Lukjanov, T. K., and Holopainen, I. E. (2007) Changes in microtubule-associated protein-2 (MAP2) expression during development and after status epilepticus in the immature rat hippocampus. *Int. J. Dev. Neurosci.* **25,** 121–131
3. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000) Intrinsic protein disorder in complete genomes. *Genome Inform.* **11,** 161–171
4. Dunker, A. K., Oldfield, C. J., Meng, J., Romero, P., Yang, J. Y., Chen, J. W., Vacic, V., Obradovic, Z., and Uversky, V. N. (2008) The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics* **9,** S1
5. Dyson, H. J., and Wright, P. E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6,** 197–208
6. Tompa, P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.* **579,** 3346–3354
7. Fink, A. L. (2005) Natively unfolded proteins. *Curr. Opin. Struct. Biol.* **15,** 35–41
8. Dehmelt, L., and Halpain, S. (2005) The MAP2/Tau family of microtubule-associated proteins. *Genome Biol.* **6,** 204
9. Sündermann, F., Fernandez, M. P., and Morgan, R. O. (2016) An evolutionary roadmap to the microtubule-associated protein MAP Tau. *BMC Genomics* **17,** 264
10. Grundke-Iqbal, I., Iqbal, K., Quinlan, M., Tung, Y. C., Zaidi, M. S., and Wisniewski, H. M. (1986) Microtubule-associated protein Tau: a component of Alzheimer paired helical filaments. *J. Biol. Chem.* **261,** 6084–6089
11. Viereck, C., Tucker, R. P., and Matus, A. (1989) The adult rat olfactory system expresses microtubule-associated proteins found in the developing brain. *J. Neurosci.* **9,** 3547–3557
12. Vallee, R. (1980) Structure and phosphorylation of microtubule-associated protein 2 (MAP2). *Proc. Natl. Acad. Sci. U.S.A.* **77,** 3206–3210
13. Yamauchi, T., and Fujisawa, H. (1983) Disassembly of microtubules by the action of calmodulin-dependent protein kinase (kinase II) which occurs only in the brain tissues. *Biochem. Biophys. Res. Commun.* **110,** 287–291
14. Burns, R. G., Islam, K., and Chapman, R. (1984) The multiple phosphorylation of the microtubule-associated protein MAP2 controls the MAP2: tubulin interaction. *Eur. J. Biochem.* **141,** 609–615
15. Hoshi, M., Akiyama, T., Shinohara, Y., Miyata, Y., Ogawara, H., Nishida, E., and Sakai, H. (1988) Protein-kinase-C-catalyzed phosphorylation of the microtubule-binding domain of microtubule-associated protein 2 in-

hibits its ability to induce tubulin polymerization. *Eur. J. Biochem.* **174,** 225–230

16. Ainsztein, A. M., and Purich, D. L. (1994) Stimulation of tubulin polymerization by MAP-2: control by protein kinase C-mediated phosphorylation at specific sites in the microtubule-binding region. *J. Biol. Chem.* **269,** 28465–28471
17. Illenberger, S., Drewes, G., Trinczek, B., Biernat, J., Meyer, H. E., Olmsted, J. B., Mandelkow, E. M., and Mandelkow, E. (1996) Phosphorylation of microtubule-associated proteins MAP2 and MAP4 by the protein kinase p110mark: phosphorylation sites and regulation of microtubule dynamics. *J. Biol. Chem.* **271,** 10834–10843
18. Drewes, G., Ebneth, A., and Mandelkow, E. M. (1998) MAPs, MARKs and microtubule dynamics. *Trends Biochem. Sci.* **23,** 307–311
19. Sánchez, C., Díaz-Nido, J., and Avila, J. (2000) Phosphorylation of microtubule-associated protein 2 (MAP2) and its relevance for the regulation of the neuronal cytoskeleton function. *Prog. Neurobiol.* **61,** 133–168
20. Illenberger, S., Zheng-Fischhöfer, Q., Preuss, U., Stamer, K., Baumann, K., Trinczek, B., Biernat, J., Godemann, R., Mandelkow, E. M., and Mandelkow, E. (1998) The endogenous and cell cycle-dependent phosphorylation of Tau protein in living cells: implications for Alzheimer's disease. *Mol. Biol. Cell* **9,** 1495–1512
21. Avila, J., Domínguez, J., and Díaz-Nido, J. (1994) Regulation of microtubule dynamics by microtubule-associated protein expression and phosphorylation during neuronal development. *Int. J. Dev. Biol.* **38,** 13–25
22. Aitken, A., Collinge, D. B., van Heusden, B. P., Isobe, T., Roseboom, P. H., Rosenfeld, G., and Soll, J. (1992) 14-3-3 proteins: a highly conserved, widespread family of eukaryotic proteins. *Trends Biochem. Sci.* **17,** 498–501
23. Skoulakis, E. M., and Davis, R. L. (1998) 14-3-3 proteins in neuronal development and function. *Mol. Neurobiol.* **16,** 269–284
24. Hashiguchi, M., Sobue, K., and Paudel, H. K. (2000) 14-3-3ζ is an effector of Tau protein phosphorylation. *J. Biol. Chem.* **275,** 25247–25254
25. Sadik, G., Tanaka, T., Kato, K., Yamamori, H., Nessa, B. N., Morihara, T., and Takeda, M. (2009) Phosphorylation of Tau at Ser214 mediates its interaction with 14-3-3 protein: implications for the mechanism of tau aggregation. *J. Neurochem.* **108,** 33–43
26. Sluchanko, N. N., Seit-Nebi, A. S., and Gusev, N. B. (2009) Effect of phosphorylation on interaction of human Tau protein with 14-3-3ζ. *Biochem. Biophys. Res. Commun.* **379,** 990–994
27. Sluchanko, N. N., Seit-Nebi, A. S., and Gusev, N. B. (2009) Phosphorylation of more than one site is required for tight interaction of human Tau protein with 14-3-3ζ. *FEBS Lett.* **583,** 2739–2742
28. Ozer, R. S., and Halpain, S. (2000) Phosphorylation-dependent localization of microtubule-associated protein MAP2c to the actin cytoskeleton. *Mol. Biol. Cell* **11,** 3573–3587
29. Alexa, A., Schmidt, G., Tompa, P., Ogueta, S., Vázquez, J., Kulcsár, P., Kovács, J., Dombrádi, V., and Friedrich, P. (2002) The phosphorylation state of threonine-220, a uniquely phosphatase-sensitive protein kinase A site in microtubule-associated protein MAP2c, regulates microtubule binding and stability. *Biochemistry* **41,** 12427–12435
30. Nováček, J., Janda, L., Dopitová, R., Žídek, L., Sklenář, V. (2013) Efficient protocol for backbone and side-chain assignments of large, intrinsically disordered proteins: transient secondary structure analysis of 49.2-kDa microtubule-associated protein 2c. *J. Biomol. NMR* **56,** 291–301
31. Obenauer, J. C., Cantley, L. C., and Yaffe, M. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs, *Nucleic Acids Res.* **31,** 3635–3641
32. Xue, Y., Liu, Z., Cao, J., Ma, Q., Gao, X., Wang, Q., Jin, C., Zhou, Y., Wen, L., and Ren, J. (2011) GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. *Protein Eng. Des. Sel.* **24,** 255–260
33. Wong, Y. H., Lee, T. Y., Liang, H. K., Huang, C., Yang, Y. H., Chu, C. H., Huang, H. D., Ko, M. T., and Hwang, J. K. (2007) KinasePhos 2.0: a Web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.* **35,** 588–594
34. Louša, P., Nedozrálová, H., Župa, E., Nováček, J., and Hritz, J. (2017) Phosphorylation of the regulatory domain of human tyrosine hydroxylase 1 monitored using non-uniformly sampled NMR. *Biophys. Chem.* **223,** 25–29
35. Hritz, J., Byeon, I. J., Krzysiak, T., Martinez, A., Sklenar, V., and Gronenborn, A. M. (2014) Dissection of binding between a phosphorylated tyrosine hydroxylase peptide and 14-3-3ζ: a complex story elucidated by NMR. *Biophys. J.* **107,** 2185–2194
36. von Bergen, M., Friedhoff, P., Biernat, J., Heberle, J., Mandelkow, E. M., and Mandelkow, E. (2000) Assembly of Tau protein into Alzheimer paired helical filaments depends on a local sequence motif ((306)VQIVYK(311)) forming beta structure. *Proc. Natl. Acad. Sci. U.S.A.* **97,** 5129–5134
37. Xie, C., Miyasaka, T., Yoshimura, S., Hatsuta, H., Yoshina, S., Kage-Nakadai, E., Mitani, S., Murayama, S., and Ihara, Y. (2014) The homologous carboxyl-terminal domains of microtubule-associated protein 2 and Tau induce neuronal dysfunction and have differential fates in the evolution of neurofibrillary tangles. *PLoS One* **9,** e89796
38. Mukrasch, M. D., Bibow, S., Korukottu, J., Jeganathan, S., Biernat, J., Griesinger, C., Mandelkow, E., and Zweckstetter, M. (2009) Structural polymorphism of 441-residue Tau at single residue resolution. *PLOS Biol.* **7,** e34
39. Joo, Y., Schumacher, B., Landrieu, I., Bartel, M., Smet-Nocca, C., Jang, A., Choi, H. S., Jeon, N. L., Chang, K. A., Kim, H. S., Ottmann, C., and Suh, Y. H. (2015) Involvement of 14-3-3 in tubulin instability and impaired axon development is mediated by Tau. *FASEB J.* **29,** 4133–4144
40. Sluchanko, N. N., and Gusev, N. B. (2010) 14–3-3 proteins and regulation of cytoskeleton. *Biochemistry* (*Mosc.*) **75,** 1528–1546
41. Ackmann, M., Wiech, H., and Mandelkow, E. (2000) Nonsaturable binding indicates clustering of Tau on the microtubule surface in a paired helical filament-like conformation. *J. Biol. Chem.* **275,** 30335–30343
42. Kennelly, P. J., and Krebs, E. G. (1991) Consensus sequences as substrate specificity determinants for protein kinases and protein phosphatases. *J. Biol. Chem.* **266,** 15555–15558
43. Landrieu, I., Lacosse, L., Leroy, A., Wieruszeski, J. M., Trivelli, X., Sillen, A., Sibille, N., Schwalbe, H., Saxena, K., Langer, T., and Lippens, G. (2006) NMR analysis of a Tau phosphorylation pattern. *J. Am. Chem. Soc.* **128,** 3575–3583
44. Sillen, A., Barbier, P., Landrieu, I., Lefebvre, S., Wieruszeski, J. M., Leroy, A., Peyrot, V., and Lippens, G. (2007) NMR investigation of the interaction between the neuronal protein Tau and the microtubules. *Biochemistry* **46,** 3055–3064
45. Yoshida, H., and Goedert, M. (2006) Sequential phosphorylation of tau protein by cAMP-dependent protein kinase and SAPK4/p38delta or JNK2 in the presence of heparin generates the AT100 epitope. *J. Neurochem.* **99,** 154–164
46. Brandt, R., Lee, G., Teplow, D. B., Shalloway, D., and Abdel-Ghany, M. (1994) Differential effect of phosphorylation and substrate modulation on Tau's ability to promote microtubule growth and nucleation. *J. Biol. Chem.* **269,** 11776–11782
47. Itoh, T. J., Hisanaga, S., Hosoi, T., Kishimoto, T., and Hotani, H. (1997) Phosphorylation states of microtubule-associated protein 2 (MAP2) determine the regulatory role of MAP2 in microtubule dynamics. *Biochemistry* **36,** 12574–12582
48. Madeira, F., Tinti, M., Murugesan, G., Berrett, E., Stafford, M., Toth, R., Cole, C., MacKintosh, C., and Barton, G. J. (2015) 14-3-3-Pred: improved methods to predict 14–3-3-binding phosphopeptides. *Bioinforma* **31,** 2276–2283
49. Mukrasch, M. D., Biernat, J., von Bergen, M., Griesinger, C., Mandelkow, E., and Zweckstetter, M. (2005) Sites of Tau important for aggregation populate β-structure and bind to microtubules and polyanions. *J. Biol. Chem.* **280,** 24978–24986
50. Mukrasch, M. D., von Bergen, M., Biernat, J., Fischer, D., Griesinger, C., Mandelkow, E., and Zweckstetter, M. (2007) The "jaws" of the Tau-microtubule interaction. *J. Biol. Chem.* **282,** 12230–12239
51. Sluchanko, N. N., and Gusev, N. B. (2011) Probable participation of 14–3-3 in Tau protein oligomerization and aggregation. *J. Alzheimers Dis.* **27,** 467–476
52. Gómez-Ramos, A., Díaz-Hernández, M., Rubio, A., Miras-Portugal, M. T., and Avila, J. (2008) Extracellular Tau promotes intracellular calcium increase through M1 and M3 muscarinic receptors in neuronal cells. *Mol. Cell. Neurosci.* **37,** 673–681

53. Gardiner, J., Overall, R., and Marc, J. (2011) The microtubule cytoskeleton acts as a key downstream effector of neurotransmitter signaling. *Synapse* **65,** 249–256
54. Ovsepian, S. V., O'Leary, V. B., and Zaborszky, L. (2016) Cholinergic mechanisms in the cerebral cortex: beyond synaptic transmission. *Neuroscientist* **22,** 238–251
55. Busceti, C. L., Di Pietro, P., Riozzi, B., Traficante, A., Biagioni, F., Nisticò, R., Fornai, F., Battaglia, G., Nicoletti, F., and Bruno, V. (2015) 5-HT(2C) serotonin receptor blockade prevents tau protein hyperphosphorylation and corrects the defect in hippocampal synaptic plasticity caused by a combination of environmental stressors in mice. *Pharmacol. Res.* **99,** 258–268
56. Wang, H., and Zhang, M. (2012) The role of Ca^{2+}-stimulated adenylyl cyclases in bidirectional synaptic plasticity and brain function. *Rev. Neurosci.* **23,** 67–78
57. Vossler, M. R., Yao, H., York, R. D., Pan, M. G., Rim, C. S., and Stork, P. J. (1997) cAMP activates MAP kinase and Elk-1 through a B-Raf- and Rap1-dependent pathway. *Cell* **89,** 73–82
58. Kim, H. A., DeClue, J. E., and Ratner, N. (1997) cAMP-dependent protein kinase A is required for Schwann cell growth: interactions between the cAMP and neuregulin/tyrosine kinase pathways. *J. Neurosci. Res.* **49,** 236–247
59. Blanco-Aparicio, C., Torres, J., and Pulido, R. (1999) A novel regulatory mechanism of MAP kinases activation and nuclear translocation mediated by PKA and the PTP-SL tyrosine phosphatase. *J. Cell Biol.* **147,** 1129–1136
60. Ambrosini, A., Tininini, S., Barassi, A., Racagni, G., Sturani, E., and Zippel, R. (2000) cAMP cascade leads to Ras activation in cortical neurons. *Brain Res. Mol. Brain Res.* **75,** 54–60
61. Mohan, R., and John, A. (2015) Microtubule-associated proteins as direct crosslinkers of actin filaments and microtubules. *IUBMB Life* **67,** 395–403
62. Elie, A., Prezel, E., Guérin, C., Denarier, E., Ramirez-Rios, S., Serre, L., Andrieux, A., Fourest-Lieuvin, A., Blanchoin, L., and Arnal, I. (2015) Tau co-organizes dynamic microtubule and actin networks. *Sci. Rep.* **5,** 9964
63. Wiśniewski, J. R., Ostasiewicz, P., and Mann, M. (2011) High recovery FASP applied to the proteomic analysis of microdissected formalin fixed paraffin embedded cancer tissues retrieves known colon cancer markers. *J. Proteome Res.* **10,** 3040–3049
64. Wiśniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009) Universal sample preparation method for proteome analysis. *Nat. Methods* **6,** 359–362
65. Kazimierczuk, K., Zawadzka, A., and Koźmiński, W. (2008) Optimization of random time domain sampling in multidimensional NMR. *J. Magn. Reson.* **192,** 123–130
66. Bodenhausen, G., and Ruben, D. J. (1980) Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chem. Phys. Lett.* **69,** 185–189
67. Sklenar, V., Piotto, M., Leppik, R., and Saudek, V. (1993) Gradient-tailored water suppression for ^{1}H-^{15}N HSQC experiments optimized to retain full sensitivity. *J. Magn. Reson.* **102,** 241–245
68. Schanda, P., and Brutscher, B. (2005) Very fast two-dimensional NMR spectroscopy for real-time investigation of dynamic events in proteins on the time scale of second. *J. Am. Chem. Soc.* **127,** 8014–8015
69. Bermel, W., Bertini, I., Felli, I. C., and Pierattelli, R. (2009) Speeding up (13)C direct detection biomolecular NMR spectroscopy. *J. Am. Chem. Soc.* **131,** 15339–15345
70. Kay, L. E., Ikura, M., Tschudin, R., and Bax, A. (1990) Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *J. Magn. Reson.* **89,** 496–514
71. Sattler, M., Schleucher, J., and Griesinger, C. (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucleic Magn. Reson. Spectrosc.* **34,** 93–158
72. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6,** 277–293
73. Stanek, J., and Koźmiński, W. (2010) Iterative algorithm of discrete Fourier transform for processing randomly sampled NMR data sets. *J. Biomol. NMR* **47,** 65–77
74. Kazimierczuk, K., Zawadzka, A., and Koźmiński, W. (2009) Narrow peaks and high dimensionalities: exploiting the advantages of random sampling. *J. Magn. Reson.* **197,** 219–228
75. Marsh, J. A., Singh, V. K., Jia, Z., and Forman-Kay, J. D. (2006) Sensitivity of secondary structure propensities to sequence differences between α- and γ-synuclein: implications for fibrillation. *Protein Sci.* **15,** 2795–2804
76. Valencia, R. G., Walko, G., Janda, L., Novacek, J., Mihailovska, E., Reipert, S., Andrä-Marobela, K., and Wiche, G. (2013) Intermediate filament-associated cytolinker plectin 1c destabilizes microtubules in keratinocytes. *Mol. Biol. Cell* **24,** 768–784

VOLUME 292 (2017) PAGES 6715–6727
DOI 10.1074/jbc.A116.771097

Quantitative mapping of microtubule-associated protein 2c (MAP2c) phosphorylation and regulatory protein 14-3-3ζ-binding sites reveals key differences between MAP2c and its homolog Tau.

Séverine Jansen, Kateřina Melková, Zuzana Trošanová, Kateřina Hanáková, Milan Zachrdla, Jiři Nováček, Erik Župa, Zbyněk Zdráhal, Jozef Hritz, and Lukáš Žídek

PAGE 6720:

There was an error in Fig. 6. *Panels A* and *B* were inadvertently exchanged. The correct legend should read as follows. **Microscale thermophoretic analysis of interaction between 5 μM E52C/C348S MAP2c and 14-3-3ζ in 50 mM Tris buffer.** *A*, phosphorylated MAP2c; *B*, unphosphorylated MAP2c. The plots show interactions in the 14-3-3ζ monomer concentration range from 36.6 nM to 1.2 mM. The mean values ± S.D. for each concentration point were calculated from triplicate measurements.

The results are described correctly in the text, and the error does not affect the results or conclusions of the work.

7 Conclusions

The 15 selected articles of the applicant are included in this habilitation thesis within four scientific chapters that are preceded by a general theoretical chapter. The theoretical chapter focuses on the various aspects of MD that are relevant to the selected scientific topics. In the general theoretical chapter, the applicant also incorporated some of his observations from the ongoing teaching process and thus this chapter, as well as the whole thesis, can be useful for students interested in the given methodologies. The main applicant's aim went beyond the simple proclamation that biomolecules and their complexes are dynamic. It is shown here how this fact, when properly implemented in a variety of computational approaches leads to much more reliable results in terms of the agreement with experimental data.

The most valuable outcome from the first scientific chapter is the observed large sensitivity of docking reliability, not only due to conformational changes but also due to thermal motion within the protein target. This fact was presented and applied in the docking results for cytochrome P450 2D6 combining MD and molecular docking. The available crystal structure of cytochrome P450 2D6 was only in the apo form where the active site was too small for the majority of known substrates, leading to only 20% reliability of structural binding poses. The described approach allowed efficient incorporation of induced fit effects when combined with the designed decision tree which provided a high degree of robustness. The designed decision tree, divides potential drug-like molecules, based on their size and number of hydrophobic regions, into three classes. When each class is then docked into the provided three different structures of cytochrome P450 2D6, the prediction reliability increased to 80% without any additional computation costs. The speed and simplicity of approaches are particularly important for the pharmaceutical application and indeed this approach became very popular in the medicinal chemistry field and was applied to a large variety of protein targets. The applicant wants to emphasize that while described modifications provide improved reliability of molecular docking in terms of structural information, they were still absolutely unsatisfactory in terms of the prediction of binding affinities from docking simulations dependent on the scoring function.

Binding affinities along with other properties such as conformer populations, protein stability, solubility, etc. are governed by the enthalpic and entropic components of the Gibbs free energy differences. This is the reason why free energy is central to physical chemistry and related thermodynamically oriented computational chemistry. It is known from statistical physics that free energy is directly related to the states within statistical ensembles and therefore the capability to sample sufficiently by particular computational tools is so important. Because the sampling capability of standard MD is quite small, dozens of different alternative approaches have been developed in the community over the last decades. The second scientific chapter presents the applicant's contributions within the scope of Hamiltonian replica exchange molecular dynamics. Here two particular modifications of the hamiltonian within the H-REMD scheme are stressed. The first modification used so-called increased softness of non-bonded interactions between particular sets of atoms. Its applicability was shown on GTP and 8-Br-GTP molecules where non-bonded interactions between base and sugar parts were softened to allow the faster transition between anti and syn conformations. The resulting sampling enhancement was over three orders of magnitude faster than regular MD while providing the production of a statistical ensemble of the same quality. The second modification of Hamiltonian within H-REMD was inspired by umbrella sampling methods, in which the ligand was pulled from and into the binding site by using un-directional distance-field distance-restraints. It was applied to the binding/unbinding phenomena of phosphopeptides to the 14-3-3 protein. In addition to the preferred binding pathways, the binding affinities were predicted with a reasonable agreement to the experimental data. On the other hand, it should be also pointed out that this approach is computationally very demanding because at every distance both the ligand and protein degrees of freedom need to be sampled sufficiently.

In many applications (e.g. when addressing a set of similar molecules differing by a functional group) it is sufficient to calculate relative rather than absolute binding affinities. In such cases, alchemical free energy perturbation approaches can be used as is described in the third scientific chapter. However, these calculations may suffer a failure of convergence if molecules contain high intramolecular energy barriers. Here, the applicant designed an enhanced sampling-one step perturbation (ES-OS) method to tackle this problem in a highly efficient way. The core idea is to design a very specific and completely artificial reference state (using two different sets of soft-core non-bonded interactions). If the reference state is well designed then its single MD run is sufficient to calculate relative free energies between similar compounds in the studied set. The practical applicability of the ES-OS method was shown for the set of C8-analogs of GTP for which conformational probabilities and relative lipophilicity, solubility and binding affinities with respect to the FtsZ protein target were calculated and their comparison with the available experimental data shows good agreement despite the very limited computational costs concerning other computational methods. The wider use of this method in computational medicinal chemistry is prohibited by the need to design a specific reference state that is generally far from trivial.

In the fourth scientific chapter focused on structural and interaction properties of IDPs, the applicant applied primarily solution NMR rather than computational methods. The reasons for this are manifold. In the case of monitoring kinetic profiles of IDPs phosphorylation, comparable data obtained by simulations would require an advanced combination of MD with time-dependent QM methods with still quite questionable outcomes. On the other hand, NMR spectroscopy allows measuring these data in a few days when using ^{15}N labeled IDP of interest in sufficient concentration and the kinase of interest. In the case of binding affinities – computational methods are applicable in a similar manner to that described in the second and third scientific chapters but the situation starts to differ when having longer IDP phosphorylated on multiple sites where each of them is capable of interacting with globular proteins such as 14-3-3 proteins. In such cases, avidity effects start to be important and very difficult to be incorporated within the computational predictions. Such situations are also quite challenging for the proper interpretation of NMR data as the applicant showed for the doubly phosphorylated fragment (each phosphorylated site is different) of TH interacting with 14-3-3zeta homodimer having two binding sites. Similar studies are very rare in literature and therefore more data should be produced to have good experimental references for future computational studies. The last significant issue is the fact that biomolecular force-field parameters were optimized and tested for globular proteins over a half-century ago. Recently, it turned out that most of them simply overstabilize protein structures and when applied to IDPs – artificial collapse tends to be observed. In our recent publication[P13], we tested several popular biomolecular force-fields and paying attention to particular parametrizations of water models and when comparing with a whole range of experimental, mostly NMR data, it turned out that most of them are inadequate for selected IDP proteins. For only very few combinations did we observe decent but far from perfect agreement with the experimental data. In addition to more reliable force-field parameters, there is the obvious challenge for the enhanced sampling of larger IDP proteins when attempting to simulate them in an explicit water environment. The applicant with his students and colleagues continues to work on addressing both of those aspects. In the longer-term perspective, we plan to apply such improved approaches to studies of conformation changes of the Tau protein from its native to the pathological fibrils observed in Alzheimer's disease patients.

Bibliography

1. Ponder JW, Case DA. Protein Simulations. *Adv Protein Chem*. 2003;66:27-85. http://www.sciencedirect.com/science/article/pii/S006532330366002X

2. Berendsen HJC. *Simulating the Physical World: Hierarchical Modeling from Quantum Mechanics to Fluid Dynamics*. Vol 9780521835.; 2007. doi:10.1017/CBO9780511815348

3. LEACH AR. *Molecular Modelling : Principles and Applications.* Essex: Longman; 1996.

4. Frenkel, D.; Smit B. *Understanding Molecular Simulation: From Algorithms to Applications. Academic Press, Boston*. Academic Press; 2002.

5. Schlick T. *Molecular Modeling and Simulation: An Interdisciplinary Guide.* Springer-Verlag, New York; 2002.

6. Zuckerman DM. *Statistical Physics of Biomolecules; CRC Press*.; 2010.

7. Allison JR. Computational methods for exploring protein conformations. *Biochem Soc Trans*. 2020;48(4):1707-1724. doi:10.1042/BST20200193

8. Swendsen, R.H.; Wang JS. Replica Monte Carlo simulation of spin glasses. *Phys Rev Lett*. 1986;57:2607–2609.

9. Sugita, Y.; Okamoto Y. Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett*. 1999;314:141–151.

10. Hansmann UHE. Parallel tempering algorithm for conformational studies of biological molecules. *Chem Phys Lett*. 1997;281:140-150.

11. Fukunishi, H.; Watanabe, O.; Takada S. On the Hamiltonian replica exchange method for efficient sampling of biomolecu-lar systems: Application to protein structure prediction. *J Chem Phys*. 2002;116:9058.

12. Sugita, Y.; Kitao, A.; Okamoto Y. Multidimensional replica-exchange method for free-energy calculations. *J Chem Phys*. 2000;113:6042.

13. Abrams, C.; Bussi G. Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy*. 2014;16:163-199.

14. Hritz J, Oostenbrink C. Hamiltonian replica exchange molecular dynamics using

soft-core interactions. *J Chem Phys*. 2008;128(14):1-10. doi:10.1063/1.2888998

15. Prakash, M.K., Barducci, A.; Parrinello M. Replica temperatures for uniform exchange and efficient roundtrip times in explicit solvent parallel tempering simulations. *J Chem Theory Comput*. 2011;7:2025–2027. doi:10.1021/ct200208h

16. Spiwok, V.; Sucur, Z.; Hosek P. Enhanced sampling techniques in biomolecular simulations. *Biotech Adv*. 2015;33:1130-1140.

17. Affentranger, R.; Tavernelli, I.; Di Iorio EE. A novel Hamiltonian replica exchange MD protocol to enhance protein conformational space sampling. *J Chem Theory Comput*. 2006;2:217–228. doi:10.1021/ct050250b

18. Meli, M.; Colombo GA. Hamiltonian Replica Exchange Molecular Dynamics (MD) Method for the Study of Folding, Based on the Analysis of the Stabilization Determinants of Proteins. *Int J Mol Sci*. 2013;14:12157-12169.

19. de Ruiter, A.; Oostenbrink C. Protein–Ligand Binding from Distancefield Distances and Hamiltonian Replica Ex-change Simulations. *J Chem Theory Comput*. 2013;9:883-892.

20. Hritz, J.; Oostenbrink C. Optimization of Replica Exchange Molecular Dynamics by Fast Mimicking. *J Chem Phys*. 2007;127:204104.

21. Rowland, P.; Blaney, F. E.; Smyth, M. G.; Jones, J. J.; Leydon VR., Oxbrow, A. K.; Lewis, C. J.; Tennant, M. G.; Modi S. E, D.S.; Chenery, R.J.; Bridges AM. Crystal structure of human cytochrome P450 2D6. *J Biol Chem*. 2006;281:7614–7622.

22. Lui VWY, Peyser ND., Ng PKS, et al. Frequent mutation of receptor protein tyrosine phosphatases provides a mechanism for STAT3 hyperactivation in head and neck cancer. *Proc Natl Acad Sci U S A*. 2014;111(3):1114-1119. doi:10.1073/pnas.1319551111

23. Oostenbrink C.; de Ruiter A.; Hritz J.; Vermeulen NPE. Malleability and versatility of Cytochrome P450 active sites studied by molecular simulations. *Curr Drug Metab*. 2012;13:190-196.

24. Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren WF. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chem Phys Lett*. 1994;222:529-539. doi:10.1016/0009-2614(94)00397-1

25. Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman PA. The

weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J Comput Chem*. 1992;13:1011–1021.

26. Torrie, G.M.; Valleau JP. Nonphysical sampling distributions in monte carlo free-energy estimation: umbrella sampling. *J Comput Phys*. 1977;23:187–199. doi:10.1016/0021-9991(77)90121-8

27. Kaestner J. Umbrella sampling. *Comput Mol Sci*. 2011;1:932-942. doi:10.1002/wcms.66

28. Tembe, B.; McCammon JA. Ligand-receptor interactions. *Comput Chem*. 1984;8:281-283.

29. Zwanzig RW. High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J Chem Phys*. 1954;22:1420-1426.

30. Beveridge, D. L.; DiCapua FM. Free energy via molecular simulation: applications to chemical and biomolecualr systems. *Ann Rev Biophys Biophys Chem*. 1989;18:431-492.

31. Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon JA. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys J*. 1997;72:1047-1069.

32. Kollman PA. Free energy calculations: applications to chemical and biochemical phenomena. *Chem Rev*. 1993;93:2395-2417.

33. Tompa P. Intrinsically disordered proteins: a 10-year recap. *Trends Biochem Sci*. 2012;37:509-516. doi:10.1016/j.tibs.2012.08.004

34. Uversky VN. A decade and a half of protein intrinsic disorder: biology still waits for physics. *Protein Sci*. 2013;22:693-724.

35. Dunker AK, Lawson JD, Brown CJ, et al. Intrinsically disordered protein. *J Mol Graph Model*. 2001;19(1):26-59. doi:10.1016/S1093-3263(00)00138-8

36. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J Mol Biol*. 2004;337(3):635-645. doi:10.1016/j.jmb.2004.02.002

37. Robustelli P, Piana S, Shaw DE. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc Natl Acad Sci U S A*. 2018;115(21):E4758-E4766. doi:10.1073/pnas.1800690115

38. Piana, S.; Lindorff-Larsen, K.; Shaw DE. How robust are protein folding

simulations with respect to force field parameterization? *Biophys J*. 2011;100:L47–L49.

39. Motáčková V, Nováček J, Zawadzka-Kazimierczuk A, et al. Strategy for complete NMR assignment of disordered proteins with highly repetitive sequences based on resolution-enhanced 5D experiments. *J Biomol NMR*. 2010;48(3):169-177. doi:10.1007/s10858-010-9447-3

40. Novacek, J.; Janda, L.; Dopitova, R.; Zidek, L.; Sklenar V. Efficient protocol for backbone and side-chain assignments of large, intrinsically disordered proteins: transient secondary structure analysis of 49.2 kDa microtubule associated protein 2c. *J Biomol NMR*. 2013;56:291–301.

41. Demo, G; Papouskova, V; Komarek, J; Kaderavek, P; Otrusinova, O; Srb, P; Rabatinova, A; Krasny, L; Zidek, L; Sklenar V; Wimmerova M. X-ray vs. NMR structure of N-terminal domain of delta-subunit of RNA polymerase. *J Struct Biol*. 2014;187:174–186.