From peptides to transmembrane proteins: helix versus kink formations in highly dynamical systems

Inauguraldissertation

zur
Erlangung der Würde
eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel
von

Olivier Daniel Bignucolo
Stäfa/ZH, Schweiz

Basel, 2016

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag der Herren Professoren

Prof. Dr. phil. Fakultätsverantwortlicher

Torsten Schwede

Prof. Dr. phil. Dissertationsleiter

Simon Bernèche

Prof. Dr. phil. Korreferent

Timm Maier

Basel, den 24. Mai 2016

Prof. Dr. Jörg Schibler

Dekan

MERCI

First and foremost, I would like to express my gratitude to my mentor and tutor Simon Bernèche. I always very appreciated his open-minded approach, his enthusiasm and encouragements. I will always remember when, during my master thesis, Navratna, looked at me staring at puzzling signals, and proposed to meet “a guy who might help us”. And, indeed, I learnt a lot with him in the following years. I appreciated his confidence. When I expressed my wish to pursue my education through a PhD, maybe just part time, he immediately decided to look for funding. At this moment, I knew I should do it full time.

I would further like to thank Professor Stephan Grzesiek and Dr. Navratna Vajpai, who opened the door to the fascinating world of molecular mechanics.

Extended thanks to Professor Dr. Joachim Seelig, who funded my studies and challenged us with the question of the membrane response to charged molecules as a possible mechanism of voltage gating modifiers.

A sincere thanks to Professor Dr. T. Schwede for support during the final phase of the thesis writing.

I would like to mention and thank three additional brave readers of this thesis, namely Niklaus Johner, Jürgen Haas, and Annaïse Jauch.

I wish to thank the present and past members of the group headed by Simon. Among those, a sweeping “tank you” to Yanyan: you always give the best of yourself, Florian: it is really nice to be your officemate, Niklaus: the ones who will work with you will be happy, Sefer: I will never forget our trip to Philadelphia and New York, Wojtiech: brilliant, Chungwen: you were a real great help when I started to construct my “double bilayers” with GROMACS. I knew that I wanted them, and you knew how I could get them.

A big hug to Yvonne, Rita and Sarah, for your smiles, your endless help in this computing world.

Many thanks, Marc, for productive scientific discussions with a cup of coffee in the hand. It is a pity that we did not start this earlier.

I would like to offer thanks to the Torsten Schwede group, for the joined group meetings and interesting discussions.

I would like to thank the sciCORE (Center for Scientific Computing, University of Basel), the CSCS (Swiss National Supercomputing Centre), the KTH (Swedish PDC center for High Performance Computing) for endless calculations.

Pour toi, Annaïse, moi qui suis toujours si bavard avec toi, les mots manquent. Tu as été mon plus grand soutien, et je t’en suis infiniment reconnaissant. La voilà :
„Trouver une super phrase pour Annaïse“.
Aussi toute ma gratitude pour ma famille et mes amis, aussi pour essayer de comprendre ce curieux monde du scientifique, qui consacre autant d’enthousiasme à étudier arbres, forêts et insectes, qu’à regarder ses « petites molécules » sur un écran, et qui se met aux études à l’âge où les autres organisent leur retraite.

Merci du fond du cœur
My apologies for the few I might have forgotten

Basel, 2016
Olivier Bignucolo

Abstract

This thesis describes investigations of the relationships between the sequence of small peptides and their folding propensities and the conformational changes of membrane lipids upon interactions with proteins, within the context of varying membrane potentials. In addition, a novel conformational change of a membrane protein will be presented.

The determination of structures of folded proteins has progressed remarkably, notably due to outstanding techniques like crystallography, nuclear magnetic resonance or cryo-electron microscopy. However, proteins are highly dynamic and, under physiological conditions, their behavior depends on the chemical and physical environment. On the other hand, a better understanding of the intrinsically disordered proteins requires approaches, which consider their dynamical nature. All-atom molecular dynamics simulations constitute a tool of choice to capture the conformational changes of peptides as well as larger systems involving bilayers and membrane proteins. The first part of this thesis is dedicated to the structural propensities of peptides explored at the amino acid level. The investigations have shown how subtle interactions with the solvent affect their fate towards helical conformations. These findings are further validated through a procedure aimed at reducing the differences between predicted and experimental values while maximizing the entropy of the ensemble. The short-lived conformations found along transition paths are difficult to observe experimentally. Consequently, a statistical approach to investigate at the picosecond timescale the dynamics of the folding events in relation to the surrounding molecules is introduced and successfully tested on a β–hairpin of known structure. These successful results lead to a proposal of a systematic study to elucidate the sequence-conformation(s) relationships at a larger scale.

The second project describes the interactions between spider toxins, the cell membrane and a voltage sensor domain in the context of ion channel gating modification. Spider toxins have contributed substantially to the understanding of ion channels. Most of them are gating modifiers, thus affecting the energy level required by ion channels to open or close. Because these molecules are capable of fine-tuning the function of ion channels, they represent very attractive candidates in the field of drug discovery, and some successes have been achieved in this regard. The initial objective of the study was to explore whether the toxin-induced perturbation of the

membrane affect consequently the voltage-gated ion channels without any direct binding to the target. A demanding statistical approach was chosen, which takes the high specificity of spider toxins observed *in vivo* into account. Although the inserted toxins altered noticeably several membrane features, the results support the idea that an indirect, lipid-mediated mode of action of spider toxins on the voltage-sensor domain is not the main driver of the voltage-gated modifier mechanism. However, the investigations led to unexpected discoveries. The strategy employed to investigate an indirect mechanism of spider toxin involved more than 100 replicated simulations of independent bilayers and voltage-sensor domains exposed to a wide range of membrane potentials. The analyses showed surprisingly that the membrane perturbation, induced by the voltage sensor domain, is voltage-dependent. In addition, a novel conformational change of the voltage sensor upon polarization was observed, namely a kink in the S4 helix.

The results discussed here aim to contribute to a better understanding in three domains: 1) The interplay between water and the amino acid side chains during conformational changes, precisely the hydration fluctuations of just a few amide or carbonyl functional groups are shown to affect the helix formation propensities of a small peptide. 2) The lipid-mediated gating modifier mechanism is not supported by the simulations. 3) A novel conformational change of the voltage-sensor domain is described as a response to variation of the membrane potential. Precisely, a kink in the middle of the S4 helix occurs only upon polarization. This kink formation allows gating charges to move across the membrane without exposing any hydrophobic residues to the cytoplasm.

Contents

List of Figures

1 Introduction

Molecules of living organisms are marked by fluctuations and conformational changes over time. Under physiological conditions, linear small peptides, intrinsically disordered proteins and lipids in a bilayer sample a basin of relatively low energy conformations. The features of the conformational ensemble depend largely on the chemical and physical environment. The intrinsically disordered proteins (IDP) constitute an extreme example. IDPs are found extended in the cytoplasm but some are thought to fold into a proper 3D structure upon binding with their cognate target through an induced-fit mechanism (*1*). Alternatively, the IDPs interaction with their target was described in terms constantly changing conformations between a folded state and a plethora of unfolded states, whereas only the appropriately folded conformation binds to the cognate target. This second view is thus called conformational selection (*2*). It is also possible that the relative contribution of each mechanism varies for different IDPs. Clearly, a better understanding of the mechanisms related to the conformational changes of IDPs is required. Although ordered proteins generally do have a low energy, native structure, they also undergo conformational changes to exert their function. A few examples include the conformational changes occurring upon ligand binding, enzyme catalysis, transport of small molecules through membranes, or opening/closing/inactivation of an ion channel as response to a varying membrane potential. The structure of membrane lipids is also subject to fluctuations related to interacting molecules, or during phase transition upon temperature change. In this thesis, I discuss my strategies to decipher conformational propensities and structural changes of small peptides on the one hand and the modifications of a membrane structure upon interaction with toxins, which are typically extracellular peptides, on the other hand. In addition, novel conformational changes of transmembrane proteins triggered by the membrane electrostatic potential will be presented. Precisely, a kink formation in the longest helix of a voltage-sensor domain is described for the first time as a response to membrane polarization. Common to all these themes is the emphasis on the conformational changes of biological molecules and the resort to statistical analyses to discriminate significant, largely replicated, conformational changes, from rare observations. This chosen methodological approach has the advantage to clearly deliver reliable results, but the risk is to dismiss rarely occurring events. On the other hand, the analysis of replicated and independent simulations can provide estimates of accuracy.

This introduction outlines the protein folding and the intrinsically disordered proteins, both areas described from the point of view of systems with a continuum of

conformational changes. Then, the biological membrane, the voltage-sensor domain and the principles of gating modifiers, especially with focus on spider toxins, are introduced. The construct of systems with a wide range of membrane potential is described in the material and methods section 3.2, because it did not belong to the initial aims of the thesis. It was introduced as a “tool” to elucidate the lipid-mediated mechanism. Finally, in section 1.3, I will discuss the motivations for my research on the mechanisms inducing protein conformational changes and the methodology used to assess a lipid-mediated mechanism of voltage gating modifiers.

1.1 The dynamics of proteins

1.1.1 The protein folding problem

Proteins are chains of 20 different α-L-amino acids covalently linked by an amide bond, formed upon condensation of the carboxyl and amine functional groups. Synthesized by the ribosome, most proteins fold spontaneously into their native state, which is encoded by the sequence of residues, also called the primary structure of the proteins.

Upon folding, two main elements of secondary structures will form: α-helices or β-sheets, both maintained by a typical pattern of intramolecular hydrogen bonds formed between their main-chain carbonyl and amide functional groups. In an α-helix, hydrogen bonds link carbonyls of residues *i* with the amides of residues *i+4* in a right-handed spiral. Other, less common, helices found in proteins are the 3_{10}-helix, where the hydrogen bonds link residues *i* with residues *i+3*, and the π-helix (*i* and *i+5*). In the 3_{10} - and π-helices, the hydrogen bonds are slightly weaker, partly because of a less favorable orientation. They are nevertheless often observed at the level of the terminal of an α-helix or they form short one-turn helices. β-sheets are formed by several strands, in a parallel or antiparallel manner, connected by hydrogen bonds between carbonyl and amide functional groups. The three-dimensional arrangement of secondary structures, connected by loops, forms the tertiary structure of proteins. In addition, the thiol functional groups of two cysteine residues can form a disulfide bridge upon oxidation. This additional covalent bond stabilizes further the tertiary structure. The quaternary structure is formed by the arrangement of several folded proteins in a complex.

Two complementary experiments paved the way of the protein-folding problem. In their pioneering work, Anfinsen and colleagues (*3*) completely denatured a protein, ribonuclease A, by exposing it to urea and a strong reducing agent, 2-mercaptoethanol, which broke the intramolecular disulfide bridges. They observed that, as soon as the denaturing conditions and the reducing agent were removed, the protein could spontaneously refold into a fully functional structure. This experiment highlights one fundamental principle of protein science, namely that the required information for folding to the native structure is encoded in the sequence itself. This experiment also showed that the chemical environment, in addition to the sequence, affects profoundly the behavior of the molecule. The second observation was purely statistical and based on many reasonable assumptions. Focusing on the dihedral angles governing the shape of the polypeptide, Levinthal (*4*) based his reasoning on a minimal number of degrees of freedom: 300 for a protein of 150 residues. Assuming that an accuracy of a tenth of a radian could suffice to reasonably describe the different structures, a single native state represents one out of 10^{300} possible conformations. Citing Levinthal: “We feel that protein folding is speeded and guided by the rapid formation of local interactions which then determine the further folding of the polypeptide”. This view can also be expressed in terms of a free energy landscape that would have a funnel shape, the minima corresponding to the lowest energy conformation, or native state. In this representation, one can easily imagine an ensemble in which some proteins are trapped in one or several local minima of “almost-folded”, yet not native states, and the changing physico-chemical environment around a protein would be described in terms of modulation of the energy surface.

Although all different aspects of protein structures, from secondary to quaternary, need to be explained in order to fully understand the mechanisms of protein folding, the folding nucleation, specifically the effect of a single amino acid substitution on the folding propensities of a small peptide, will be the subject of the peptide folding part of this thesis.

As stated above, a solution may generally contain folded and unfolded protein conformations forming an ensemble in a dynamic equilibrium. This means that a biologically relevant understanding of protein function considers, in addition to a description of the native state, the conformational changes of the molecules and the factors affecting the equilibrium between different states. For this reason, the aim of an investigation of protein folding from sequence is not only the prediction of the most energetically favorable state, or native state, but rather to capture the dynamics of the system. The experimental investigation of all the states occurring during folding is very challenging, especially transition states, because they are visited only transiently and

are sparsely populated. Generally, these very rare conformations have a negligible impact on averaged observables, making them experimentally almost invisible. In support to this description, a work published in April 2016 (*5*) used high-resolution force spectroscopy to specifically investigate transition paths of nucleic acids and proteins. Precisely, the folding-unfolding events of a DNA hairpin and a protein were extracted through measurements of the lengths of single molecules tethered to handle and beads. Using the lifetimes of such length measurements as reaction coordinates, the authors estimated that the transition state conformation of the DNA hairpin would represent roughly 0.001 of a trajectory. Unbiased classical molecular dynamics simulations, despite their limitations in terms of accuracy exposed in section 2.1.1, constitute a tool of choice in this context. There is no technical problem to isolate "rare" frames out of a complete trajectory.

On the other hand, using all-atom molecular dynamics, one can explore the precise interactions between the molecules of interest and the solvent or other molecules, as developed in the sections 2.1 and 2.3.

In a few words, while the beauty of folded proteins can be impressive, the roots of these organized structures are to be found in the unfolded and transition states.

1.1.2 The intrinsically disordered proteins

The biological importance of intrinsically disordered proteins (IDP) and of disordered regions in otherwise fully folded proteins has been recognized only progressively and recently (*6*). During the 20th century, the paradigm sequence-structure-function overshadowed largely the discovery of disordered proteins or disordered segments of proteins. For many years, "not so well" defined proteins, with non-classical conformational features were considered rare exceptions that contradict the paradigm mentioned above. IDPs are mainly found among signaling proteins, transcription and cell cycle control. In these classes of proteins, the conformation versatility increases the recognition possibilities without paying the entropic cost of folding. Often, the IDPs bound to their cognate ligand are found folded. Two views are generally exposed to explain the transition of IDPs from an unfolded to a folded conformation. One states that folding follows binding to the appropriate ligand, and is thus due to the complex formation. The other one was postulated by Pauling in 1940 in the context of antibody-antigen recognition (*7*). This is a selection process within interconverting structures, whereas the cognate antigen would bind to the appropriate

one. However, in other cases, IDPs remain devoid of structure even if bound to the target (*8*, *9*).

Whatever the process(es) underlying the conformational behavior of IDPs, we are far from being able to predict exactly which states will be populated under which conditions. For the same reasons as the ones exposed above for the protein folding study, all-atom molecular dynamics simulations, despite the current computational power limitations, constitute an appropriate tool to explore the principles underlying the conformational behavior of IDPs. Particularly, one would aim, not to identify a single, low energy native structure, but rather to know which physicochemical factors alter the ensemble distribution of conformations. The question of the role of cellular crowding, which has been proposed to increase the folding propensities of IDPs in the cytoplasm, is also relatively difficult to access experimentally, since many methods require the use of diluted solution containing an unique protein species (*10*). This question can also be explored through MD trajectory analyses.

1.2 The bilayer: between toxins and the voltage-sensor domain

1.2.1 The membrane bilayer

Cell and organelle membranes are among the essential structures of biological systems, playing important roles as interfaces between compartments of different contents, chemical or electrostatic potentials. Phospholipids, the primary molecules found in the plasma membrane, are characterized by their amphiphilic structure constituted of a hydrophilic head group and a hydrophobic acyl chain. The consequence of this amphiphilic character and of their rather cylindrical shape is the formation of a stable bilayer, where the head groups face the solvent and the hydrocarbon tails form a hydrophobic barrier. Since the interior of the membrane is highly hydrophobic, water, ions and other hydrophilic molecules are essentially prevented from leaking through, and the result is a chemically isolated environment inside the cytoplasm or in an organelle. This notably allows for the storage of energy in the form of an electrochemical gradient.

Biological membranes are extremely diverse, with various proportions of phosphatidylcholine, phosphatidylserine, phosphatidylethanolamine, cholesterol,

sphingomyelin, glycolipids and others. Three of these lipids, namely phosphatidylserine, phosphatidylcholine and cholesterol are depicted in Figure 1.1. The mass percentage of phosphatidylcholine, for example, varies from 10% in myelin to 40% in endoplasmic reticulum (*11*). In a recent study, Quehenberger et al. (*12*) identified different lipid species in human plasma. They documented more than 500 different lipid molecular species, while an even much higher diversity may be expected in cell membranes. This diversity of phospholipids arises from chemical modification of the acyl chain (length and degree of saturation) or modification of the head groups or glycerol linkage. Alterations of membrane lipid composition are related to pathogenic processes. For example, decreased levels of phosphatidylinositol and phosphatidylethalonamine are implicated in Alzheimer disease (*13*, *14*). A very common feature of Golgi, endosomal and plasma membranes is the asymmetric distribution of phospholipids, where the majority of anionic lipids are found in the cytosolic leaflet. The functional importance of this asymmetry is evidenced by its active maintenance by phospholipid scramblases, ATP-binding cassette transporters and aminophopspholipid translocases (*15*). Membrane asymmetry may enhance mechanical stability and is involved in cell fusion and apoptotic stage recognition (*15-17*). Nevertheless, due to experimental and computational limitations, most studies have generally concentrated on single lipid systems, whereas most of the principles arising from this lipid diversity and membrane asymmetry are still unknown. Current computational power and MD force field accuracy have reached the point where one can progressively construct more realistic membranes, with the aim to capture the consequences of the complexity of the biological systems.

1.2.2 The voltage-sensor domain

The concentration and charge gradients between both sides of a membrane introduced above would be useless if they remained static. Indeed, this stored energy is further used by cells and organelles to do work or for information transfer. This is one of the main functions of membrane proteins, which mediate permeability or transport. Embedded proteins, which represent about 50% of the membrane mass ratio (*11*), have one or more membrane spanning domains, generally α-helical, except for the β-barrel topology found in bacteria and mitochondria. Structurally, the surface of membrane proteins displays hydrophobic side chains in the middle of the bilayer, whereas aromatic residues are generally found at the membrane-water interface. It is thought that they serve as anchor for the appropriate orientation of the protein. This

rule of thumb does not imply that polar residues are completely absent, and indeed, polar residues constitute about 20% of the transmembrane helices (*18*).

Membrane proteins include transporters, receptors, which mediate communication between the cytoplasm and the extracellular compartments, enzymes, and proteins involved in cell junctions, which anchor the extracellular matrix to the intracellular cytoskeleton. As an example of a transporter, the sodium-potassium pump moves three sodium ions out of the cell, while importing two potassium ions. This active process involves the hydrolysis of Adenosine triphosphate (ATP). Thus, a unique mechanism maintains an electrochemical gradient. The potassium concentration is high inside of the cell, whereas sodium accumulates outside of the cell. In addition, calcium and chloride ions tend to be concentrated outside of the cell. Typical values for a neuron at rest are given in Table I. Passive transport of these ions, following their electrochemical gradient, occurs through ion selective channels. Approximately 400 genes encode human transmembrane ion channels, which can be roughly classified as either voltage or ligand gated, depending on the factors determining channel activation. Ligand gated channels are gated by second messengers, light, pressure or stretch, cyclic nucleotides or temperature (*19-21*). Ion channels are vital in the functioning of humans. Their dysfunction is implicated in various diseases, including multiple sclerosis, cardiac arrhythmia, hypertension and chronic pain, and thus they constitute major drug targets (*22-24*).

Voltage-gated potassium channels are tetramers that open and close as a function of the membrane electrostatic potential, and regulate action potential in nerve, muscle, and cardiomyocytes (*25-27*). Each subunit of a tetramer is composed of six transmembrane helices. Four helices (S1-S4) form an anti-parallel helical bundle and constitute the voltage-sensor domain (VSD), which is linked to the pore domain (S5-S6) by a short linker. A much-conserved structural feature of the voltage-sensor domains is that the S4 helix contains four to six basic residues (generally Arg), each followed by two hydrophobic residues. The voltage-sensing properties are attributed to these positively charged residues, which move in response to variation of the membrane potential (Vm). The resulting current is called gating charge transport, or gating current (*28-30*). In 2003, the full length structure of the KvAP channel from *Aeropyrum pernix* was resolved at atomic resolution (*31*).

Isolated voltage-sensor domains are independent functional units, as evidenced by several findings. First, it was discovered that other proteins than voltage-gated ion channels contain voltage-sensor domains. The voltage-sensing phosphatase, for example, contains a VSD, which activates the phosphatase activity upon membrane depolarization (*32*). Moreover, in 2001, Lu et al. (*33*) attached a VSD

to a voltage-insensitive channel. As a result, the channel became responsive to membrane potential fluctuations. Finally, voltage-dependent proton currents were measured in 1982 (*34*), and the relevant channel was sequenced in 2006 (*35*). A particularity of the voltage-gated proton channels is their responsiveness to the trans-membrane pH gradient. Specifically, the trans-membrane pH gradient modifies the gating properties of the VSD. Whereas the onset of outward currents lies at approximately 20 mV in the absence of any trans-membrane pH gradient, the onset is shifted by about -20 mV upon intracellular acidification by half of a pH unit (*35*, *36*) (Figure 1.3B). The next section will describe other instances of gating modifications, which were investigated in this work.

These observations indicating that VSD are independent functional units show that ion channels are composed of clearly separated modules: the pore domain and the voltage-sensor domain. A useful consequence is that investigation of voltage sensing can be performed on a voltage-sensor domain alone, reducing the computational costs. Consequently, a voltage-sensor domain alone was introduced in the simulated systems, instead of a whole tetramer.

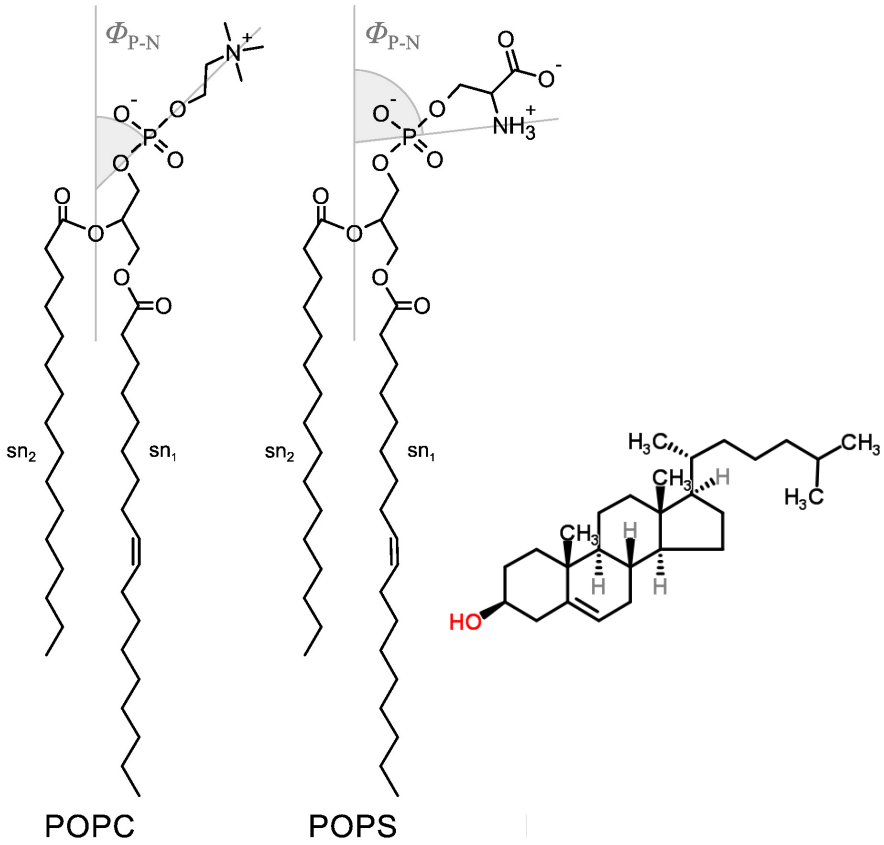


Figure 1.1. Lipids involved in this study
Molecular representations of POPC and POPS, adapted from (*37*) and cholesterol (downloaded from http://www.chemspider.com/Chemical-Structure.5775.html CSID:5775, , 2016). For the phospholipids, the figure also shows the orientation of the P-N vector with respect to the normal.

Table I: Neuron ion concentrations at resting potential, adapted from (*38*).

Ion	[ion] $_{intracellular}$ (mM)	[ion] $_{extracellular}$ (mM)	Vm_{ion} (mV)
K^+	96	4	-85
Na^+	10	145	+71
Ca^{2+}	0.070	2	+137
Cl^-	7	145	-80
pH[a]	7.2	7.4	-13

[a] For the hydronium ion, the pH values are given, instead of the concentration.

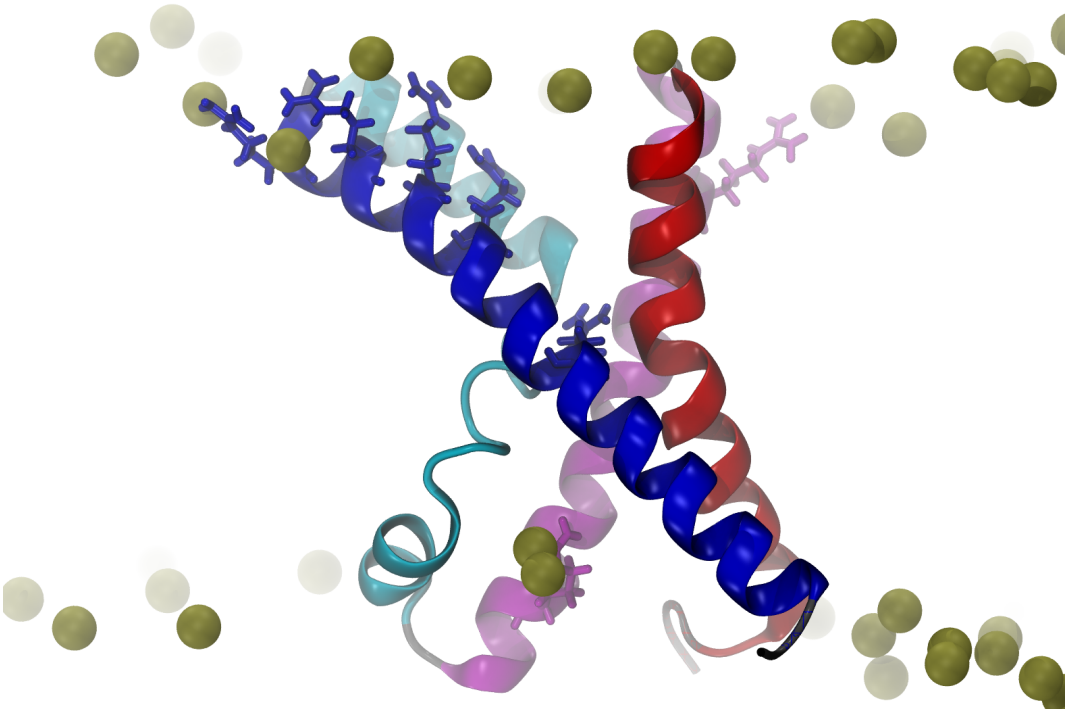


Figure 1.2. KvAP voltage-sensor domain highlighting the four helices and the Arg residues.
The protein is shown in cartoon representation, the Arg residues are shown as licorice colored according to the helix they belong to: S1: red, S2: magenta, S3a and S3b: cyan, S4: blue.

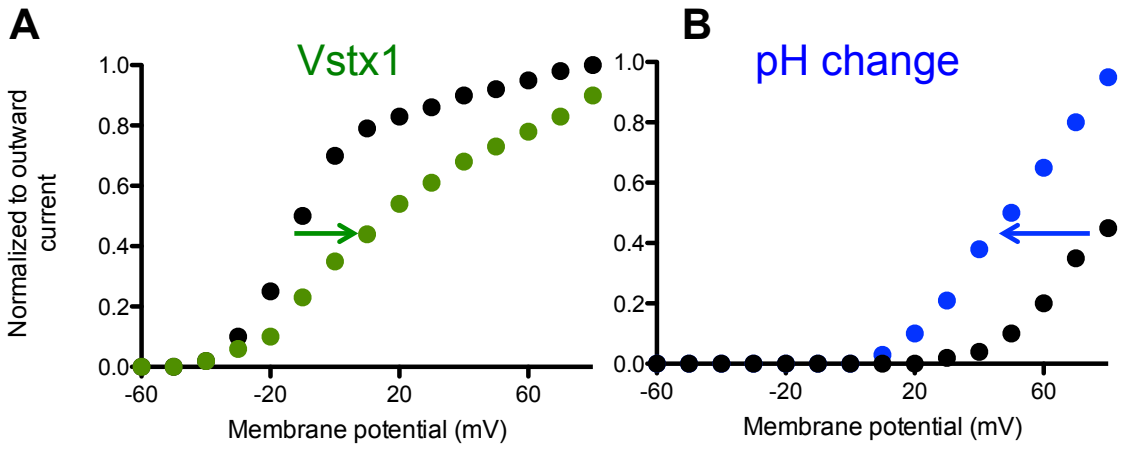


Figure 1.3. Voltage gating modulation.
A) Upon interaction with 4 μM Vstx1 (green), the voltage-current response curve of the Kv2.1/KvAP chimera is shifted to the right (data taken from (*39*)). B) In a voltage-gated proton channel, a transmembrane pH gradient change from -0.5 (black) to 0.0 (blue) induces a left shift of the voltage-current response curve (data taken from (*35*)).

1.2.3 The spider toxins

Animal toxins have been fine-tuned through millions of years of evolution for specific purposes, like immobilization of a prey. For this task, most of them affect ion channel functions. The target and functional diversity of animal toxins led several authors to advocate a likely therapeutic usage of these molecules (*40*, *41*). Some successes have already been achieved. For example, as a consequence of its high specificity for particular potassium channels, a slightly engineered version of a sea anemone toxin, ShK, has entered Phase Ia of clinical trials in 2011 (*42*). Another toxin, MVIIA, produced by a cone snail species, was approved in 2004 by the U. S. Food and Drug Administration for the treatment of chronic pain. On the other hand, animal toxins have been invaluable tools in the study of ion channels. Investigation of scorpion and spider toxins led to the discovery of ion channel subunit stoichiometry (*43*) or to the identification of diverse receptor sites in ion channels (*44*). However, a precise understanding of their mechanisms of action on ion channels is still lacking.

Spiders are, with more than 42,000 described species, among the most successful groups of animals. The 400 million years of improvements in term of venom diversity and specificity have contributed to this success. However, only about 100 species have been studied for their venom (*45*). The function of venom spans from killing to paralyzing a prey, or is directed against aggressors. Venoms can be sorted in two broad categories: necrotic or neurotoxic. Necrotic substances contain phospholipases, proteases and lytic factors, which induce tissue necrosis or hemolytic effects. Neurotoxic peptides are typically fast acting. They target nerve tissue and neurotransmitters, either through degradation of neurotransmitters or by interfering directly with voltage-gated ion channels (VGIC). Most neurotoxins affect potassium or sodium channels, and they can be, again, broadly sorted into two groups. First, channel blockers, which bind to the outer part of the conduction pore and stop ion flow (*46*). Second, gating modifiers, for which the mechanism is not well understood. Similarly to the pH induced gating of the proton channel, as introduced in the section 1.2.2, the gating modification outcome is the channel requirement of either a higher or a smaller depolarization to open, depending on the specific channel-toxin interaction pair. If the channel requires a higher depolarization upon interaction with the toxin, the voltage-current response curve of the channel is shifted to the right (Figure 1.3A).

A common feature of most spider toxin structures is the folding into a so-called inhibitor cystine knot (ICK), characterized by two or three beta-strands, and in which three disulfide bridges are arranged in order to form a “knot”, where a ring formed by

two disulfide bonds is penetrated by a third disulfide bond (Figure 1.4). This structure enhances significantly the stability of these 30-40 AA long peptides. Another feature of several ICK spider toxins resides in a hydrophobic prominent cluster, mostly with four to seven aromatic residues, surrounded by polar and charged residues. This arrangement leads to a precise insertion depth and orientation within the membrane, in which the hydrophobic cluster lies at the level of the hydrocarbon chains and the polar and charged residues interact with the lipid head groups (*47*). Another conserved characteristic of the ICK toxins is a global positive charge (+2 to +4). The chemical diversity of spider venoms is impressive: several hundreds of different ICK peptides were isolated from a single venom (*45*). As of March 2016, for the spider species Chilian rose tarantula (*Grammostola rosea*), 60 toxins are deposited in the ArachnoServer database (*48*), of which 37 contain three disulfide bridges, and thus are expected to fold into an ICK. In this list, four have bona fide deposited PDB structures: GsMTx4, Hanatoxin-1, Vstx1, and ω-grammotoxin SIA. These four sequences contain a large proportion of aromatic amino acids and carry a positive charge between +2 and +4 (Table II) and all peptides form an ICK, with a hydrophobic cluster surrounded by polar residues (Figure 1.4).

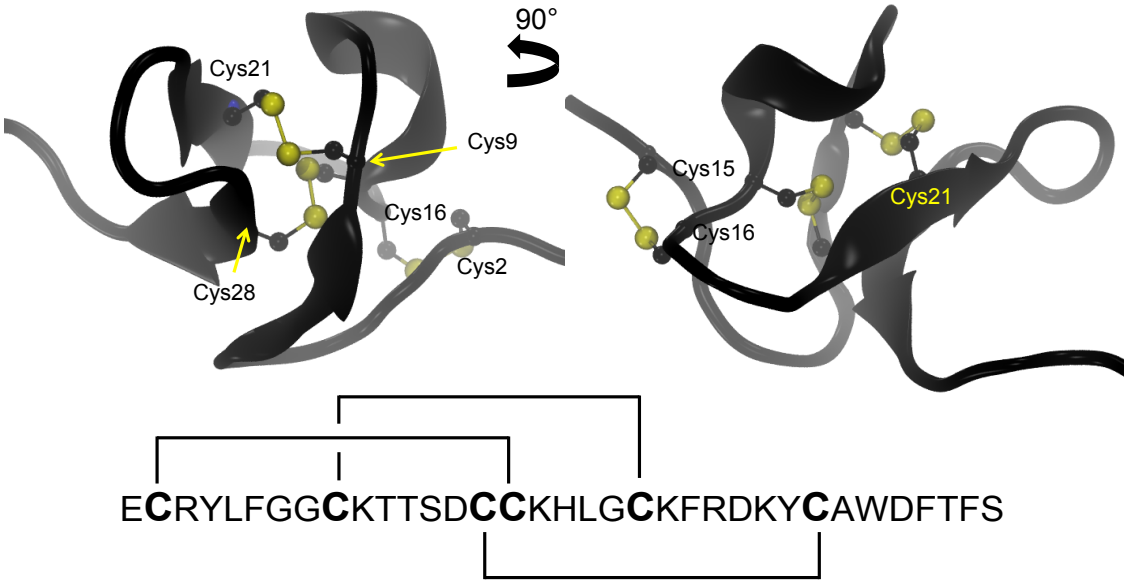


Figure 1.4. The ICK motif exemplified in the case of Hanatoxin.
Above: two molecular representations of the toxin are shown in ribbons, with the Cys residues labeled and the S atoms colored in yellow. Below: the lines show the arrangement of disulfide bonds resulting in a knot motif.

Table II: Structurally solved inhibitory cystin knot toxins from Chilian rose Tarantula. Basic, acidic and aromatic residues are shown in blue, red, and green.

Toxin	Sequence	Charge
GsMTx4	GCLEFWWKCNPNDDKCCRPKLKCSKLFKLCNFSF	4
ω-grSIA	DCVRFWGKCSQTSDCCPHLACKSKWPRNICVWDGSV	2
Hanatoxin-1	ECRYLFGGCKTTSDCCKHLGCKFRDKYCAWDFTFS	2
Vstx1	ECGKFMWKCKNSNDCCKDLVCSSRWKWCVLASPF	3

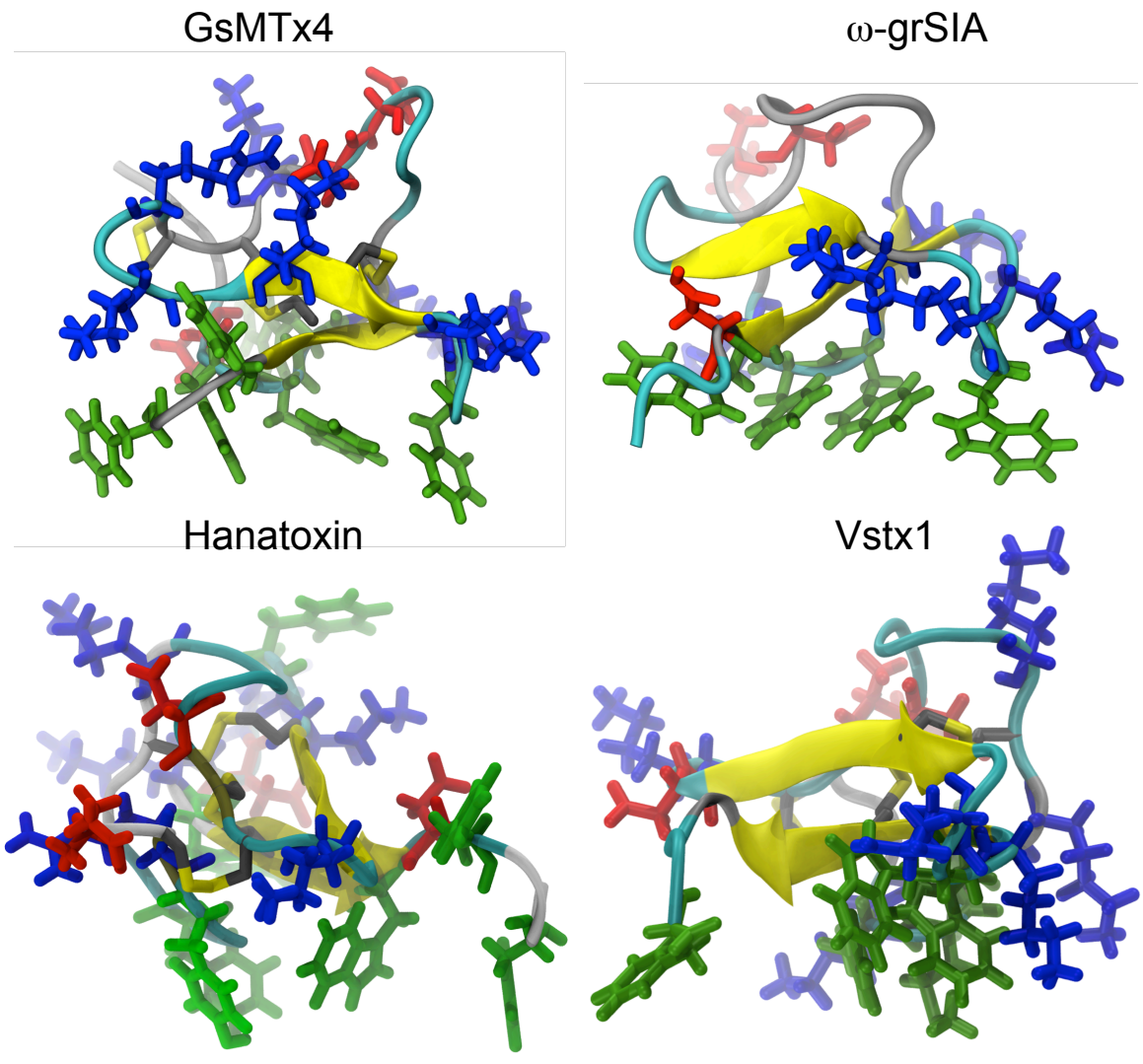


Figure 1.5. Molecular representations of toxins highlighting the hydrophobic cluster and the charged residues.
GsMTx4, ω-grSIA, Hanatoxin and Vstx1. are shown with beta-strands in yellow, turns in cyan, loops in silver, acidic residues in red, basic residues in blue, hydrophobic residues in green and disulfide bridges in yellow.

1.2.4 A lipid hypothesis of voltage-gating modifiers

Several previous experiments demonstrated that charged molecules affect specific features of a lipid bilayer, particularly the orientation of the head group dipoles (*49*, *50*). It was also shown that the insertion of electric dipoles at the membrane-water interface suffices to reorient the lipid head groups (*51*). These works showed that several different types of molecules, like metals, local anesthetics in the charged form, salts, or charged amphiphiles had similar effects. In the presence of positively charged species located at the level of the (negatively charged) phosphate groups, the positively choline groups are repelled toward the water phase, reducing the value of the angle formed by the phosphocholine segments and the membrane normal. The hypothesis exposes that, since the phospholipids themselves have a dipole moment, their reorientation should consequently affect the membrane surface potential. Seelig et al (*50*) calculated that a reorientation of all the phosphocholine groups in a membrane by about 20° would modify the overall membrane surface potential by ≈ 90 mV, which is obviously large enough to modify the gating of a voltage-sensor domain (to be compared to Figure 1.3). Since gating modifiers are amphiphilic charged molecules (section 1.2.3), this hypothesis could explain, in principle, their effect on voltage-sensor domains.

On the other hand, the conformational changes of any transmembrane proteins require work to be performed within the membrane. Consequently, a lateral pressure increase, due to the insertion of several amphiphilic molecules in the bilayer surrounding a target protein, could favor a given conformation. Applying this idea to the case of voltage-gated ion channels, gating modifiers could affect the opening/closing dynamics through modification of the mechanical properties of the bilayer.

In the same line of arguments, there is increasing evidence that the structures or functions of membrane proteins can be modulated by the lipid environment (*52*) and this has notably been shown for several members of the ion channel family (*53*). This means that the lipid environment *per se* plays an important role in the structural conformations and the dynamics of voltage-gated ion channels. Enzymatic removal of the phospholipid head groups was shown to restrain movements of the embedded Kv2.1 channel, making the channel totally irresponsive to Vm changes. Reconstitution of KvAP channels in artificial membranes exhibiting various types of lipid head groups showed that the proper functioning of the channel requires phosphodiester (*54*, *55*).

Such experiments suggest that the interactions between the basic residues of the VSD and the head groups play essential roles in voltage-dependent ion channels.

On the other hand, several spider toxins were shown experimentally to interact strongly with the membrane, opening the way to the idea of a membrane-access mechanism of gating modification (*56*).

All these observations together lead to ask the question whether ICK spider toxins, which, like the salts and metals investigated by Seelig and colleagues, carry an electric charge and interact strongly with the membrane, may consequently affect the gating of target channels through a lipid-mediated mechanism.

1.3 Motivations of my dissertation

When I started working with the peptides of sequence EGAAXAASS, all that was known about them, was a hypothesized tendency to form a kink when a Trp or Tyr residue was substituted at position X in the sequence. The other 12 experimentally tested substitutions were thought to produce rather extended peptides (*57*). I elucidated the conformational propensities of four of these short peptides and the results are described in section 2.1 of this thesis. The simulations *per se* did not reveal the conformations of the aromatics containing peptides, because they never converged to a stable structure, and because the ensemble average did not reproduce adequately the experimental data. However, the comparison between predicted and experimental values combined to statistical tools demonstrated that the substitution of X by Trp or Tyr induced the folding of the short peptide into a turn or a one-turn helix. A significant amount of time of the thesis was dedicated to ascertain this conclusion. A second point of interest was the exploration of the underlying mechanism. Why do bulky side chains increase the folding propensities of this peptide? The computation of the hydration of specific atomic groups in extended, disordered structures, provided an answer to this question and the results are also exposed in section 2.1. Although the importance of hydration in protein folding has been recognized for a long time (*58*), its exact function remains unresolved. The question whether hydration levels fluctuate before or after folding (*59*, *60*) is still under debate. In section 2.3, a statistical method to test the order of these events is introduced. First, the folding of the fully extended peptide into its native β–hairpin structure within 600 ns is validated. On this basis, a cross-correlation function analysis

is then performed between folding events and hydration level fluctuations. The results suggest that the hydration fluctuations occur before the protein conformational changes. All these results lead to the proposal, in section 2.4, of a large-scale project, aimed at a better understanding of the mechanisms underlying intrinsically disordered protein conformational changes and the prediction of structural ensembles from sequence.

The initial purpose of the second project was to investigate the interactions between charged molecules known to affect the gating of ion channels and a lipid bilayer. To be more precise, the question whether these molecules would modify the gating indirectly, without binding to the target (as explained more in details in section 1.2.4), but through a lipid-mediated mechanism, had to be elucidated. This lipid-mediated effect could take the form of a modification of the global properties of the membrane around the toxin, like the acyl chain ordering, the orientation of the head groups or the membrane thickness, and these modified properties would affect the functioning of the VSD. This indirect effect could also affect locally the VSD interaction pattern required by the voltage-sensor for proper functioning. Several spider toxins are known to modify the gating of ion channels, so that we decided to study members of this family of peptides. The principle of voltage-gating modification is introduced in Figure 1.3, and the spider toxins in section 1.2.3. However, the following rationales have led to extend the number of variables and replications. First, to increase the level of confidence, a voltage-sensor domain was added to the systems. Theoretically, one could imagine that a modification of the membrane, even if it is clearly documented in a system without any membrane protein, does not finally affect significantly the mechanism of a VSD *per se*. The addition of a voltage-sensor domain imposes strong constraints, since one must observe membrane structural changes and additionally their correlated effect on the VSD. On the other hand, the problem of the specificity of spider toxins is exposed in section 3.1. Concisely, there is experimental evidence that some toxins are effective against a given ion channel, while other toxins are not. Consequently, two different toxins were chosen, which have different known experimental outcomes when tested on the specific VSD, one being active and the other not. The toxin known to be ineffective thus serves as a negative control. The constraints become then stronger: the induced perturbation of the membrane due to a toxin must lead to measurable effects on the VSD with one of the two toxins, but not with the other one.

After having first validated the mode of insertion of the toxins in light of experimental results (section 3.3.1.1), the following sections show that both toxins modified notably several structural features of the membrane. However, the similarity

between their effects on the membrane and the absence of any correlated effect on the VSD do not support an indirect mechanism of action of gating modifiers.

Yet, a lucky choice, leading to unexpected discoveries, was to perform a large number of independent simulations in which the systems were exposed to a wide spectrum of membrane potentials. In section 3.3.2.2, it is not only shown that the VSD affects the membrane, but that this VSD induced modification depends on the sign of the membrane potential. Finally, in section 3.3.2.4, a kink in the S4 helix of the VSD is described for the first time as a response to membrane polarization. It is further shown that some hints of this kink in the S4 helix have been mentioned in previous experimental works. The novelty here is to link this conformational change with the membrane polarization.

A common methodological theme of both projects is the effort to use an approach in which several independent simulations of similar systems are analyzed with the aim of identifying features linked to reproducible observations. It is hoped that with the increasing computing capacities, it will become possible to progressively conduct the MD simulations more “test-tube”-like, extracting averages or other descriptors out of a representative molecular ensemble. The modest attempt performed here in this direction, with a little bit more than 100 replicated systems in the second project, led to the identification of unexpected relationships between the membrane potential, the bilayer and the voltage-sensor domain.

2 Driving force of peptide folding nucleation

2.1 Backbone hydration determines the folding signature of amino acid residues

2.1.1 Introduction

In the course of evolution, proteins have become a very versatile class of biomolecules, playing a central role in biological processes. The properties of proteins are related to their three-dimensional structure. Protein conformational changes are involved in biological function, and defects in folding are associated to severe disorders, like Creutzfeld-Jacob disease, type II diabetes, Alzheimer's, Parkinson's and Huntington's diseases (*61*, *62*). Several investigations on the relation between protein sequence and their conformational tendencies have been developed (*6*, *63*, *64*). Nevertheless, the fundamental principles that underlie the folding of proteins remain poorly understood (*3*, *65*, *66*). The detailed mechanism driving the protein folding process is unknown, and notably its dependency on amino acid side chains (*62*).

As of May 2016, more than 118,000 protein structures are available in the RSCB Protein Data Bank (PDB) (http://www.rcsb.org/) (*67*). While our understanding of the folded state of ordered proteins has largely increased, the structural propensities of intrinsically disordered proteins, introduced in the section 1.1.2, remain largely unknown. Intrinsically disordered proteins are fully functional, and play indispensable biological roles, despite lacking a stable three-dimensional structure (*10*). Folding is not a permanent condition. In solution, even folded proteins are in a dynamic equilibrium with unfolded or partially unfolded conformations. This equilibrium depends on a delicate balance of weak, non-covalent and competing interactions involving the peptide chain as well as the surrounding molecules, solvent and ligands. In the case of unbound intrinsically disordered proteins, this balance disfavors a single folded state. However, in order to understand the relation between the sequence and this equilibrium, a better understanding of the mechanisms by which individual amino acid side chains impact the conformational dynamics of the protein is particularly required (*62*, *68*).

Experimental and computational techniques provide high-resolution three-dimensional structure of folded proteins. Protein crystallography produces atomic resolution structures, even for large systems. Nuclear Magnetic Resonance (NMR)

spectroscopy, which investigates the molecules in solution, is not restricted to folded proteins but provides access to information about dynamical molecules, intrinsically disordered proteins included. Cryo-electron microscopy requires relatively small amounts of material, and has recently proven to be very accurate, with a structure solved at 2.2 Å (*69*). Homology modeling, or template-based modeling, uses an experimentally determined structure as template. In addition, the rational assumes that evolutionary related proteins will fold into similar structures, if they also share similar amino acid sequences. All-atom molecular dynamics (MD) simulations complement these methods because they provide the most detailed computational description of peptide structural dynamics with high spatial and temporal resolution. Recent works show that simulations of 500 ns or more can trace the complete folding of fast-folding proteins up to 80 amino acids long (*70*) or reproduce NMR parameters of folded proteins, like residual dipolar couplings (RDC) (*61*). Despite these successes, it is important to remember that MD simulations have limited accuracy, which is due, for example, to the approximations needed to perform classical physics computations at the atomic level. For this reason, making longer trajectories may not necessarily provide more accurate results. As stated above, other methods, experimental and even computational, perform better in the identification of the native state of folded proteins. One of the domains in which MD fully complete these tools is in the investigation of conformational changes and interactions between the molecules of interest and their chemical and physical environment. In this context, analysis of MD trajectories can provide new insights in the understanding of the mechanisms at play. Particularly, intrinsically disordered proteins, because they populate a large number of conformations, with fluctuations involving the formation and release of local secondary structures, cannot be described by a single, low energy state, but rather by an heterogeneous conformational ensemble (*63*). Also, in order to decipher the relevant mechanisms and driving forces of protein folding, an investigation of conformations situated at or close to a folding transition state is needed (*71*).

2.1.2 Methods

In order to investigate the role of individual amino acids on the conformational propensities of a peptide, peptides of sequence EGAAXAASS were investigated in explicit solvent, where residue X was mutated with residues Gly, Ile, Tyr or Trp and the trajectories were validated through comparison with data from NMR experiments. As stated above, NMR is adapted for the study of unstructured proteins in solution, because even for disordered states, resonance is still measurable. While most NMR derived constraints provide short-range information, residual dipolar couplings (RDCs), which result from partial alignment of molecules with respect to the magnetic field, provide a quantitative, long range information about a dipole orientation with respect to the magnetic field. The RDCs of 14 peptides of the sequence mentioned above were previously published (*57*). This particular sequence was designed with hydrophilic ends to ensure solubility, with a neutral N terminus to avoid strong Coulomb interactions between the termini, and a neutral environment around residue X provided by nonpolar residues. Whereas most substituted peptides produced a relatively flat $^{1}D_{C\alpha H\alpha}$ and $^{1}D_{NH}$ RDC pattern, the substitutions with the aromatic amino acids Tyr and Trp resulted in a considerably contrasted pattern, suggesting the presence of a kink in the structure (*57*). The analysis of MD data was performed through systematic comparison between simulated and experimental parameters: RDCs, but also secondary chemical shifts, $^{3}j_{HN\text{-}HA}$, $^{3}j_{HA\text{-}N}$ couplings and the χ1 dihedral angle of Trp as extracted from $J_{HN\text{-}CG}$ and $J_{CO\text{-}CG}$ values. The most informative parameters proved to be the RDCs. However, as mentioned in section 1.3, the averaged values of the predicted RDCs from peptides with an aromatic residue at position X did not reproduce exactly the experimental pattern, and the fluctuations did not show any hint of convergence. Therefore, a principal component analysis (PCA) involving the radius of gyration, the distances between the termini, the coulomb and Lennard-Jones interactions and the total number of hydrogen bonds within the peptide and between the peptide and the solvent was performed using the R environment. This analysis highlighted the number of intramolecular hydrogen bond as a key parameter to describe the most populated conformations reproducing the experimental RDC pattern. Consequently, the multifactorial linear regression analyses performed after this PCA addressed specifically the intramolecular hydrogen bonds.

2.1.3 Results

The main achievements of these investigations are summarized in the publication reproduced in the section 2.1.5. Briefly, it was found that the substitution of residue X with a Tyr or Trp significantly increased the folding propensities of these peptides into a turn or a one-turn α-helix. Additionally, the peptides with the Gly, Ile, Tyr, and Trp substitutions could be sorted as a function of their folding propensity: the peptide with Trp at position X being the most folded. This finding is in line with statistical studies of IDPs compared to ordered globular proteins and proteins found in the Swiss-Prot database (*6*, *10*, *72*, *73*). This comparison showed that folded proteins are particularly enriched in specific amino acids, and the four above-mentioned amino acids were sorted in the same order as in the MD trajectory analyses. Other bioinformatics investigations led authors to propose to call Trp, Cys, Phe, Ile, Tyr, Val, and Leu “order-promoting” amino acids. Interestingly, in both bioinformatics studies as well as in the simulations, Trp is expected to have the strongest folding promoting effect.

On the other hand, the investigations showed that the lack of hydration of the carbonyl and amide groups on either side of the bulky hydrophobic side chain was a key driving force increasing the folding propensity of peptides containing aromatic residues arises. These observations imply that the well-known fundamental function of the solvent, in terms of folding rate or stabilization of the folded conformations, may exert its effect in the immediate proximity of a single residue. These local interactions help reduce the size of conformational search space, thus speeding up protein folding.

2.1.4 Statement of contribution

I performed the simulations and designed the statistical approaches. I performed all the data analyses. The minimization and clustering were performed also, in addition, by Dr. Hoi Tik Alvin Leung.

I wrote a complete draft of the manuscript.

2.1.5 Original publication

Olivier Bignucolo, H.T. Alvin Leung, Stephan Grzesiek and Simon Bernèche

Backbone hydration determines the folding signature of amino acid residues

Journal of the American Chemical Society 2015 vol. 137 (13) pp. 4300-4303

JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Communication

pubs.acs.org/JACS

Backbone Hydration Determines the Folding Signature of Amino Acid Residues

Olivier Bignucolo,[†,‡] Hoi Tik Alvin Leung,[‡] Stephan Grzesiek,[‡] and Simon Bernèche*[,†,‡]

[†]SIB Swiss Institute of Bioinformatics and [‡]Biozentrum, University of Basel, Klingelbergstrasse 50/70, CH-4056 Basel, Switzerland

Supporting Information

ABSTRACT: The relation between the sequence of a protein and its three-dimensional structure remains largely unknown. A lasting dream is to elucidate the side-chain-dependent driving forces that govern the folding process. Different structural data suggest that aromatic amino acids play a particular role in the stabilization of protein structures. To better understand the underlying mechanism, we studied peptides of the sequence EGAAXAASS (X = Gly, Ile, Tyr, Trp) through comparison of molecular dynamics (MD) trajectories and NMR residual dipolar coupling (RDC) measurements. The RDC data for aromatic substitutions provide evidence for a kink in the peptide backbone. Analysis of the MD simulations shows that the formation of internal hydrogen bonds underlying a helical turn is key to reproduce the experimental RDC values. The simulations further reveal that the driving force leading to such helical-turn conformations arises from the lack of hydration of the peptide chain on either side of the bulky aromatic side chain, which can potentially act as a nucleation point initiating the folding process.

The prediction of the structural properties of a protein from its amino acid sequence remains a major challenge.[1−3] The detailed mechanism driving the protein folding process is unknown, and specifically its dependence on amino acid side chains.[4] The functional importance of intrinsically disordered proteins has stimulated investigation of the relation between their sequence and their conformational tendency.[5,6] In order to improve predictions about the structure of proteins (folded or disordered), a better understanding of the mechanisms by which individual amino acid side chains impact the conformational dynamics of the protein is required.[4,7]

NMR spectroscopy is particularly appropriate to investigate the structural dynamics of peptides in disordered and folded states.[8] Particularly, residual dipolar couplings (RDCs), which arise when molecules are dissolved in anisotropic liquid phases,[9] provide local as well as long-range quantitative structural information on individual chemical bonds.[10−12] The RDC between two nuclei is proportional to the ensemble and time average $\langle(3\cos^2\theta - 1)/2\rangle$, where θ is the instantaneous angle between the internuclear vector and the magnetic field.

In order to investigate the role of individual amino acids on the conformational propensities of a peptide, Dames et al.[13] engineered a series of 14 peptides of sequence EGAAXAASS. The hydrophilic ends ensured solubility, while the nonpolar adjacent residues provided a neutral environment for the systematically single-mutated residue X. They recorded $^1D_{C\alpha H\alpha}$ and $^1D_{NH}$ RDCs of these peptides, performing the alignment measurement with polyacrylamide gels.[14] Most peptides (with X = G, I, V, L, N, Q, T, D, E, or K) produced a relatively flat pattern consistent with a rather extended average conformation with little specific local structure. However, the substitutions with the aromatic amino acids Tyr and Trp resulted in a strong reduction of the RDCs or even changes in their signs at the center of the peptide (black lines in Figure S1 in the Supporting Information (SI)), suggesting the presence of a kink at this position.

All-atom molecular dynamics (MD) simulations provide the most detailed description of peptide structural dynamics with high spatial and temporal resolution. The NMR data can be used to validate the MD simulation trajectories, which can potentially reveal the mechanistic specificity of aromatic amino acids. Here we show through a systematic comparison between simulated and experimental RDCs that the conformations that best reproduce the experimental data correspond to dynamical ensembles of short helices or turns stabilized by backbone hydrogen bonds. We find that one key driving force that increases the folding propensity of peptides containing aromatic residues arises from the lack of hydration of the carbonyl and amide groups on either side of the bulky hydrophobic side chain.

We performed MD simulations in explicit solvent to reproduce previously measured residual dipolar couplings and chemical shifts of peptides of sequence EGAAXAASS.[13] In order to produce adequate sampling, we carried out 7−12 replicated simulations per investigated peptide, each lasting 100 ns. We calculated the $^1D_{C\alpha H\alpha}$ and $^1D_{NH}$ RDCs as well as the ^{1}HN chemical shifts from the coordinates (see Methods in the SI). Figure S1 compares the $^1D_{C\alpha H\alpha}$ and $^1D_{NH}$ RDCs, averaged over all of the replicated simulations, and the experimental values published previously. The predicted RDC patterns of the peptide with X = Gly or Ile are relatively flat, accurately reproducing the experimental values, whereas the profiles obtained for X = Tyr or Trp only partially show the RDC variations along the peptide sequence that are observed experimentally. Both the $^1D_{C\alpha H\alpha}$ and $^1D_{NH}$ RDCs of the peptides with X = Tyr and Trp fluctuate a lot between the replicated simulations as well as within a given simulation, reflecting the fact that the peptides adopt an ensemble of conformations.

Received: January 20, 2015
Published: March 20, 2015

DOI: 10.1021/jacs.5b00660
J. Am. Chem. Soc. 2015, 137, 4300−4303

We took advantage of these conformational variations to investigate the relations between structural order parameters and the RDCs. Time series analysis over 700 ns (seven 100 ns simulations) showed that the conformations of the X = Trp peptide that best reproduced the experimental RDC profile are characterized by a rather compact conformation, as evidenced by structural parameters such as the radius of gyration and the number of intramolecular hydrogen bonds (Figures 1 and S2).

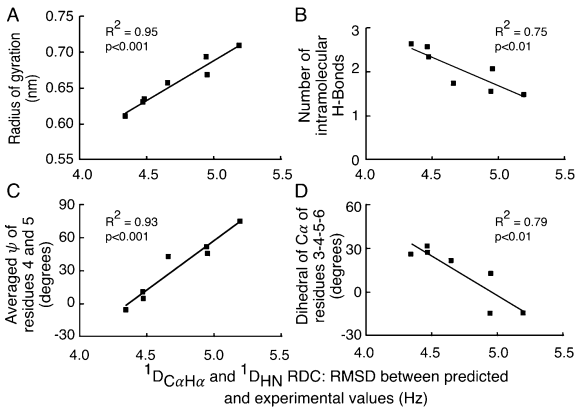


Figure 1. Correlation between different structural parameters and the root-mean-square deviation (RMSD) with respect to experimental RDCs with X = Trp. The RMSD is correlated to (A) the radius of gyration, (B) the formation of intramolecular hydrogen bonds, (C) the ψ angle of residues 4 and 5, and (D) the pseudodihedral angle formed by the Cα atoms of residues 3−6. Each point represents the average over one simulation of 100 ns.

We performed a stepwise regression analysis (see Methods) to identify which hydrogen bonds are statistically relevant. The analysis showed that the conformations that better reproduce the experimental RDC profile involve hydrogen bonds typical of a helix or a turn in the middle of the chain. Precisely, the structurally relevant atomic pairs involve hydrogen bonds between backbone carbonyls of Ala3 or Ala4 and amide groups of residue Ala6, Ala7, or Ser8 (Table S1 in the SI). Analysis of the backbone dihedral angles showed that residues 4 and 5 mainly occupy conformations around $\psi = 150°$ or $-30°$, the latter corresponding to a turn or an α-helical conformation. Clustering of the conformations on the basis of the (ϕ, ψ) dihedral angles revealed that about one-third of all simulation frames contain a short α-helix centered on the X = Trp residue and form the main cluster (Figure S3). The RDC values of this cluster reproduce the characteristics of the experimental profile, as illustrated in Figure S1. The next three clusters account together for about one-third of all conformations and contain β-turns, notably of types I and VIII. Reweighting of the individual conformations to reproduce the RDC and additional J-coupling data[13] while maximizing the entropy (see Methods) confirmed that the first cluster is the most prominent one (Figure S4). Only small readjustments of the cluster population (maximal change per cluster is 4% of the total population) were necessary to reproduce the experimental data within their experimental error, suggesting the overall good accuracy of the MD sampling.

Taken together, these results are consistent with the idea that the conformations of the X = Trp peptide that best reproduce the experimental RDC profile correspond to a turn or a short helix toward the middle of the chain. An analogous analysis of the peptide with X = Tyr showed that this substitution also favors similar conformations, although to a slightly lesser extent than X = Trp (Table S2 and Figures S5 and S6).

We clustered the simulated conformations of the four peptides according to their numbers of intramolecular backbone hydrogen bonds. The left panels of Figure 2 show that the

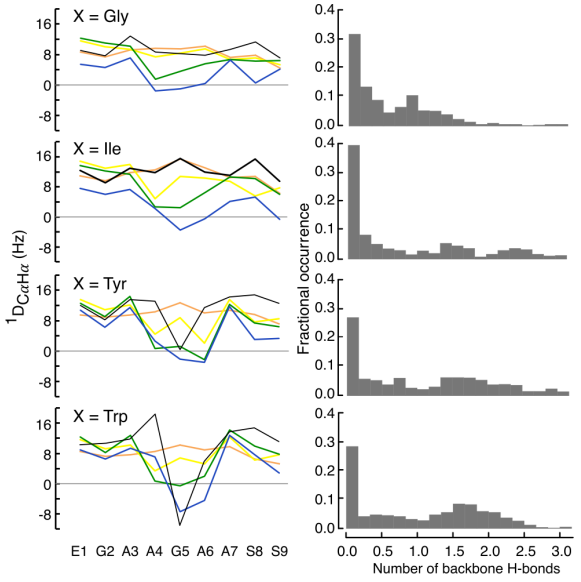


Figure 2. The RDC pattern depends on intramolecular backbone hydrogen bonds. The left panels show for the different peptides the experimental $^1D_{C\alpha H\alpha}$ RDCs (black) and the predicted RDCs for clusters of conformations containing different numbers of backbone hydrogen bonds: 0 (orange), 1 (yellow), 2 (green), 3 (blue). On the right is shown the conformational probability distribution for each peptide as a function of the number of backbone hydrogen bonds (averaged over intervals of 1 ns).

number of hydrogen bonds determines the peptide's RDC profile. Although the averaged values obtained for the peptides with X = Gly or Ile produce rather flat RDC profiles, corresponding to extended peptides with little local structural preference, the conformations that have two or three hydrogen bonds reproduce the distinct dips in the center of the $^1D_{C\alpha H\alpha}$ and $^1D_{NH}$ RDC patterns. On the other hand, the conformations of peptides with X = Tyr or Trp that have only one or no hydrogen bonds produce flat RDC patterns. These different conformations are dynamically visited by the peptides, and thus it is important to consider their probability distribution. On the right side of Figure 2 are shown histograms of the numbers of backbone hydrogen bonds averaged over intervals of 1 ns. The peptides with X = Tyr or Trp both present a maximum around 1.8 bonds (Hartigan's dip test for unimodality, p(Tyr) < 0.01 and p(Trp) < 0.001). The X = Ile peptide also shows some conformations containing more than one hydrogen bond, but with a lower probability than for X = Tyr and Trp. The peptide with X = Gly adopts a non-negligible number of conformations with one hydrogen bond, but the probability of observing more bonds is small. These observations echo the facts that hydrogen bonds become more stable and backbone dihedral fluctuations decrease along the following order of substitutions: Gly → Ile → Tyr → Trp (Figures S7 and S8). The peptides containing

DOI: 10.1021/jacs.5b00660
J. Am. Chem. Soc. 2015, 137, 4300−4303

an aromatic residue adopt significantly fewer conformations with no hydrogen bond in comparison with the Gly and Ile peptides. Similar conformations are visited by all of the peptides, but there is a higher probability of observing more than one hydrogen bond in the peptides containing an aromatic residue and, to some extent, an isoleucine.

The all-atom MD simulations provide a detailed description of the structural dynamics of the peptide and thus allow us to investigate the mechanism by which aromatic residues initiate the folding process. We postulate that their bulky side chains limit the access of water molecules to nearby carbonyl and amide groups. As a consequence, in line with the general understanding of the formation of secondary structure elements,[15] it would be energetically preferable for these backbone functional groups to interact with each other, forming intramolecular hydrogen bonds and favoring peptide folding. To test this hypothesis, we computed the number of water molecules coordinating the carbonyl and amide groups of residues in the middle of the chain. To show that the observed difference in hydration is a potential driving force and not only a consequence of a folded conformation, we compared folded and extended conformations of the Tyr or Trp peptides to conformations of the Gly peptide, which is generally extended. For both the Tyr and Trp peptides, the folded and extended pools respectively contain the conformations corresponding to the 20% lowest and 20% highest RMSDs with respect to the experimental RDC values. Figure 3 shows that the backbone of the X = Trp peptide is significantly less hydrated than that of the peptide with X = Gly, even in its fully extended conformations. The strongest effect is observed for the amide groups of residues 5 and 6. Similar results were obtained for X = Tyr, but the dehydration of the backbone polar groups is significantly less.

Proteins undergoing folding and intrinsically disordered proteins are typically characterized by dynamical ensembles of conformations, with fluctuations involving the formation and release of local secondary structures.[16] Comparison of sequences of intrinsically disordered proteins with natively folded ones showed that disordered regions are generally depleted of specific residues, which were termed "order-promoting amino acids".[6] These include, in decreasing order, Trp, Tyr, and Phe followed by Ile, Leu, and Asn.[17] The bulkiness of the side chains has been proposed to have a direct impact on the local conformation and dynamics of natively unfolded proteins.[18] Our calculations more specifically suggest that bulkier side chains, notably aromatic ones, impede the hydration of neighboring carbonyl and amide groups, favoring the formation of backbone hydrogen bonds and peptide folding. Such elementary folding events are likely to act as nucleation points initiating the folding process and leading to the formation of protein secondary structure elements, without excluding that coalescence of neighboring chains might be essential to stabilize them.[19] The long-standing view that the interaction between backbone functional groups is favored by the formation of hydrophobic pockets[15,20,21] and shielding from solvent[22−25] is thus shown to hold at the scale of a single amino acid side chain. Cooperativity between adjacent side chains is expected to play a key role in defining the level of backbone hydration, suggesting that the processes involved in protein folding are even more local than previously thought.[26] This further reveals that the effective, or biased,[27] conformational search space can involve as little as a few tens of atoms per nucleation point, in line with the mechanism hypothesized by Levinthal in the 1960s as a way to circumvent the protein folding paradox.[28] Here we have provided a mechanistic view at the atomistic scale in which the level of hydration of the main chain is shown to be a determinant of the protein folding process and is defined locally by the characteristics of the lateral chains. These findings contribute to the development of an amino acid-based code to understand the interatomic driving forces defining the tridimensional structure of proteins.

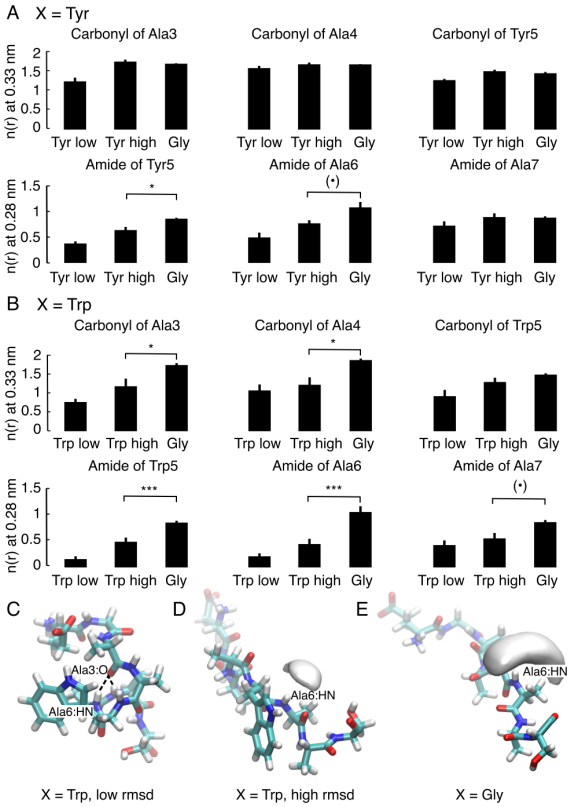


Figure 3. Backbone hydration. (A, B) Histograms showing the numbers of water molecules coordinating the amide hydrogen or carbonyl oxygen atoms of residues 3−7 for the peptides with (A) X = Tyr and (B) X = Trp. Data for X = Gly are shown as a reference. The "low" and "high" labels refer to pools of conformations that have low or high RMSDs with respect to the experimental RDC values. (C−E) Representative molecular structures for X = Trp (low and high RMSD) and X = Gly. The average water density within 3.5 Å of the amide hydrogen of residue Ala6 is shown for X = Trp with high RMSD (D) and X = Gly (E). The water density is isocontoured at 0.016 molecule/Å^3 (i.e., half of the bulk density).

■ ASSOCIATED CONTENT

Supporting Information

Methods, tables, and additional figures. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*simon.berneche@isb-sib.ch

Notes

The authors declare no competing financial interest.

DOI: 10.1021/jacs.5b00660
J. Am. Chem. Soc. 2015, 137, 4300−4303

■ ACKNOWLEDGMENTS

We are very grateful to Dr. Navratna Vajpai for help and advice in the initial stages of the project. This work was supported by grants from the Swiss National Science Foundation to S.B. (SNF Professorship 139205) and S.G. (31-149927), and a stipend from the Croucher Foundation to H.T.A.L. Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel.

■ REFERENCES

(1) Anfinsen, C. B. *Science* **1973**, *181*, 223.
(2) Seife, C. *Science* **2005**, *309*, 78.
(3) Dill, K. A.; MacCallum, J. L. *Science* **2012**, *338*, 1042.
(4) Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R. *Annu. Rev. Biophys.* **2008**, *37*, 289.
(5) Chouard, T. *Nature* **2011**, *471*, 151−153.
(6) Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. M.; Hipps, K. W. *J. Mol. Graphics Modell.* **2001**, *19*, 26.
(7) Dyson, H. J.; Wright, P. E. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197.
(8) Dyson, H. J.; Wright, P. E. *Methods Enzymol.* **2000**, *339*, 258.
(9) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111.
(10) Nodet, G.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 17908.
(11) Jensen, M. R.; Markwick, P. R. L.; Meier, S.; Griesinger, C.; Zweckstetter, M.; Grzesiek, S.; Bernado, P.; Blackledge, M. *Structure* **2009**, *17*, 1169.
(12) Meier, S.; Blackledge, M.; Grzesiek, S. *J. Chem. Phys.* **2008**, *128*, No. 052204.
(13) Dames, S. A.; Aregger, R.; Vajpai, N.; Bernado, P.; Blackledge, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2006**, *128*, 13508.
(14) Sass, H.; Musco, G.; Stahl, S.; Wingfield, P.; Grzesiek, S. *J. Biomol NMR* **2000**, *18*, 303.
(15) Dyson, H. J.; Wright, P. E.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 13057.
(16) Radivojac, P.; Iakoucheva, L. M.; Oldfield, C. J.; Obradovic, Z.; Uversky, V. N.; Dunker, A. K. *Biophys. J.* **2007**, *92*, 1439.
(17) Tompa, P. *Trends Biochem. Sci.* **2002**, *27*, 527.
(18) Cho, M.-K.; Kim, H.-Y.; Bernado, P.; Fernandez, C. O.; Blackledge, M.; Zweckstetter, M. *J. Am. Chem. Soc.* **2007**, *129*, 3032.
(19) Karplus, M.; Weaver, D. L. *Nature* **1976**, *260*, 404.
(20) Klein-Seetharaman, J.; Oikawa, M.; Grimshaw, S. B.; Wirmer, J.; Duchardt, E.; Ueda, T.; Imoto, T.; Smith, L. J.; Dobson, C. M.; Schwalbe, H. *Science* **2002**, *295*, 1719.
(21) Yang, A.-S.; Honig, B. *J. Mol. Biol.* **1995**, *252*, 351.
(22) Avbelj, F.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 10967.
(23) Avbelj, F.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 5742.
(24) Drozdov, A. N.; Grossfield, A.; Pappu, R. V. *J. Am. Chem. Soc.* **2004**, *126*, 2574.
(25) Cho, S. S.; Reddy, G.; Straub, J. E.; Thirumalai, D. *J. Phys. Chem. B* **2011**, *115*, 13401.
(26) Dill, K. A.; Fiebig, K. M.; Chan, H. S. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 1942.
(27) Zwanzig, R.; Szabo, A.; Bagchi, B. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 20.
(28) Levinthal, C. In *Mossbauer Spectroscopy in Biological Systems: Proceedings of a Meeting Held at Allerton House, March 17 and 18, 1969, Monticello, IL*; University of Illinois Press: Urbana, IL, 1969; pp 22−24.

DOI: 10.1021/jacs.5b00660
J. Am. Chem. Soc. 2015, 137, 4300−4303

Supporting Information

Backbone hydration determines the folding signature of amino acid residues

Olivier Bignucolo[1,2], Hoi Tik Alvin Leung[2], Stephan Grzesiek[2] and Simon Bernèche[1,2]*

[1]Swiss Institute of Bioinformatics

[2]Biozentrum, University of Basel, Klingelbergstrasse 50/70, CH-4056 Basel, Switzerland

*Correspondence to: simon.berneche@unibas.ch

Material and Methods

Molecular dynamics

The simulations were carried out using the GROMACS simulation package[1] and the Amber03 force field[2]. The initially completely extended peptides were constructed with the MOLMOL package[3] and solvated in a dodecahedron box with about 8700 TIP4P water molecules[4]. Sodium chloride at a concentration of about 0.02 mol.L^{-1} was used to neutralize the system (3 Na^+, 2 Cl^-). The system was energy minimized by the steepest descent procedure with a tolerance of 1000 kJ•mol^{-1}, followed by a peptide position restrained run of 500 ps to allow the solvent to equilibrate. Production run lasted typically 100 ns. For the purpose of the analyses, coordinates were taken every 20 ps, so that a typical simulation is represented by 5000 conformations (an interval of 4 ps was used for the autocorrelation analyses). Long-range electrostatic interactions were calculated using particle-mesh Ewald (PME) with a grid spacing of 0.12 nm[5,6], and a cutoff of 1.4 nm was used for the Lennard-Jones interactions. All bonds were constrained with the P-LINCS algorithm[7], allowing an integration time step of 2 fs in leap-frog dynamics. The temperature was kept constant at 300 K through velocity rescaling ($\tau_t = 0.1$ ps)[8], and the pressure at 1 bar ($\tau_p = 0.2$ ps)[9].

RDC and chemical shift calculations

Theoretical RDCs for every snapshot of the MD trajectories were predicted based on a steric alignment model using PALES[10]. Many RDC calculations were repeated using an efficient algorithm described previously[11] and yielded the same results. The calculated RDC values of each conformation were scaled by a constant that was determined by a least square fit between the average RDCs of all conformations and the experimental RDC values of a given peptide. The chemical shifts were calculated using SPARTA[12]. Most of the analyses were performed on ensemble averages of the calculated RDCs or chemical shifts over the total time of the independent simulations (typically 5000 frames), or over stretches of 10 ns (500 frames) for the time series analyses. The RDCs are proportional to $\frac{\gamma_1\gamma_2}{r^3}\langle\frac{3cos^2\,\theta-1}{2}\rangle$, where γ represents a nuclei gyromagnetic ratio, and r represents the inter-nuclear distance. In order to take into account the different values of carbon and nitrogen gyromagnetic ratios and bond lengths related to the two sets of RDCs ($^1D_{C\alpha H\alpha}$ and $^1D_{NH}$), the RMSDs were pooled as follows:

$$RMSD = \sqrt{\frac{1}{2}\cdot\left((RMSD_{C\alpha H\alpha}/2)^2 + {RMSD_{NH}}^2\right)} \qquad (1)$$

The $RMSD_{C\alpha H\alpha}$ is divided by 2 to take into account that the magnitude of $^1D_{C\alpha H\alpha}$ is roughly twice that of $^1D_{NH}$, which is due to the different gyromagnetic ratios (67.23 MHz T^{-1} for 13C and -27.1 MHz T^{-1} for 15N) and bond lengths (109 pm for C-H and 104 pm for N-H).

Cluster analysis and maximum entropy

We clustered the 35,000 conformations of the X = Trp peptide according to the backbone dihedral angles (ϕ,ψ) of residues 3 to 7, independently of the RDC values. A hierarchical cluster analysis was performed on a Euclidean distance matrix with the metric:

$$d(i,j) = \sqrt{\frac{1}{\text{(num. of res.)}} \sum_{res} \left(d1\left(\phi_{i.,res}, \phi_{j,res}\right)^2 + d1\left(\psi_{i,res}, \psi_{j,res}\right)^2 \right)} \quad (2)$$

where

$$d1(\vartheta_1, \vartheta_2) = \min\left(|\vartheta_1 - \vartheta_2|, |\vartheta_1 - \vartheta_2 + 360°|, |\vartheta_1 - \vartheta_2 - 360°| \right) \quad (3)$$

Twenty clusters were requested and sorted in descending order of their number of conformations. The largest cluster contains about 12,000 conformations, i.e. 34% of the whole ensemble, and the first 8 clusters account for about 80% of all conformations (Fig. S2).

Using the software package IPOPT[13], individual weights producing the best agreement to the experimental measurements were attributed to each conformation by maximizing the information entropy of the ensemble:

$$S = -\sum_i p_i \ln p_i \quad (4)$$

while respecting the constraints:

$$\sum_i p_i = 1 \quad (5)$$

$$\sum_{res} \left(\sum_i p_i D_{calc,i,res} - D_{exp,res} \right)^2 \leq \sum_{res} \sigma_{exp}^2 \quad (6)$$

where p_i is the optimized population of conformation i. Three datasets were used for this calculations, the RDCs $^1D_{C\alpha H\alpha}$ and $^1D_{NH}$, and the J coupling $^3J_{HNH\alpha}$ from Dames et al.[14], and thus three independent constraints were defined based on Eq. 6. For each dataset, $D_{calc,i,res}$ and $D_{exp,res}$ correspond, for a given residues, to the calculated RDC (or J coupling) of conformation i and the corresponding experimental RDC (or J coupling). Eq. 6 states that χ^2 should be below or equal to the experimental error, assuming that each residue measurement has an error of σ_{exp} Hz. The following σ_{exp} were used for $^1D_{C\alpha H\alpha}$, $^1D_{NH}$, and $^3J_{HNH\alpha}$ respectively: 2 Hz, 1 Hz, and 0.3 Hz. The weight of each cluster was obtained by summing the population of all conformations found in the cluster.

Statistical analyses

Statistical analyses were conducted using the R software environment[15]. For the identification of the structurally relevant hydrogen bonds (reported in Tables S1 and S2) that could best explain the correlation observed in Fig. 1b, we performed the following analysis. First we discarded the pairs of residues for which the frequency of hydrogen bond formation was less than 2.5% of all intra-molecular hydrogen bonds formed. For the peptide with X=Trp, for example, 11 out of the 36 residue pairs account for 80% of the total number of intra-molecular backbone hydrogen bonds. For each hydrogen bond, we performed a linear regression between its observed frequency and the RMSD to experiment, using the individual simulations as replicates (n = 12 and 10 for Tyr and Trp respectively). We finally retained only the residue donor-acceptor pairs for which the regression was statistically significant. Having detected the relevant residue pairs, we repeated an identical analysis on these residues only, but at the level of the hydrogen-bond donor and acceptor chemical groups. We report in Tables S1 and S2 the donor-acceptor pairs for which the regression was statistically significant. Furthermore, some of these analyses were repeated with a stricter criteria for counting the hydrogen bonds within the structures, namely with distance between acceptor and donor of 3.0 Å instead of GROMACS default value of 3.5 Å. Despite a strongly reduced number of detected hydrogen bonds, they produced essentially the same trends.

The superposition of the structural parameter values and the RMSD to experimental RDCs in Figs. 2 and S4 was minimized by linear regression. The procedure allows the quantification of the relation between any calculated structural parameter and the RMSD to experimental measurements by a single RMSD° value, as reported on the figures. This quantity is written RMSD° to emphasize that it refers to a RMSD between different quantities, e.g. radius of gyration (nm) and RDCs (Hz). The quality of the regressions was tested through the R "Fitting Linear Models" module. After having verified that the time had no significantly measurable effect on any parameter, we computed the regressions without implied intercept term. All the presented correlations were statistically significant with $p < 0.01$, and all the models explained more than 90% of the variance (n = 70 for 700 ns of simulations).

Hydration of backbone polar groups

We compared the number of water molecules within the first hydration shell of chosen atoms in folded or extended peptides. For each simulation with X = Tyr or X = Trp, we selected two groups of peptide conformations: A first group was formed with the 20% of structures with the lowest RMSD to the experimental RDC values, thus representing mainly folded conformations, and a second group with the 20% of structures with the largest RMSDs, thus with mainly extended conformations. The remaining

structures are intermediate cases that are partially folded. Analysis of the hydration of these structures would lead to inconclusive results and were thus not considered. Since the same RMSD cutoff values were used for the different simulations, the proportion of structures retained in the two groups varied slightly around 20% for each of the simulations. For the X = Gly peptide, all structures were retained since they are mainly extended.

For each group of conformations of each simulation, we computed the radial distribution function of the solvent oxygen atoms around functional groups of the backbone. The number of water molecules in the first hydration shell was then obtained by integration from 0 to a cutoff value of d=0.28 and 0.33 nm for amide and carbonyl groups respectively. The cutoff was chosen based on normalized radial distribution function, in line with previous studies[16,17]. Standard errors were calculated on the basis of the independent simulations with n = 7, 12, and 10 for Gly, Tyr and Trp respectively. The Tukey honest significant differences test[18] was used to perform a pairwise comparison of the average number of water molecules. In Fig. 3 the Tukey's Honesty Significance Difference test is only applied to the comparison between X = Gly and the extended conformations of peptides with X = Tyr or Trp.

Autocorrelation

The stability of backbone hydrogen bonds and dihedral angles was addressed through the calculation of autocorrelation functions. In order to get smoother autocorrelation curves, conformational sampling was increased by taking coordinates every 4 ps. The existence function of all hydrogen bonds involving the backbone of residues 3 to 8 was used for the autocorrelation calculation, as defined by the g_hbond module of Gromacs[19]. The ψ dihedral angles of the four alanine residues next to the substituted residue were considered in two groups: direct neighbors (Ala4 and Ala6) and second neighbors (Ala3 and Ala7). To take into account the periodicity of the dihedral angles, the following autocorrelation function defined in the g_chi module of Gromacs was used[20]:

$$C(t) = \langle \cos [\theta(t) - \theta(t+\tau)] \rangle \qquad (7)$$

In both cases, we fitted the sum of a stretched exponential (in the picosecond time range) and a single exponential (in the nanosecond time range)[21] over the first 20 ns, and retained the value of the single exponential as the structurally relevant lifetime. The number of replicates was 7, 7, 7 and 9 simulations of 100 ns for X = Gly, Ile, Tyr and Trp respectively. Following the ANOVA, the Tukey honest significant differences test[18] was used to perform pairwise comparisons.

Supplementary Results

Stability of the structure

We computed the autocorrelation function of hydrogen bond existence between backbone atoms of residues 3 to 8, as well as the ψ dihedral angles of the alanine residues next to the substituted residue. The stabilization of the structure, as inferred through larger autocorrelation times, increases with the size of the side-chains (Figs. S6 and S7). Particularly, for the analysis of the dihedral angles, we postulated that, if the residue at position X would affect the stability of the backbone dihedrals, we should detect a larger impact on the direct neighbors (positions 4 and 6), and a smaller one on the second neighbors (positions 3 and 7). Effectively, whereas all autocorrelation times of the second neighbors are identical with values of about 2.5 ns, they increase for the direct neighbors to 2.8, 4.2, and 4.7 ns for Ile, Tyr, and Trp respectively, and decrease to 1.5 for Gly. These results support the idea that larger amino acids not only promote the formation of a turn or short helix, but also stabilize this conformation.

Tables

Table S1. Atom pairs for which the formation of a hydrogen bond correlates significantly with the experimentally measured RDCs (Peptide with X=Trp)

Atom pair	Correlation coefficient [1)]	p-value	Slope of the regression
A_3-O – A_6-NH	- 0.85	0.002	-15.9
A_3-O – A_7-NH	- 0.92	0.001	-16.7
A_4-O – S_8-NH	- 0.74	0.02	-16.8
W_5-O – S_8-NH	- 0.62	0.06	-31.8

[1)] Coefficient of correlation between the hydrogen bonding probability and the RMSD between experimental and predicted RDCs. Regression calculated over the averaged values of 10 simulations of 100 ns each.

Table S2. Atom pairs for which the number of hydrogen bonds correlates significantly with the experimentally measured RDCs (Peptide with X=Tyr)

Atom pair	Correlation coefficient [1)]	p-value	Slope of the regression
A_3-O – A_6-NH	- 0.60	0.03	-3.21
A_3-O – A_7-NH	- 0.71	0.005	-4.54
A_4-O – A_7-NH	- 0.69	0.006	-7.50
A_4-O – S_8-NH	- 0.65	0.02	-4.82
Y_5-O – S_8-OH	- 0.63	0.02	-6.95

[1)] Coefficient of correlation between the hydrogen bonding probability and the RMSD between experimental and predicted RDCs. Regression calculated over the averaged values of 12 simulations of 100 ns each.

Figures

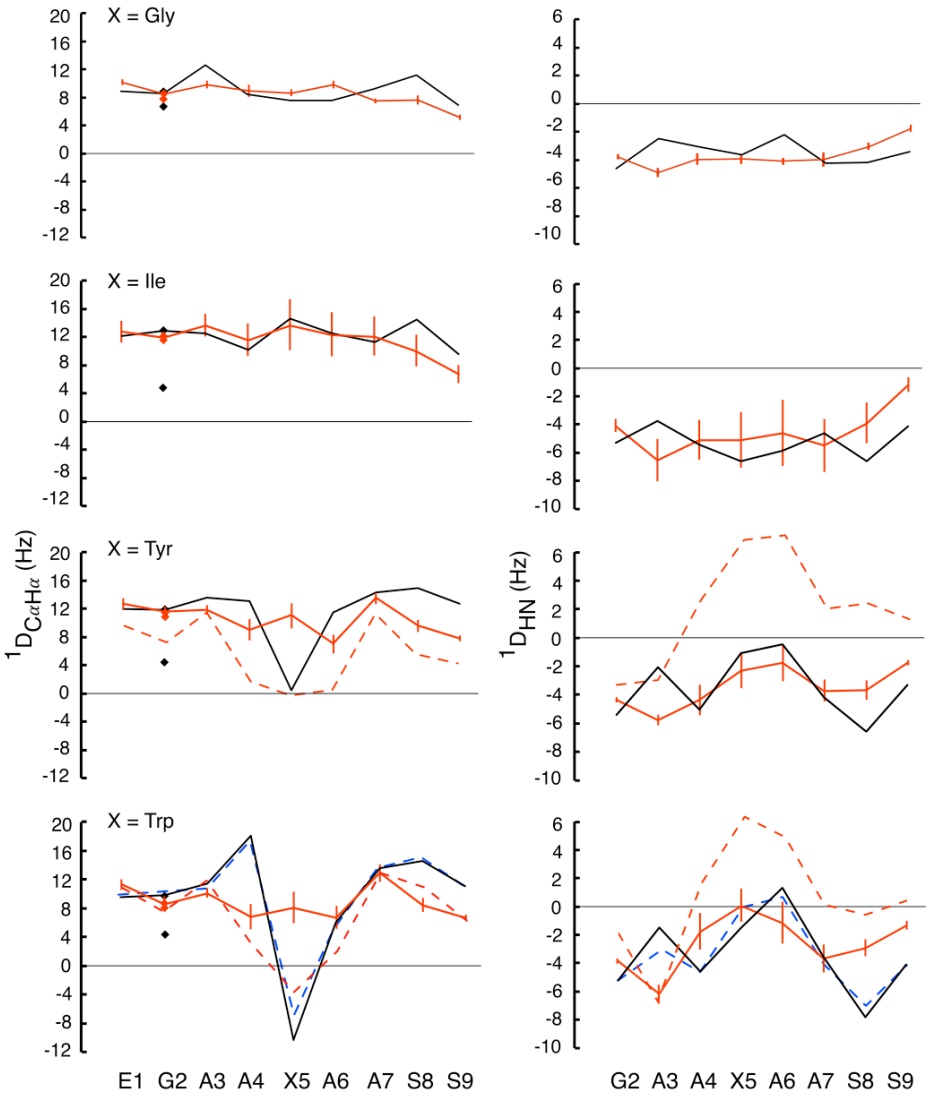


Figure S1. Comparison of experimental and predicted RDCs. For each peptide, the experimental RDC patterns as published by Dames et al.[13] are shown (black lines), with the MD predicted RDC patterns (solid red lines). The averages and standard errors were taken over 7 (Gly), 7 (Ile), 12 (Tyr) and 10 (Trp) simulations of 100 ns. The average RDCs of the conformations found in the first cluster are shown for X = Tyr and Trp (dashed red lines). A perfect agreement with the experimental data can be obtained after limited reweighting of the different sampled conformations (dashed blue lines).

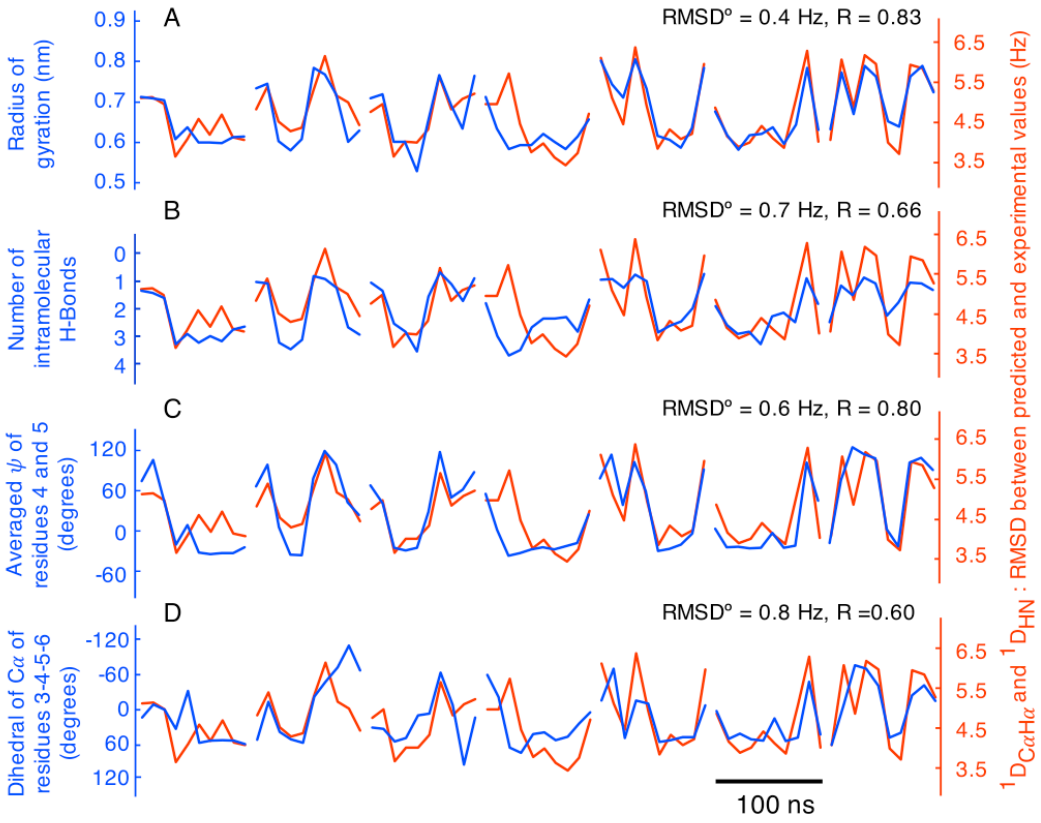


Figure S2. Correlation between structural parameters and the RMSD between predicted and experimental RDCs of the peptide with X = Trp. Time series of seven 100 ns simulations are shown side by side. The different structural parameters are: (**A**) Radius of gyration, (**B**) Total number of intra-molecular hydrogen bonds, (**C**) Averaged ψ angle of residues 4 and 5, (**D**) Torsion angle as defined by the Cα atoms of residues 4, 5, 6 and 7. Each point represents the average over 10 ns. The curves were aligned by minimization through linear regression. The term RMSD° expresses the quality of the alignment fits (see Methods).

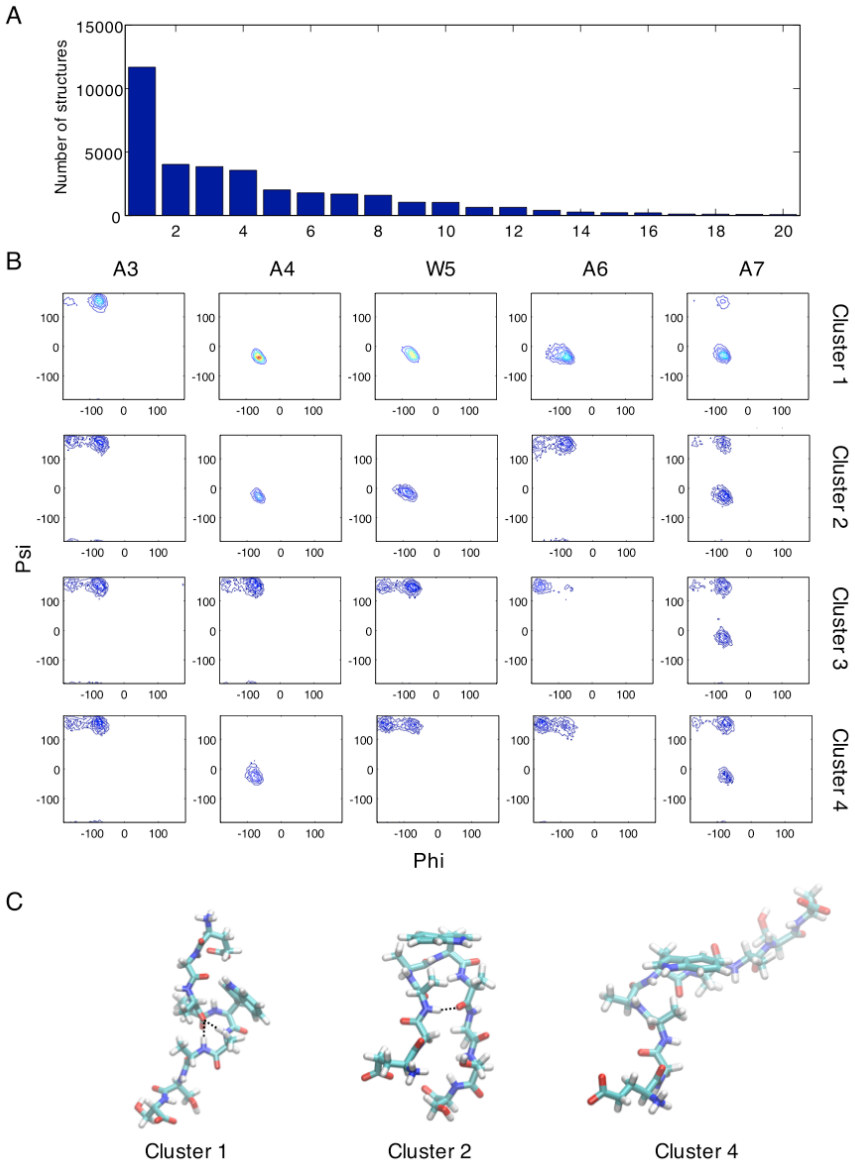


Figure S3. Clustering of conformations for the X = Trp peptide. (**A**) Distribution of the 35,000 conformations over 20 clusters defined on the basis of the dihedral angles of residues 3 to 7. (**B**) Ramachandran plots for these residues in the four first clusters. (**C**) Molecular graphics of representative conformations of clusters 1, 2, and 4. Hydrogen bonds are indicated by dashed lines.

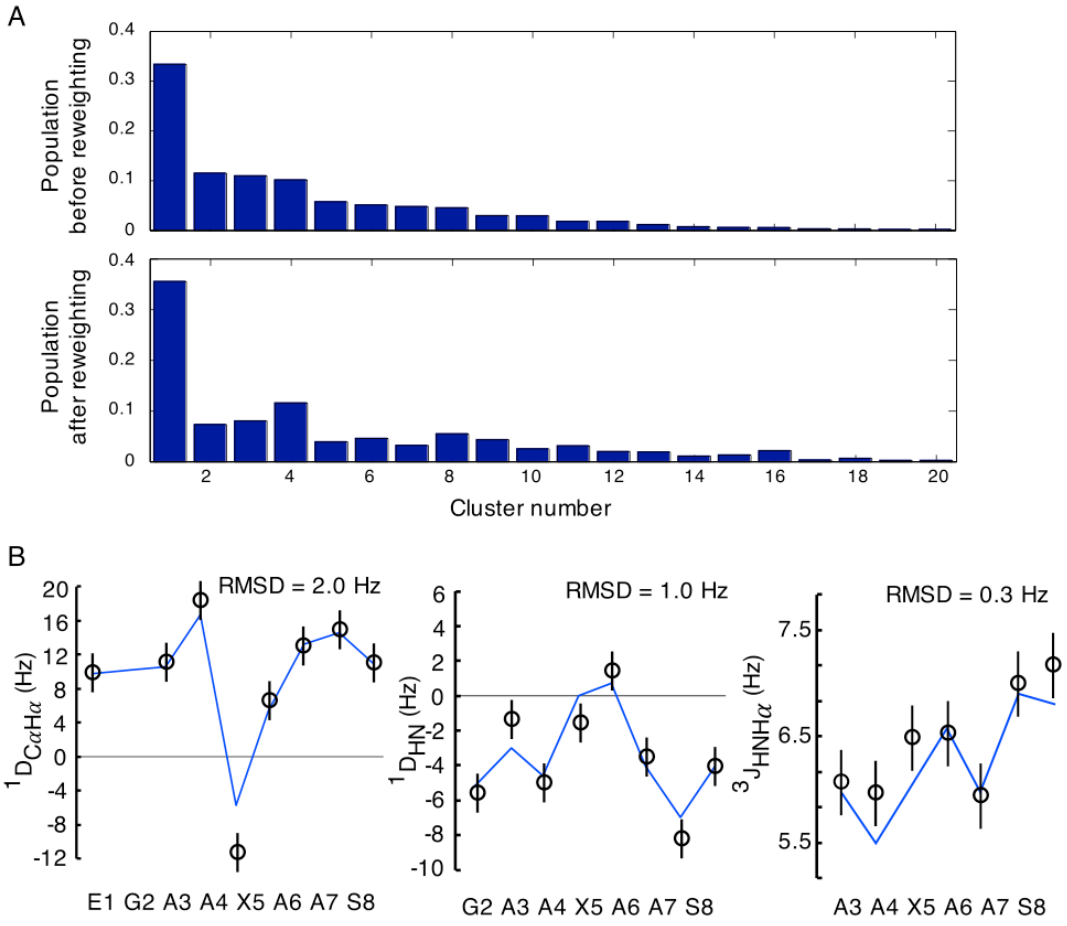


Figure S4. Reweighting of population clusters using a maximum entropy approach for X = Trp. (A) The population of the different clusters extracted from the MD simulations are corrected to reproduce the data shown in (B) while maximising the entropy of the distribution. For all clusters, the required correction is relatively small. **(B)** With the corrected distribution, the RDC and J-coupling data from Dames et al. [14] are reproduced within their experimental errors. The experimental data and associated error bars are shown in black and the fitted data in blue.

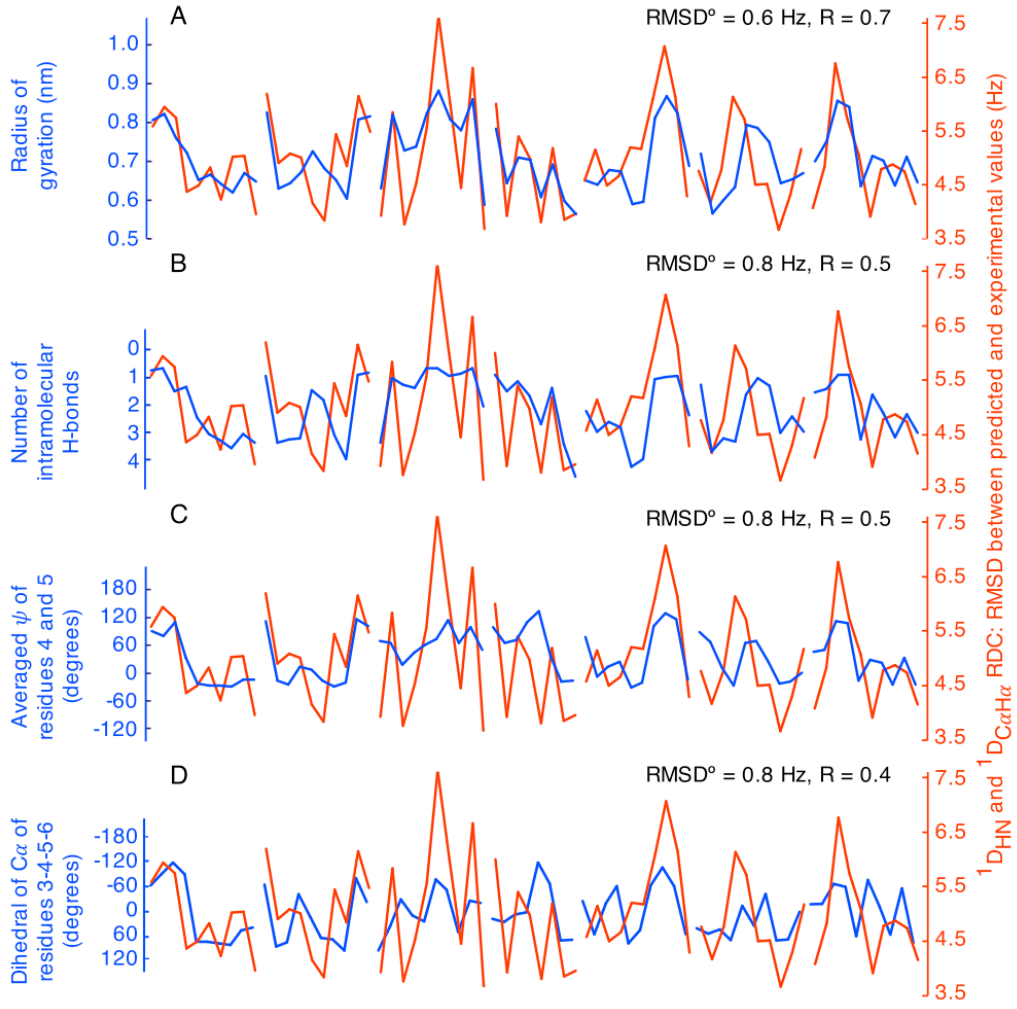


Figure S5. Correlation between structural parameters and the RMSD between predicted and experimental RDCs of the peptide with X = Tyr. Time series of seven 100-ns simulations are shown side by side. The different structural parameters are: (**A**) Radius of gyration, (**B**) Total number of intra-molecular hydrogen bonds, (**C**) Averaged ψ angle of residues 4 and 5, (**D**) Torsion angle as defined by the Cα atoms of residues 3, 4, 5 and 6. Each point represents the average over 10 ns. The curves were aligned by simple linear regression. The term RMSD° expresses the quality of the alignment fits (see Methods).

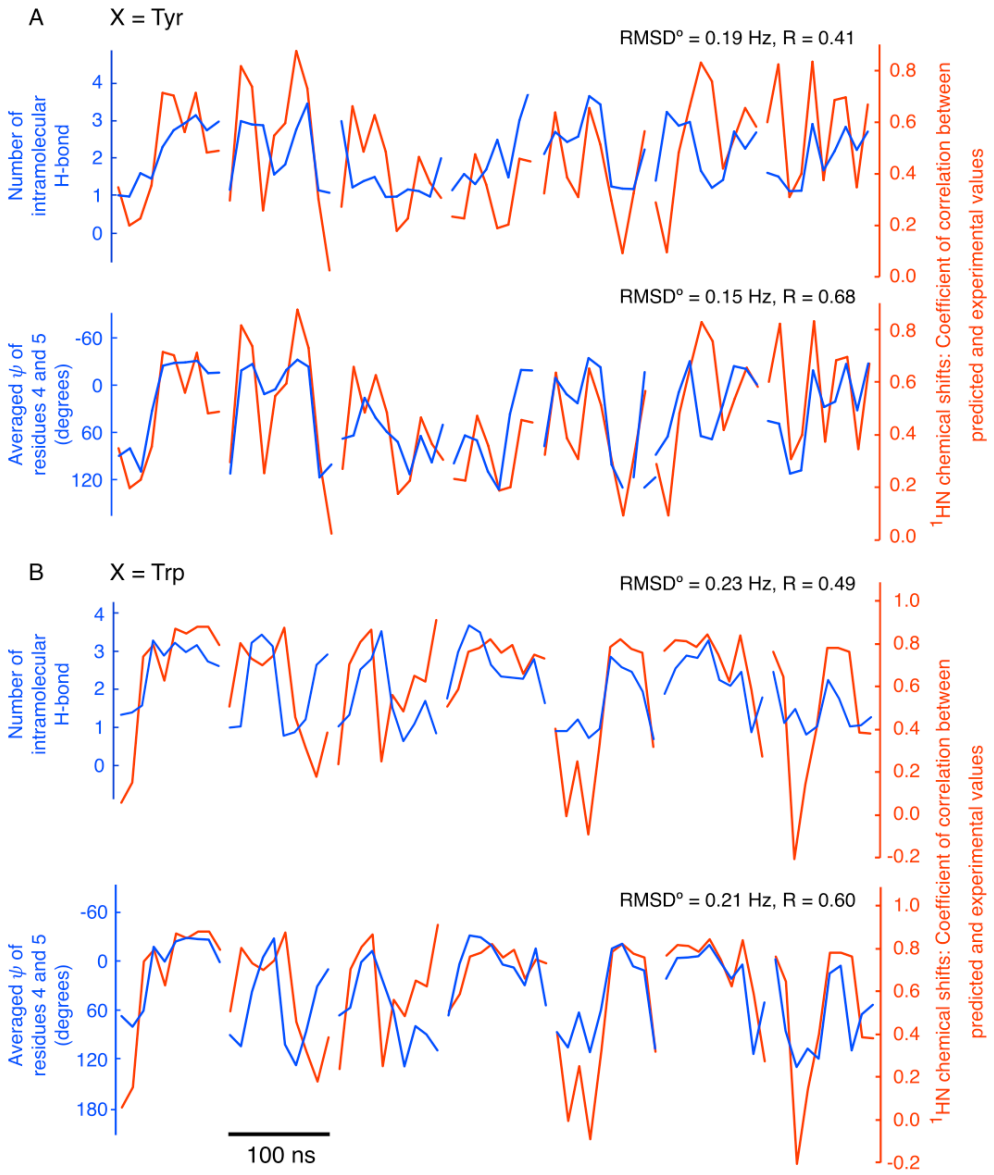


Figure S6. Correlation between ^{1}HN chemical shifts and structural parameters. Time series of seven 100-ns simulations are shown side by side for peptides X = Tyr (**A**) and X = Trp (**B**). The coefficient of correlation between the experimental and predicted ^{1}HN chemical shifts are compared to structural parameters: the number of intra-molecular hydrogen bonds and the average of ψ angle of residues 4 and 5. Each point represents the average over 10 ns.

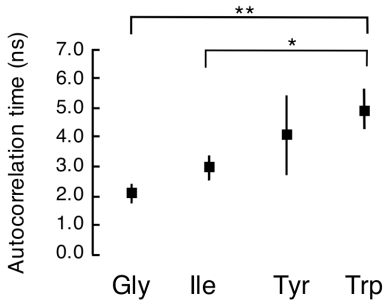


Figure S7. Larger amino-acids stabilize helical conformations. We report the average and SE of the backbone hydrogen bond autocorrelation time for different substitutions at position X.

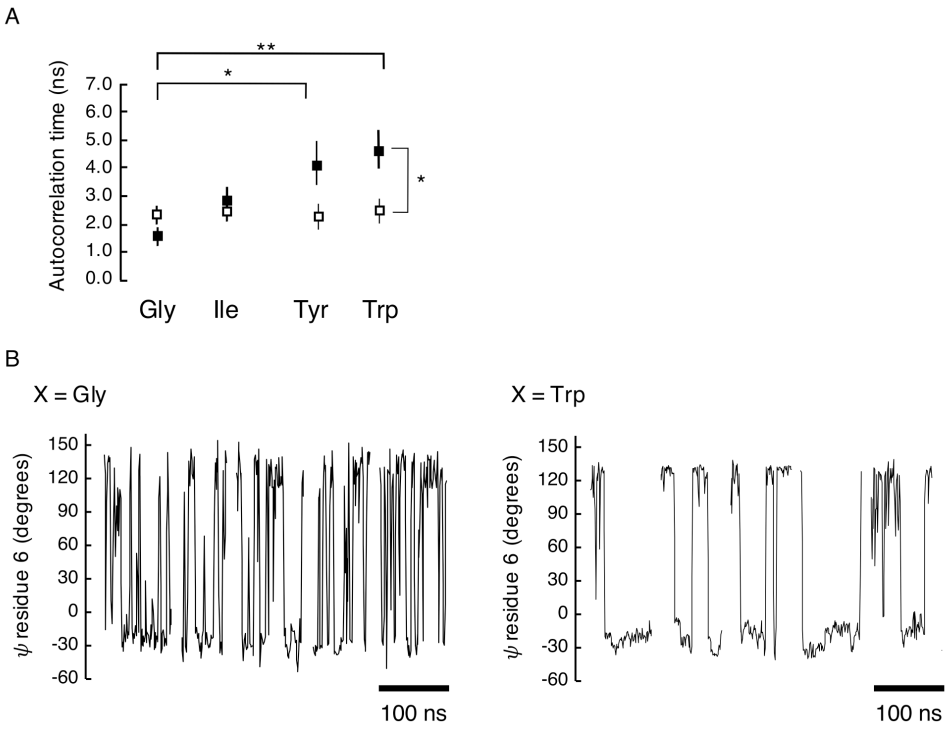


Figure S8. Larger amino-acids stabilize backbone dihedral angles of neighboring residues. (**A**) The averages and SE of the cosine autocorrelation time of the ψ dihedral angles of residues 3 and 7 (white squares) and residues 4 and 6 (black squares) are shown. (**B**) Time series of the ψ angle of residue 6, over 5 simulations of 100 ns each for X = Gly and X = Trp. The aromatic amino acid stabilizes the dihedral of residues in the middle of the chain to values typical of helices or turns, i.e. $\psi \approx -30^{o}$.

References

(1) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701.
(2) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999.
(3) Koradi, R.; Billeter, M.; Wüthrich, K. *J. Mol. Graph.* **1996**, *14*, 51.
(4) Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. *J. Chem. Phys.* **1983**, *79*, 926.
(5) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577.
(6) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.
(7) Hess, B.; Bekker, H.; Berendsen, H.; Fraaije, J. *J. Comput. Chem.* **1997**, *18*, 1463.
(8) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.
(9) Berendsen, H. J.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
(10) Zweckstetter, M. *Nat. Protoc.* **2008**, *3*, 679.
(11) Huang, J.-R.; Grzesiek, S. *J. Am. Chem. Soc.* **2010**, *132*, 694.
(12) Shen, Y.; Bax, A. *J. Biomol. NMR* **2007**, *38*, 289.
(13) Wächter, A.; Biegler, L. T. *Math. Program.* **2005**, *106*, 25.
(14) Dames, S. A.; Aregger, R.; Vajpai, N.; Bernado, P.; Blackledge, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2006**, *128*, 13508.
(15) R core Team. *R: A Language and Environment for Statistical Computing, 2014*.
(16) Busch, S.; Pardo, L. C.; O'Dell, W. B.; Bruce, C. D.; Lorenz, C. D.; McLain, S. E. *Phys. Chem. Chem. Phys.* **2013**, *15*, 21023.
(17) Cho, S. S.; Reddy, G.; Straub, J. E.; Thirumalai, D. *J. Phys. Chem. B* **2011**, *115*, 13401.
(18) Miller Rupert, G. *Simultaneous statistical inference*; Springer Verlag, 1981.
(19) Van Der Spoel, D.; van Maaren, P. J.; Larsson, P.; Tîmneanu, N. *J. Phys. Chem. B* **2006**, *110*, 4393.
(20) Van Der Spoel, D.; Berendsen, H. J. *Biophys. J.* **1997**, *72*, 2032.
(21) Daidone, I.; Neuweiler, H.; Doose, S.; Sauer, M.; Smith, J. C. *PLoS Comput. Biol.* **2010**, *6*, e1000645.

2.2 Validation through COPER: convex optimization for ensemble reweighting

2.2.1 Introduction

In the investigations of the peptides of sequence EGAAXAASS presented above, the most probable conformation of the peptides was deduced from a systematic comparison between the experimental and predicted RDCs and chemical shifts (*74*). Theoretical RDCs for every snapshot of the MD trajectory were calculated based on a steric alignment model using PALES (*75*), and chemical shifts were estimated using SPARTA (*76*). However, the large and fast conformational fluctuations of these mostly disordered peptides required to consider ensemble averages in order to determine the most favored conformation as a function of the sequence. A strong relationship was identified between the number of intramolecular hydrogen bonds and the agreement with the experimental RDCs (Fig. 1 of the paper). However, more than hundred different combinations of hydrogen bond patterns could theoretically be formed in this peptide. Thus, as reported in the supplementary material of the publication reproduced in the section 2.1.5, the identification of the structurally relevant hydrogen bonds was performed in a multistep regression analysis. First, the pairs of residues for which the probability of hydrogen bond formation represented less than about 2% of the total amount of hydrogen bonds were removed from the hydrogen-bond list. Using this scoring procedure, about 30% of the possible theoretical pairs could be removed, while preserving maximal information. Because of the large and fast conformational changes of the peptides, the averages over complete trajectories were used for the linear regressions analyzes. The residue pairs for which the regression was highly significant were retained for a second procedure performed at the level of hydrogen bond donor and acceptor chemical groups. The relationship between the agreement to the experiment and the number of hydrogen bonds was only highly statistically significant in the case of a one-turn helix, thus supporting the idea that the experimentally measured RDC is related to a one-turn helix or a turn. However, using this procedure, some frames were discarded. The next paragraph and the following paper show that, using a slightly different approach, which incorporates all the frames, the same overall conclusions could be drawn.

Recently, Leung et al. (*77*) reassessed the identification of the conformers of the same peptide EGAAXAASS which reproduce at best the experimental RDCs for two of the four substitutions, namely Ile and Trp. For this second analysis, the $^{3}J_{H\alpha N}$ couplings of the two short peptides were included. The newly introduced method, which reweights every single frame as a function of its agreement to experiment while maximizing the ensemble entropy, was called COPER, for convex optimization for ensemble reweighting. The results

corroborate the previous findings that, in the case of a substitution with Trp, the formation of a main cluster forming a one-turn helix was key to reproduce the experimental RDC pattern (*77*). The reweighting factors were less than 3 kT, consistent with errors in MD force fields (*78*).

2.2.2 Statement of contribution

I performed all the simulations and the clustering analysis as well. I wrote an evolutionary based algorithm in the Python programming language, which reweights the frames according to the criteria described in the paper. The Appendix 2.2.3.1 contains the code and the Appendix 2.2.3.2 presents an illustrative example. Dr. Hoi Tik Alvin Leung, using a different algorithm, calculated the results reported in the paper.

2.2.3 Original publication

Hoi Tik Alvin Leung, Olivier Bignucolo, Regula Aregger, Sonja A. Dames, Adam Mazur, Simon Bernèche, and Stephan Grzesiek

A Rigorous and Efficient Method to Reweight Very Large Conformational Ensembles Using Average Experimental Data and To Determine Their Relative Information Content

Journal of Chemical Theory and Computation 2016 vol. 12 (13) pp. 383-394

JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

A Rigorous and Efficient Method To Reweight Very Large Conformational Ensembles Using Average Experimental Data and To Determine Their Relative Information Content

Hoi Tik Alvin Leung,[†] Olivier Bignucolo,[‡] Regula Aregger,[§] Sonja A. Dames,[∥,¶] Adam Mazur,[†] Simon Bernèche,[‡] and Stephan Grzesiek*[,†]

[†]Focal Area Structural Biology and Biophysics, Biozentrum, [‡]SIB Swiss Institute of Bioinformatics, University of Basel, CH-4056 Basel, Switzerland

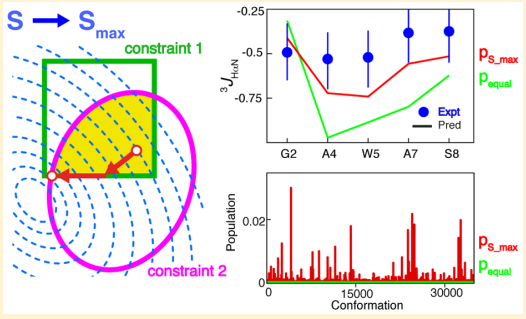
[§]Institut für Biochemie, University of Leipzig, D-04103 Leipzig, Germany

[∥]Department of Chemistry, Technische Universität München, D-85748 Garching, Germany

[¶]Institute of Structural Biology, Helmholtz Zentrum München, D-85764 Neuherberg, Germany

Supporting Information

ABSTRACT: Flexible polypeptides such as unfolded proteins may access an astronomical number of conformations. The most advanced simulations of such states usually comprise tens of thousands of individual structures. In principle, a comparison of parameters predicted from such ensembles to experimental data provides a measure of their quality. In practice, analyses that go beyond the comparison of unbiased average data have been impossible to carry out on the entirety of such very large ensembles and have, therefore, been restricted to much smaller subensembles and/or nondeterministic algorithms. Here, we show that such very large ensembles, on the order of 10^4 to 10^5 conformations, can be analyzed in full by a maximum entropy fit to experimental average data. Maximizing the entropy of the population weights of individual conformations under experimental χ^2 constraints is a convex optimization problem, which can be solved in a very efficient and robust manner to a unique global solution even for very large ensembles. Since the population weights can be determined reliably, the reweighted full ensemble presents the best model of the combined information from simulation and experiment. Furthermore, since the reduction of entropy due to the experimental constraints is well-defined, its value provides a robust measure of the information content of the experimental data relative to the simulated ensemble and an indication for the density of the sampling of conformational space. The method is applied to the reweighting of a 35 000 frame molecular dynamics trajectory of the nonapeptide EGAAWAASS by extensive NMR ^3J coupling and RDC data. The analysis shows that RDCs provide significantly more information than ^3J couplings and that a discontinuity in the RDC pattern at the central tryptophan is caused by a cluster of helical conformations. Reweighting factors are moderate and consistent with errors in MD force fields of less than $3kT$. The required reweighting is larger for an ensemble derived from a statistical coil model, consistent with its coarser nature. We call the method COPER, for convex optimization for ensemble reweighting. Similar advantages of large-scale efficiency and robustness can be obtained for other ensemble analysis methods with convex targets and constraints, such as constrained χ^2 minimization and the maximum occurrence method.



INTRODUCTION

Proteins exist as ensembles of interchanging conformations. Obviously, unfolded polypeptide chains, such as chemically or physically denatured proteins and intrinsically disordered proteins (IDPs), can access an extremely large number of conformations.[1] A comprehensive description of their structural preferences is a prerequisite for understanding protein folding and the function of IDPs in health and disease.[2] However, native, folded proteins also usually adopt many conformations close to the global free energy minimum,[3] and their interchange is a hallmark of protein function, such as catalysis[4] or signal transduction.[5]

A detailed experimental determination of individual structures in such protein ensembles becomes impossible as soon as their number exceeds a few, since the number of conformational degrees of freedom quickly outpaces the number of measurable parameters.[6] To make progress,

Received: August 7, 2015
Published: November 2, 2015

DOI: 10.1021/acs.jctc.5b00759
J. Chem. Theory Comput. 2016, 12, 383−394

ensembles containing tens of thousands of conformers are often simulated and compared to experimental data. Simulated ensembles can be obtained by many methods, e.g., the simulation of a random chain according to the coil model of the unfolded state,[7−9] coarse-grained simulations of protein domain motions,[10,11] or all-atom molecular dynamics (MD) simulations with varying degrees of complexity.[12−16] The quantitative analysis of such very large ensembles presents a formidable challenge. An initial analysis needs to establish the accuracy and information content of the predicted ensemble relative to any experimental knowledge, and, if necessary, the ensemble needs to be refined to reproduce the experimental data. Only then are more detailed predictions of unobserved parameters warranted. Due to the very large size, so far analyses of entire large ensembles have been limited to the comparison of unbiased averages over the ensemble to measured experimental average values. Thus, e.g., unbiased averages derived from even the most advanced MD force fields still fail to accurately predict the experimental data without further adjustments.[15]

Due to computational intractability, more detailed analyses of simulated ensembles have been restricted to much smaller size ensembles, i.e., typically on the order of, at most, several hundred conformers. Procedures such as sample-and-select (SAS),[17] the ensemble optimization method (EOM),[18] ASTEROIDS,[19] and sparse ensemble selection (SES)[20] select smaller subsets by various strategies from initially created large ensembles that satisfy the measured parameters. Similarly, maximum entropy (ME) reweighting of individual conformers,[11,21] Bayesian estimation of individual conformer weights,[22] and maximum occurrence (MO),[10] which estimates the maximal possible occurrence of a conformer within an ensemble, have been used only on smaller selected subsets or clusters but not on entire very large ensembles. Minimal-size ensembles compatible with experimental data may also be generated by constrained ensemble structure calculations.[6] Besides the recently proposed SES,[20] all proposed methods use stochastic mathematical procedures such as genetic algorithms or simulated annealing, and their solutions are not guaranteed to be optimal and unique.

Here, we show that very large ensembles can be analyzed in full and very efficiently by a maximum entropy approach that reweights all individual populations in the ensemble such that the average over the ensemble reproduces the experimental data within the experimental error ($\chi^2 \leq 1$). This constrained search for the maximum entropy S^{max} falls into the class of convex optimization problems, which can be solved in a very efficient and deterministic manner even for very large data sets. As the population weights are calculated in a robust manner on the entire ensemble, the reweighted large ensemble represents the most accurate representation of the combination of simulation and experiment in an information-theoretical sense. Furthermore, since S^{max} is a well-defined parameter, its reduction relative to an unconstrained ensemble presents the true measure of the information content of experimental data relative to the simulated ensemble. We call the method COPER, for convex optimization for ensemble reweighting. As an example, we analyze an ensemble of 35000 snapshots of a 700 ns MD trajectory of the nonapeptide EGAAWAASS in water, for which we had previously obtained extensive residual dipolar coupling (RDC), J coupling, and chemical shift data.[23,24] The results show that the unconstrained MD simulation overestimates the α-helical content. However, reweighting factors are moderate, corresponding to free energy changes of $2.6kT$, which are within the expected inaccuracy of MD force fields. A very strong discontinuity observed in the RDCs around the central tryptophan residue can be explained by a cluster of helical conformations of the central residues.[24] Not surprisingly, reweighting of a 35 000 member ensemble generated from a random coil model of the unfolded state by the program Flexible-Meccano[8] requires a larger free energy change of $3.7kT$, consistent with its coarser nature of approximation. In contrast, a similar analysis carried out for the nonapeptide EGAAIAASS indicates largely extended conformations and much smaller necessary reweighting factors for its MD trajectory. As a corollary, we show that reweighting populations for χ^2 minimization and the maximum occurrence method[10] are also convex optimization problems that can be solved in an equally efficient, deterministic manner.

THEORY

Maximum Entropy Reweighting as a Convex Optimization Problem. We consider an ensemble of N members with populations p_i ($0 \leq p_i \leq 1$, $\sum_i p_i = 1$, $i = 1, ..., N$). Its entropy S in the sense of Shannon[25] is given as

$$S = -\sum_{i=1}^{N} p_i \ln p_i \tag{1}$$

Let d_j^{exp} be one of M measured experimental parameters ($1 \leq j \leq M$) and $d_{i,j}^{pred}$, its predicted value for the ith member of the ensemble. Its predicted weighted average d_j^{pred} over the ensemble is then given as

$$d_j^{pred} = \sum_{i=1}^{N} p_i d_{i,j}^{pred} \tag{2}$$

Equation 2 holds for cases where each conformation can be treated individually and the experimental parameter is a population-weighted average. Many NMR parameters, such as chemical shifts, residual dipolar couplings, paramagnetic relaxation enhancements, and J couplings, fulfill this condition. Assuming that the experimental system is ergodic, the ensemble average also equals the time average of an individual member. We apply this here to the prediction of NMR parameters from the average over the time frames from a MD trajectory, for which we assume that it is long enough for convergence.

The quality of the agreement between predicted average and the experimental data is judged by χ^2

$$\chi^2 = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{d_j^{pred} - d_j^{exp}}{\sigma_j}\right)^2 \tag{3}$$

where $\chi^2 \leq 1$ signifies agreement within the error limits σ_j. The error σ_j for the parameter j in eq 3 presents the total error, e.g., composed of the error of the measurement $\sigma_{j,expt}$ and the error of the model $\sigma_{j,model}$, i.e., $\sigma_j = (\sigma_{j,expt}^2 + \sigma_{j,model}^2)^{1/2}$.

The maximum entropy search problem can now be formulated as the following optimization problem

$$\text{maximize } S(\boldsymbol{p}) \tag{4a}$$

$$\text{subject to } \chi^2(\boldsymbol{p}) \leq 1 \tag{4b}$$

DOI: 10.1021/acs.jctc.5b00759
J. Chem. Theory Comput. 2016, 12, 383−394

$$0 \le p_i \le 1,\ i = 1, \ldots, N \tag{4c}$$

$$\sum_{i=1}^{N} p_i = 1 \tag{4d}$$

Here, the vector $\boldsymbol{p} = (p_1,\ldots,p_N)$ is the optimization variable of the problem. An optimal solution $\boldsymbol{p}^{S\text{-max}}$ is found when the object function $S(\boldsymbol{p})$ has its maximal value among all vectors $\boldsymbol{p}$ that satisfy the inequality constraints (eqs 4b and 4c) and the equality constraint (eq 4d).

A convex optimization problem is one where both the objective function and the inequality constraint functions are convex, whereas the equality constraint functions are affine.[26] A function f is convex when its epigraph, epi f (the set of points above or on the graph of f, epi $f = \{(x,t)|x \in \text{dom } f, f(x) \le t\}$) is a convex set. A set is convex if for any two points of the set the connecting straight line segment between the two points is also in the set (Figure 1A). Thus, the

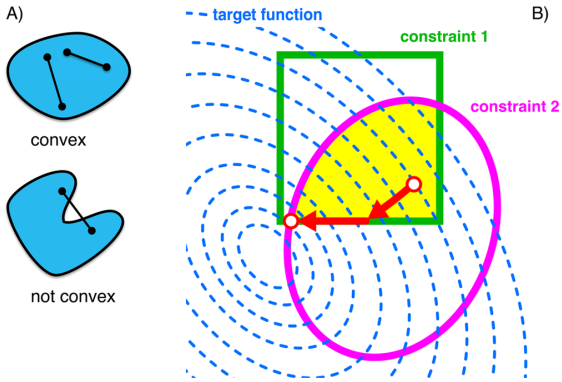


Figure 1. Solution of constrained convex optimization problem by interior point method. (A) Example of convex and nonconvex sets. (B) Illustration of interior point method. The intersection of all convex constraints (constraint 1 (green), constraint 2 (magenta), ...) defines the convex set of feasible points (yellow). The value of the convex objective function is shown by the blue, dashed contour lines. The search starts from an interior point (red circle) within the set of feasible points and follows the gradient of the objective function until the boundary of the set of feasible points is reached. The search then continues along the boundary following the gradient of the objective function until the optimal solution (red circle) is attained (see text).

convex inequality constraints define convex sets of feasible points. Similarly, the affine equality constraints define affine sets. A set is affine if for any two points of the set their entire connecting straight line is also in the set. Since intersections of convex and affine sets are convex, the combined conditions imposed by convex inequality and affine equality constraints define a set of feasible points, which is also convex[26] (Figure 1B). Convex optimization problems can be solved very efficiently by interior point (IP) methods.[26,27] Figure 1B illustrates how the optimal solution can be reached from an interior point within the intersection of the feasible regions of all constraints. Starting from the interior point, the search follows the gradient of the objective function until the boundary of the set of feasible points is reached, from where the search continues along the boundary until the optimal solution is attained. If the optimum is located at an interior point, then the problem reduces to an unconstrained optimization. The convex nature of the objective function ensures that the solution is unique in both cases.

It is easy to show that the negative of the entropy $-S(\boldsymbol{p})$ (eq 4a) and the constraining functions of the inequality constraints (eqs 4b and 4c) are convex since their Hessians are positive semidefinite.

$$\frac{\partial^2}{\partial p_a \partial p_b} - S(\boldsymbol{p}) = \frac{\delta_{ab}}{p_a} \ge 0 \tag{5a}$$

$$\frac{\partial^2}{\partial p_a \partial p_b}\chi^2(\boldsymbol{p}) = \frac{2}{M}\sum_{j=1}^{M}\frac{d_{a,j}^{\text{pred}}}{\sigma_j}\frac{d_{b,j}^{\text{pred}}}{\sigma_j} \tag{5b}$$

where δ_{ab} is the Kronecker delta.

We also note that the constrained χ^2 minimization problem

$$\text{minimize } \chi^2(\boldsymbol{p}) \tag{6a}$$

$$\text{subject to } 0 \le p_i \le 1,\ i = 1, \ldots, N \tag{6b}$$

$$\sum_{i=1}^{N} p_i = 1 \tag{6c}$$

is a convex optimization problem and hence can be solved very efficiently.

In order to find the maximum entropy according to eqs 4 by interior point methods, we use the following two-step procedure:

(1) Search for the population vector $\boldsymbol{p}^{\chi^2\text{-min}}$ that minimizes χ^2 under the constraints of eqs 6b–6c starting from an interior point such as $p_i^{\text{equal}} = 1/N$. If $\chi^2(\boldsymbol{p}^{\chi^2\text{-min}}) \le 1$, then $\boldsymbol{p}^{\chi^2\text{-min}}$ is an interior point for the constrained maximum entropy problem eqs 4; otherwise, it has no solution.

(2) Search for the population vector $\boldsymbol{p}^{S\text{-max}}$ that maximizes the entropy under the constraints of eqs 4b–4d starting from the interior point $\boldsymbol{p}^{\chi^2\text{-min}}$.

Change of Entropy and Free Energy under Reweighting. When no experimental information is present, i.e., the chi-square condition (eq 4b) is dropped from the optimization problem of eqs 4, the maximum entropy is achieved when all populations are equal and $p_i^{\text{equal}} = 1/N$. In this situation, the entropy takes the value $S(\boldsymbol{p}^{\text{equal}}) = \ln(N)$. The change in population weights due to the experimental information under maximum entropy principle leads to a decrease in entropy ΔS from this value

$$\begin{aligned}\Delta S &= S(\boldsymbol{p}^{S\text{-max}}) - S(\boldsymbol{p}^{\text{equal}}) \\ &= -\sum_{i=1}^{N} p_i^{S\text{-max}} \ln p_i^{S\text{-max}} - \ln(N)\end{aligned} \tag{7}$$

This decrease in entropy coincides with the definition of the relative entropy (Kullback–Leibler divergence)[28] ΔS_{AB} of two populations $\boldsymbol{p}^A$ and $\boldsymbol{p}^B$

$$\Delta S_{AB} = -\sum_{i=1}^{N} p_i^A \ln\left(\frac{p_i^A}{p_i^B}\right) \tag{8}$$

for the case $p_i^B = p_i^{\text{equal}} = 1/N$. The negative of the relative entropy ΔS_{AB} presents the mean information $I(A{:}B)$ for

DOI: 10.1021/acs.jctc.5b00759
J. Chem. Theory Comput. 2016, 12, 383–394

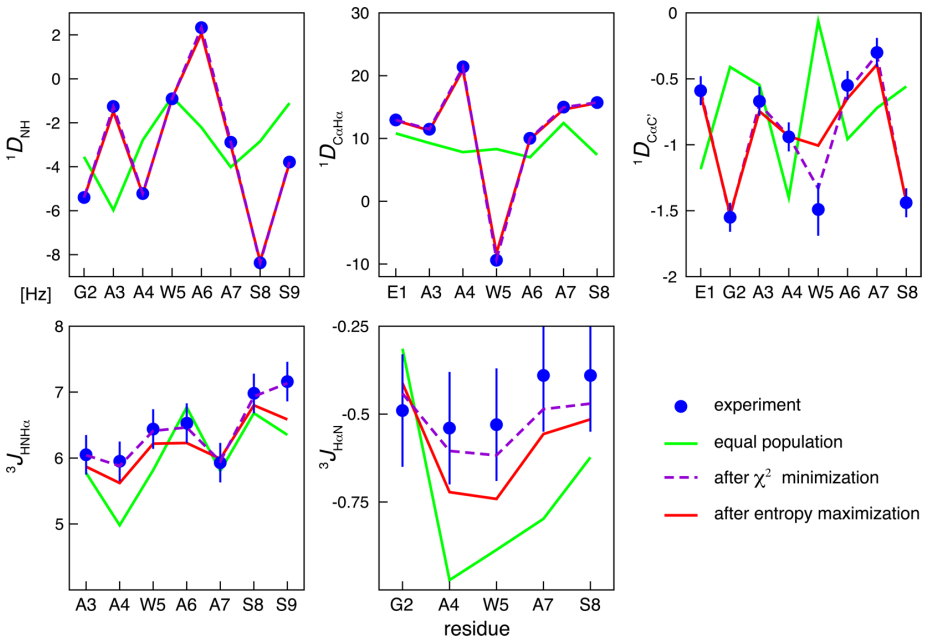


Figure 2. Comparison of experimental RDCs and backbone scalar couplings obtained on the nonapeptide EGAAWAASS to values back-calculated from its 35 000 frame MD trajectory. Experimental data are shown as blue circles, and the unbiased (equal population) average of the predicted observables from the trajectory, as green lines. The reweighted averages after χ^2 minimization and the COPER entropy maximization are indicated as dashed magenta lines and red lines, respectively.

discrimination in favor of $\boldsymbol{p}^A$ against $\boldsymbol{p}^B$. Thus, $-\Delta S$ in eq 7 is the information content of experimental data for discrimination against an equal population.

To quantify the reweighting of individual populations under the experimental constraints and the maximum entropy principle, we define the reweighting factor r_i and its associated free energy change ΔG_i

$$r_i = p_i^{\text{S-max}}/p_i^{\text{equal}} \tag{9a}$$

$$\Delta G_i = -kT\ \ln(r_i) \tag{9b}$$

where k is the Boltzmann constant and T the absolute temperature. Using eqs 7−9, it is obvious that $kT\Delta S$ presents the mean free energy change $\langle \Delta G \rangle$

$$\begin{aligned}\langle \Delta G \rangle &= \sum_{i=1}^{N} p_i^{\text{S-max}} \Delta G_i \\ &= -kT \sum_{i=1}^{N} p_i^{\text{S-max}} \ln\left(\frac{p_i^{\text{S-max}}}{p_i^{\text{equal}}}\right) \\ &= kT\Delta S\end{aligned} \tag{10}$$

RESULTS AND DISCUSSION

Experimental NMR Data and MD Simulations on the EGAAWAASS Nonapeptide. Previously, we had systematically investigated the influence of single amino acid substitutions X on the conformation of unfolded model peptides EGAAXAASS, as monitored by $^1D_{NH}$ and $^1D_{C\alpha H\alpha}$ RDCs, $^3J_{HNH\alpha}$ scalar couplings, and $^{13}C^{\alpha}$ secondary shifts.[23] Homogeneous RDC, chemical shift, and J coupling values along the peptide sequence indicated extended peptide conformations for most amino acid types X. However, substitutions by the aromatic amino acids tryptophan and tyrosine led to a kink in the center of the peptide, as was evident from a discontinuity in the NMR data. The original NMR data were obtained on peptides at natural abundance of ^{13}C and ^{15}N. To obtain access to further NMR parameters, the tryptophan-substituted EGAAWAASS peptide was ^{13}C and ^{15}N isotope-labeled in a bacterial expression system.[29] Figure 2 shows sequential RDC ($^1D_{NH}$, $^1D_{C\alpha H\alpha}$, $^1D_{C\alpha C'}$) and J coupling ($^3J_{HNH\alpha}$, $^3J_{H\alpha N}$) data acquired on this isotope-labeled EGAAWAASS peptide (a complete list of experimental data is provided in Supporting Information, Table S1). The discontinuity is evident in the sequence profile of the $^1D_{NH}$ and $^1D_{C\alpha H\alpha}$ RDCs. Whereas they are negative and positive, respectively, for almost all amino acids, consistent with an extended conformation of the peptide in horizontally compressed polyacrylamide gels,[30] they change sign at the central residues A6 ($^1D_{NH}$) and W5 ($^1D_{C\alpha H\alpha}$), indicative of a kink. Figure 2 also shows experimental statistical error estimates for the J coupling and RDC values (Supporting Information, Table S1). The error estimates for the $^3J_{HNH\alpha}$ and $^3J_{H\alpha N}$ couplings are very close to RMSD values found previously between experimental data and data predicted from structural knowledge by the respective Karplus parameters.[31,32] For the RDC data, a true error estimate is much harder to establish due to the lack of detailed knowledge on the interactions of the peptide with the alignment medium and possible induced conformational changes during this interaction. We have previously observed a similar discontinuity in the RDC pattern with a different alignment medium (Pf1 phages),[23] which indicates that the kink in the peptide is not induced by interactions with the medium. Nevertheless,

DOI: 10.1021/acs.jctc.5b00759
J. Chem. Theory Comput. 2016, 12, 383−394

Table 1. χ^2 and Entropy Values[a] before Reweighting, after χ^2 Minimization, and after Entropy Maximization of the Frames of the EGAAWAASS Nonapeptide's MD Trajectory or Its Flexible-Meccano Data Set Using Different NMR Observables as Constraints

		before reweighting[b]	after χ^2 minimization		after entropy maximization			
constraints	N[c]	$\sum_\alpha \chi^2_\alpha$[d]	$\sum_\alpha \chi^2_\alpha$	ΔS[e]	$\sum_\alpha \chi^2_\alpha$	ΔS[e]	ΔS average[f]	ΔS SD[f]
$^3J_{HNH\alpha}$	7	3.52	0.00	−0.201	1.00	−0.045	−0.044	0.000
$^3J_{H\alpha N}$	5	4.41	0.00	−0.878	1.00	−0.215	−0.199	0.006
$^3J_{NC\gamma}$	1	4.42	0.00	−0.043	0.00	−0.043	−0.008	0.001
$^3J_{C'C\gamma}$	1	11.77	0.00	−0.277	0.98	−0.195	−0.210	0.070
$^1D_{NH}$	8	498.97	0.00	−1.415	1.00	−0.668	−0.652	0.008
$^1D_{C\alpha H\alpha}$	7	383.92	0.00	−0.905	1.00	−0.540	−0.540	0.013
$^1D_{C\alpha C'}$	8	35.56	0.01	−1.387	1.00	−0.459	−0.459	0.000
all backbone ^3J[g]	12	4.41	0.00	−1.037	1.00	−0.206	−0.206	0.006
all ^1D[h]	23	959.94	0.31	−4.194	3.01	−2.297	−2.311	0.068
all ^1D + all backbone ^3J[i]	35	967.87	0.37	−4.870	4.94	−2.512	−2.581	0.158
all ^1D + all backbone ^3J (FM ensemble)[j]	34	1158.08	0.98	−5.738	4.94	−3.611	−3.685	0.035
all ^1D + all ^3J[k]	37	1024.19	0.20	−5.203	6.82	−2.633	−2.656	0.057

[a]Unless noted otherwise, all values correspond to a calculation of the 35 000 frame MD data set. [b]The entropy value of the equal population distribution for the 35 000 conformations is $S(\boldsymbol{p}^{equal}) = \ln(35\,000) = 10.463$. [c]Number of experimental constraints. [d]The values correspond to the sum of individual χ^2 values for the different data types. [e]The entropy difference is calculated as the deviation from $S(\boldsymbol{p}^{equal})$. [f]The 35 000 conformation data set was randomly divided into two mutually exclusive data sets, each containing 17 500 conformations. The calculation was repeated on both data sets, and entropy differences were calculated as the deviation from $S(\boldsymbol{p}^{equal}) = \ln(17\,500) = 9.770$. ΔS average (SD) corresponds to the average (standard deviation) of both entropy differences. [g]The constraints consist of the backbone scalar couplings $^3J_{HNH\alpha}$ and $^3J_{H\alpha N}$. [h]The constraints consist of the RDCs $^1D_{NH}$, $^1D_{C\alpha H\alpha}$, and $^1D_{C\alpha C'}$. [i]The constraints consist of $^3J_{HNH\alpha}$, $^3J_{H\alpha N}$, $^1D_{NH}$, $^1D_{C\alpha H\alpha}$, and $^1D_{C\alpha C'}$. [j]The calculation was carried out on a 35 000 conformation ensemble generated by Flexible-Meccano using $^3J_{HNH\alpha}$, $^3J_{H\alpha N}$, $^1D_{NH}$, $^1D_{C\alpha H\alpha}$, and $^1D_{C\alpha C'}$ constraints. Due to the nature of the Flexible-Meccano simulation, the φ angle of the last residue is fixed and its $^3J_{HNH\alpha}$ constraint is not meaningful. [k]The constraints consist of $^3J_{HNH\alpha}$, $^3J_{H\alpha N}$, $^3J_{NC\gamma}$, $^3J_{C'C\gamma}$, $^1D_{NH}$, $^1D_{C\alpha H\alpha}$, and $^1D_{C\alpha C'}$.

the RDC model error is unknown for flexible peptides. Since better estimates are not available, the total RDC error was taken as the experimental error.

To identify the structural reason for the kink in the peptide, we have carried out a total of seven, 100 ns MD simulations on the EGAAWAASS peptide under full hydration. Theoretical RDCs and J couplings were then calculated for every 20 ps frame using a steric exclusion model[6] and available Karplus parameters.[31,32] Figure 2 shows the equally weighted averages of the different observables over the total of 35 000 conformations (green solid lines). Clearly, the unbiased averages do not reproduce the experimental data (blue) within the indicated error, which is particularly noticeable for the kinks observed in the experimental RDC data in the region around residues W5 and A6. These deviations lead to a total χ^2 value (eq 3) of 116.

COPER Procedure: χ^2 Minimization Followed by Entropy Maximization. The COPER procedure was then applied to reweight the individual conformations. In practice, using a single total ($\chi^2 \leq 1$) constraint for all different data types in the maximum entropy search led to a very uneven distribution of deviations among the different RDC and J coupling data types. Therefore, we used individual $\chi^2_\alpha \leq 1$ constraints for each of the different data types α (RDC or J coupling). To find a feasible inner point for this maximum entropy search, the initial χ^2 minimization was then carried out on the sum $\sum_\alpha \chi^2_\alpha$, which differs from the original χ^2 definition in eq 3 only by reweighting via the number of data points M_α in the individual data sets. Minimization of $\sum_\alpha \chi^2_\alpha$ within the usual constraints on population weights (eqs 6b and 6c) of the 35 000 conformations led to very good agreement of the average data predicted from the minimizing population vector $\boldsymbol{p}^{\chi^2\text{-min}}$ with the experimental data (Figure 2, dashed magenta lines). As compared to the equal population entropy $S(\boldsymbol{p}^{equal}) = \ln(35\,000) = 10.46$, the entropy for the $\boldsymbol{p}^{\chi^2\text{-min}}$ vector is significantly reduced to a value $S(\boldsymbol{p}^{\chi^2\text{-min}}) = 5.59$ ($\Delta S = -4.87$; Table 1).

The minimized $\sum_\alpha \chi^2_\alpha$ value of 0.37 (Table 1) guarantees that the individual χ^2_α values are also smaller than 1; hence, $\boldsymbol{p}^{\chi^2\text{-min}}$ presents a feasible starting point for the maximum entropy search within the χ^2 and population constraints (eqs 4b–4d). The subsequent maximum entropy search then yielded average predicted data that agree less well with the experimental data than the minimal χ^2 prediction, but the data are still within the error limits (Figure 2, red lines). Consequently, the entropy is again increased to a value $S(\boldsymbol{p}^{S\text{-max}}) = 7.95$, but it is still reduced relative to the equal population situation by $\Delta S = -2.51$. This reduction in entropy corresponds to the minimal restriction of the accessible conformational space needed to satisfy the experimental information. In the simplest case, it may be pictured as making certain conformations completely inaccessible, whereas the accessible conformations remain equally likely. In this situation, $S(\boldsymbol{p}^{S\text{-max}})$ would correspond to the logarithm of the number of accessible conformations; hence, $e^{-\Delta S}$ (12.3 for the current case) would correspond to the factor by which the number of accessible conformations is reduced due to the experimental information.

Due to the efficiency of the inner point method, the entire COPER procedure of constrained χ^2 minimization followed by constrained entropy maximization took only about 9 min to complete on a single core of a 2.6 GHz Intel Xeon CPU for the 35 000 member EGAAWAASS peptide ensemble. Tests with different ensemble sizes showed that this time increased approximately linearly with ensemble size for ensembles of up to 70 000 members.

DOI: 10.1021/acs.jctc.5b00759
J. Chem. Theory Comput. 2016, 12, 383−394

Robustness Test, Effect of Error Uncertainty, and Information Content of Individual Data Types. To estimate the information content of different types of NMR constraints on the MD ensemble, we systematically determined the entropy loss induced by these constraints via COPER fitting relative to the equilibrium population. Table 1 lists these losses for different combinations of scalar and dipolar couplings. In order to estimate the errors and robustness of the method, the set of 35 000 conformations from the MD trajectory was further subdivided into two randomly chosen subsets of 17500 conformations, for which the COPER fit procedure was repeated and the entropy was calculated. Table 1 lists the average and standard deviation values of the resulting entropy reductions for the two subpopulations relative to their equal population entropy $S(\boldsymbol{p}^{\text{equal}}) = \ln(17\,500) = 9.77$. It is obvious that the entropy losses are highly reproducible, with relative standard deviations of less than 7%, and very close to the losses calculated for the 35 000 conformation data set. This indicates that the sampling of the conformational space is dense, since significant variations of the entropy reduction would be expected for a sampling of conformational space that is too low. In this manner, the comparison of the entropy reduction within different subsets of an ensemble provides both a test for the robustness of the reweighting and for the density of sampling.

To assess the effect of the errors used on the total entropy reduction, we have also varied the limits for χ^2_α from 0.25 to 4, corresponding to a scaling of the errors by factors between 0.5 and 2 (Supporting Information, Figure S1A). Consistent with the expectation that weaker constraints allow larger conformational entropy and with previous findings by Hummer and colleagues,[11] the entropy reduction decreases monotonously with increasing χ^2_α limits. A limit of $\chi^2_\alpha \leq 4$ decreases the entropy reduction to 1.75 from its value of 2.51 for $\chi^2_\alpha \leq 1$. Thus, a 2-fold increase of the error size has an effect of less than 1 kT unit on the free energy change.

The entropy reduction induced by the experimental constraints may range from zero, for which the reweighted population is identical to the equally populated state, to $\ln(N)$, for which the constrained ensemble reduces to a single conformation (equivalent to a final entropy value of zero). The entropy losses due to $^3J_{\text{HNH}\alpha}$ (ϕ angle) or $^3J_{\text{H}\alpha\text{N}}$ (ψ angle) constraints are 0.05 and 0.22, respectively (Table 1). Thus, the $^3J_{\text{HNH}\alpha}$ data carry about four times less information than the $^3J_{\text{H}\alpha\text{N}}$ data. This is in agreement with the fact that among the different conformations accessible to the polypeptide, i.e., helical vs extended, variations in ϕ angle are much smaller than in ψ-angle. The entropy losses for individual $^1D_{\text{NH}}$, $^1D_{\text{C}\alpha\text{H}\alpha}$, and $^1D_{\text{C}\alpha\text{C}'}$ constraints are 0.67, 0.54, and 0.46 respectively, which indicates a significantly higher information content of the dipolar couplings relative to the scalar couplings. Combining all three dipolar couplings constraints increases the entropy loss to 2.30, which is a more than additive effect on the restriction of conformational space. Finally, when both dipolar and scalar coupling constraints are applied simultaneously, the total entropy loss of 2.51 approximately equals the sum of their individual contributions. This shows that the dipolar and scalar coupling constraints each contain information not captured by the other data type. Since the entropy loss times the thermal energy kT represents the mean free energy change (eq 10), an adjustment of the MD force field by 2.51 kT units would be necessary to bring the ensemble in agreement with the experiment.

Comparison to Flexible-Meccano Random Coil Ensemble. The entropy reduction presents a measure of the accuracy of the model ensemble. This can be used to quantitatively compare different types of ensembles. For this, we created a 35 000 conformation ensemble based on a random coil model of the unfolded with residue-specific ϕ/ψ propensities using the program Flexible-Meccano.[8] Reweighting its populations by COPER using the same backbone J coupling and RDC constraints as those for the MD ensemble caused an entropy reduction by 3.61 (Table 1), which is more than the corresponding value of 2.51 for the MD ensemble and consistent with the considerably coarser nature of approximation used in Flexible-Meccano.

It is interesting to note that the $\sum\chi^2_\alpha$ difference between the experimental and back-calculated data for the MD ensemble before reweighting amounts to 968 and is therefore only slightly smaller than the corresponding $\sum\chi^2_\alpha$ value of 1158 for the Flexible-Meccano ensemble (Table 1). However, the minimized $\sum\chi^2_\alpha$ decreases to 0.37 for the MD, but it decreases only to 0.98 for the Flexible-Meccano ensemble. Thus, the MD ensemble contains conformations that, as population-weighted combinations, better represent the experimental data than the Flexible-Meccano ensemble. Since the $\sum\chi^2_\alpha$ minimum is lower for the MD ensemble, it is expected that the $\chi^2_\alpha \leq 1$ conditions lead to a larger allowed space for the population weights and, as a consequence, to a smaller reduction in entropy.

Cross-Validation of COPER ME Populations. Cross-validation of ME-derived populations with additional, independent experimental data is problematic, since the ME solution is, per definition, underdetermined. If the predictions agree with the additional data, then their information is redundant and they would not have constrained the original fit. In contrast, if the additional data deviate from the predictions of the original fit, then they contain independent information. Using COPER, the information content of the additional data can be estimated from the entropy reduction that results from including these data in the fit.

Using this quantitative concept, populations of the 35 000 frame MD data set obtained by the COPER ME fit of the $^1D_{\text{NH}}$, $^1D_{\text{C}\alpha\text{H}\alpha}$, $^1D_{\text{C}\alpha\text{C}'}$, $^3J_{\text{HNH}\alpha}$, $^3J_{\text{H}\alpha\text{N}}$ couplings (Figure 2) were cross-validated by experimentally determined χ_1 angle populations of the W5 side chain. Assuming staggered conformers, the χ_1 angle populations were derived by a simple linear transformation (linear least-squares fit) from experimental $^3J_{\text{NC}\gamma}$ and $^3J_{\text{C}'\text{C}\gamma}$ couplings[33,34] (Supporting Information, Table S1), which had not been used as constraints (Figure 3). These experimental populations of the $\chi_1 + 60°$, $\chi_1 + 180°$, and $\chi_1 - 60°$ rotamers are 22, 46, and 31%, respectively. The χ_1 populations from the MD simulation (58, 34, 6%) deviate strongly, but they get closer (44, 45, 12%) to the experimental values after COPER reweighting by the $^1D_{\text{NH}}$, $^1D_{\text{C}\alpha\text{H}\alpha}$, $^1D_{\text{C}\alpha\text{C}'}$, $^3J_{\text{HNH}\alpha}$, and $^3J_{\text{H}\alpha\text{N}}$ data, thereby confirming the correct trend of the independent fit. Obviously, including the $^3J_{\text{NC}\gamma}$ and $^3J_{\text{C}'\text{C}\gamma}$ scalar couplings in the COPER procedure leads to the best agreeing χ_1 populations (31, 40, 28%) at a cost of reducing the entropy by 2.63 relative to the equal population situation (Table 1). However, this reduction is only 0.12 larger than for the fit without the side chain $^3J_{\text{NC}\gamma}$ and $^3J_{\text{C}'\text{C}\gamma}$ scalar couplings. Therefore, their additional information content is rather small

DOI: 10.1021/acs.jctc.5b00759
J. Chem. Theory Comput. 2016, 12, 383−394

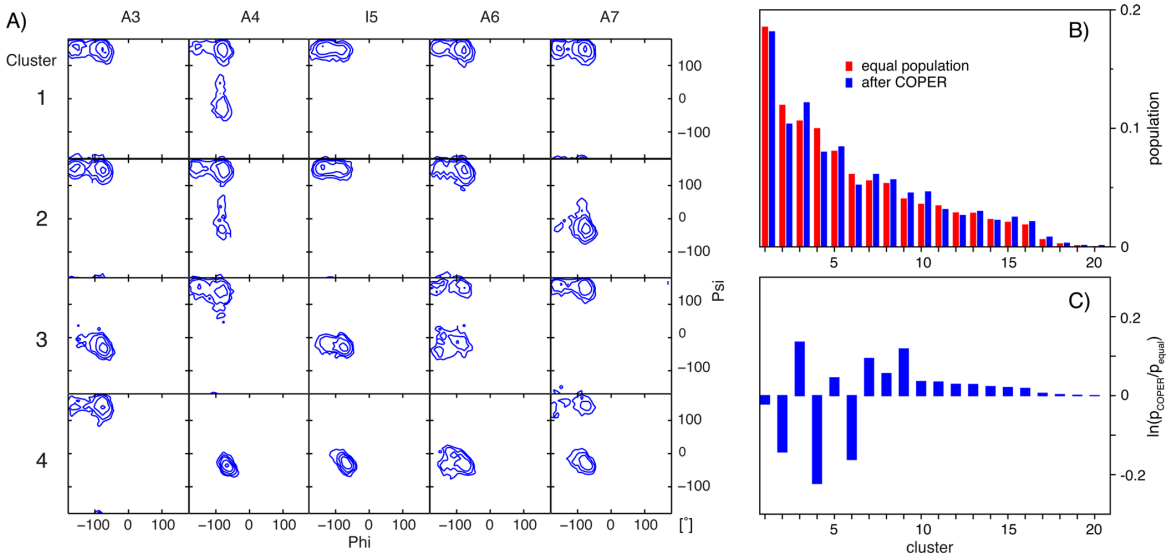


Figure 6. Analysis of the 10 000 conformations from the MD trajectory of the EGAAIAASS peptide. The conformations were clustered into 20 subsets according to the φ/ψ angles of its central five residues (A3–A7). (A) Ramachandran population plots of the four most highly populated clusters are shown with contour levels spaced by a factor of 2.5. The most highly populated cluster 1 has extended conformations for residues A3 and I5 to A7 and partially α-helical conformations for residue A4. (B) Populations of the 20 clusters before reweighting are shown in red, and those after COPER reweighting, in blue. (C) Reweighting factors for the cluster populations are indicated as $\ln(p_{\mathrm{COPER}}/p_{\mathrm{equal}})$, where p_{equal} and p_{COPER} represent the populations before and after COPER reweighting, respectively.

Thus, Hummer and colleagues[11] use the free energy function $G = \chi^2 - \Theta S$, where Θ is a tunable temperature parameter that balances the agreement between experimental and back-calculated data with conformational diversity. Θ is then varied until the corresponding free energy change matches an expected error in the force field. In COPER, we define the error of the parameters, i.e., χ^2, and obtain as a result the entropy change ΔS and the concomitant free energy change ΔG. Thus, as shown in Supporting Information Figure S1, the relation between ΔS and χ^2 can be easily established. This relation can also be used to achieve a certain ΔG ($=kT\Delta S$) that matches expected force field errors. Using our error estimates, the free energy changes of less than $3kT$ were in the range of the expected force field errors.

In contrast to the maximum entropy approaches, the Bayesian[22] ensemble reweighting algorithm determines the population weights from assumed prior distributions of the weights and likelihood functions of the parameters based on experimental, theoretical, or assumed errors. This approach also provides estimates of the uncertainties in the weights, which are not easily obtained by other methods. However, the computational cost is rather high, and, so far, it has been applied only to small ensembles of hundreds of conformations.

Extension of the Inner Point Convex Optimization to Maximum Occurrence. Bertini and colleagues have previously introduced the method of maximum occurrence (MO)[10] for the analysis of ensembles of flexible macromolecules. The method tries to determine the maximum time or occurrence that a molecule can spend in a given conformation k such that the weighted average over all conformations of a theoretical ensemble is still compatible with the experimental average data. The problem can thus be formulated as

$$\text{maximize } p_k \tag{11a}$$

$$\text{subject to } \chi^2(\boldsymbol{p}) \leq 1 \tag{11b}$$

$$0 \leq p_i \leq 1,\ i = 1, \ldots, N \tag{11c}$$

$$\sum_{i=1}^{N} p_i = 1 \tag{11d}$$

where the populations p_i and the constraining function χ^2 are defined as in eqs 4. Previously, this problem could be solved only by using a nondeterministic, simulated annealing procedure on smaller subsets (480 families of 50 members) of a large ensemble (56 000 structures).[10] However, since the target function p_k (eq 11a) and the constraints (eq 11b–11d) are convex or affine, the entire problem is a convex optimization problem that can be solved efficiently by the described inner point method.

While it is beyond the scope of the present work to perform a detailed analysis of the EGAAWAASS peptide conformations by the MO method, we have tested the efficiency of this inner point solution to the MO problem on ensembles of random conformations generated for this peptide by the program Flexible-Meccano.[8] The ensembles ranged in size from 10 000 to 70 000 members and were subjected to the MO optimization using the experimental RDC and J coupling backbone constraints described in Figure 2. The CPU time necessary to calculate one MO population increased approximately linearly with the ensemble size and amounted to 850 s on a single core of a 2.6 GHz Intel Xeon CPU for the 70 000 member ensemble. This compares very favorable with the 6 h reported previously for subsets of a 56 000 member ensemble.[10]

DOI: 10.1021/acs.jctc.5b00759
J. Chem. Theory Comput. 2016, 12, 383–394

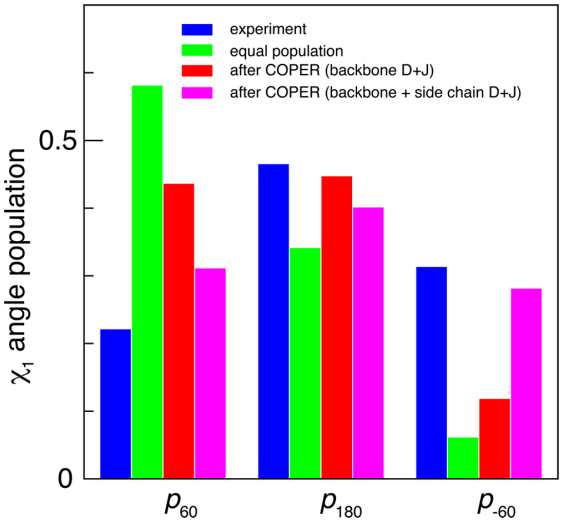


Figure 3. Cross-validation of the COPER-reweighted populations by χ_1 rotamer populations determined independently from $^3J_{NC\gamma}$ and $^3J_{C'C\gamma}$ scalar couplings for the side chain of W5 in the EGAAWAASS peptide. Experimentally determined populations are shown in blue, unbiased populations from the 35 000 frame MD trajectory, in green, COPER-reweighted populations according to backbone RDCs and J couplings ($^1D_{NH}$, $^1D_{C\alpha H\alpha}$, $^1D_{C\alpha C'}$, $^3J_{HNH\alpha}$, $^3J_{H\alpha N}$), in red, and COPER-reweighted populations according to backbone and side chain RDCs and J couplings ($^1D_{NH}$, $^1D_{C\alpha H\alpha}$, $^1D_{C\alpha C'}$, $^3J_{HNH\alpha}$, $^3J_{H\alpha N}$, $^3J_{NC\gamma}$, $^3J_{C'C\gamma}$), in magenta.

and reduces the conformational space only by an additional 11%.

Structural Interpretation by φ/ψ Cluster Analysis. To obtain structural insights into the effects of the ME reweighting, the 35 000 conformations were clustered into 20 clusters based on the similarity of the ϕ and ψ torsion angles of the central five residues (A3−A7) using a hierarchical clustering algorithm. The clusters were then ordered according to the size of their populations in the original MD trajectory. Figure 4A shows the ϕ/ψ angle distributions of the four most highly populated clusters, accounting for 66% of all conformations. The largest cluster 1 has α-helical conformations for residues A4−A6 and partially α-helical conformations for residues A3 and A7, whereas the other clusters contain more extended conformations. A representative set of conformations of cluster 1 is shown in Figure 4B. It is obvious that residues A3−A7 form a turn with backbone hydrogen-bond contacts. These contacts are protected from water by the bulky aromatic side chain of residue W5, as shown in a recent analysis[24] of the full MD trajectory, which explains the tendency of the aromatic groups to induce kinks in the unfolded peptide chain.

Figure 5A shows the 20 cluster populations before and after reweighting by the $^1D_{NH}$, $^1D_{C\alpha H\alpha}$, $^1D_{C\alpha C'}$, $^3J_{HNH\alpha}$, and $^3J_{H\alpha N}$ COPER fit. Before reweighting, cluster 1 has a population of about 33%, whereas the other clusters have populations of less than 12%. After reweighting, the population of cluster 1 decreases significantly to about 15%, the population of cluster 2 decreases, and those of 3 and 4 increase. The rest of the cluster populations remain below 10%. To test the statistical significance of this result, the cluster populations after reweighting were also determined for the two randomly selected subsets of 17 500 conformations. Figure 5A also

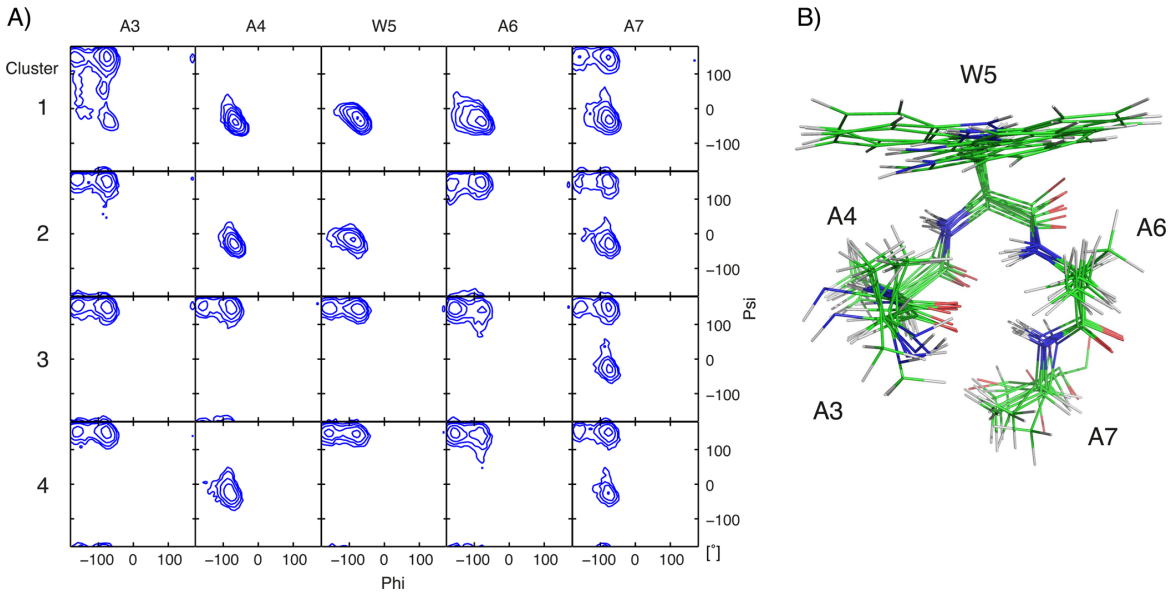


Figure 4. Clustering of the 35 000 conformations from the MD trajectory of the EGAAWAASS peptide according to the φ/ψ angles of its central five residues (A3−A7). (A) Ramachandran population plots of the four most highly populated clusters are shown with contour levels spaced by a factor of 2.5. The most highly populated cluster 1 has α-helical conformations for residues A4−A6 and partially α-helical conformations for residues A3 and A7. (B) Overlay of eight representative conformations from cluster 1 where residues A3−A7 form a helical turn. Backbone hydrogen contacts in this turn are shielded from external water by the side chain of W5.[24]

DOI: 10.1021/acs.jctc.5b00759
J. Chem. Theory Comput. 2016, 12, 383−394

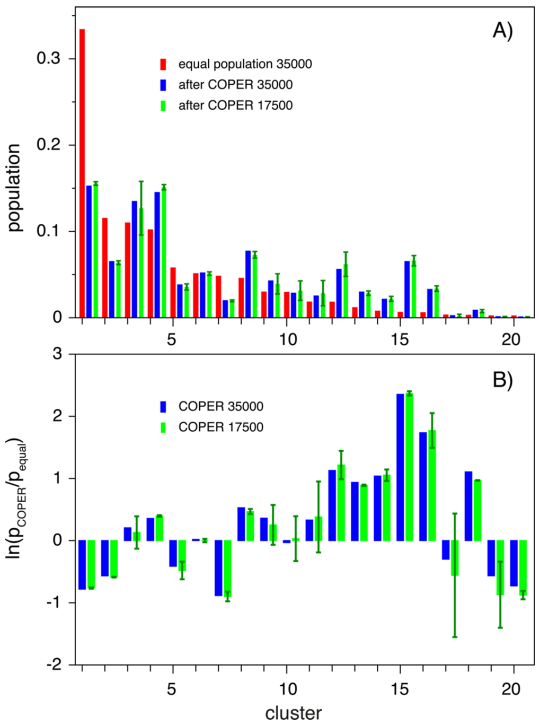


Figure 5. Reweighting of populations for the 20 clusters from the 35 000 MD conformations of the EGAAWAASS peptide. (A) Populations of the clusters before reweighting are shown in red, and those after COPER reweighting, in blue. For testing robustness, the 35 000 conformations were split into two 17 500 conformation sets. Averages and standard deviations of the cluster populations of these two subsets after COPER reweighting are shown in green. (B) Reweighting factors for the cluster populations indicated as $\ln(p_{COPER}/p_{equal})$, where p_{equal} and p_{COPER} represent the populations before and after COPER reweighting, respectively. Data for the COPER analysis of the 35 000 conformations and of the two 17 500 conformation subsets are shown in blue and green, respectively.

shows their averages and standard deviations. The maximal standard deviation of populations in the 17 500 conformation sets is only 3%, and their averages agree within this limit to the results from the 35 000 conformation set. Thus, the reproducibility of the COPER-derived populations is very high.

We have also assessed the effect of the used errors on the cluster populations by varying the limits for χ^2_α from 0.25 to 4 (Supporting Information, Figure S1B). Again, as for the induced entropy changes, the cluster populations vary monotonously with the χ^2_α limits. This is a further indication of the robustness of the results. For a change of the χ^2_α limits from 1 to 4, the populations for most clusters vary by less than 2-fold, with cluster 1 always remaining the dominant cluster.

The reduction of the population of cluster 1 caused by the COPER reweighting with experimental data indicates an overestimation of helical content by the AMBER03 force field, which is in agreement with findings by Best, Lindorff-Larssen, and colleagues.[15,35] It is noted that this reduction of the helical cluster 1 is stronger than that reported previously.[24] This is caused by the $^3J_{H\alpha N}$ couplings that were not present in the previous study, which increased the content of extended conformations. However, even after reweighting by COPER with these additional data, helical cluster 1 remains the most highly populated cluster, albeit closely followed by clusters 3 and 4 (Figure 5A).

The relative changes in the cluster populations due to the COPER reweighting are shown in Figure 5B as $\ln(p_{COPER}/p_{equal})$, where p_{equal} and p_{COPER} represent the cluster populations before and after reweighting, respectively. Values for $\ln(p_{COPER}/p_{equal})$ range between about −0.9 and 2.4, corresponding to errors on the order of less than $3kT$ in the free energy of the individual clusters.

Results for the EGAAIAASS Nonapeptide. As indicated, in contrast to the kinked form of EGAAXAASS peptides with aromatic amino acids X in their center, peptides with other amino acids besides proline and glycine showed extended conformations from the sequence profile of their NMR parameters.[23] We further tested the reweighting of a 10 000 conformation trajectory of the prototypical extended EGAAIAASS peptide, for which the published $^1D_{NH}$, $^1D_{C\alpha H\alpha}$, and $^3J_{HNH\alpha}$ values were used as input for the COPER ME method. As for EGAAWAASS, the conformations were clustered into 20 clusters based on the ϕ and ψ torsion angles of residues A3−A7. Figure 6A shows the ϕ/ψ distributions of the four most highly populated clusters before reweighting. In this case, the most highly populated (18%) cluster 1 has almost completely extended conformations, with only a slight admixture of helical conformations for residue A4. Clusters 2−4 have about 10% populations and are mostly extended (cluster 2), mixed extended/helical (cluster 3), and mostly helical (cluster 4). COPER reweighting reduced the total χ^2 value from 8.8 to 1.0, but it changed the individual cluster populations by less than 2% (Figure 6B). Accordingly, the reweighting factors $\ln(p_{COPER}/p_{equal})$ ranged only from about −0.2 to 0.1 (Figure 6C), showing that the free energy adjustment is less than 0.2 kT. The total entropy loss due to the reweighting was only 0.11. Apparently, the AMBER03 force field in conjunction with the TIP4P water model reproduced the extended conformations of the EGAAIAASS peptide almost quantitatively, whereas it significantly exaggerated the more helical conformations of the EGAAWAASS peptide.

Comparison of COPER with Other Ensemble Reweighting Algorithms. The COPER approach may be compared to the previously proposed maximum entropy[11,21] and Bayesian[22] ensemble reweighting algorithms. These previous methods all contained nondeterministic random sampling algorithms and were limited to smaller subsets (at most several thousand structures) from computed ensembles of tens of thousands of structures. In contrast, due to the efficiency of the inner point convex optimization method and the use of only gradients of the objective and constraining functions,[36] COPER can calculate globally optimized weights in a very efficient, numerically stable, and deterministic manner for very large ensembles of, so far, up to 70 000 structures. We note that this limit is dictated rather by numerical precision than by computational speed.

Besides this advantage in efficiency and the well-defined nature of the solution, the underlying mathematical target of COPER also differs from that in previous approaches. The described maximum entropy approaches[11,21] minimize a free energy, in which an entropy term was subtracted from χ^2.

DOI: 10.1021/acs.jctc.5b00759
J. Chem. Theory Comput. 2016, 12, 383−394

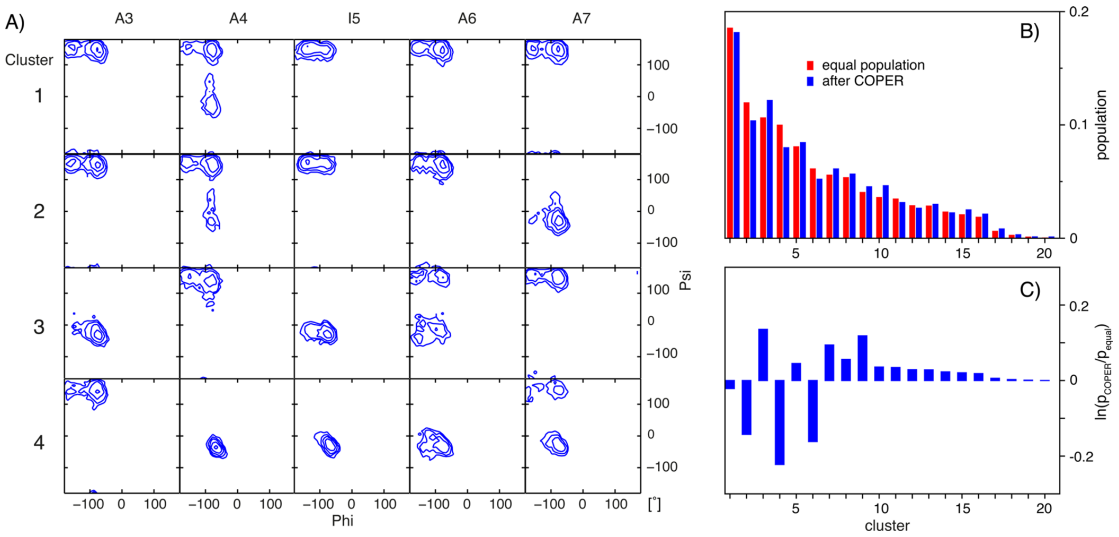


Figure 6. Analysis of the 10 000 conformations from the MD trajectory of the EGAAIAASS peptide. The conformations were clustered into 20 subsets according to the φ/ψ angles of its central five residues (A3−A7). (A) Ramachandran population plots of the four most highly populated clusters are shown with contour levels spaced by a factor of 2.5. The most highly populated cluster 1 has extended conformations for residues A3 and I5 to A7 and partially α-helical conformations for residue A4. (B) Populations of the 20 clusters before reweighting are shown in red, and those after COPER reweighting, in blue. (C) Reweighting factors for the cluster populations are indicated as $\ln(p_{COPER}/p_{equal})$, where p_{equal} and p_{COPER} represent the populations before and after COPER reweighting, respectively.

Thus, Hummer and colleagues[11] use the free energy function $G = \chi^2 - \Theta S$, where Θ is a tunable temperature parameter that balances the agreement between experimental and back-calculated data with conformational diversity. Θ is then varied until the corresponding free energy change matches an expected error in the force field. In COPER, we define the error of the parameters, i.e., χ^2, and obtain as a result the entropy change ΔS and the concomitant free energy change ΔG. Thus, as shown in Supporting Information Figure S1, the relation between ΔS and χ^2 can be easily established. This relation can also be used to achieve a certain ΔG ($=kT\Delta S$) that matches expected force field errors. Using our error estimates, the free energy changes of less than $3kT$ were in the range of the expected force field errors.

In contrast to the maximum entropy approaches, the Bayesian[22] ensemble reweighting algorithm determines the population weights from assumed prior distributions of the weights and likelihood functions of the parameters based on experimental, theoretical, or assumed errors. This approach also provides estimates of the uncertainties in the weights, which are not easily obtained by other methods. However, the computational cost is rather high, and, so far, it has been applied only to small ensembles of hundreds of conformations.

Extension of the Inner Point Convex Optimization to Maximum Occurrence. Bertini and colleagues have previously introduced the method of maximum occurrence (MO)[10] for the analysis of ensembles of flexible macromolecules. The method tries to determine the maximum time or occurrence that a molecule can spend in a given conformation k such that the weighted average over all conformations of a theoretical ensemble is still compatible with the experimental average data. The problem can thus be formulated as

$$\text{maximize } p_k \tag{11a}$$

$$\text{subject to } \chi^2(\boldsymbol{p}) \leq 1 \tag{11b}$$

$$0 \leq p_i \leq 1,\ i = 1, \ldots, N \tag{11c}$$

$$\sum_{i=1}^{N} p_i = 1 \tag{11d}$$

where the populations p_i and the constraining function χ^2 are defined as in eqs 4. Previously, this problem could be solved only by using a nondeterministic, simulated annealing procedure on smaller subsets (480 families of 50 members) of a large ensemble (56 000 structures).[10] However, since the target function p_k (eq 11a) and the constraints (eq 11b−11d) are convex or affine, the entire problem is a convex optimization problem that can be solved efficiently by the described inner point method.

While it is beyond the scope of the present work to perform a detailed analysis of the EGAAWAASS peptide conformations by the MO method, we have tested the efficiency of this inner point solution to the MO problem on ensembles of random conformations generated for this peptide by the program Flexible-Meccano.[8] The ensembles ranged in size from 10 000 to 70 000 members and were subjected to the MO optimization using the experimental RDC and J coupling backbone constraints described in Figure 2. The CPU time necessary to calculate one MO population increased approximately linearly with the ensemble size and amounted to 850 s on a single core of a 2.6 GHz Intel Xeon CPU for the 70 000 member ensemble. This compares very favorable with the 6 h reported previously for subsets of a 56 000 member ensemble.[10]

DOI: 10.1021/acs.jctc.5b00759
J. Chem. Theory Comput. 2016, 12, 383−394

CONCLUSIONS

We have presented the ME method COPER using inner point convex optimization to reweight large simulated conformational data sets by average experimental data. Compared to previous methods, COPER can analyze full, very large ensembles of 10^4 to 10^5 conformers, not just smaller subsets thereof, in a deterministic, fast, and robust manner. The convex optimization guarantees a global unique optimal solution and, hence, a reliable determination of the final population weights for the full ensembles. Therefore, such reweighted ensembles constitute the best representation of the information contained in both the simulated ensemble and the experimental data. Since the final entropy is determined reliably, its loss relative to the unconstrained ensemble can be used as a quantitative measure of the information content of experimental data relative to the theoretical ensemble. A large reduction in entropy will indicate that the theoretical ensemble is not a good representative of the real-world situation and, hence, that the simulation needs to be improved. However, the measure can also be used to judge the information content of individual data types, e.g., a comparison of the entropy reduction induced by the different NMR data types clearly revealed the much higher information content of RDCs relative to three-bond J couplings. Furthermore, the reproducibility of the entropy reduction on different subsets of the large ensembles provides an estimate for its density of sampling of the conformational space. Thus, if the reproducibility becomes low, then a larger number of structures needs to be generated in the initial ensemble to cover the space adequately.

The application to the reweighting of the MD trajectories of small peptides by NMR data showed that the AMBER03 force field overestimated the helical content for the turn-forming EGAAWAASS peptide but not for the extended EGAAIAASS peptide. The reduction in entropy was, in all cases, smaller than 3, indicating that adjustments of the force field of less than 3 kT units would be needed to bring the MD trajectory into agreement with the experimental data. An ensemble created by the Flexible-Meccano statistical coil model of the EGAAWAASS peptide needed stronger reweighting than the MD-derived ensemble to fit the experimental data, consistent with the cruder nature of this model. Eventually, such COPER-reweighted populations may be used via projection onto some essential coordinates to improve existing MD force fields by free energy perturbation methods.[37] Compared to pure χ^2 minimization for force field optimization,[38,39] this may have the advantage of reducing the risk of overfitting,[40] since the entropy is maximized.

While the application of COPER was shown here for average NMR data, it is, in fact, applicable to any experimental average data that can be predicted from a set of molecular conformations, such as small-angle X-ray scattering[41] or Förster resonance energy transfer[42] data. Furthermore, convex optimization can provide similar advantages of well-defined, robust solutions and large-scale efficiency for other ensemble analysis methods with convex target functions and constraints such as constrained χ^2 minimization and MO.[10]

MATERIALS AND METHODS

Sample Preparation. Uniformly $^{15}N/^{13}C$-labeled peptide EGAAWAASS was prepared by expression in *Escherichia coli* as a C-terminal fusion with the immunoglobulin-binding domain of streptococcal protein G as described previously.[43] The peptide was cleaved bluntly from the fusion by factor Xa. NMR samples were prepared as 1 mM (0.25 mM) peptide, 25 mM acetate, pH 4.5, in 5/95% D_2O/H_2O for measurement under isotropic (anisotropic) conditions. Residual alignment of peptides was achieved by introducing the peptide solutions into 10% (w/v) polyacrylamide gels and horizontal compression.[44,45]

NMR Experiments. All NMR experiments were carried out at 298 K on a Bruker Advance III 600 MHz spectrometer equipped with a TXI probe. Spectra were processed using NMRPipe.[46] $^3J_{NH\alpha}$ couplings were obtained from a quantitative-J version of the $^3J_{NH\beta}$-HNHB experiment using a 27 ms $^{15}N-^1H^\alpha$ dephasing delay.[34,47] The resonance line shapes were fitted with the NLINLS program contained in NMRPipe, and $^3J_{NH\alpha}$ coupling constants were determined from the ratios of cross and reference peak heights as described.[34] The $^3J_{HNH\alpha}$ values were taken from the work by Dames et al.[23] $^3J_{NC\gamma}$ and $^3J_{C'C\gamma}$ scalar couplings of the central W5 residue were determined by quantitative-J 2D constant-time $^{15}N-\{^{13}C^\gamma\}$ and $^{13}C'-\{^{13}C^\gamma\}$ spin−echo difference experiments.[33] Error estimates for the quantitative-J measurements were obtained from the noise of the spectra.

$^1D_{C\alpha C'}$ RDCs were calculated as the difference in $^{13}C'-^{13}C^\alpha$ doublet splittings observed under anisotropic and isotropic conditions, which had been measured with a modified version of HNCO experiment, in which the 180° C^α decoupling pulse in the C′ evolution was removed. Similarly, $^1H-^{15}N$ RDCs were obtained from $^1H-^{15}N$ HSQCs without ^1H decoupling during the ^{15}N evolution. A modified version of the HN(CO)CA experiment without ^1H decoupling in the $^{13}C^\alpha$ evolution period was used to detect $^1H^\alpha-^{13}C^\alpha$ RDCs. Each RDC experiment was carried out twice, and the reported values and the error estimates refer to mean and standard deviation values derived from such repeated experiments.

MD Simulations. MD simulations were carried out with the GROMACS simulation package[48] using the AMBER03 force field.[49] Extended input starting structures of the peptides EGAAXAASS were generated using MOLMOL[50] and solvated in a dodecahedron box containing about 8700 TIP4P water molecules, three sodium ions, and two chloride ions. The energy of the system was first minimized by the steepest descent method, followed by a 500 ps simulation for equilibration of solvent molecules, with the position of the peptide kept fixed. Electrostatic interactions were implemented by particle-mesh Ewald (PME) summation with a grid spacing of 0.12 nm,[51] while the Lennard-Jones interactions had a cutoff at 1.4 nm. The integration time step was 2 fs. Production runs for 100 ns were carried out at a constant temperature of 300 K and pressure of 1 bar. 35 000 (X = W) or 10 000 (X = I) conformations were obtained as 20 ps frames sampled uniformly from seven (X = W) and two (X = I) 100 ns trajectories started with different random seeds.

Back Calculation of NMR Parameters. For every snapshot of the MD trajectory, theoretical RDCs were predicted based on a steric alignment model using an efficient algorithm described previously.[6] The RDC values of each conformation were scaled by a constant determined by a least-square fit between the average RDCs of all conformations and the experimental RDC values of the peptide. Theoretical ^3J values (in Hertz) were calculated using the following Karplus

DOI: 10.1021/acs.jctc.5b00759
J. Chem. Theory Comput. 2016, 12, 383−394

relations: ${}^3J_{HNH\alpha} = 8.40\cos^2(\phi - 60°) - 1.36\cos(\phi - 60°) + 0.33$;[31] ${}^3J_{H\alpha N} = -1.00\cos^2(\psi - 120°) + 0.65\cos(\psi - 120°) - 0.15$;[32] ${}^3J_{NC\gamma}(WS) = 1.29\cos^2(\chi_1) - 0.49\cos(\chi_1) + 0.34$; and ${}^3J_{C'C\gamma}(WS) = 2.31\cos^2(\chi_1 - 120°) - 0.87\cos(\chi_1 - 120°) + 0.49$.[52]

Clustering of MD Conformations. To obtain structural insights, the ensemble of MD conformations for the EGAAXAASS peptides were divided into 20 clusters using the hierarchical clustering function of MATLAB (MathWorks, Inc.) and a ϕ/ψ angle distance metric $d(i,j)$ between individual conformations i and j

$$d(i,j) = \sqrt{\sum_{res} d_{ang}^2(\phi_{res}(i), \phi_{res}(j)) + d_{ang}^2(\psi_{res}(i), \psi_{res}(j))}$$

where the summation runs over central residues A3−A7 of the peptide to emphasize their conformation and the periodic angular distance metric d_{ang} is defined as

$$d_{ang}(\alpha, \beta) = \min(|\alpha - \beta|, 360° - |\alpha - \beta|)$$

The distance between two clusters was defined as the average of all of the individual distances of their members.

Implementation of COPER. COPER was implemented using the IPOPT[36] open-source software package written in C++ for large-scale nonlinear optimization. The IPOPT algorithm utilizes primal-dual interior point methods[26] to find local solutions of optimization problems. COPER objective functions, constraints, and their derivatives, as well as data input and output, were coded in C and linked to IPOPT. To speed up the search for the maximum entropy solution, the optimization was implemented as a minimization of the convex function e^{-S} rather than as a maximization of the entropy S. COPER source code and compiled executables for several platforms are available from the authors upon request.

Default tolerances and the maximum numbers of iterations for the chi square minimization (entropy maximization) were set to 1×10^{-3} and 20 000 (1×10^{-5} and 80 000), respectively. Using these parameters, the total reweighting of the 35 000 member EGAAWAASS peptide ensemble with 35 constraints took 560 s on a single core of a 2.6 GHz Intel Xeon CPU.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00759.

Table of experimental NMR constraints used for the conformational analysis of the EGAAWAASS peptide and analysis of the 35 000 conformations from the MD conformations of the EGAAWAASS peptide with different χ_α^2 limits (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: ++41 61 267 2100. Fax: ++41 61 267 2109. E-mail: stephan.grzesiek@unibas.ch.

Funding

This work was supported by a stipend from the Croucher Foundation (H.T.A.L) and Swiss National Science Foundation grant 31-149927 (S.G.). Computational resources were provided by the Basel Computational Biology Center (http://www.bc2.ch/).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledge Prof. Olaf Schenk for initially pointing out the IPOPT algorithm to us and for very helpful discussions.

■ REFERENCES

(1) Zwanzig, R.; Szabo, A.; Bagchi, B. Levinthal's Paradox. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 20−2.

(2) Tompa, P. Unstructural Biology Coming of Age. *Curr. Opin. Struct. Biol.* **2011**, *21*, 419−425.

(3) Frauenfelder, H.; Sligar, S.; Wolynes, P. The Energy Landscapes and Motions of Proteins. *Science* **1991**, *254*, 1598−1603.

(4) Boehr, D. D.; McElheny, D.; Dyson, H. J.; Wright, P. E. The Dynamic Energy Landscape of Dihydrofolate Reductase Catalysis. *Science* **2006**, *313*, 1638−1642.

(5) Manglik, A.; Kobilka, B. The Role of Protein Dynamics in GPCR Function: Insights From the β2AR and Rhodopsin. *Curr. Opin. Cell Biol.* **2014**, *27*, 136−143.

(6) Huang, J.-R.; Grzesiek, S. Ensemble Calculations of Unstructured Proteins Constrained by RDC and PRE Data: a Case Study of Urea-Denatured Ubiquitin. *J. Am. Chem. Soc.* **2010**, *132*, 694−705.

(7) Feldman, H. J.; Hogue, C. W. A Fast Method to Sample Real Protein Conformational Space. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 112−131.

(8) Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R.; Blackledge, M. A Structural Model for Unfolded Proteins From Residual Dipolar Couplings and Small-Angle X-Ray Scattering. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 17002−17007.

(9) Jha, A.; Colubri, A.; Freed, K.; Sosnick, T. R. Statistical Coil Model of the Unfolded State: Resolving the Reconciliation Problem. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 13099−13104.

(10) Bertini, I.; Giachetti, A.; Luchinat, C.; Parigi, G.; Petoukhov, M. V.; Pierattelli, R.; Ravera, E.; Svergun, D. I. Conformational Space of Flexible Biological Macromolecules From Average Data. *J. Am. Chem. Soc.* **2010**, *132*, 13553−13558.

(11) Różycki, B.; Kim, Y. C.; Hummer, G. SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. *Structure* **2011**, *19*, 109−116.

(12) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(13) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562−12566.

(14) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated Molecular Dynamics: a Promising and Efficient Simulation Method for Biomolecules. *J. Chem. Phys.* **2004**, *120*, 11919−11929.

(15) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Systematic Validation of Protein Force Fields Against Experimental Data. *PLoS One* **2012**, *7*, e32131.

(16) Lindorff-Larsen, K.; Trbovic, N.; Maragakis, P.; Piana, S.; Shaw, D. E. Structure and Dynamics of an Unfolded Protein Examined by Molecular Dynamics Simulation. *J. Am. Chem. Soc.* **2012**, *134*, 3787−3791.

(17) Chen, Y.; Campbell, S. L.; Dokholyan, N. V. Deciphering Protein Dynamics From NMR Data Using Explicit Structure Sampling and Selection. *Biophys. J.* **2007**, *93*, 2300−2306.

(18) Bernadó, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. Structural Characterization of Flexible Proteins Using Small-Angle X-Ray Scattering. *J. Am. Chem. Soc.* **2007**, *129*, 5656−5664.

(19) Nodet, G.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M. R.; Blackledge, M. Quantitative Description of Backbone Conformational

DOI: 10.1021/acs.jctc.5b00759
J. Chem. Theory Comput. 2016, 12, 383−394

Sampling of Unfolded Proteins at Amino Acid Resolution From NMR Residual Dipolar Couplings. *J. Am. Chem. Soc.* **2009**, *131*, 17908−17918.
(20) Berlin, K.; Castañeda, C. A.; Schneidman-Duhovny, D.; Sali, A.; Nava-Tudela, A.; Fushman, D. Recovering a Representative Conformational Ensemble From Underdetermined Macromolecular Structural Data. *J. Am. Chem. Soc.* **2013**, *135*, 16595−16609.
(21) Choy, W. Y.; Forman-Kay, J. D. Calculation of Ensembles of Structures Representing the Unfolded State of an SH3 Domain. *J. Mol. Biol.* **2001**, *308*, 1011−1032.
(22) Fisher, C. K.; Huang, A.; Stultz, C. M. Modeling Intrinsically Disordered Proteins with Bayesian Statistics. *J. Am. Chem. Soc.* **2010**, *132*, 14919−14927.
(23) Dames, S. A.; Aregger, R.; Vajpai, N.; Bernadó, P.; Blackledge, M.; Grzesiek, S. Residual Dipolar Couplings in Short Peptides Reveal Systematic Conformational Preferences of Individual Amino Acids. *J. Am. Chem. Soc.* **2006**, *128*, 13508−13514.
(24) Bignucolo, O.; Leung, H. T. A.; Grzesiek, S.; Bernèche, S. Backbone Hydration Determines the Folding Signature of Amino Acid Residues. *J. Am. Chem. Soc.* **2015**, *137*, 4300−4303.
(25) Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379−423.
(26) Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
(27) Karmarkar, N. A New Polynomial-Time Algorithm for Linear Programming. *Combinatorica* **1984**, *4*, 373−395.
(28) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79−86.
(29) Koenig, B. W.; Kontaxis, G.; Mitchell, D. C.; Louis, J. M.; Litman, B. J.; Bax, A. Structure and Orientation of a G Protein Fragment in the Receptor Bound State From Residual Dipolar Couplings. *J. Mol. Biol.* **2002**, *322*, 441−461.
(30) Meier, S.; Blackledge, M.; Grzesiek, S. Conformational Distributions of Unfolded Polypeptides From Novel NMR Techniques. *J. Chem. Phys.* **2008**, *128*, 052204.
(31) Vogeli, B.; Ying, J.; Grishaev, A.; Bax, A. Limits on Variations in Protein Backbone Dynamics From Precise Measurements of Scalar Couplings. *J. Am. Chem. Soc.* **2007**, *129*, 9377−9385.
(32) Lohr, F.; Schmidt, J. M.; Maurer, S.; Rüterjans, H. Improved Measurement of 3J(Hαi, Ni+1) Coupling Constants in H_2O Dissolved Proteins. *J. Magn. Reson.* **2001**, *153*, 75−82.
(33) Hu, J.-S.; Grzesiek, S.; Bax, A. Two-Dimensional NMR Methods for Determining χ^1 Angles of Aromatic Residues in Proteins From Three-Bond $J_{C'C\gamma}$ and $J_{NC\gamma}$ Couplings. *J. Am. Chem. Soc.* **1997**, *119*, 1803−1804.
(34) Vajpai, N.; Gentner, M.; Huang, J.-R.; Blackledge, M.; Grzesiek, S. Side-Chain χ^1 Conformations in Urea-Denatured Ubiquitin and Protein G From ^3J Coupling Constants and Residual Dipolar Couplings. *J. Am. Chem. Soc.* **2010**, *132*, 3196−3203.
(35) Best, R. B.; Hummer, G. Optimized Molecular Dynamics Force Fields Applied to the Helix-Coil Transition of Polypeptides. *J. Phys. Chem. B* **2009**, *113*, 9004−9015.
(36) Wächter, A.; Biegler, L. T. On the Implementation of an Interior-Point Filter Line-Search Algorithm for Large-Scale Nonlinear Programming. *Math. Programming* **2006**, *106*, 25−57.
(37) Zwanzig, R. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420−1426.
(38) Norgaard, A. B.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Experimental Parameterization of an Energy Function for the Simulation of Unfolded Proteins. *Biophys. J.* **2008**, *94*, 182−192.
(39) Li, D.-W.; Brüschweiler, R. NMR-Based Protein Potentials. *Angew. Chem., Int. Ed.* **2010**, *49*, 6778−6780.
(40) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ^1 and χ^2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257−3273.
(41) Svergun, D.; Barberato, C.; Koch, M. H. J. CRYSOL− a Program to Evaluate X-Ray Solution Scattering of Biological Macromolecules From Atomic Coordinates. *J. Appl. Crystallogr.* **1995**, *28*, 768−773.
(42) Kalinin, S.; Peulen, T.; Sindbert, S.; Rothwell, P. J.; Berger, S.; Restle, T.; Goody, R. S.; Gohlke, H.; Seidel, C. A. M. A Toolkit and Benchmark Study for FRET-Restrained High-Precision Structural Modeling. *Nat. Methods* **2012**, *9*, 1218−1225.
(43) Koenig, B. W.; Rogowski, M.; Louis, J. M. A Rapid Method to Attain Isotope Labeled Small Soluble Peptides for NMR Studies. *J. Biomol. NMR* **2003**, *26*, 193−202.
(44) Sass, H.; Musco, G.; Stahl, S.; Wingfield, P.; Grzesiek, S. Solution NMR of Proteins Within Polyacrylamide Gels: Diffusional Properties and Residual Alignment by Mechanical Stress or Embedding of Oriented Purple Membranes. *J. Biomol. NMR* **2000**, *18*, 303−309.
(45) Chou, J.; Gaemers, S.; Howder, B.; Louis, J.; Bax, A. A Simple Apparatus for Generating Stretched Polyacrylamide Gels, Yielding Uniform Alignment of Proteins and Detergent Micelles. *J. Biomol. NMR* **2001**, *21*, 377−382.
(46) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. NMRPipe: a Multidimensional Spectral Processing System Based on UNIX Pipes. *J. Biomol. NMR* **1995**, *6*, 277−293.
(47) Archer, S. J.; Ikura, M.; Torchia, D. A.; Bax, A. An Alternative 3D NMR Technique for Correlating Backbone 15 N with Side Chain Hβ Resonances in Larger Proteins. *J. Magn. Reson.* **1991**, *95*, 636−641.
(48) 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.
(49) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(50) Koradi, R.; Billeter, M.; Wuthrich, K. MOLMOL: a Program for Display and Analysis of Macromolecular Structures. *J. Mol. Graphics* **1996**, *14*, 51−55.
(51) 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.
(52) Pérez, C.; Löhr, F.; Rüterjans, H.; Schmidt, J. M. Self-Consistent Karplus Parametrization of 3J Couplings Depending on the Polypeptide Side-Chain Torsion Chi1. *J. Am. Chem. Soc.* **2001**, *123*, 7081−7093.

DOI: 10.1021/acs.jctc.5b00759
J. Chem. Theory Comput. 2016, 12, 383−394

2.2.3.1 Appendix 1: Python script for the maximum entropy reweighting

The following python code performs a minimization of the RMSD between experimental and predicted values, while satisfying the criteria exposed in the publication enclosed in section 2.2.3. Blue text that follows a '#' indicates comments which are used to annotate the code.
This code was tested in Python 2.7 and requires an appropriate initial RDC table as indicated in the notes. An illustrative example of the output is presented in the appendix 2.2.3.2.

```
# Purpose: minimize RMSD between predicted and experimental RDCs with the #following constraints:
# sum(pi) = 1
# 0 <= pi <= 1
# X2 <= 1
# NOTES:
#1) the RDC table must contain the reference (i.e. experimental) values in the first line
#2) if the number of frames (or clusters) exceeds the number of measured parameters, there is "overfitting"
#3) the coefficient optimization is performed here through an evolutionary based algorithm.
# The evolutionary algorithm is slower and less efficient than the inner point descent, but it converges to the X2 value given before the name of the function.
#The outputs of the iterations are
#1) a vector of length = number of frames (or clusters), containing the pi values #attributed to each frame (or cluster) in order to reduce the RMSD to experiment while #maximizing the entropy.
#Name of the vector: name_opt.txt, where "name" is the name of the table containing the RDC for all frames
#2) a table of 6 columns and n rows, where n is the number of performed iterations
#Column 2 => RMSD, column 4 => iteration number, column 6 => S = -sum(pi*LN(pi)
from math import sqrt, log
from random import uniform, randrange
import commands

#Target value of X2
X2 = 1

def minimize(data):
        f_data = open(data + ".txt","r")
        f_pi = open(str(data) + '_opt.txt','w')
        txt_all = f_data.readline().splitlines()
        HN_exp, HA_exp = [], []
        #these 4 lines take the experimental values from the first line of the RDC table
        for i_meas in range(2, 10):
                HN_exp.append(float(txt_all[1].split("\t")[i_meas]))
        for i_meas in range(10,20):
                HA_exp.append(float(txt_all[1].split("\t")[i_meas]))
        i, n = 2, len(txt_all)
        frames, HN_MD, HA_MD = [], [], []
        while i < n:
                text_i = txt_all[i].split("\t")
                frames.append(float((text_i)[0])/1000)
                i+ = 1
        #List_pi is the list of pi values, with length = number of frames or clusters
        list_pi = frames
        rmsd, percent, j, entr = 1000, 0.5, 0, 0
        #select randomly one of the clusters
```

```
cl = randrange(1, 1 + len(list_pi))
commands.getoutput("rm -rf "+str(data) + "_follow")
counter = 5000000
while rmsd > X2 and j < counter:
        f_follow = open(str(data)+'_follow','a')
        rmsd_bef = rmsd
        entr_bef = entr
        #modify the pi of the selected cluster using uniform
        p = -1
        #store the value of pi before the modification
        p_bef = list_pi[cl - 1]
        while p <0:
                p = list_pi[cl-1]+uniform(-percent, percent)
        list_pi[cl-1] = p
        l_entr = []
        # here calculate the new entropy
        for value in list_pi:
                l_entr.append(float(value)*log(float(value)))
        entr = -sum(l_entr)
        HN_MD_RDC, HA_MD_RDC = [], []
        for HN in range(2, 10):
                l_hn, l_fr = [], []
                for i in range (0, len(list_pi)):
                        hn = float(txt_all[i+2].split("\t")[HN])
                        pi = list_pi[i]
                        l_hn.append(hn*pi)
                        l_fr.append(pi)
                val = sum(l_hn)/sum(l_fr)
                HN_MD_RDC.append(float(val))
        for HA in range(10,20):
                l_ha, l_fr = [], []
                for i in range(0, len(list_pi)):
                        ha = float(txt_all[i+2].split("\t")[HA])
                        pi = list_pi[i]
                        l_ha.append(ha*pi)
                        l_fr.append(pi)
                val = sum(l_ha)/sum(l_fr)
                HA_MD_RDC.append(float(val))
        # scaling of the MD RDCs
        DobsDcalc,Dcalc2=0,0
        for k in range(0,8):
                DobsDcalc+ =HN_exp[k]*HN_MD_RDC[k]
                Dcalc2+ =HN_MD_RDC[k]**2
        for k in range(0,10):
                DobsDcalc+ =HA_exp[k]*HA_MD_RDC[k]
                Dcalc2+ =HA_MD_RDC[k]**2
        Scaling = DobsDcalc/Dcalc2
        HN_MD_scl, HA_MD_scl = [], []
        for i in range(0, 8):
                HN_MD_scl.append(float(HN_MD_RDC[i]*scaling))
        for i in range(0, 10):
                HA_MD_scl.append(float(HA_MD_RDC[i]*scaling))
        HN_rmsd, HA_rmsd = [], []
        for i in range(0, 8):
                HN_rmsd.append(float(HN_MD_scl[i]-HN_exp[i])**2)
        HN_rmsd = sqrt(sum(HN_rmsd)/len(HN_rmsd))
        for i in range(0, 10):
```

```
                HA_rmsd.append(float(HA_MD_scl[i]-HA_exp[i])**2)
            HA_rmsd = sqrt(sum(HA_rmsd)/len(HA_rmsd))
            Rmsd = 0.5*(HA_rmsd+HN_rmsd)
            #increasing entropy under condition that RMSD is close to target
            if rmsd - X2 <= 0.5:
                tester = [rmsd-rmsd_bef, entr_bef-entr]
            else:
                tester = [rmsd-rmsd_bef]
            #rejection of the new value:
            if any(item >=0 for item in tester):
                list_pi[cl-1] = p_bef
                entr = entr_bef
                rmsd = rmsd_bef
                #for next round: select a cluster or frame
                cl = randrange(1, 1+len(list_pi))
            #record the parameters every 200th iteration
            if j%200 == 0:
                f_follow.write("Rmsd: "+str(rmsd)+" j: "+str(j)+" and Entr: "+str(entr)+"\n")
            f_follow.close()
            #acceptation of new pi value
            normalize(list_pi)
            j+ = 1
    f_follow = open(str(data)+'_follow','a')
    for i in range(0, len(list_pi)):
        f_pi.write(str(list_pi[i]) + "\n")
    f_follow.write("Done " + str(j) + " iterations\nobtain a rmsd of " + str(rmsd) + "\nand Entr =
"+str(entr) + "\n")
    f_data.close()
    f_pi.close()

def normalize(list_of_pi):
    a = sum(list_of_pi)
    for i in range(0, len(list_of_pi)):
        list_of_pi[i] = list_of_pi[i]/a
    return list_of_pi

#uncomment the next line to execute the code
#name refers to a table.txt containing the RDCs and the experimental values in the #first line

#minimize(str(name))
```

2.2.3.2 Appendix 2: RDC minimization and entropy maximization

The purpose of this section is to illustrate that the python code of the previous section solves the RMSD minimization while maximizing the entropy, as further explained in the paper enclosed in section 2.2.3. Precisely, the code performs the following optimization (equation 4 of the JCTC paper):

- maximizes

$$S = -\sum_{i=1}^{N} p_i \cdot Ln\, p_i$$

- minimizes

$$RMSD = \sqrt{\frac{1}{2} \cdot \left((RMSD_{C\alpha H\alpha}/2)^2 + RMSD_{NH}{}^2 \right)}$$

This expression for the RMSD takes into account the different values of carbon and nitrogen gyromagnetic ratios and bond lengths related to the two sets of RDCs ($^1D_{C\alpha H\alpha}$ and $^1D_{NH}$), as reported in the supplementary material of the paper enclosed in section 2.1.5.

- normalizes the sum of all pi values, so that their sum is equal to 1, in order to represent the full probability space of the chosen set of values (frames or clusters).

Two sets of 500 frames were chosen, one for which the unbiased average was extremely far from the experimental data (first few ns of a simulation, containing essentially extended structures) and a second one for which the agreement with the experimental data was much better. Figure 2.1 shows that both sets of values could be brought to a very close agreement to the experimental data. The RMSD, coefficients of correlation, and entropy values are shown in table III.

Table III: RMSD, coefficients of correlations and entropy values before and after reweighting.

	Before reweighting			After reweighting		
Trajectory	RMSD RDC	R	Entropy	RMSD RDC	R	Entropy
A	5.22	-0.60	6.21	1.28	0.91	3.07
B	4.46	0.61	6.21	0.67	0.95	3.84

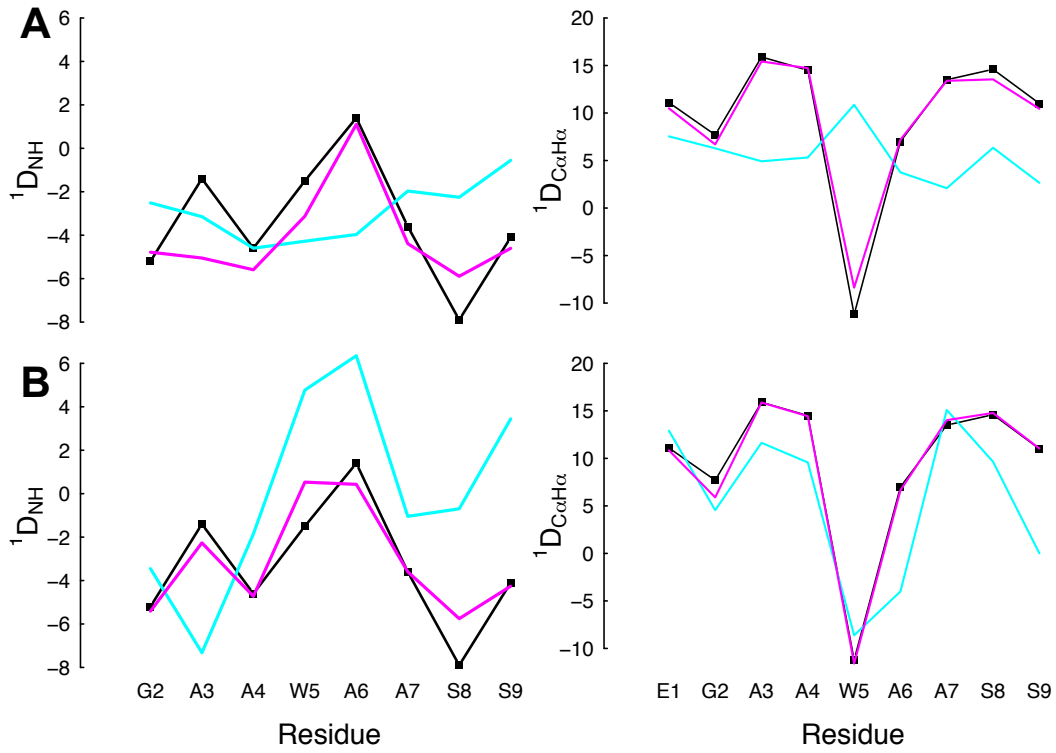


Figure 2.1. Reweighting of predicted populations under the constraints of the experimental RDCs and the maximum entropy principle

A) $^1D_{NH}$ and $^1D_{C\alpha H\alpha}$ RDCs for a representative ensemble of 500 frames containing essentially extended structures. B) $^1D_{NH}$ and $^1D_{C\alpha H\alpha}$ RDCs for an ensemble of 500 frames containing a large proportion of helical structures. Experimental data are shown as black squares. MD data before and after reweighting are shown as cyan and magenta lines, respectively.

The results presented here show that the Python code given in the appendix 2.2.3.1 performs a minimization satisfying the criteria of the equation 4 in the JCTC paper. The entropy decrease of the “essentially extended” population was larger then the one of the “almost-folded” population. This indicates that the algorithm can provide a metric to describe the starting population as a function of its agreement with the experimental values. Therefore, an appropriate usage of this algorithm requires considering the minimization and also the entropy loss.

2.3 Using cross-correlation function analysis to study protein conformational changes

2.3.1 Introduction

The exact role of explicit water in protein dynamics is not completely understood, although its has been emphasized already in 2004 (*79*). For example, a recent study described the interactions between trimethylamine N-oxide (TMAO), dipeptides and hexapeptides and the solvent. TMAO is an osmolyte known to promote protein folding. In mammalians, it is mainly found in kidneys, where it counteracts the denaturating effect of urea (*80*). The investigations led to the hypothesis that the TMAO molecules reduce the access of water from the surface of the peptide, which further enhance the folding propensities (*60*). However, other studies suggest that the hydrophobic collapse precede the desolvation (*59*). In these studies, Cho et al. studied small helices whereas Cheung and colleagues tested their model on a β-barrel, so that the apparent contradiction may be related to the specific secondary structure investigated, as well as to the size of the molecule. The purpose of the method introduced below is to contribute to elucidate which of these two mechanisms is the more likely to occur in a β-turn. The test provides a precise determination of the succession of events during folding and unfolding at the level of a few hundreds of picoseconds.

The lack of hydration of backbone carbonyls and amides was postulated as a key driving force triggering folding nucleation in the investigations of peptides of sequence EGAAXAASS presented in section 2.1. It was shown that bulky side chains limit the access of water molecules to the polar groups of adjacent residues, thus increasing their energy. The consequence is the formation of intramolecular hydrogen bonds form, which potentially may initiate further folding. To test this hypothesis, the number of water molecules coordinating the carbonyl and amide groups of the residues involved in the helix formation was computed. To show that the observed difference in hydration was a potential driving force and not only a consequence of a folded conformation, extended conformations of the Tyr or Trp substituted peptides were compared to conformations of the Gly peptide, which was generally extended. Focusing on these unfolded conformations allowed deciphering how the level of hydration of the backbone carbonyl and amide groups might affect the folding propensities of the peptides. In the case of the Trp substituted peptide, there was a significant reduction of water access to the amide and the carbonyl groups. These calculations supported the view that bulky hydrophobic side chains, because they prevent the access of water molecules to the adjacent functional groups, increase the folding propensity locally (Fig. 3 of the JACS paper). Similar trends were observed for the Tyr substituted peptide.

Here an additional way to investigate whether locally reduced hydration may be a driving force sustaining folding nucleation is introduced. This hypothesis implies that the concerned atomic groups are dehydrated before forming the native hydrogen bonds. Here, the expression "concerned atomic groups" refers to the carbonyl and amide atoms that initiate the folding into a helical structure. It is expected that, in a larger protein, the following helix elongation would occur cooperatively and the related mechanisms are not the subject of this work.

2.3.2 Materials and methods

The succession of the folding events, namely hydration of specific atomic groups, formation of intramolecular hydrogen bonds and native contacts, was investigated through MD simulations of a well-known fast folding β-hairpin for which a folded structure has been deposited in the PDB (*67*). Since the native structure of the protein is known, it is possible to score the conformations along the folding pathways through comparison between the predicted and deposited structures during an MD simulation. In addition, all-atom simulations allow computing the time evolution of the hydration of the polar groups involved in the formation of the native structure hydrogen bonds. If the lack of hydration plays a role as a folding driving force, the hydration fluctuations at the level of the key polar groups are expected to occur before these groups interact with each other and form or break native intramolecular hydrogen bonds.

The model molecule chosen to test the order of events during folding is a fast folding small peptide. Chignolin is a ten residue peptide of sequence GYDPETGTWG, which was designed through fragment assembly from a set of more than 10,000 short segments (*81*). This peptide was shown to fold in several hundreds of nanoseconds to a few microseconds into a unique β-hairpin structure, characterized by four intramolecular hydrogen bonds. The structure of chignolin has been solved by solution NMR. 185 restraints, mostly NOE (nuclear Overhauser effect), were used to determine the most probable structure. The 20 lowest energy conformations are deposited under the PDB code 1UAO (*81*) and one representative structure is shown in Figure 2.2. Since the solution structure of this fast-folding peptide has been deposited, the folding and unfolding events of this β-hairpin have been thoroughly investigated by coarse-grain and all–atom MD simulations (*70*, *82*-*86*). The molar fraction of folded peptide in solution has been determined to be about 60% at 300K (*70*, *83*).

The all-atom MD simulations of chignolin were performed in explicit water TIP3P (*87*) using the GROMACS software package version 4.5 (*88*) and the CHARMM36 force field (*89*). The simulation started from a fully extended structure generated with PyMOL, imposing an all-trans conformation to all backbone dihedrals (*90*). In order to quantify the folding of the peptide, the time course of the RMSD of the backbone atoms from residues two to nine with respect to the deposited NMR structure was computed. The experimentally determined backbone hydrogen bonds and the distances between hydrophobic residues shown through NOE determination to be

close to each other in the folded structure were also investigated. The hydration was accessed through computation of the number of water molecules within the first hydration shell of specific atomic groups. The radius of the hydration shell was extracted from the analysis of several radial distribution functions of the concerned atomic functional groups.

2.3.3 Results

According to the NMR solution structure, the β-hairpin is stabilized by hydrogen bonds involving the following backbone amide and carbonyl groups: Asp3CO-Gly7NH, Thr8CO-Asp3NH and Gly7CO-Asp3NH. In addition, the side chain carbonyl group of Asp3 was involved in a fourth hydrogen bond formation with the amide of Gly5, termed Asp3sc-Gly5NH in the following. This pattern suggests a key function in terms of hydrogen bond formation for Asp3. The inspection of the experimentally determined NOEs, summarized in Table 3 of (*81*), shows that the following atom pairs involving the aromatic residues were closer to each other than 3.5 Å in the folded structure: Tyr2Hα-Trp9Hα, Tyr2Hα-Trp9Hε3, Tyr2Hδ#-Trp9Hα. Consequently, the distances between these groups were computed too, in order to fully describe the folding/unfolding events of the peptide.

The backbone RMSD in reference to the deposited structure, three of the four native intramolecular hydrogen bonds and the NOE distances show that, while visiting some local minima at 150 and 200 ns, the protein adopted the native conformation at around 600 ns, and remained stable for the next 100 ns. However, one of the four native hydrogen bonds, Gly7CO-Asp3NH, was not formed during the simulations. An explanation for this discrepancy was not found. Precisely, at t = 600 ns, the backbone RMSD was approximately 1.5 Å, all the tested NOE distances between the hydrophobic residues were smaller or close to 5 Å and three native hydrogen bonds were formed (Figure 2.3). In order to test the reduced hydration hypothesis, the time course of the number of water molecules within the first hydration shell of the carbonyl oxygen or amide nitrogen of the following atoms was computed: Asp3CO, Asp3NH, Gly7NH, the side chain of Asp3-Gly5NH and Thr8CO. These atoms are involved in the formation of hydrogen bonds in the native structure. Additionally, the hydration of Gly7CO, which does not form any hydrogen bond in the native structure, was also extracted. In other words, Gly7CO was used as a negative control. The average hydration of the atoms involved in the native hydrogen bonds followed the same fluctuations than the RMSD to the native structure (Figure 2.4B). The Gly7CO-Asp3NH hydrogen bond was not formed in the MD trajectory. As expected, the time course of hydration of Gly7CO oxygen did not change significantly during the 700 ns of simulation (Figure 2.4A). In Figure 2.4C, ensembles of potentially unfolded and extended structures and the cluster of natively folded structures are easily identifiable.

Because this trajectory shows transitions between ensembles of extended and folded structures, it provides the required conditions to elucidate whether the fluctuations of functional

groups hydration precede the conformational changes. Precisely, a cross-correlation analysis between the RMSD to the native structure and the hydration was performed. This first calculation involved 7500 structures, calculated every 100 ps. The cross-correlation plot suggests that the hydration and dehydration events occur ahead of the folding/unfolding events by a few hundreds of ps (Figure 2.5A). This is indicated by the fact that the maximum of the function is not at $\Delta t = 0$, marked by the dashed lines in the figure, but a few hundreds of picoseconds to the right. The asymmetry of the correlation distribution also suggests that the folding/unfolding events follow the hydration fluctuations. A second, more precise investigation was performed, in which the RMSD to the deposited structure and the hydration values were computed every 4 ps. Interestingly, in this cross-correlation function, only the amine group of Asp3 was clearly dehydrated ahead of the folding into the native structure, which may suggest that this functional group would initiate the folding. Precisely, the cross-correlation function displays a prominent “shoulder” about 1 ns to the right of the values with $\Delta t = 0$ (Figure 2.5B). Similar analyses were performed on a second simulation, lasting 850ns, but in which the peptide was trapped two times into a low energy basin, without folding exactly into the deposited structure. The RMSD to the deposited structure and the level of hydration were recorded every 40 ps. This second cross-correlation function analysis also suggested that the level of hydration precedes the variations of RMSD.

2.3.4 Conclusion

Cross-correlation functions are standard analysis tools in the area of computational time series analysis. The results presented in this section indicate that a cross-correlation analysis can be applied on MD trajectories to determine the order of events in the context of peptide folding. Here, the role of hydration in the mechanism of protein folding was investigated at the time scale of a few picoseconds, and these preliminary results suggest that hydration fluctuations occur a few hundreds of picoseconds ahead of folding. The same method could be used to investigate the relations between hydration and other conformational changes, particularly to better understand the dynamics of IDPs or to study the interactions between proteins and other molecules. Especially, the interactions with denaturants as well as small molecules that promote folding like TMAO, trimethylglycine, glycerol, trehalose or sucrose, (*91*) could be investigated.

Interestingly, the sequence of Chignolin provides an additional support to the idea that aromatic amino acids deserve to be called “order-promoting” residues, as was exposed in the section 2.1.3. Two out of the ten residues of this peptide are aromatic residues. For comparison, aromatic residues represent about 8.5% of the sequences deposited to the Swiss-Prot server (http://www.uniprot.org/) (*72*). In the sequence engineering of chignolin, only residues two to eight were optimized, whereas the terminal Gly residues were not. As stated in the paper describing the engineering of this peptide, they were introduced to reduce electrostatic effects between the central eight residues (*81*). However, subsequent optimization of this peptide in order to further

improve its folded stability led to the sequence YYDPTGTWY, in which the aromatic amino acids represent 40% of the residues (*92*). Not surprisingly, the replacement of Tyr at position two by Ala was shown to strongly destabilize the β-hairpin structure (*82*).

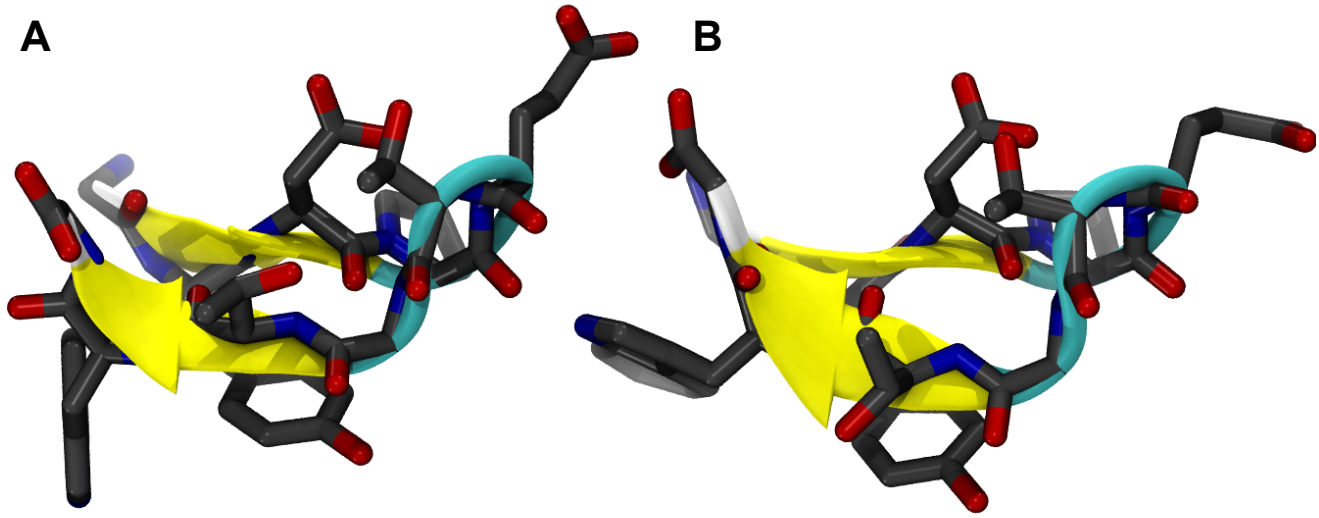


Figure 2.2. β-hairpin structure of chignolin from NMR experiment and from simulation. Molecular representation of the PDB deposited structure (A) and a snapshot of the molecular dynamics trajectory at t=625 ns (B). The secondary structure is highlighted with beta-strands in yellow, turns in cyan, and loops in white, Residues are shown in licorice with carbon: grey, nitrogen: blue, and oxygen: red. Hydrogen atoms are not shown.

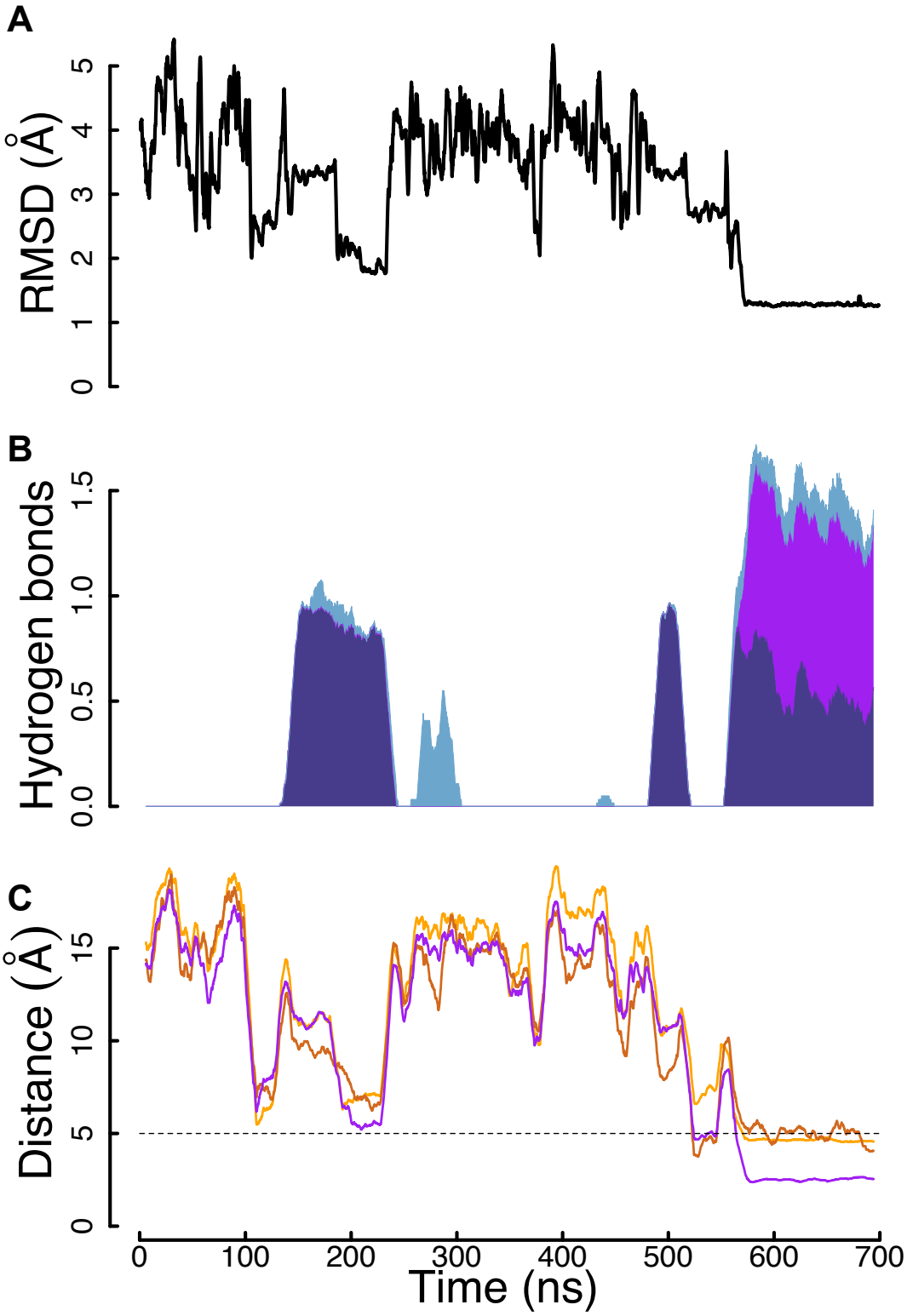


Figure 2.3: Folding trajectory of Chignolin
A) Time course of the RMSD of the backbone atoms in reference to the deposited structure. B) Cumulative numbers of hydrogen bonds. Asp3O-Gly7NH: dark blue, Thr8O-Asp3NH: purple, Asp3-sc-carboxyl-Gly5NH: light blue. C) Main hydrophobic contacts as determined from the NOEs. Tyr2Hα-Trp9Hα: purple Tyr2Hα-Trp9Hε3: brown, Tyr2Hδ#-Trp9Hα: orange. The dashed line highlights the 5Å distance.

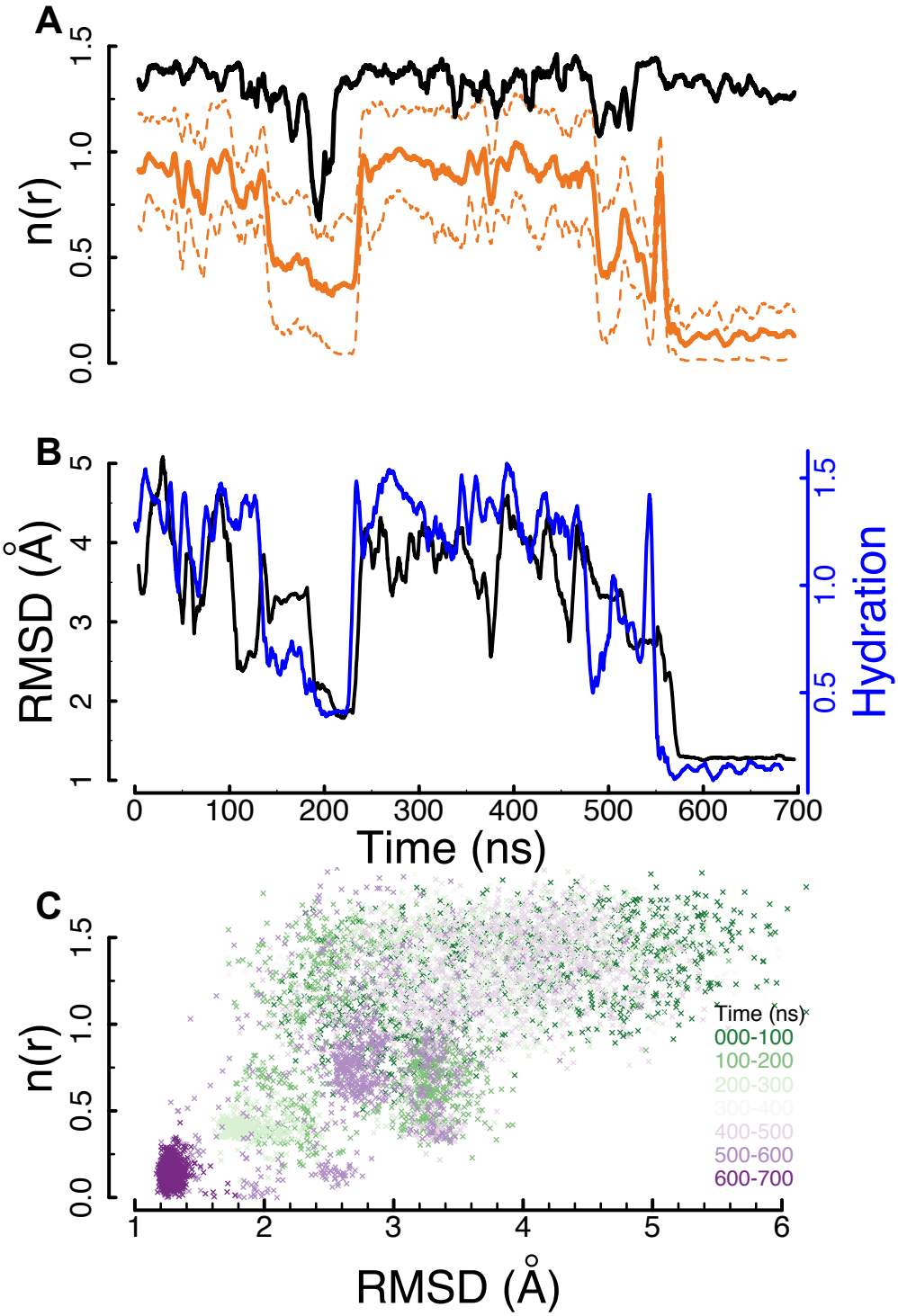


Figure 2.4.Specific atomic group dehydration upon folding.
A) Time course of the hydration of Asp3CO, Asp3NH, Gly7NH and Thr8CO (orange). The average (solid line) and the standard errors (dashed lines) of the hydration level are shown. Black: Time course of the hydration of Gly7CO. B) Time course of the RMSD to the deposited structure (black, left Y-axis) and hydration of Asp3CO, Asp3NH, Gly7NH and Thr8CO (blue, right Y-axis). C) Average number of water molecules hydrating Asp3CO, Asp3NH, Gly7NH and Thr8CO versus the RMSD to the deposited structure. Each data point corresponds to an average over 100 ps of simulation, colored according to the time.

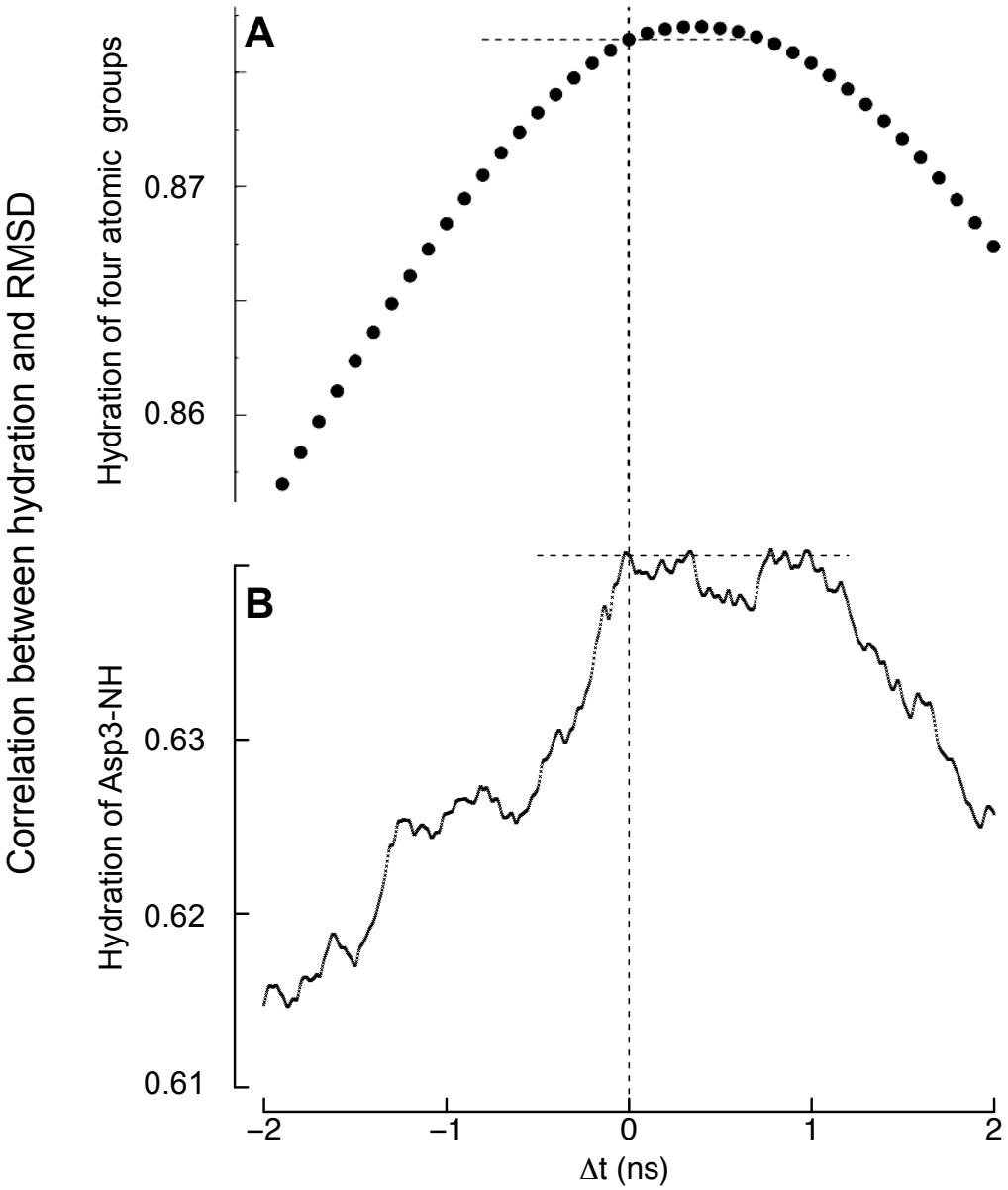


Figure 2.5: Hydration fluctuations occur ahead of conformational changes
A) Cross-correlation function of the RMSD to the deposited structure versus the number of water molecules within the first hydration shell of the for the atomic groups involved in the four experimentally observed hydrogen bonds computed every 100 ps. B) Cross-correlation function of the RMSD to the deposited structure versus the number of water molecules within the first hydration shell of the Asp3 amide group recorded every 4 ps. In both figures, the dashed lines mark the level of correlation for $\Delta t=0$, highlighting the right shift of the correlation distribution.

2.4 A project to systematically explore the relationships between sequence and conformational changes

The observations described above have shown that molecular dynamics investigations may provide useful insights, not only on the conformations adopted by peptides, but also on the fundamental mechanisms directing the transitions between such conformations. Without a profound understanding of the factors affecting the conformational behavior of proteins, we may not be able to predict biologically relevant interactions among proteins and between proteins and other molecules. As mentioned above, the low energy conformation of a protein can be determined to a very high resolution through crystallography, NMR spectroscopy, Cryo-electron microscopy, and computational techniques (homology modeling). MD might not execute this particular task much better and faster, given the present accuracy of force fields. However, MD trajectory analysis can increase our understanding of the mechanisms leading to conformational changes. We still do not have a deep understanding of the mechanisms by which amino acid chains change their conformations to fold or to interact with solvent, denaturants, and ligands, and how this dynamic may be predicted from sequence. This knowledge is required for accurate prediction of protein conformations.

The lack of backbone hydration in the vicinity of large side chains was shown to impact on the folding propensity of peptides of sequence EGAAXAASS, where 14 substitutions at position X were experimentally tested. The RDCs of these peptides could be sorted into two groups according to their RDC pattern. The substitutions with the two aromatics Tyr and Trp resulted in a considerably contrasted pattern, with some residues even changing sign, whereas the other RDC patterns were rather flat. The MD simulations focused on four substitutions: the two aromatics Trp and Tyr, and two residues used as controls: Gly and Ile. The role of hydration was also investigated in the folding process of the β-hairpin chignolin, of sequence GYDPETGTWG. Thus, a combination of MD simulations, NMR experiments, and statistical analyses constitutes an appropriate tool for the investigation of peptide conformational changes.

A more complete screening would be required in order to be able to predict conformational changes occurring during folding and unfolding. An initial step would be to fully assess the effect of nearest-neighbor residues in a peptide. An experimental layout would be to mutate amino-acid triplets, which would require the investigation of 8,000 sequences. The use of tripeptide sequences was reported for the estimation of the effect of nearest-neighbor on the chemical shifts of the central residue via bioinformatics analysis (*93*). In line with the rational that led to the sequence EGAAXAASS in the previous work, the study of the conformational behavior of peptides of sequence $S_1S_2A_pA_pX_1X_2X_3A_pA_pS_3S_4$ would elucidate the role of nearest-neighbors in short peptides. Here, S stands for spacers, A_p for apolar residues with relatively small side chains, and the substitutions by $X_1X_2X_3$ ensure the exploration of all possible occurrences of

amino acid triplets. In the previous work, Ala residues were used around the mutated residue, and the spacers were Glu, Gly, and Ser. Structure determination by NMR offers insight into the dynamics of molecular systems, and it is tailorable to answer specific questions at atomic resolution. Validation of MD trajectories and putative conformations would be performed through NMR determination of NOEs, RDCs and chemical shifts for selected peptides.

In 2015, I supervised a research project performed by Tomas Tomka (Computational Sciences, Department Mathematics and Informatics, University of Basel). A small set of peptides with sequence $EGAAX_1X_2X_3AASS$ has been simulated for 200 ns with the CHARMM27 force field. Each simulation was performed three times, with a different set of initial particle speeds at 300 K using the GROMACS package 4.5.3. The following triplets were investigated: EAE, FAF, GAG, KAK, QAQ, and WAW. Since most of the previous work focused on the role of large hydrophobic residues on the conformational behavior of small peptides, several hydrophilic residues were included in this short pilot study. Phe and Trp display stronger folding propensities than Gly, confirming the results of the previous study and the subsequent work introduced in section 1.2 (Figure 2.6). Interestingly, the folding propensities into helices of peptides with sequences EGAAEAEAASS and EGAAKAKAASS are even stronger. Precise analyses of the interactions between the side chains, the backbone and the solvent in this sort of trajectories may have help decipher the mechanisms involved in the conformational behavior of peptides.

In relation to the hypothetical differences between helices and β-barrels exposed in section 2.3, it would be fascinating to use the statistical tool exposed therein to explore if different driving forces induce the folding into different secondary structures, and how similar mechanisms could underlie the conformational changes of intrinsically disordered proteins.

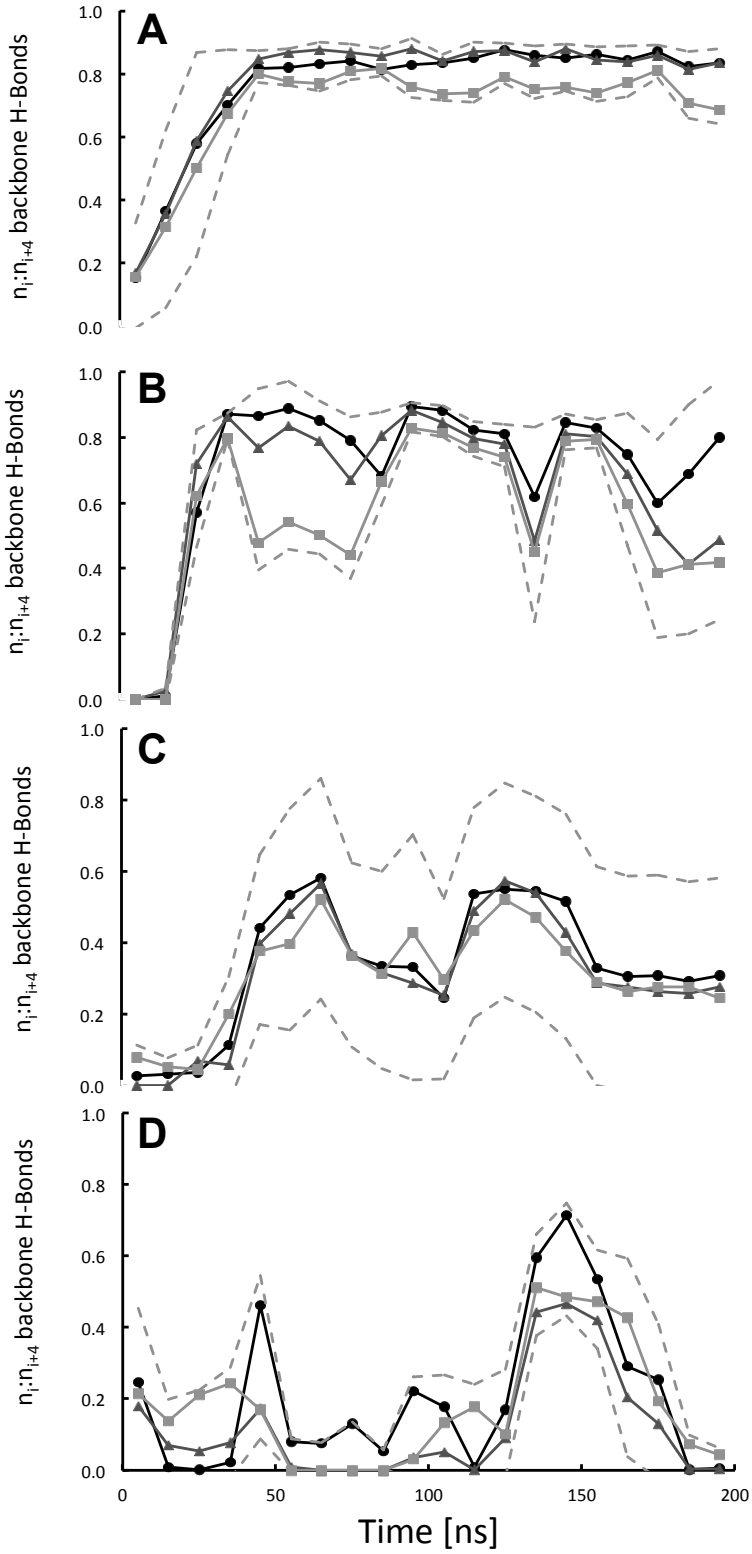


Figure 2.6: Folding propensities of four peptides of sequence EGAAXAXAASS
X is the substituted residue Glu (A), Lys (B), Trp (C) and Gly (D). The averages and standard deviation of the frequencies of hydrogen bond formation between the backbone carbonyls of residues *i* and the backbone amides of residue *i+4*, during three independent simulations of 200 ns each. (Circles: A3-A7, triangles: A4:A8, squares: X5:A9)

3 Membrane perturbations induced by toxins, a voltage-sensor domain and the membrane potential

3.1 Introduction

Most spider toxins, as was described in the introduction section 1.2.3, carry an electric charge, generally +2 to +4, and their structure is essentially amphiphilic. For this reason, they were chosen as tools to investigate the lipid-mediated hypothesis of gating modification. However, despite the overall structural similarity of the inhibitory cysteine knot (ICK) evidenced in Figure 1.4, these toxins are highly specific. Hanatoxin (Hatx), for example, inhibits several voltage-gated ion channels, among them the potassium channel Kv2.1 and the calcium channel Cav2.1 (*94*). However, according to the literature, it does not affect KvAP (*39*). Other toxins, like Vstx1 or GsMTx4, have no effect on the K^+ channels Kv2.1 or Shaker, but induce a right shift in the response curve of KvAP (*95*). Vstx1 inhibits, in addition, the Nav1.7 channel (*96*).

On the other hand, by transfer experiments between channels, Alabi et al. (*39*) determined precisely which parts of the VSD are mediating the toxin sensitivity. When the S4 helix of Shaker, which is normally insensitive to Vstx1, was replaced by the corresponding KvAP-S4 helix, the chimera became sensitive to Vstx1. This experiment indicates that S4 of KvAP is important for the sensitivity to Vstx1. In another experiment, Kv2.1 became insensitive to Hatx after replacement of its S3b segment by the corresponding S3b segment of KvAP. (*39*). These findings, summarized in Figure 3.1, raise questions about the lipid-mediated mechanism of action: if the mechanism is mediated by the membrane, this specificity should appear in the membrane perturbations. Thus, an approach intended to study spider toxin interactions with the membrane and their effect on the target protein requires the identification of similarities and differences in the membrane perturbations they induce. Considering several toxins known experimentally to inhibit different channels, any difference between them in terms of effects on the membrane would be a hint to explain their specificity, while similar effects would potentially describe common mechanisms of actions. This does not exclude the combination of some shared mechanisms reflecting the similar overall 3D structure, and additionally more specific interactions explaining the experiments exposed above. With the aim to take the specificity into account, the two toxins Hanatoxin and Vstx1 were selected. Both structures have been determined to a high-resolution and other experimental data are available, and they were shown experimentally to inhibit different targets.

On the other hand, since the targets of gating modifiers are VSDs, some features of the VSD were investigated to. The indirect effect hypothesis may be explored through the responses of the VSD to the membrane perturbations. Since Vstx1 was shown experimentally to inhibit KvAP, but Hanatoxin not, the VSD of KvAP (PDB code 1ORS) was inserted in the bilayers. Any

differences between the interactions of the two toxins with the VSD of KvAP could be regarded as hints for their specificity. However, an additional challenge is that the exact conformational changes of a VSD as a function of the membrane voltage are unknown (*97*, *98*). According to these considerations, the following minimum requirements are needed to investigate an indirect mechanism of action of spider toxins on ion channels:

- two toxins (at least), known experimentally to inhibit different targets,
- the specific target of one of the two toxins,
- a molecular system, which enables the explicit tuning of the membrane potential.

Interestingly, in support of this approach, in which a wide range of membrane potentials were studied, it was shown experimentally in 2015 that the binding of Vstx1 to KvAP depends on the membrane potential, namely that Vstx1 affects the VSD only under depolarized conditions (*99*). Yet, the membrane asymmetry and relative diversity were added in order to generate trajectories of realistic models, but they are not directly linked to the lipid-mediated gating modifier hypothesis. Interestingly, it was recently demonstrated that PIP_2 has a profound effect on the VSD (*1*). This negatively charged phospholipid is found essentially in the intracellular membrane leaflet.

The amphipathic character of the spider toxin structure suggests that they partition into the membrane, and this partitioning was shown experimentally (*56*). However, at the beginning of the thesis, the orientation of most toxins in the membrane was unknown (*47*, *99*). One important part of the work was then to investigate the orientation of each toxin within a membrane, before assessing their effect on the bilayer. Fortunately, experimental studies of the orientation Vstx1 were published in 2014 (*100*) and 2015 (*99*), giving the required information to validate the orientation of the toxins within the bilayer.

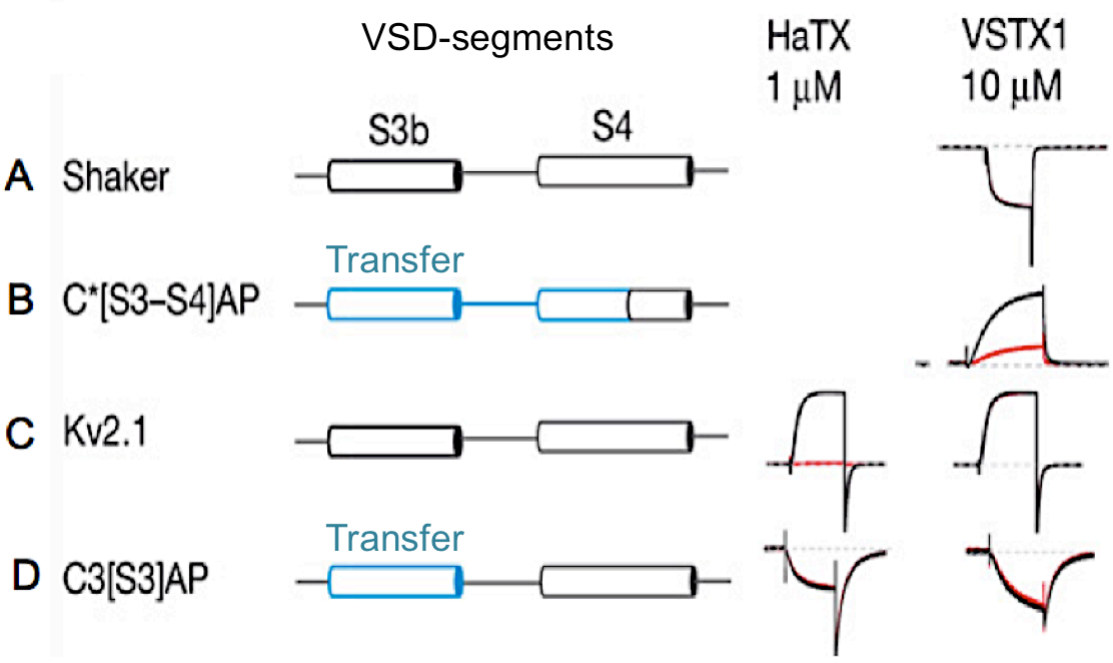


Figure 3.1. The toxin sensitivity of the VSD is determined by unique VSD subsets

A) Vstx1 does not inhibit Shaker. B) Vstx1 inhibits Shaker, if the S3b-S4a segment of Shaker has been replaced by the KvAP residues Pro99 – Arg126. C) Hanatoxin inhibits Kv2.1, but Vstx1 does not. D) Neither Vstx1 nor Hanatoxin inhibit Kv2.1, if its S3b segment was replaced by the KvAP-S3b segment. Left panel: channel constructs, transferred segments are highlighted in blue (KvAP segments were transferred either in Shaker or Kv2.1). Right panel: Voltage-current response curves in the absence (black) or in the presence of toxin (red). The Shaker-Hanatoxin interaction was not reported in the original study. Figure adapted from (*39*).

3.2 Material and methods

The all-atoms molecular dynamics simulations were performed using the GROMACS software package version 4.5 (*88*) with the CHARMM36 force field (*89*).

The initial protein structures were obtained from the coordinates deposited in the RSCB Protein Data Bank (PDB) (http://www.rcsb.org/) (*67*), explicitly:

- Hanatoxin, code 1D1H, NMR solution structure (*101*)
- Vstx1, code 1S6X, NMR solution structure (*102*)
- KvAP VSD, code 1ORS, crystal structure (*31*)

In order to investigate the responses of the lipids, toxins and VSD to an applied membrane voltage, the systems were built in two steps. First, a simple initial system with an asymmetric membrane was generated (Figure 3.2A), using the CHARMM-GUI web site (*103*). A typical membrane contained approximately 100 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) and 80 cholesterol molecules on the “extra-cellular” leaflet, and 50 POPC, 50 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoserine (POPS) and 80-90 cholesterol molecules on the “intra-cellular” leaflet. Around 25,000 water TIP3P molecules (*87*) were added. Finally, KCl was included by molecular replacement of water to a final concentration of 0.15 M. Two such systems were combined in an antiparallel way to obtain a double bilayer simulating an asymmetric membrane. A double bilayer system is the most biologically relevant way to simulate a cell membrane separating two different water compartments (*104-107*). Through antiparallel orientation of the bilayers, the construct contained in its center a water slab simulating the “intracellular” compartment, which is in contact with the POPS enriched leaflets. The “extracellular” compartment was constituted of two water slabs, linked together through the periodic boundary condition, and was in contact with the leaflet without POPS. Each double bilayer system contained in addition a VSD in each bilayer and in most cases two toxins in the extra-cellular compartment, for a total number of ≈ 235,000 atoms.

A membrane potential (Vm) was then simulated through charge imbalance between the compartments (Figure 3.2B). The displacement of a single ion affected the membrane potential by approximately 200 mV, which is almost twice the sensitivity of larger systems used in other

studies (*108*). Luckily, this high sensitivity led to the easy identification of gating charge transport events (section 3.3.2.4).

Standard periodic boundary conditions were used in an ensemble held at 1 bar (compressibility 4.5 x 10^{-5} bar and time constant 1ps) using the Berendsen algorithm (*109*). The temperature was kept at 310 K and to ensure a proper canonical ensemble, the velocities were rescaled using a stochastic term (time constant 0.2 ps) (*110*). Simulations were run with a 2 fs integration time step. Electrostatic interactions were computed using the particle mesh Ewald method (*111*), and for the van der Waals interactions a cut-off of 12 Å was used, and the neighbor list was updated every 10 steps.

The number of constructed systems was as follows:

- Hanatoxin: 34 double bilayers, initial Vm between -1.29 and +0.42 V. lengths between 200 and 400 ns, total 7400 ns.

- Vstx1: 26 double bilayers, initial Vm between -1.75 and 0.46 V, lengths between 200 and 740 ns, total 7160 ns.

- Controls, with a VSD in each bilayer, but no toxin: 6 double bilayers, initial Vm between -1.33 and +0.39 V, lengths between 200 and 400 ns, total 1600 ns.

The analysis of the trajectories was performed using a combination of GROMACS utilities and in-house R and Python codes, and the statistical analyzes were performed using the R environment (*112*).

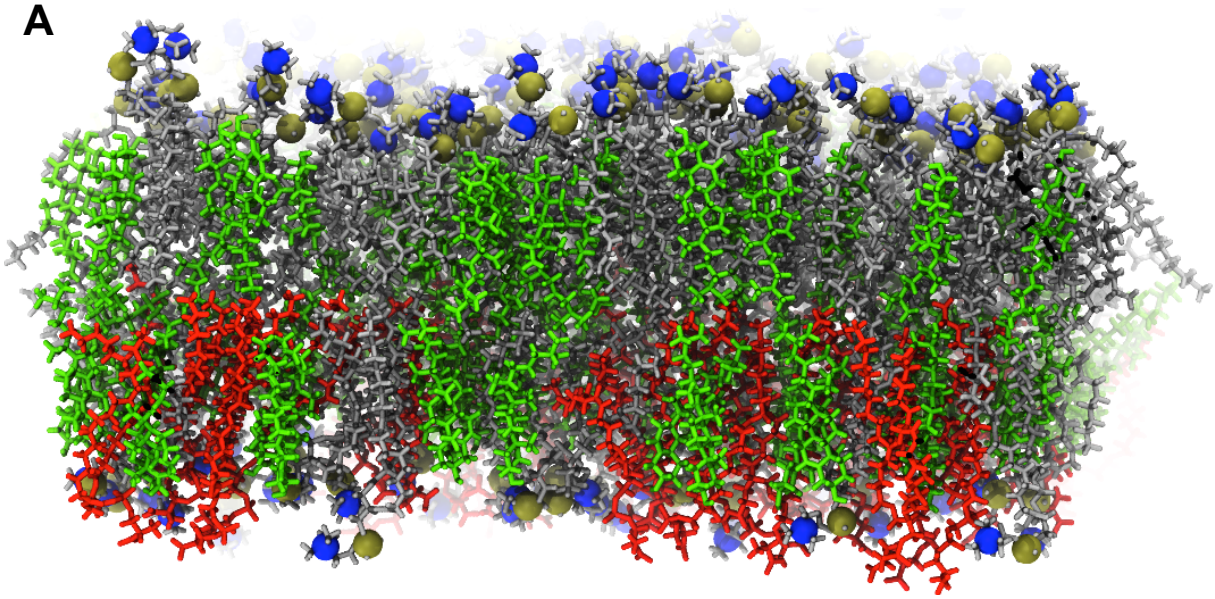
3.3 Results

3.3.1 Perturbation of the bilayer upon toxin insertion

In the next sections, the insertion of the two selected toxins will be described first. For each of them, the analysis are restrained to the simulations in which the toxin, which was in the water phase at the beginning of the simulation, inserted at least at the level or below the phosphate groups of the membrane within the 200 to 740 ns of simulations. Vstx1 is described first, because the orientation of this toxin has been recently experimentally determined (*99*). It is found that, as expected, the hydrophobic residues insert deeply into the membrane, while the charged residues form hydrogen bonds with the lipid head groups. A similar approach is then applied for the orientation of Hanatoxin, for which such precise experimental data are not available.

Next, the toxin induced perturbations of acyl chains of the POPC molecules are described. A second analysis shows that the insertion of the toxins induces a reorientation of the choline head groups, so that the positively charged choline group moves toward the water phase. An explanation for this effect, based on the literature and on our results, is provided. The

simulations suggest that the head groups respond to a subtle combination of electric repulsion, since both the choline head group and the toxin are positively charged, and, in addition, a steric effect consecutive to the formation of hydrogen bonds between the Arg residues of the toxin and the phosphate groups. It is also found that the bilayer thickness is reduced near the toxin.



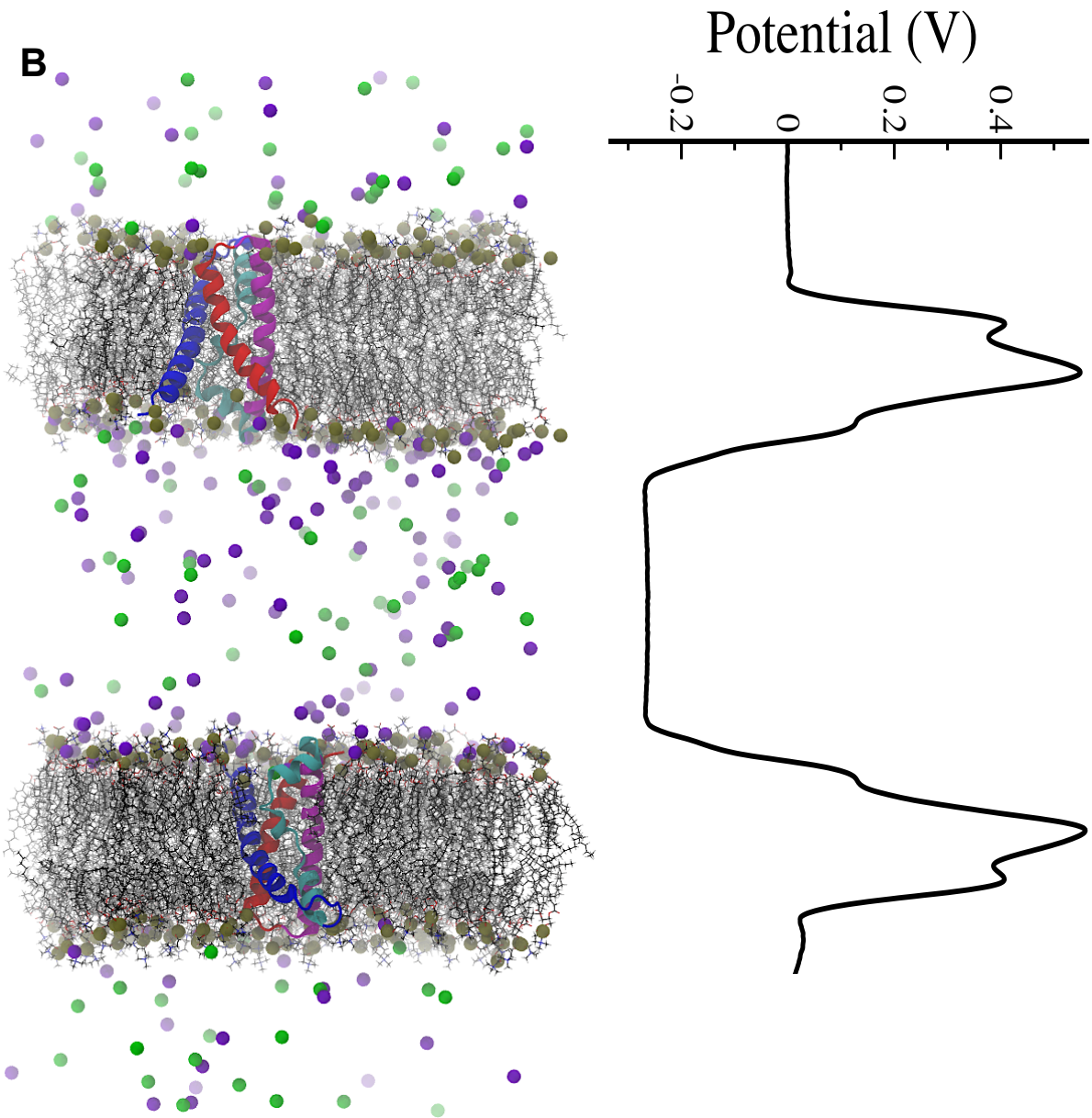


Figure 3.2, previous page: The double bilayer system enables the explicit tuning of the membrane potential.
A) The simulated asymmetric bilayer contains 100 POPC and 80 cholesterol molecules in the upper leaflet, and 50 POPC, 50 POPS and 85 cholesterol molecules in lower leaflet. The lipids are colored as follows: POPC: grey; POPS: red; Cholesterol: green. The phosphate and nitrogen atoms of the phospholipids are shown as tan and blue spheres, respectively.
B) Left: Molecular representation of two bilayers with embedded VSDs. Proteins are represented in cartoons, with the helices colored as follows: S1: red, S2: magenta, S3: cyan, S4: blue. Lipids are represented with acyl chains in grey, and the phosphorus atoms as tan spheres. Potassium and chloride ions are shown as yellow and green spheres, respectively. Right: In this illustrative configuration, the imbalance in the number of charged species on both sides of the membrane induces a negative potential of about 250 mV in the middle water slab, as compared to the top and lower slabs.

3.3.1.1 Orientation of the membrane bound toxins

Hanatoxin and Vstx1 have been shown to partition into lipid membranes (*56*, *113*) and their toxin activity is thought to depend on their amphipathic character (*114*). However, although some experiments assessed the depth of the Trp residues of these toxins upon interaction with the membrane, very little is known about their exact orientation. Luckily, during the time of this thesis, two experimental studies investigating the orientation of Vstx1 in micelles and in membranes were published, allowing a comparison with the computational results.

3.3.1.1.1 Vstx1

To date, two experimental studies have provided a detailed description of Vstx1 membrane partitioning, so that a validation of the molecular dynamics trajectories may be performed by comparison with experiment. Ozawa et al. reported a residue level description of the Vstx1 partitioning in n-decyl-β-D-maltopyranoside (DM) micelles (*99*). In the work of Mihailescu et al., the orientation of the toxin in lipid bilayers was obtained more indirecty through rigid body rotations and translations of the solution structure in order to fit measured scattering-length density profiles (*100*). Ozawa et al. (*99*) reported the chemical shift perturbations of each residue of Vstx1 upon interaction with the DM micelles. In the same study, they identified the membrane binding residues by a cross-saturation experiment. Both experimental procedures provided, independently, a residue-based quantification of interaction with DM micelles and, in both the intensity of the signal is expected to increase linearly with the insertion depth. The orientation of Vstx1 found in the MD simulations was compared to a combination of the two experimental value series, obtained through normalization and pooling. This procedure was chosen, as it takes into account the discrepancies between the two series of experimental results. For the analysis of the toxins' orientation in the MD trajectories, the position of the center of mass (COM) of each residue side chain along the normal to the membrane (Z axis) was computed during the last 20 ns of simulation. The position of the lipid phosphate groups was calculated as a reference and the trajectories for which the average computed insertion over a whole toxin was below the level of the phosphate groups were retained. Toxins that were not yet well inserted were excluded. The simulations started with the toxin, either in the water phase (22 replications) or the toxin placed randomly at the surface of the bilayer (30 replications), and a set of 5 simulations was retained, based on the mentioned average depth criteria. The distribution of residue depth within the membrane of these simulations matches the experimental measurements within the statistical error of the experimental determination (Figure 3.3). This is also indicated by the coefficient of correlation (R = -0.67) calculated between the depth of the Vstx1 side chains and the NMR measurements (Figure 3.4)

Mihailescu et al. described a “relatively superficial position” (*100*) of Vstx1 on the membrane bilayer, and the orientation of the toxin differed slightly from that measured by Ozawa et al. The later reported a deeper insertion at the level of the C-terminus and of some charged residues (K17, D18). In Figure 3.6, membrane interacting residues determined by NMR, scattering-length density profile, and MD are mapped on the molecular structure of the toxin. Interestingly, the surface identified through MD simulations seems closer to the NMR identified one, but it additionally contains some features observed only by the scattering experiment of Mihailescu et al (*100*). Particularly, they reported that the three Trp residues “form a ridge that aligns along the water-hydrocarbon interface”. Figure 3.3 shows that, in the MD simulations, Trp7, Trp25, and Trp27 all insert at the same depth, about 4-6 Å under the phosphate level. However, while the scattering-length density profile suggests a deeper insertion of the C-terminus, the MD simulations show only a particularly high variability at this terminus.

In almost all the 53 simulated systems containing Vstx1, the toxin formed hydrogen bonds with the membrane within the first 20-40 ns. In order to estimate any large conformational changes upon binding to the membrane, a simulation in which a toxin remained far from the membrane for at least 80 ns before binding to the membrane was selected. The RMSD of the backbone atoms in reference to the deposited structure did not change much more when the toxin interacted with the membrane (Figure 3.5), which was also observed through NMR spectroscopy measurements (*47*). Thus, NMR measurements and the MD trajectories suggest that Vstx1 does not undergo large conformational changes upon insertion.

3.3.1.1.2 Hanatoxin

Although Hanatoxin has been investigated much more than Vstx1, an experimentally determined description of its insertion into the membrane, at the level of the side-chains, is missing. The agreement between the orientation of Vstx1 towards the end of our simulations and the NMR studies exposed above (Figures 3.3 & 3.4) strengthens our confidence in the MD approach. An experimental reference is nevertheless provided by the measured insertion of the Trp indole group. Contrary to Vstx1, which has three Trp residues, Hanatoxin displays a single Trp residue, resulting in a one-to-one correspondence between measurement and depth. Accordingly, several authors have determined the depth of the toxin by a fluorescence-quenching approach (*115*, *116*). Upon insertion into a membrane, the Trp fluorescence is shifted toward shorter wavelengths. By measuring quenching profiles of brominated lipid tails, the depth of the Trp indole can be estimated, and values of about 7-9 Å from the center of the bilayer were reported. These values correspond to approximately 10 Å below the level of the phosphate groups. In the MD trajectories, the depth of the Trp30 side chain depth displayed a large variability among simulations, with values ranging from z = 1 to 12 Å below the phosphate groups. In Figure 3.7B, a snapshot is depicted, highlighting the insertion of Hanatoxin into the

membrane at the end of a 200 ns long simulation. In the same way as in the analysis of Vstx1 insertion, the insertion of Hanatoxin over the last 20 ns of simulations was computed. The agreement between the experimental and computational orientations found for Vstx1 indicates that Hanatoxin probably inserts into the membrane as illustrated in Figure 3.7. The membrane interacting surface of the toxin displays a cluster of mostly hydrophobic residues (Y4, L5, F6, C28, A29, W30, F32) and a higher variability at the C-terminus in comparison to the N-terminus. With this orientation, most charged side chains (E1, R3, K10, K17, K22, D25, K26) sit at the level of the phosphate groups (dotted line in Figure 3.7A).

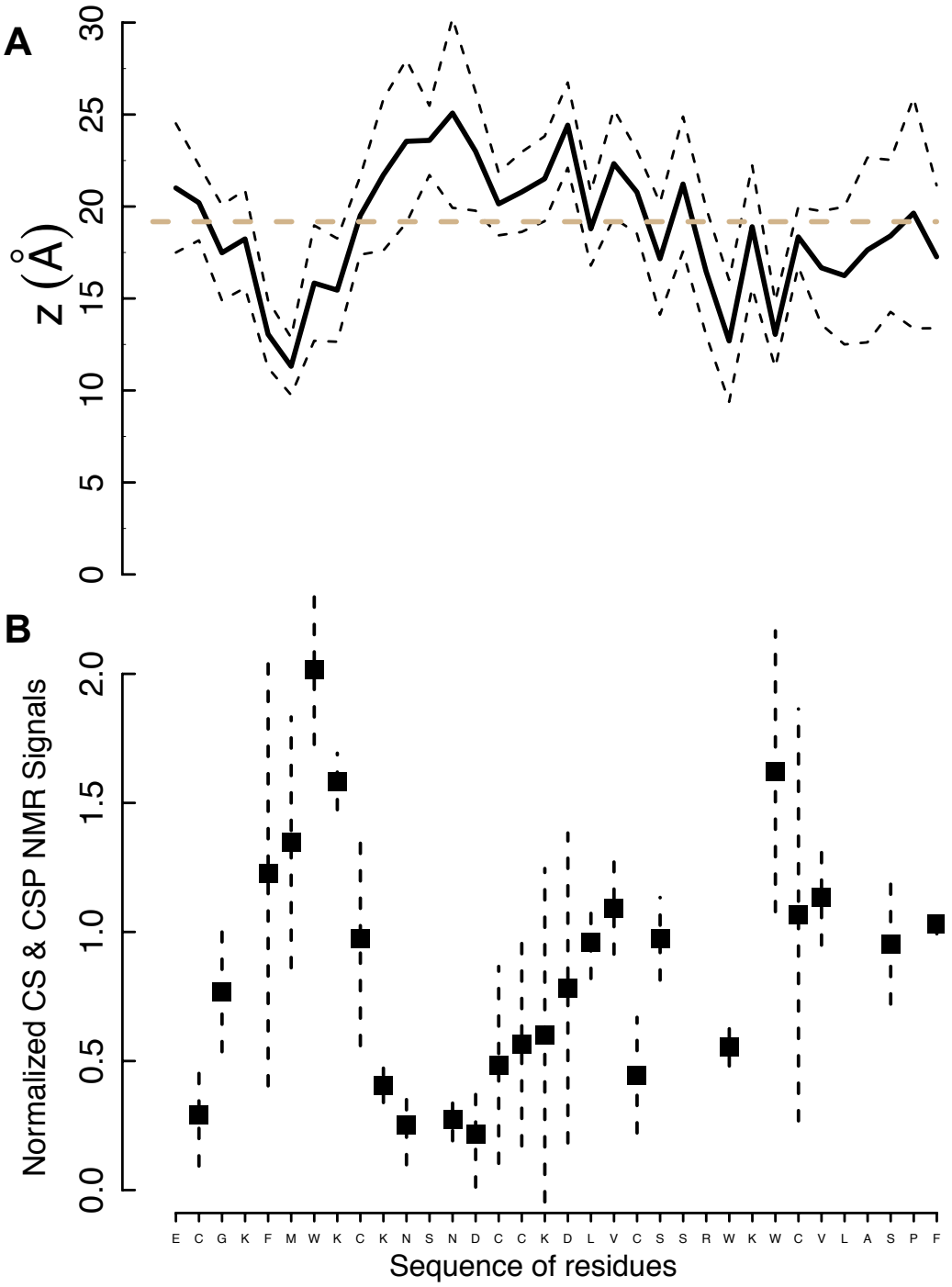


Figure 3.3: Orientation of Vstx1 upon interaction with the membrane and its correspondence with experimental data.

A) Position of residue side chain along the z-axis, averaged over the last 20 ns of 5 simulations. Solid lines: averages. Dashed lines: average ± standard deviation. The level of the phosphate groups is represented by tan dashed lines. z = 0 Å corresponds to the center of the membrane. B) Normalized chemical shift perturbations and cross-saturation experiment values from Ozawa et al. (*99*). CS: cross-saturation, CSP: chemical shift perturbation. The averages ± standard deviations are shown. Residues for which at least one experimental value was missing were omitted.

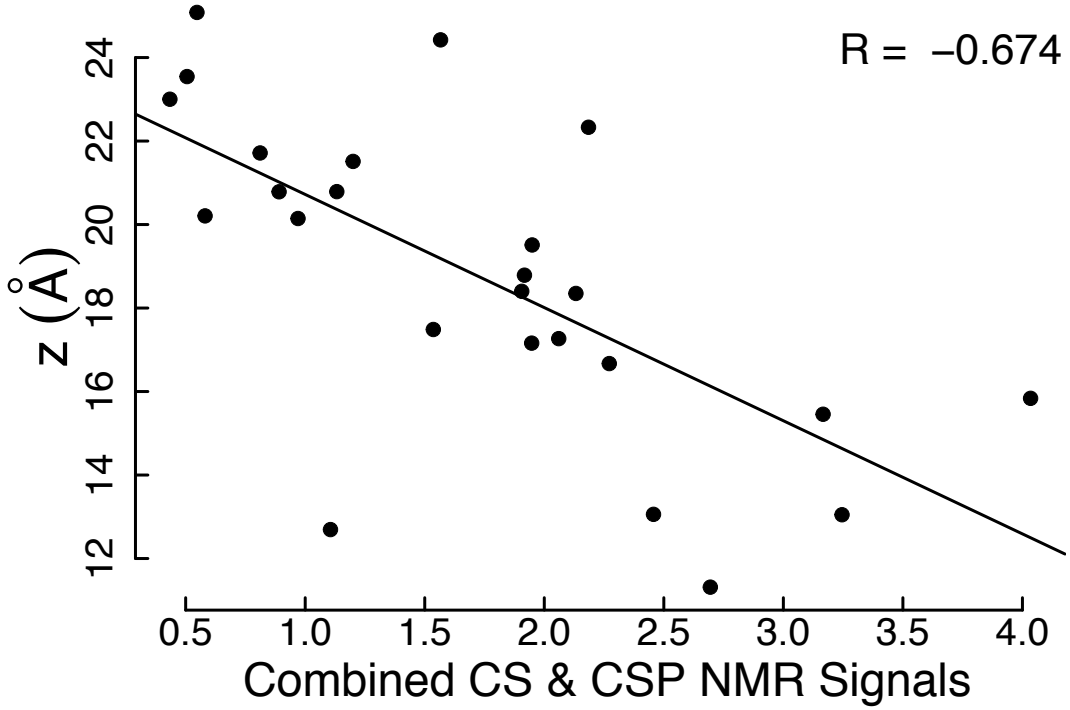


Figure 3.4: Relationship between the NMR membrane interaction signals and the depth of Vstx1 residue side chains.
MD measurements were averaged over the last 20 ns of 5 simulations.

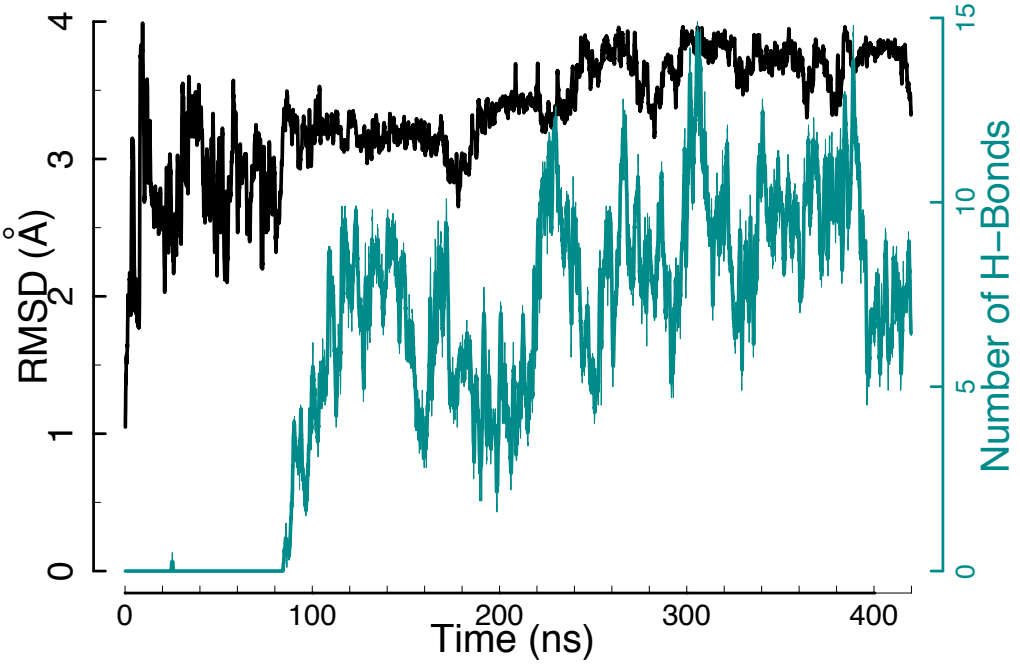


Figure 3.5. The overall structure of Vstx1 is only slightly affected upon binding to the membrane.
The time course of the RMSD to the deposited structure (black) does not correlate with the binding to the membrane, which is assessed by the number of hydrogen bonds between the toxin and the lipid head groups (cyan).

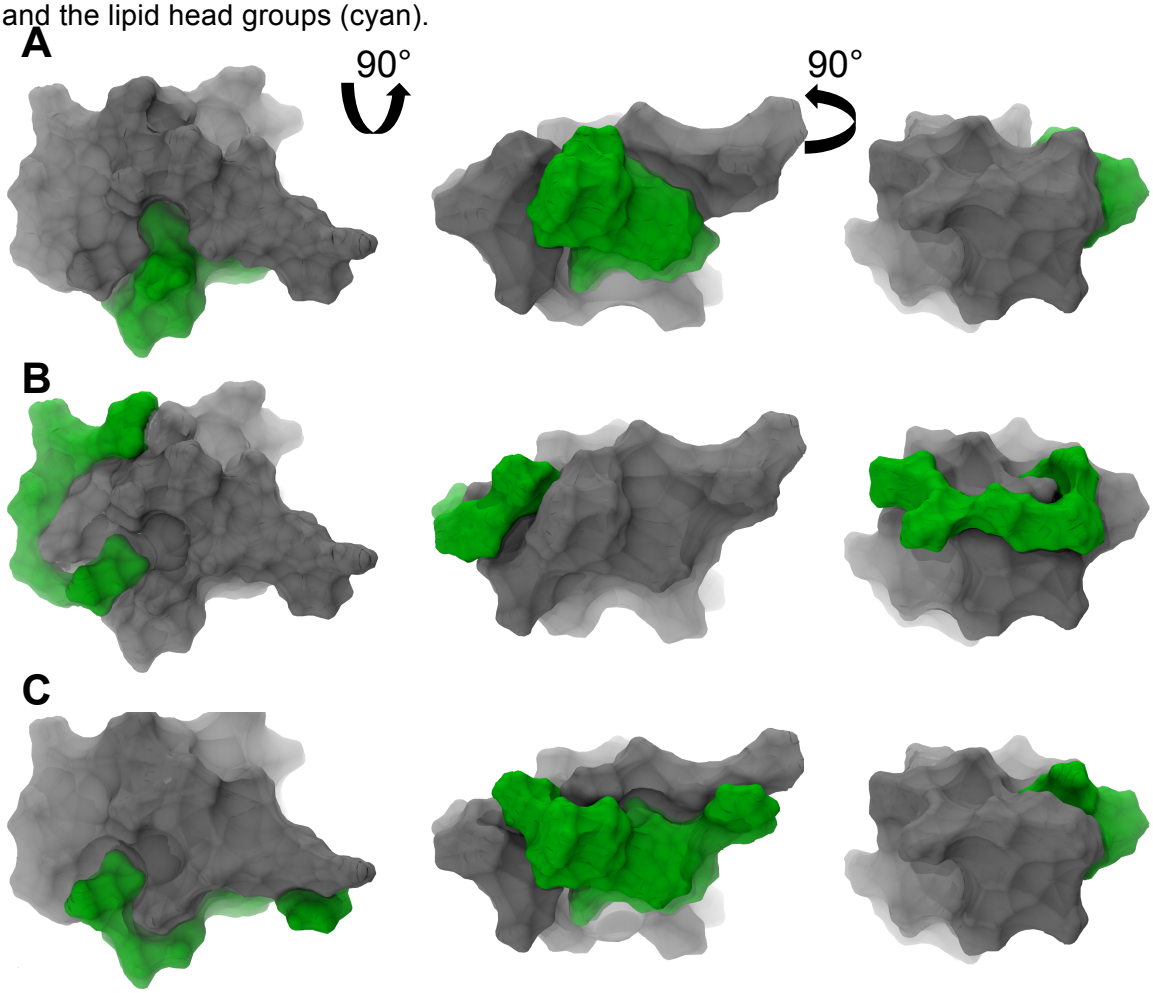


Figure 3.6. Similar NMR determined and MD predicted clusters of Vstx1 residues interacting with the membrane.
The surfaces of residues interacting with the membrane, colored in green, were determined by three methods. A) NMR chemical shifts perturbation. B) Neutron diffraction data. C) MD simulation: Residues for which the side chains inserted more than 4 Å below the phosphate groups are colored in green.

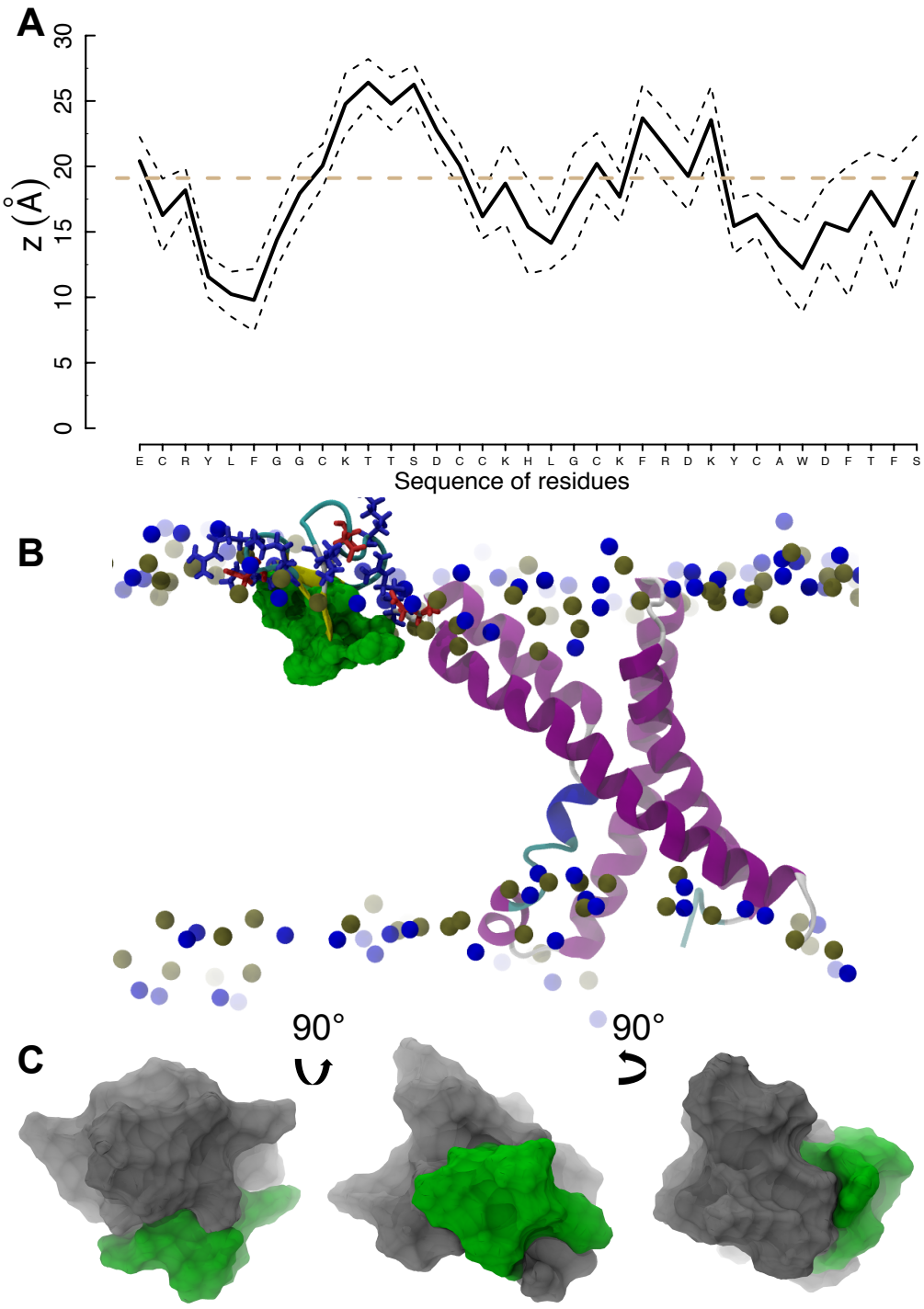


Figure 3.7. Orientation of Hanatoxin upon interaction with the membrane.
A) Position of residue side chains along the z-axis, averaged over the last 20 ns of 6 simulations. Solid lines: averages. Dashed lines: average ± standard deviation. The level of the phosphate groups is represented by a tan dashed line. z = 0 Å corresponds to the center of the membrane. B) Snapshot of an inserted toxin after 200 ns of simulation. The proteins are shown with helices in purple, β-strands in yellow, turns in cyan and loops in silver. Tyr4, Leu5, Phe6, Gly7, Trp25 and Phe27 of Hanatoxin are shown as green surfaces. The phosphate and nitrogen atoms of the phospholipids are shown in tan and blue spheres, respectively. C) Surface representations of Hanatoxin with Tyr4, Leu5, Phe6, Gly7, Trp25 and Phe27 colored in green, and the other residues in grey.

3.3.1.2 Disordering of the lipid chains near the toxins

The toxin may be expected to locally decrease the order of the lipid tails upon insertion in the membrane. Lipid order parameters can be extracted from ^{2}H NMR measurements or calculated from the atom coordinates in an MD trajectory. The order parameter S_{CD} is defined by the following time and ensemble average:

$$SCD = \langle \frac{3\cos\theta^2 - 1}{2} \rangle$$

where θ is the instantaneous angle between the C-D (carbon-deuterium) vector and the bilayer normal. More detailed descriptions of the geometry of order parameters are given in (*117*).

The preceding section showed that a subset of the simulations displayed a toxin insertion into the membrane, which was comparable with experimental values in terms of orientation. Accordingly, the order parameters of the POPC tails for the last 20 ns of these simulations were calculated. Since the purpose of this investigation was to quantify the effect of the toxins on the membrane, lipids closer than 10 Å from a VSD were excluded. The remaining lipids were grouped in rings around a toxin, depending on their distance to the toxin, in order to pool the calculated parameters from molecules experiencing a similar environment. The order parameters of POPC Sn-1 and Sn-2 decreased significantly upon toxin insertion, as illustrated in Figure 3.8 (Vstx1) and 3.9 (Hanatoxin). The observed trends were reproducible for several independent simulations (n=5 for Vstx1, n = 6 for Hanatoxin). The relationship between the average distance to a toxin and the S_{CD} parameters was fitted by an asymptotic hyperbolic function, since the toxin is not expected to exert any measurable effect on lipids further than a given distance. Beside a slightly better fit and smaller variability among the different simulations in the case of Hanatoxin, which was also observed to insert slightly deeper than Vstx1 in the membrane, there was no apparent difference between the effects of the two toxins on the lipid tail order parameters.

An apparent contradiction deserves an explanation. Since it was evidenced above that the toxin insertion reduces the values of the order parameters in its vicinity, the fact that values of the lipids far from any toxin are slightly higher than the values of lipids in control simulations, without any toxin, is not expected (Figures 3.8 and 3.9). The insertion of the toxin increases the number of particles in the extracellular leaflet. Because the number of atoms in the intracellular leaflet does not change, and as the system is too small to respond with a significant curvature change, the lateral pressure of the extracellular leaflet increases as a consequence of the higher number of particles. This higher lateral pressure restraints the acyl chains and consequently increases the values of the order parameters. In other words, these observations would be a consequence of the finite and relatively small size of the computed system, and would not be biologically relevant.

In a recent investigation of the insertion of a small molecule (S-methyl methanethiosulfonate) in membrane, Miguel et al. described a complex effect on the ordering of 1,2-dipalmitoyl-sn-glycero-3-phosphocholine (DPPC) bilayers. Four MD simulations were performed, each at a different concentration. While the small concentrations (0.09 and 0.18 M) induced a slight disordering of the S_{CD} parameters (about 0.02 smaller values for positions C6

and higher), the two simulations at higher concentrations (0.33 and 0.44 M) suggested an ordering effect (*118*). Disordering of lipid chain upon peptide insertion has also been reported by Kandasamy and Larson (*119*). However, in a NMR study, Dave et al. investigated the membrane insertion of a 28 residues long α-helix. This helix contains mainly hydrophobic residues and one Arg at its N-terminus. The measured Sn-1 order parameters of POPC did not vary upon interaction with a 4 mol solution of the peptide (*120*). In a recent NMR investigation of the interactions between Vstx1 and POPC (Vstx1: lipid ratio 1:100), Mihailescu et al. recorded through ^{2}H solid-state NMR the Sn-1 S_{CD} parameters of the lipid and reported a decrease of about 0.03 of the order parameters (*100*). The analysis performed in this work, and presented in Figure 3.8 emphasizes the effect of the toxin as a function of the distance, which may help understanding the involved mechanisms. However, in order to compare the results from the MD with the ^{2}H solid-state NMR measurements, Figure 3.10 presents a comparison of the average Sn-1 order parameters in simulations containing an inserted Vstx1 toxin with control simulations, which did not contain any toxin. The magnitude of the effect is very close to the decrease found in the ^{2}H solid-state NMR experiment (*100*).

All together, these experimental and computational results show that the interactions between a protein and a membrane do not always induce a disordering of the acyl chains. In the case of Vstx1 and Hanatoxin, however, NMR experimental measurements and computational observations suggest a clear disordering at the vicinity of the toxin. In addition, the effects are similar for both toxins.

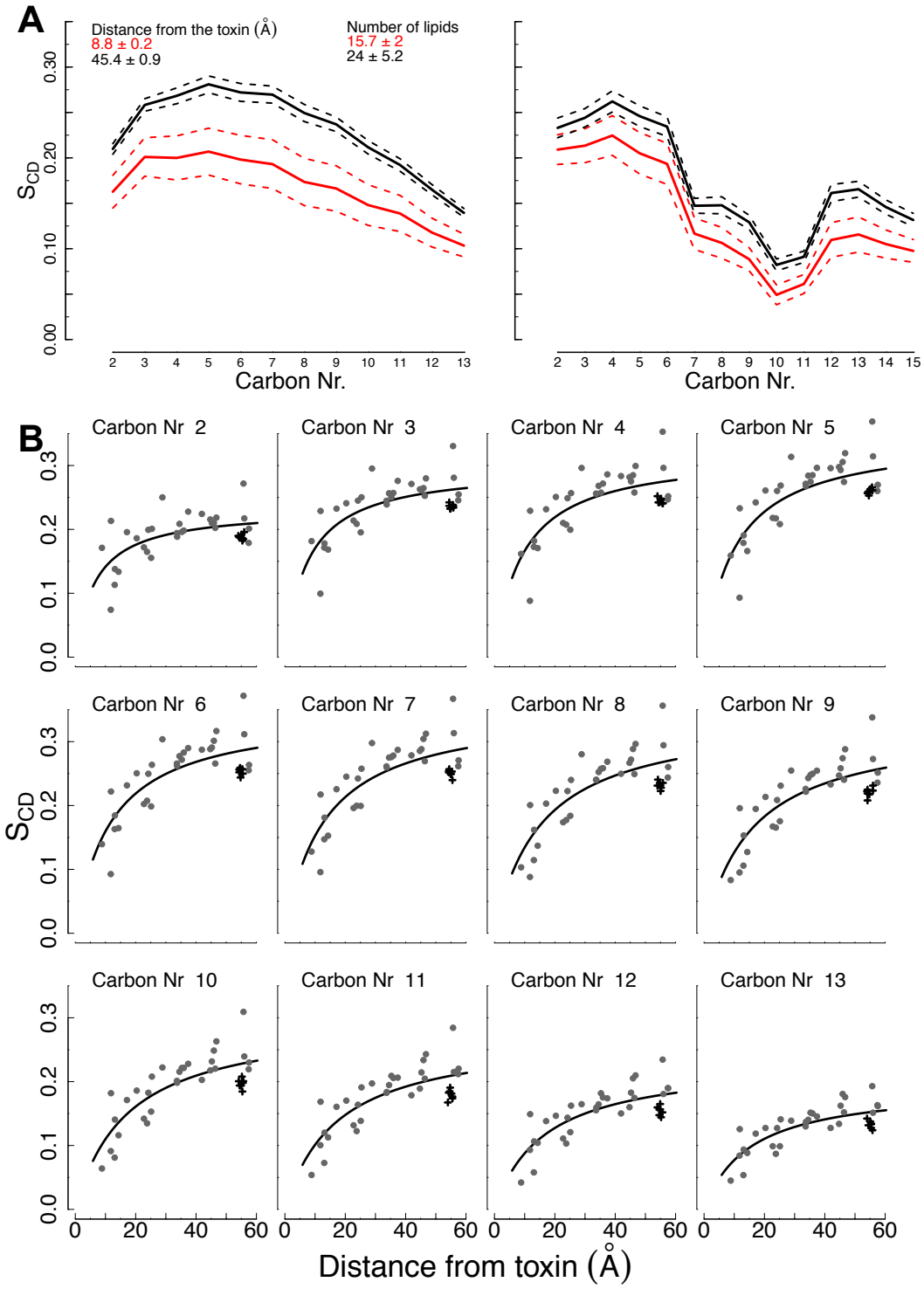


Figure 3.8. The S_{CD} lipid order parameters decrease near Vstx1.
A) S_{CD} order parameters (left: Sn-1, right: Sn-2) of POPC lipids pooled as a function of their distance to the inserted peptide. Solid lines: averages. Dashed lines: average ± standard error of the mean, n = 6 independent simulations. The inset gives the averages and standard deviations of the distance to the toxin and the number of lipids. Acyl-chain carbons on the x-axis are numbered according to their serial position. B) Sn-1 S_{CD} of each carbon atom as a function of the distance to the toxin for 6 independent simulations. The smooth curves correspond to asymptotic hyperbolic functions. The black crosses correspond to the S_{CD} values of 7 different simulations without toxin (included at an arbitrary distance of around 55 Å).

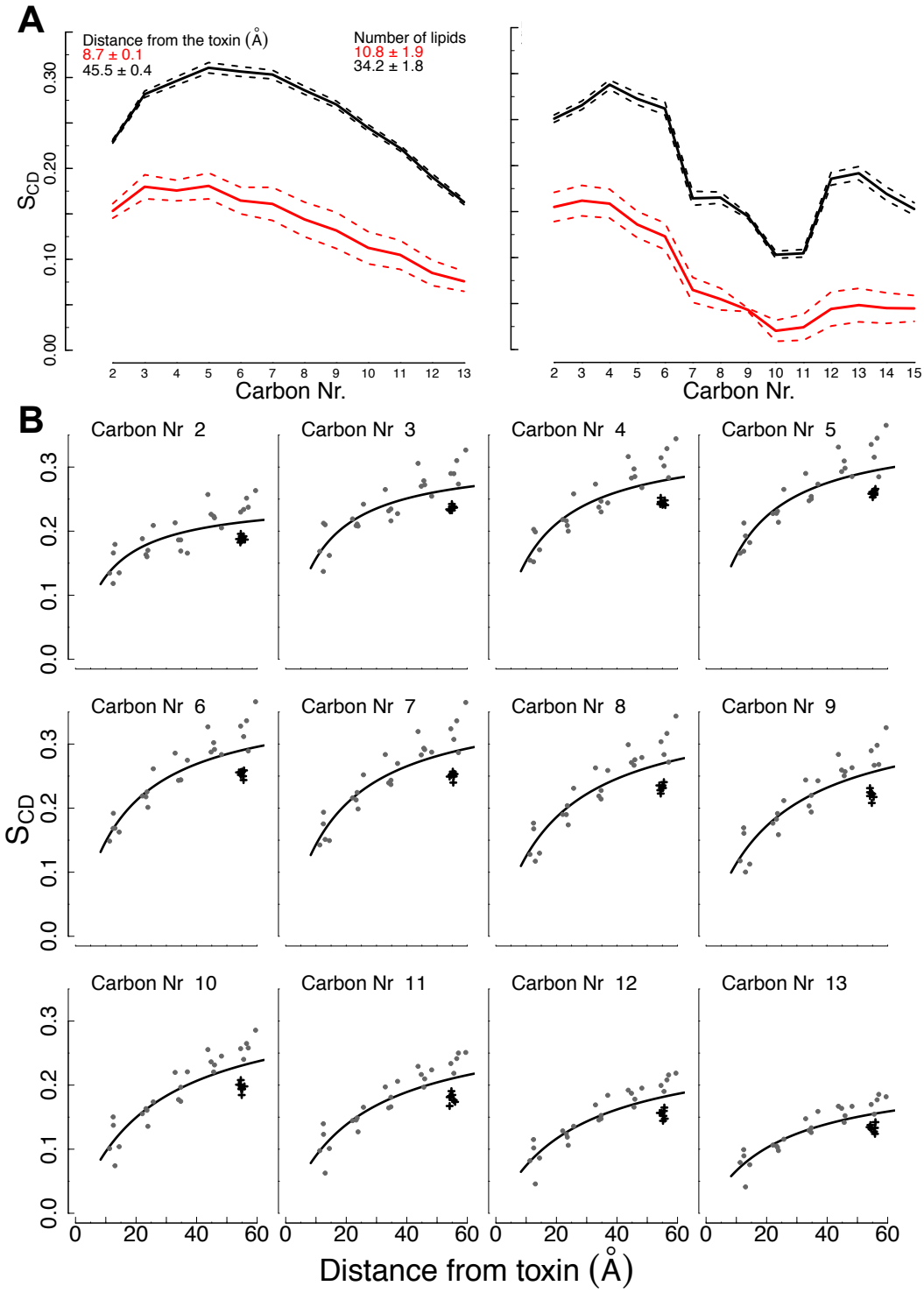


Figure 3.9. The S_{CD} lipid order parameters decrease near Hanatoxin.
A) S_{CD} order parameters (left: Sn-1, right: Sn-2) of POPC lipids pooled as a function of their distance to the inserted peptide. Solid lines: averages. Dashed lines: average ± standard error of the mean, n = 5 independent simulations. The insets give the averages and standard deviations of the distance to the toxin and the number of lipids. B) Sn-1 S_{CD} of each carbon atom as a function of the distance to the toxin for 5 independent simulations Smooth curves correspond to asymptotic hyperbolic functions. The black crosses correspond to the S_{CD} values of 7 different simulations without toxin (included at an arbitrary distance of around 55 Å).

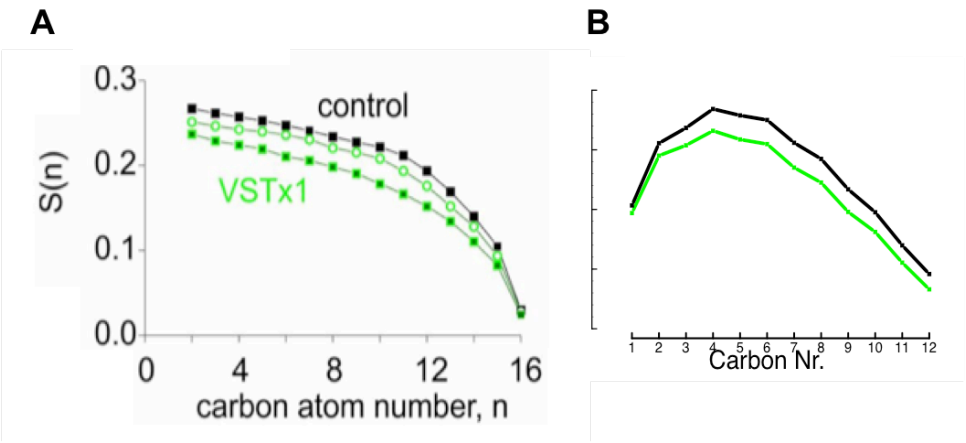


Figure 3.10. Reproducibility of experimental data with regard to the disordering effect.
A) Disordering effect observed in NMR measurements (taken from Mihailescu et al.(*100*)). B) SCD order parameters of POPC lipids in simulations containing a membrane interacting toxin (green, n = 6 simulations) and control simulations without toxin (black, n = 3 simulations) are shown.

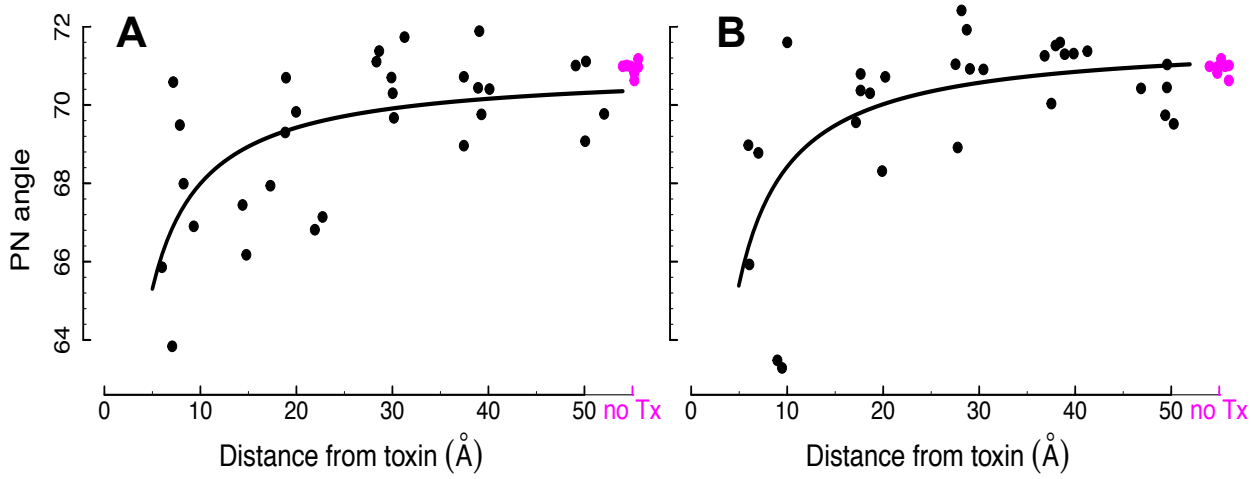


Figure 3.11. The angle of the POPC P-N vectors relative to the normal of the bilayer decreases near the toxins.
The lipid molecules from 6 different simulations were clustered as a function of their average distance to the toxin during the last 20 ns of simulations. The smooth curves correspond to hyperbolic functions with asymptotes at ≈ 70-71 degrees in both panels. The values corresponding to 7 different simulations where no toxin was included are shown in magenta at an arbitrary distance around 55 Å from any toxin. A) Vstx1. B) Hanatoxin.

3.3.1.3 Reorientation of the phosphocholine head groups

As shown above, experimental and computational methods suggest that Vstx1 and Hanatoxin partition into the membrane through a cluster of hydrophobic residues while most of the charged residues remain close to the surface of the membrane, where they interact with the lipid head groups. Most spider toxins carry a global positive electric charge. The global charge of Hanatoxin and Vstx1 are +2 and +3, respectively. Scherer and Seelig have measured alterations of the orientation of phosphocholine head groups upon interaction with electrically charged amphiphile molecules (*49*, *50*). Their measurements suggested that the choline group of the P-N dipole reorients toward the water phase upon addition of positively charged amphiphiles. The same type of effect was observed independently of the nature of the electrically charged molecules (peptides, lipids, metal ions). The rational explaining this P-N response relies on the insertion of the charged amphiphile molecule in such a way that the effective positive charge interacts with the phosphodiesters and so inserts below the choline functional groups. Consequently, the positive choline group would be repelled. A similar electrostatic effect may be expected upon interaction of the spider toxin with the membrane. According to the MD simulations, most of the arginine and lysine side chains, particularly for Vstx1, were positioned just below or close to the level of the phosphate groups (Figures 3.3 and 3.7).

Similarly to the calculations described in the section 3.3.1.2, the cosines of the angle formed by the P-N vector of each individual lipid and the normal to the bilayer were computed for the last 20 ns. The same simulations as in the last section were used for this analysis. The strategy was to sample over several simulations in order to disentangle the variability among simulations from variations due to a specific interaction with the toxin. The lipids closer than 10 Å from the VSD were excluded for this analysis. The lipids were then clustered within concentric circles around an inserted toxin.

Vstx1 insertion into the membrane induced a significant change of the POPC head group orientation ($p < 0.001$). The relationship between the average distance to a toxin and the angle relative to the normal of the bilayer was fitted by a hyperbolic asymptotic function, with an asymptote corresponding to around 70 degrees far from the toxin. However, the head group reorientation, although similar among different simulations, was rather small. The average angle of the POPC P-N vector relative to the membrane normal decreased from around 70 degrees to approximately 66 degrees for lipids close to the toxin (Figure 3.11A). Hanatoxin insertion affected the head group P-N vectors similarly to Vstx1 ($p < 0.001$, from approximately 71 to 66 degrees, Figure 3.11B).

The previous analysis focused on the distance between the lipids and the toxin. Upon toxin insertion, two different phenomena may influence the reorientation of the head groups. First, the positively charged toxin may repel the positively charged choline groups. Second, with decreasing distance between the toxin and the lipids, hydrogen bonds will form between the toxin and the phosphate groups, leading to a steric displacement of the choline head groups. The

observed reorientation might be the synergistic result of the two phenomena. The dependence of head group tilting on the formation of hydrogen bonds between acidic residues and the phosphate groups can be extracted from the trajectories, and differentiated from the Coulomb effect. In order to investigate direct contacts between the toxin and the lipids, the number of intermolecular hydrogen bonds formed between the peptide and the POPC molecules was summed. A hydrogen bond was counted if an interaction between a donor and an acceptor (defined by a distance of less than 3.5 Å and an angle of more than 30°) was observed for more than 10% of the investigated simulation time. As in the previous section, the simulations for which the average toxin insertion was below the level of the phosphate groups were selected. In most cases, hydrogen bond acceptors were the oxygen atoms of the phosphate groups. Six examined simulations involving an inserted Hanatoxin displayed 139 hydrogen bonds, of which 110 involved the phosphate groups and 29 the ester or carbonyl oxygen atoms of the bilayer. On the side of the toxin, the 4 Lys and the 2 Arg were involved in approximately half of the total number of hydrogen bonds formed with the membrane, implying a main function for these residues in terms of maintaining the toxin at the surface of the bilayer. Similar hydrogen bond proportions were observed upon insertion of Vstx1. Using the results of the toxin insertion analysis of section 3.3.1.1, it is possible to identify the residues which most likely interact with the phosphate groups. Focusing on these residues, Figure 3.12 shows that the reorientation of the lipid head groups follows a positive relationship with the number of hydrogen bonds they form with the toxin. This indicates that a significant part of the head group reorientation is due to direct contacts between those residues of the toxin and the phosphodiesters.

On the other hand, if the global charge were the only factor affecting the head groups, Vstx1 could be expected to exert a stronger influence, since it carries a charge of +3, while Hanatoxin carries a charge of +2. Yet, the relationship is slightly steeper in the case of Hanatoxin. Interestingly, Figure 3.12 shows that a lipid forms up to three hydrogen bonds with Hanatoxin. The inspection of that particular case shows that His and Arg residues of the toxin interact with the phosphate groups, pushing the choline group toward the water phase. Consequently, this phosphocholine group forms an angle of 40° with the bilayer normal, compared to the average value of about 71° observed far from any toxin. A similar head group reorientation could be observed upon interaction between Arg residues of Vstx1 and phosphodiesters (Figure 3.13).

In experiments involving the interaction of the positively charged dialkyldimethylammonium with a POPC bilayer, a P-N dipole reorientation up to 30° has been measured, with the choline group moving toward the water phase (*49*). However, this elevated value was obtained using an oversaturated mole fraction of 0.8, (50 times higher than the concentration typically used in experiments with spider toxins). This concentration implies that after partitioning almost every second molecule in the membrane would be an added amphiphile, assuming that all the added dialkyldimethylammonium molecules partitioned into the membrane.

In standard experiments involving toxins, as well as in the MD simulations performed in this work, the mole fraction is much lower, and corresponds to one toxin for approximately 200 lipid molecules, or an initial toxin concentration in water of about 4 mmol/L.

3.3.1.4 Reduced membrane thickness

Scattering-length density experiments suggest that Vstx1 thins the bilayer. At a protein-to-lipid ratio of 1:30, Mihailescu et al. (*100*) measured a decrease of about 2 Å of the membrane thickness. The same thinning of the membrane of about 2 Å was observed at the vicinity of Vstx1 in the simulations (Figure 3.14A). More specifically, the comparison between the values corresponding to the distance from the upper leaflet to the middle of the bilayer versus the distance to the opposite leaflet implies that the thinning occurs mainly between the level of the phosphate groups and the center of the bilayer.

In the case of Hanatoxin, however, almost parallel slopes are observed for the length of the phosphate groups to the center of the bilayer (0.024 Å·Å^{-1}) as to the opposite (lower) leaflet (0.021 Å·Å^{-1}). These slopes link the given membrane thickness to the distance to the insertion area of the toxin.

Additionally, in both cases the relationship appears to be linear until the furthest distance within the simulated box, indicating that the observed effect extends to a distance of more than 40 Å from the toxin. Although the construct contained 200 lipids per leaflet, this observation shows that a larger system is required to observe lipids completely relaxed from the effect of the toxin insertion. For this reason, the computed regressions do not tell whether the thinning of the membrane due to each toxin really differ from each other.

3.3.1.5 Conclusion

The purpose of the comparison of Vstx1 and Hanatoxin was to investigate the similarities and differences of the membrane bilayer perturbations induced by their insertion, in order to explore a possible indirect mode of voltage gating modification. The term indirect mode of gating modification implies any modification of voltage gating that would be exerted by the toxin without direct binding to the VSD. The experimentally observed specificity of these two toxins led to the approach used here: Vstx1 is a gating modifier of KvAP, whereas Hanatoxin does not affect this VSD. Therefore, different perturbations of the bilayer would be potential hints for further investigations of an indirect effect.

In the simulations, both toxins showed a comparable pattern of interactions with the membrane. After 200 ns, the depth of the aromatic or methionine residues was at most 8-10 Å below the phosphate groups. Vstx1 decreased the Sn-1 and Sn-2 S_{CD} order parameters of the

POPC by about 0.10 respectively 0.05, whereas the effects were more pronounced with Hanatoxin, with decreases of approximately 0.15.
However, only 11 simulations of 200 ns could be used, which could make it challenging to detect very small differences between the two toxins.

Seelig et al. (*50*) observed, that contrary to many other classes of molecules, the experimental results for peptides did not always substantiate the electrostatic effect hypothesis exposed in the section 3.3.1.3. Briefly, several converging NMR investigations of the tilting of lipid head groups upon added metals, salts, detergents and lipids, all electrically charged molecules, led Seelig and colleagues to propose that the phosphocholine head groups would reorient in response to a Coulomb effect induced by the effectors interacting with the membrane. However, when proteins interacted with the membrane, they rather mentioned a "general disordering effect of membrane proteins on the average lipid structure (…), not consistent with a specific electrostatic effect at the head group level". The MD simulations lead to similar conclusions, since a strong disordering effect of the acyl chains was observed close to the toxins, and since on average the reorientation of the head groups was rather small. However, the precise investigation of the interactions between individual amino acid side chains and phosphodiesters revealed a P-N reorientations of about 30° when positively charged side chains formed hydrogen bonds with the phosphodiester groups. This reorientation is similar to the experimental findings mentioned above. Briefly, these observations indicate that the electrostatic effect hypothesis (*49*, *50*), although not substantiated at the level of globally charged peptides, applies to peptides, but at the residue level.

According to the MD trajectories, the disordering of the acyl chains and the reorientation of the head groups induced by the toxins were similar for both toxins and thus do not explain the specificity of Hanatoxin and Vstx1 with respect to KvAP. It is possible that Hanatoxin inserts deeper in the membrane. According to these results, it is unlikely that a specific and indirect, lipid-mediated mechanism of gating modification, exerts its effect through the acyl chains or a reorientation of the head groups.

An indirect mode of action could involve an alteration of the binding network required by the VSD to function properly. As explained in the section 1.2.4, the VSD requires phosphodiesters groups in order to respond to membrane potential fluctuations. In the section 3.3.2.2.3, it will be shown that the Arg side chains of the voltage-sensor domain formed hydrogen bonds with the phosphate groups of the lipids (Figure 3.22). This hydrogen bond network involves mostly Arg residues from S1 and S4. On the other hand, the study of the spider toxins revealed that they formed hydrogen bonds with the phosphate groups of the membrane, mostly via Lys and Arg residues (Figure 3.13). Consequently, an indirect mechanism involving a local competition for the binding to the phosphate groups may be proposed. In order to verify this hypothesis, one could compare the hydrogen bond network of the VSD in presence and absence of membrane interacting toxins. For each individual Arg residue of S4, the number of hydrogen

A B

bonds formed with the membrane was computed. However, neither the toxin nor the membrane voltage affected significantly the hydrogen bond pattern formed by the phosphate groups and any individual Arg residues of S4. For comparison, using a special-purpose machine designed for high-speed simulations, Jensen et al. (*1*) performed up to 256 μs long simulations of a VSD under different membrane potentials. Despite this very long simulation, the hydrogen bond pattern formed by the individual Arg residues with the membrane was only slightly affected.

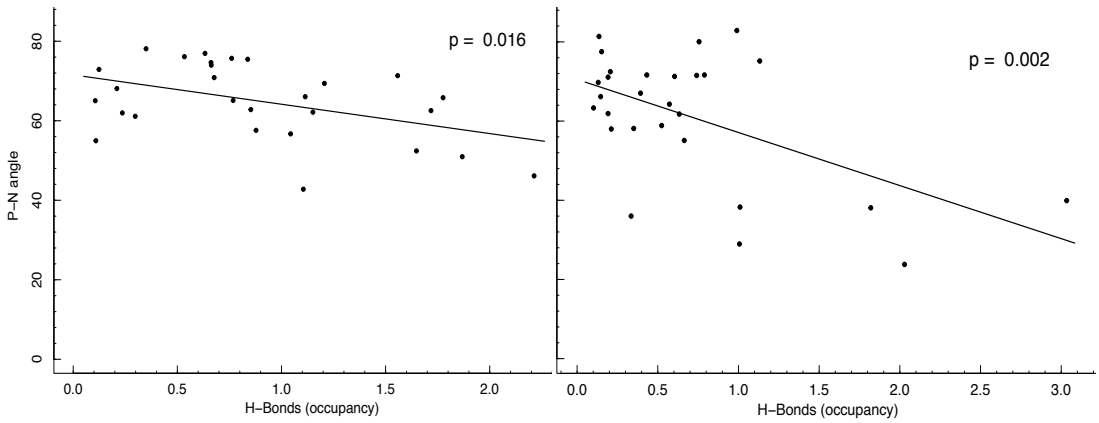


Figure 3.12. The reorientation of the head group involves hydrogen bonding with the toxins.

Lipids closer than 10 Å on average from any toxin were considered. Each point represents the average number of stable hydrogen bonds formed during the last 20 ns of a simulation between a phosphodiester and a toxin versus the cosines of the P-N vector relative to the membrane normal. A) Vstx1. B) Hanatoxin.

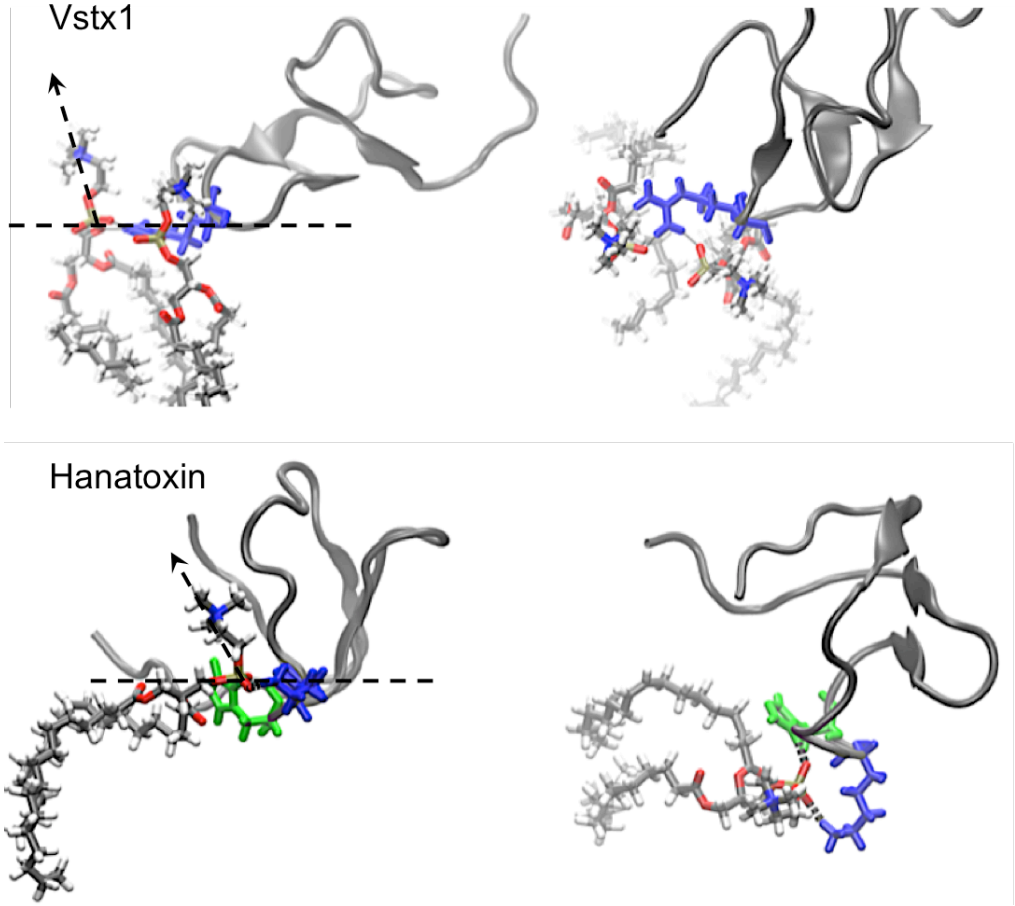


Figure 3.13. The reorientations of the P-N vectors upon direct interaction with the toxins are large.

Molecular representations of POPC molecules are shown in licorice, with toxin His (green) and Lys or Arg (blue) and backbone in cartoon representations. On the left panels (side views), the average level of the phosphate groups is highlighted (dashed line) and an arrow highlights the orientation of the P-N vector. Right panels: views from above the membrane. Above: Vstx1. Below: Hanatoxin

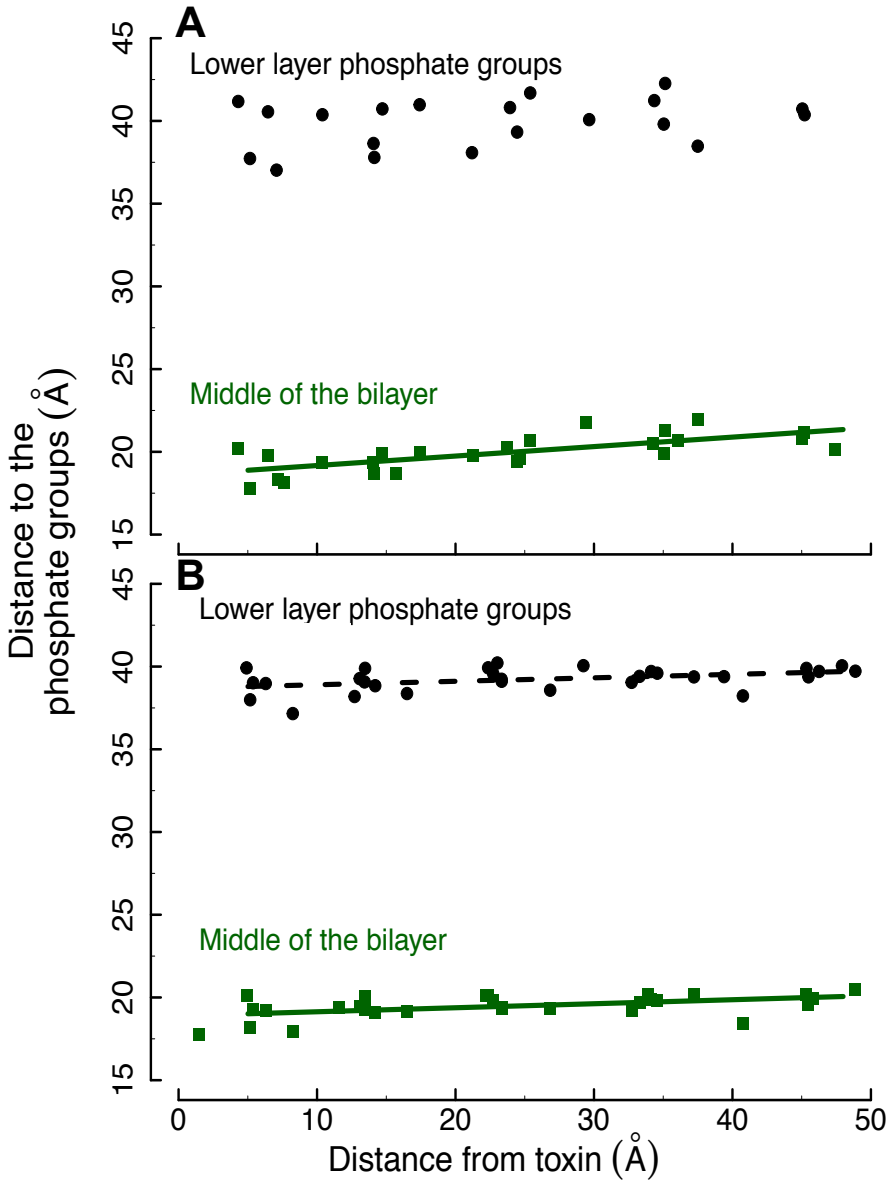


Figure 3.14. The membrane thinning induced by the toxins occurs mainly on the extracellular leaflet.

The lipid molecules of the extracellular leaflet were clustered as a function of their average distance to the toxins. The solid/dashed lines corresponds to linear regression for which $p < 0.01/0.05$. A) Vstx1. B) Hanatoxin.

3.3.2 How the membrane potential and the voltage-sensor domain affect the bilayer

3.3.2.1 Introduction

Many observations suggest that the effect of transmembrane proteins on the lipid order parameters depends on the specific proteins and lipids. Song et al. (*121*) recorded ^{2}H ss-NMR spectra along the palmitoyl chains of POPE in membranes exposed to Dermcidin, an antimicrobial 48-residue peptide that forms hexamers in membranes. They reported a disordering effect, the S_{CD} parameters being reduced by about 0.05 on average along the chain. On the other hand, an NMR study of DLPC (1,2-Dilauroyl-sn-glycero-3-phosphocholine) in bicelles reported an increase of the S_{CD} parameters when the lipids were exposed to model transmembrane peptides (*122*). In a ^{2}H and ^{31}P-NMR study of the interactions between POPC in bilayers and gramicidin, Killian et al. (*123*) observed that the S_{CD} order parameters were not affected by the transmembrane peptide. In an all-atom, explicit solvent simulation involving G-protein coupled receptors and POPC bilayers, Mondal et al. reported that the acyl chains were strongly disordered near the proteins. The amplitude of the observed disordering was similar to the one described below (*1*). Thus, similarly to the problem of inserted proteins, introduced previously, we are lacking a clear picture of the effect of transmembrane proteins on the acyl-chain order parameters, which seems to depend on the protein-lipid interaction pair.

The effects of some physical factors on the lipid tails, like the temperature (*1*) have been experimentally studied. However, despite the clear importance of the membrane voltage in the function of cells, direct investigations of the membrane structure under varying voltages are scarce.

In the following, the perturbations of the acyl-chain order parameters near the VSD will be described first, followed by the reorientation of the head groups. Contrary to the study of the toxins, the sampling is higher, since the trajectories of 134 VSDs, lasting between 200 and 740 ns (average 240 ns) were performed. The water slabs corresponding to the extra- or intracellular solution (methods) contained various ion concentrations, so that initial membrane potential between -1.7 and +0.5 V were simulated. The MD simulations revealed a complex modulation of the order parameters, where the effect of the VSD on the lipids depends on the applied membrane potential. On the other hand, similarly to the study of the toxins, the phosphocholine head groups reoriented themselves close to the VSD. In addition, there is a significant reorientation of the head groups as a function of the membrane potential, in line with previous observations.

3.3.2.2 The VSD induced perturbation of the acyl chains depend on the membrane potential

3.3.2.2.1 VSD induced perturbations of the phospholipids

The results for the extra-cellular leaflet will be described first. In the case of the KvAP VSD and a bilayer containing POPC and cholesterol, the membrane potential modulates the VSD induced disordering of the POPC Sn-1 chains. The S_{CD} values corresponding to 43 independent simulations were considered, in which the lipids were clustered as a function of their distances to the VSD during the last 20 ns of the simulations. A clear decrease of the order parameters near the VSD was detected (Figure 3.15). Lipids close to the VSD (within approximately 10 Å) displayed significantly lower order parameters, as compared to lipids in the bulk. The perturbation of the lipids followed a gradient as a function of the distance to the VSD, as shown in Figure 3.16. As for the Sn-1 order parameters, The Sn-2 S_{CD} parameters of the extra-cellular lipids were strongly reduced close to the VSD and the effect slowly decreased as a function of the distance to the VSD.

Concerning the intra-cellular leaflet, due to time constraints, the analyses focused on only 10 simulations. On the other hand, since the intracellular leaflet contains two phospholipid species, the number of individual POPC or POPS in the concentric rings around the VSD was lower. The statistical power of the intra-cellular leaflet analyses is accordingly reduced. In spite of this, a similar decrease of the order parameters of the POPC Sn-1 and Sn-2 acyl chains was observed in the intra-cellular leaflet, and the amplitude was comparable to the one observed on the extra-cellular side (Figure 3.15B). However, the Sn-1 of the POPS molecules were not affected by the VSD, and a non-significant decrease was observed in the case of the Sn-2 acyl chains (Figure 3.15C).

The disordering of both Sn-1 and Sn-2 acyl chains of the POPC molecules was thus observed in the vicinity of the VSD, on the two leaflets, over 53 simulations, whereas the POPS acyl chains may not be affected by the voltage-sensor domain.

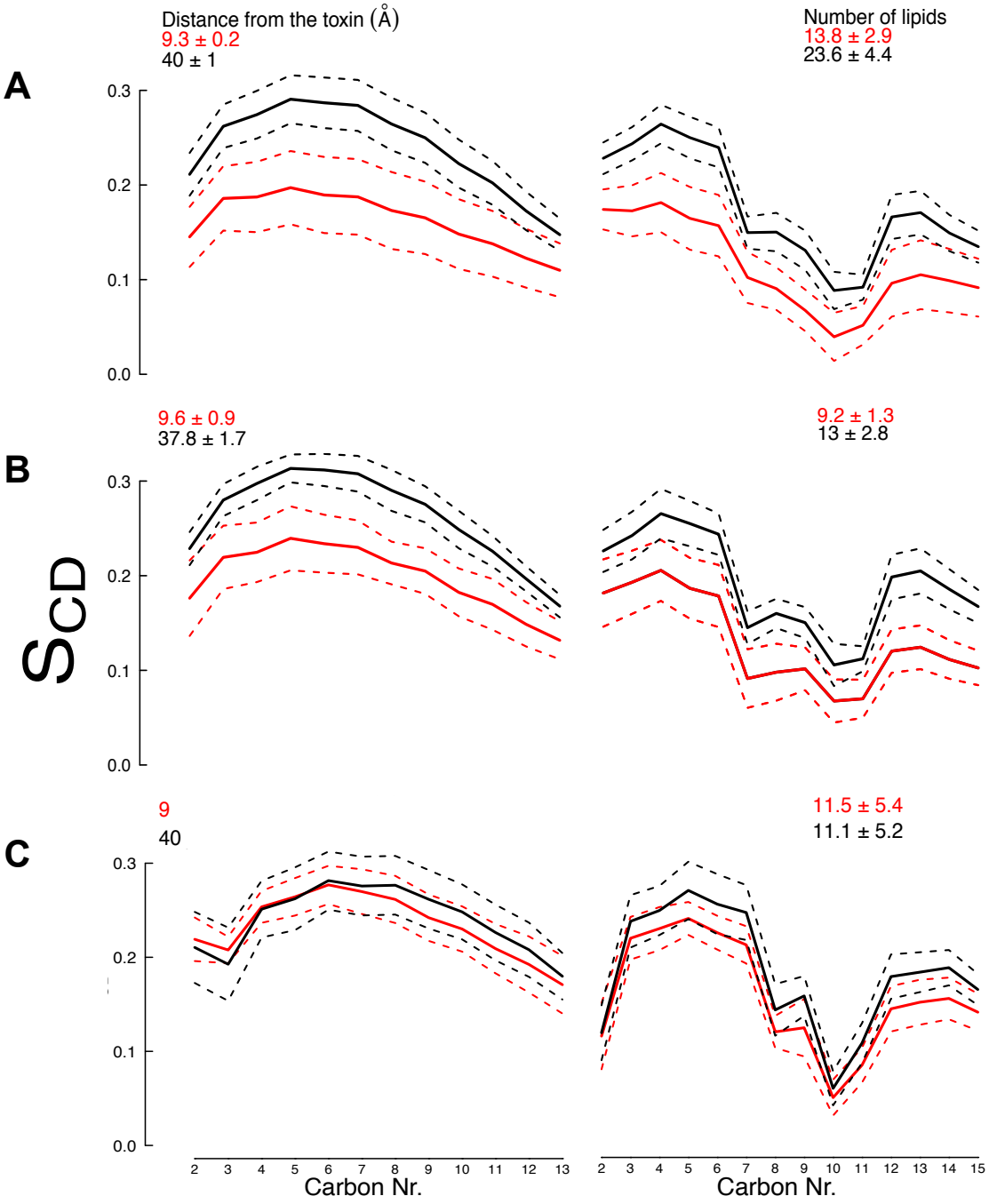


Figure 3.15. The S_{CD} lipid order parameters decrease near the VSD.
S_{CD} order parameters (A: extra-cellular leaflet; B: intra-cellular leaflet, POPC, C: intra-cellular leaflet, POPS, left: Sn-1, right: Sn-2) of lipids pooled as a function of their distance to the VSD. Solid lines: averages. Dashed lines: average ± standard deviation. Extracellular leaflet: n = 43 independent simulations, 100 POPC molecules. Intracellular leaflet: n = 10 independent simulations, 50 POPC and 50 POPS molecules. The legends give the averages and standard deviations of the distance to the VSD and the number of lipids. Note that for POPS, fixed distances of 9 and 40 Å were applied for all simulations.

3.3.2.2.2 Concerted effect of membrane potential and VSD

Focusing again on the POPC molecules of extracellular leaflet, in which there is a large number of replication, there is a relatively high variability among the 43 simulations, visible from the standard deviations in Figure 3.15A. This could be the hint that additional factors affect the features of the acyl chains. Since the conditions, and notably the membrane potential are not the same in these simulations, and since the function of the VSD is to respond to Vm, the membrane voltage could affect the interactions between the lipids and the VSD and thus explain this variability. In order to test this idea, the simulations were sorted as a function of the membrane potential and 10 simulations with strong polarization (close to -1 V) were compared with 10 simulations under depolarized conditions (from about 0 to 0.4 V). Interestingly, Figure 3.17 shows that, while the lipids in the bulk were either not at all affected by the membrane potential or were possibly slightly more ordered under a polarized potential, the lipids close to the VSD were clearly much more disordered under the polarized potential than at nearly neutral or positive membrane potential. This modulation observed for all the carbons of the chain is exemplified in Figure 3.18 for C6, which again shows that the lipids were differently affected by the membrane potential as a function of their distance to the VSD. Indeed, the membrane potential did not affect the lipids far from the VSD (the small order parameter increase upon polarization is statistically not significant). Lipids close to the VSD were significantly more disordered under polarized conditions ($p < 0.01$). This trend was not limited to the selected 20 simulations, but was observed for 43 simulations, as can be seen in Figure 3.18A. Thus, the analyses suggest a complex response of the lipid S_{CD} parameters of the Sn-1 chain: they decrease close to the VSD, but this decrease becomes much larger under a polarized membrane potential, whereas lipids far from the VSD are not affected by the membrane potential.

A comparable Vm modulation of the disordering effect induced by the VSD was not detected at this level of analysis for the Sn2 S_{CD} parameters. However, the more precise analyses showed in Figure 3.20 and discussed in the next section, indicated that the same phenomenon occurred on the Sn2 S_{CD} parameters. On the other hand, these trends were not observed on the POPS acyl chains.

3.3.2.2.3. Three possible explanations

A membrane voltage modulation of the disordering effect is observed here for the first time and one can envision dependence of the voltage-sensing mechanism in relation to the lipid environment. Experimental investigations would be required to validate this finding. Yet, despite a literature review and personal contacts with experimentalists, we could not figure out any experimental method, which investigates the lipid order parameters in electrostatically polarized

bilayers. However, in the following, three possible ways to further elucidate this phenomenon are proposed.

A first possible origin is linked to the novel conformational change of the VSD as a function of the membrane potential, which will be described in the section 3.3.2.4. Briefly, the formation of a kink and the breaking of a salt bridge, both in the middle of S4, were induced by strongly polarized potential. Since both this conformational change and the S_{CD} parameter near the VSD are related to membrane potential fluctuations, one may ask whether they might be linked. The comparison of the overall RMSD between VSD structures hold at different membrane potentials did not answer the question. Precisely, the RMSDs of the backbone atoms in reference to the deposited structure, calculated during the last 20 ns of simulations, of 14 structures of the VSD obtained from simulations performed under polarized potential (<-0.9 V) versus 18 structures of the VSD obtained from simulations performed under depolarized potential (> 0.2V), were similar, with values of about 1.5 to 3.0 Å (Figure 3.19). The section 3.3.2.4 further shows that the breaking of the salt bridge and the kink in S4 induced the movement of a positive charge across the membrane potential. The simulations performed under polarized potential were then further splitted: a group contained the four systems in which a significant charge displacement occurred, and a second group contained the nine other systems. Surprisingly, there was a significantly more important perturbation of the Sn-2 S_{CD} parameters when a gating charge transport occurred (Figure 3.20). This result indicates a link between the lipid ordering decrease and the electric charge displacement.

A second series of analyses relates the S3b helices to the acyl chain order parameters. As introduced in section 3.1, experimental investigations suggest that the C-terminus of the S3 helix, termed S3b, forms with the S4 helix a helix-turn-helix motif. The analyses revealed that the length of S3b correlated with both the lipid disorder and the membrane potential, whereas the other helices did not. Additionally, the four helices seem to display slightly different values, and particularly, the average and the variability of the rise per residue in S3b appeared to be more prominent under depolarized conditions (Figure 3.21). In order to better describe the Vm modulation of the Sn-1 S_{CD} order parameters, a specific metric is introduced. The difference between the ensemble and time averages of the S_{CD} parameter of lipids in the bulk (taken as reference) and the lipids close to the VSD is calculated:

$$\Delta S_{CD} = \langle S_{CD}(b) \rangle - \langle S_{CD}(p) \rangle$$

where the values from lipids in the bulk (b) and lipids close to the VSD (p) are considered. Using this ΔS_{CD}, it is easier to represent the perturbation of the lipid tails due to the VSD under different conditions. As shown in the Figure 3.21A, the ΔS_{CD} values increased with decreasing rise per residue of S3b under a polarized potential, while no relationship was observed under depolarized conditions. A similar interaction was not observed for S1 and S2, whereas S4 exhibits a possible

similar, but not significant, trend. In addition, the S3b helix seemed to shorten only under a polarized potential, which was not observed for the other helices. In a particular simulation started with a strong polarized potential, with Vm < -1.0 V, a gating charge displacement occurred after approximately 100 ns. At t = 200ns, the membrane potential was switched through ion displacement, so that the next 200 ns corresponded to a trajectory at Vm = +0.4 V. Interestingly, the average length of the S3b helix decreased during the first 200 ns, and increased during the next 400 ns (Figure 3.21D). The S3b helix may play a particular function in the context of the interactions with the acyl chains, because it is highly hydrophobic. Out of 13 residues, the S3b helix contains 11 hydrophobic ones. Due to the orientation of S3b at about 45° in respect to the bilayer normal, the lipids acyl chains would accordingly be reoriented in order to wrap around the helix.

The third approach describes the direct interactions between individual amino acid and POPC molecules. Using a special purpose machine designed for high-speed MD simulations, Jensen et al. performed several 14 to 256 μs long trajectories of Kv1.2 VSDs under different voltages (*1*). They observed a large translation of S4, but a slightly reduced number of salt bridges between the Arg residues of S4 and the phosphate groups in the activated state (depolarized) as compared to the resting state (polarized potential). The much shorter MD simulations performed in this work did not display any large translation of S4. The membrane potential did not alter the pattern of hydrogen bonds between any Arg residue of S4 and the lipid phosphate groups. However, a molecular representation of the interactions between the POPC molecules and the VSD shows that most of these interactions involve Arg residues making hydrogen bonds with the phosphate groups, and the lipid acyl chain tend to wrap around the hydrocarbon part of the Arg (Figure 3.23). A translation, even small, of S4 would modify the orientation of these side chains in respect to the membrane normal, which then could affect the wrapped acyl chains.

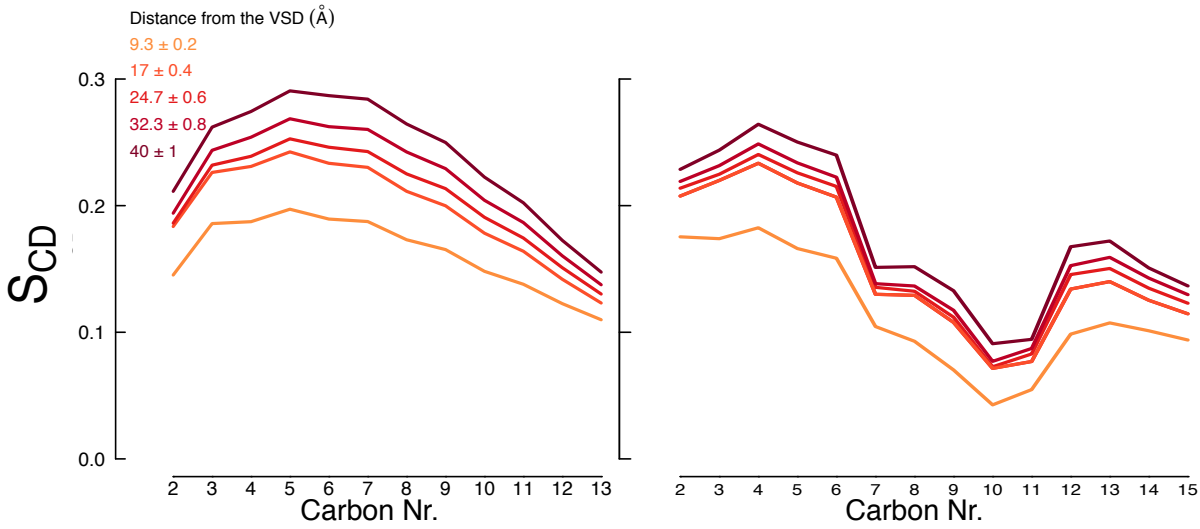


Figure 3.16: The closest to the VSD the more disordered.

SCD order parameters (Left: Sn-1; right: Sn-2) of POPC lipids clustered in rings around the VSD in the extracellular leaflet. N = 43 simulations.

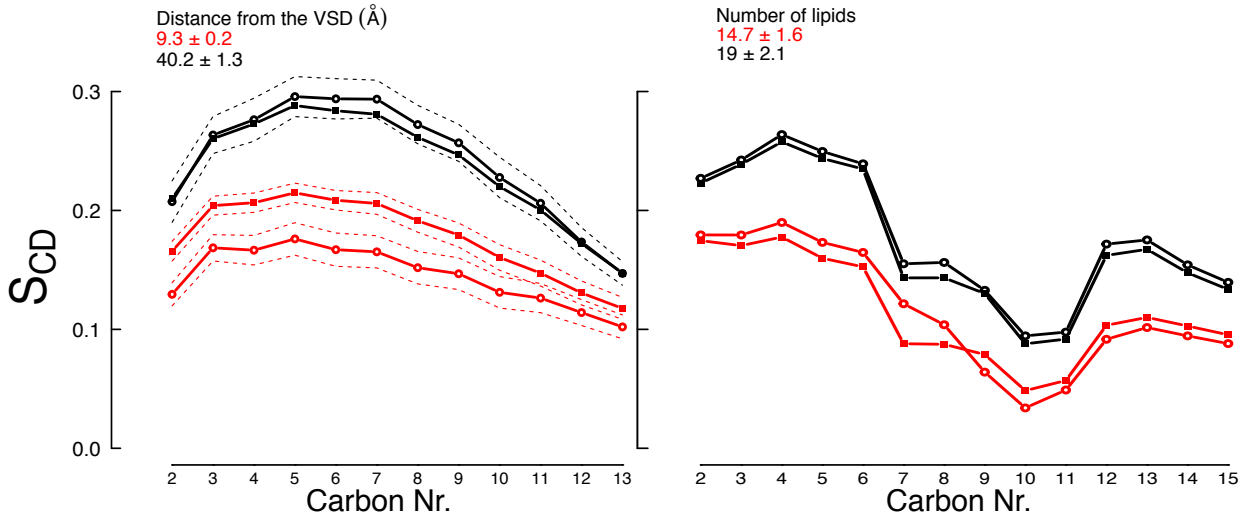


Figure 3.17. The membrane perturbations induced by the VSD depend on the membrane potential.

Order parameters SCD (Left: Sn-1; right: Sn-2) of POPC pooled as a function of their minimal distance to the VSD and of the membrane potential. (Black: lipids far from the VSD; red: lipids within 10 Å of the VSD; closed squares: depolarized potential; open circles: polarized potential). Averages of 10 simulations are shown in each case. SEM are shown for the Sn-1 order parameters.

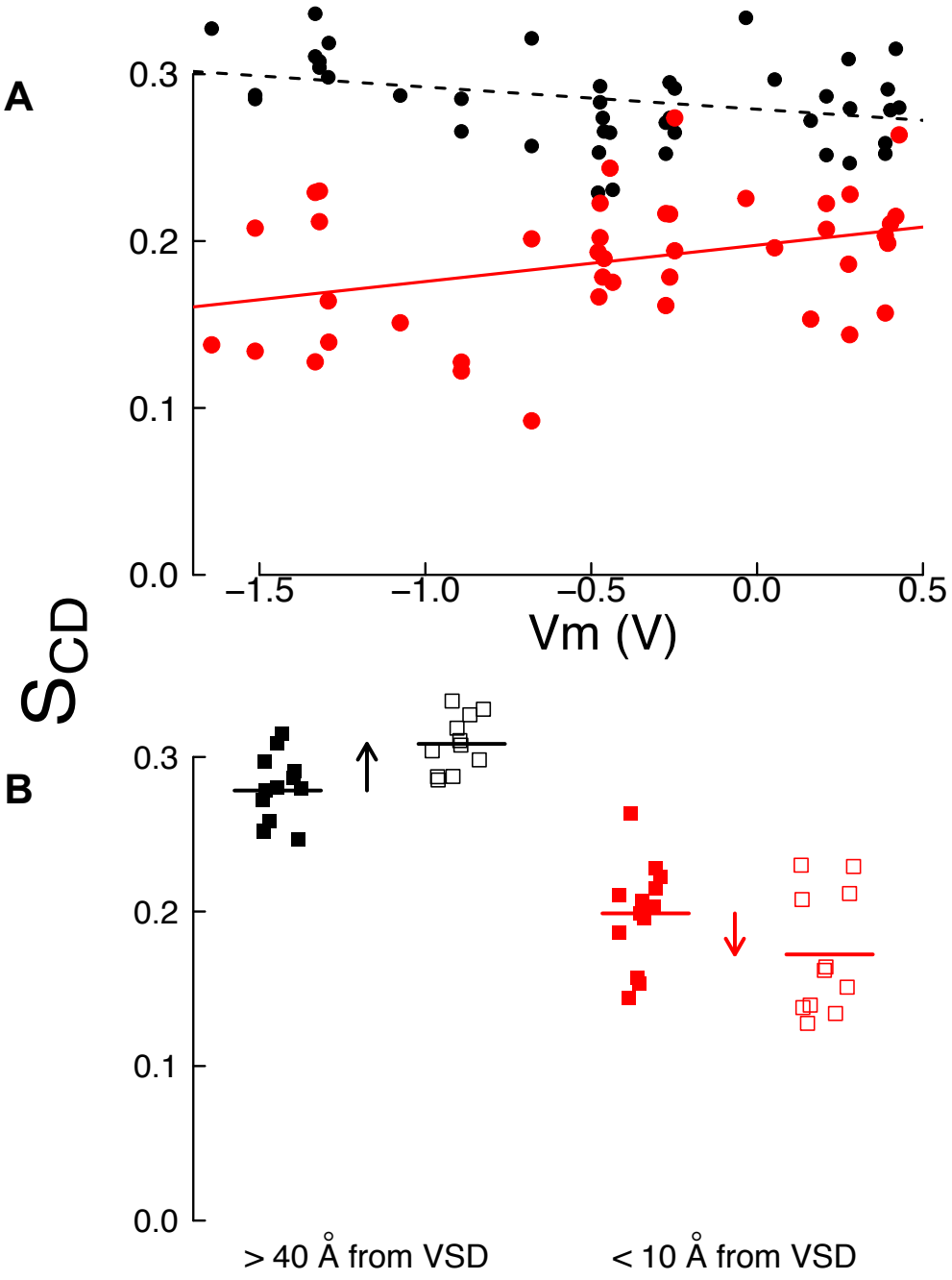


Figure 3.18. Statistical interaction between the VSD and the membrane potential
A) Sn-1 S_{CD} parameters from C6 of POPC lipids close (red) or far (black) from the VSD are shown for 43 simulations as a function of the corresponding membrane potential. Solid lines: $p < 0.05$. B) Sn-1 S_{CD} parameters from C6 of POPC lipids sorted as a function of their distance to the VSD are shown for 10 simulations under polarized potential (open squares) and under depolarized potential (closed squares). Horizontal lines: average S_{CD} values. The arrows highlight the opposite effects of the membrane potential.

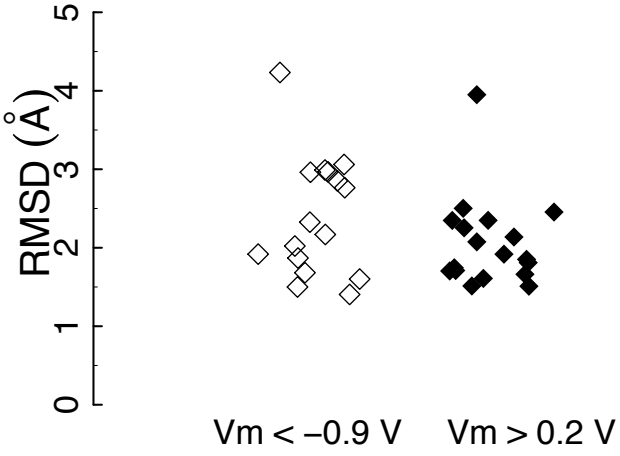


Figure 3.19. The membrane potential does not affect the RMSD of the backbone atoms in reference to the KvAP VSD deposited structure.
The values of 16 VSD under polarized potential (open squares) and 18 VSD under depolarized potential (solid squares) were computed for the last 20 ns of simulations.

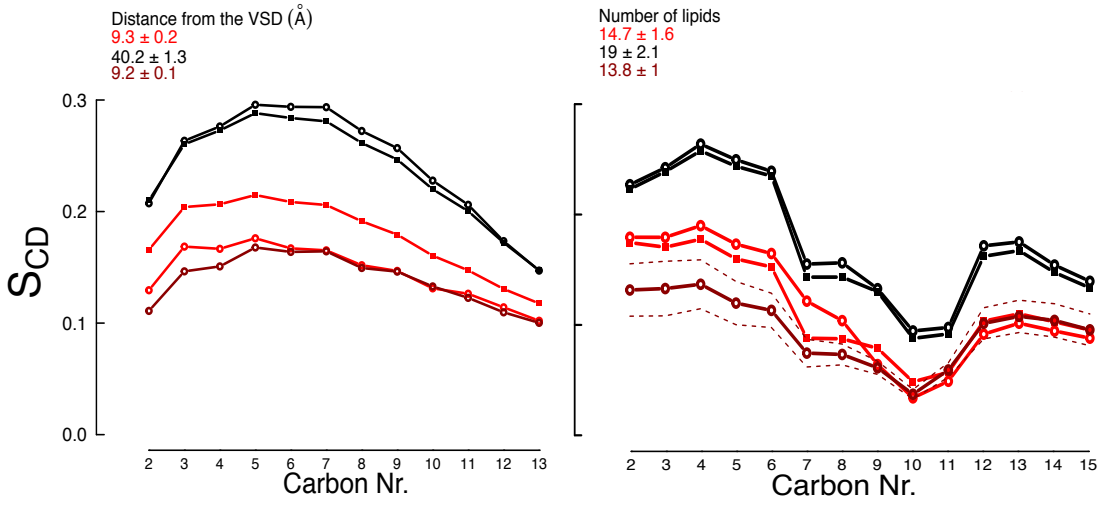


Figure 3.20: The lipid disordering decrease is associated with charge transport.
Order parameters S_{CD} of POPC pooled as a function of their minimal distance to the VSD, the membrane potential and the occurrence of a gating charge transport. Black: lipids far from the VSD; red: lipids close to the VSD, 9 simulations without any gating charge transport, dark red: lipids close to the VSD, four simulations where a gating charge transport occurred. Closed squares: Depolarized potential, Open circles: Polarized potential. Gating charge transports were only observed under polarized potential. Left: Sn1. Right: Sn-2.

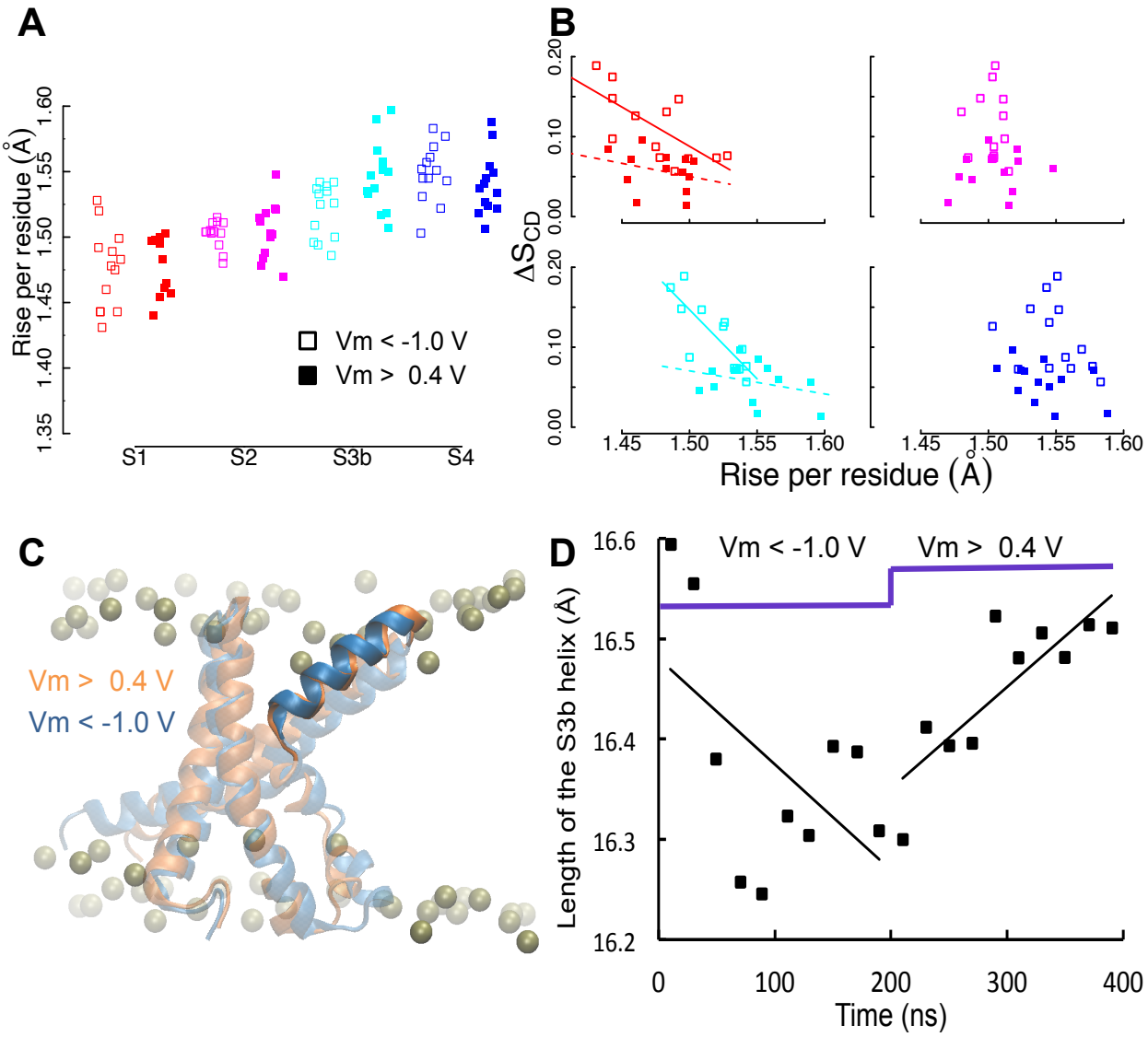


Figure 3.21. The length of the S3b helix correlates with the membrane potential and with the lipid order parameters

A) Each square corresponds to the average rise per residue of a given helix over the last 20 ns of one simulation. Red: S1; magenta: S2; cyan: S3b; blue: S4. Open symbols: polarized potential, closed symbols: depolarized potential. B) ΔS_{CD} (see text) of the four helices. Same color code as in the Figure 3.21A. C) Molecular representation of two VSDs simulated under Vm ≈ -1.0 V (blue) and +0.4 V (orange) highlighting the S3b helix. The phosphorus atoms of the POPC head groups are shown in tan. D) Time course of the length of the S3b helix in a simulation with varying Vm values. Starting at t=0 ns with Vm < -1.0 V, the membrane potential was switched to +0.4 V at t=200 ns.

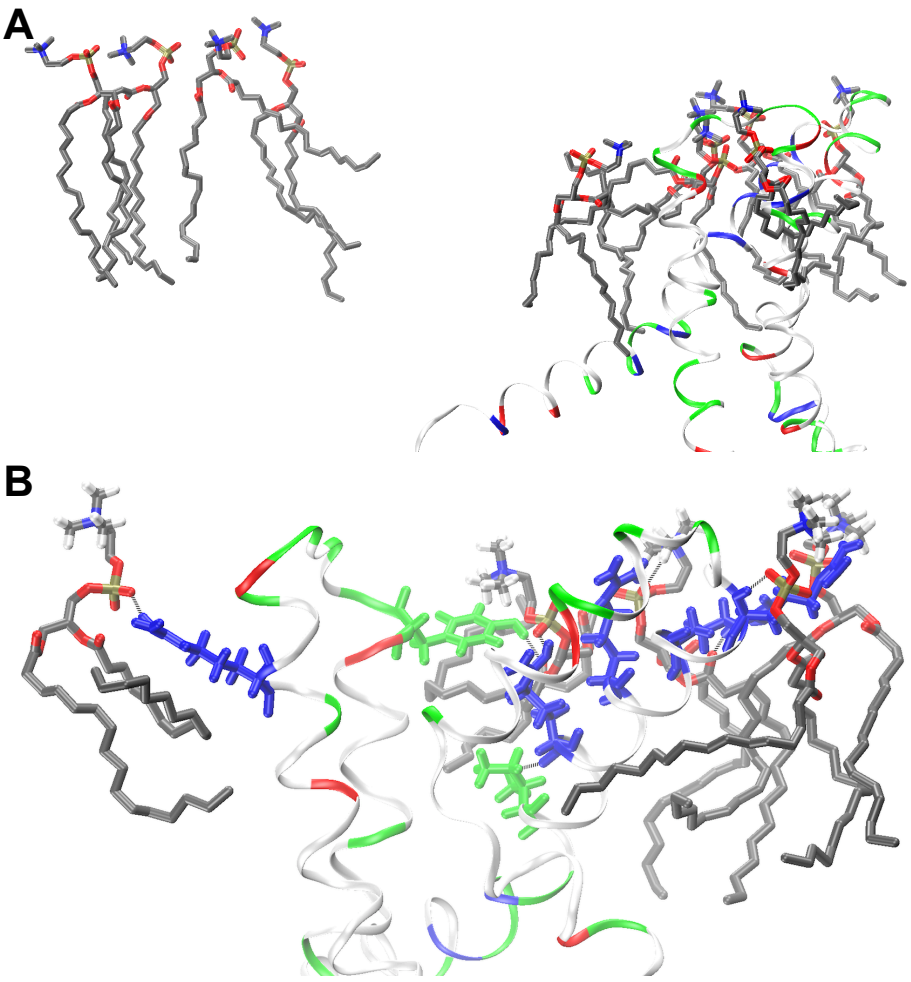


Figure 3.22. The interactions between the VSD and the lipids involve mostly hydrogen bonds with Arg side chains.
A) The tails of the POPC molecules tend to wrap around the VSD, whereas the lipids far from the VSD are well ordered. B) Interactions between phosphate groups of POPC and the following protein residues: Tyr46 (S1), Arg57 (S1), Arg117, Arg120, Arg123, Arg126 and Ile130 (all S4). The protein is shown in ribbons, highlighted residues and POPC molecules are shown in licorice. Blue: Arg. Green: Tyr or Ile. Lipids: C: grey, O: red. N: blue, H (choline group only): white.

3.3.2.3 The reorientation of the lipid head groups

The reorientation of the POPC head groups in the vicinity of Vstx1 and Hanatoxin indicated that the interaction of positively charged side chain with the phosphate groups of the lipids cause a repulsion of the choline group, which consequently moves toward the water phase. The voltage-sensor domain considered in this study carries 13 positively charged residues and 12 negatively charged ones. However, inspection of the structure embedded in a membrane shows that repartition of the charged residues along the z-direction is not homogenous. While the positively charged residues are evenly distributed across the bilayer, there are 3 Asp and 5 Glu on the intracellular leaflet, and only 3 Glu and 1 Asp on the extracellular side of the protein (Figure 3.23A). As a result, the extracellular part of the VSD carries an effective positive charge

of +2. If these basic side chains interact with lipid head groups similarly to the toxins, a similar effect as was observed for the inserted toxins may be expected.

The analyses of the MD simulations revealed a highly significant ($p < 0.001$) reorientation of the POPC head groups of the extra-cellular leaflet in the vicinity of the VSD (Figure 3.24A) The angle formed with the normal to the bilayer decreased from an average value of 70 in the bulk to about 64 degrees within 10 Å from the VSD. Similar observations could be reproduced for the POPC and the POPS molecules of the intra-cellular leaflet.

The head group orientations of the POPC and POPS molecules were clustered according to the applied trans-membrane voltage (Vm). Figure 3.25 shows that the head groups of bulk lipids displayed linear relationships with Vm, following the direction of the electric field. The reorientation amplitude was about 2 degrees over the investigated trans-membrane potential, which is in line with the recent literature. In an MD simulation involving pure POPC bilayers, Carr and MacPhee (*1*) reported a head group reorientation from 67 degrees under depolarized conditions to about 68.5 under a polarized potential of -1.95 V. In another MD simulation involving POPC only, Böckmann et al. (*1*), observed a shift from 70 degrees under depolarized conditions to 76 degrees under a polarized potential of about -2.0 V. These researchers reported larger reorientations in the intra-cellular leaflet. The results of this work also point toward larger effect in the intra-cellular leaflet, although not statistically significant. As stated above, the intra-cellular leaflet contains only 50 POPC molecules, as compared to 100 molecules in the extra-cellular leaflet, which reduces the statistical power of the test. The intra-cellular leaflet also contains 50 negatively charged POPS molecules, which may interact with the POPC. The response of the phosphatidylserine groups to the membrane potential is inversed, as compared to the phosphocholine groups, as expected because these head groups carry a negative charge (Figure 3.25B and 3.25C).

Since the slope representing the reorientation as a function of the membrane voltage, shown in Figure 3.25A, was statistically significant, the values of the angles of the POPC molecules of the extra-cellular leaflet presented on Figure 3.24A were accordingly normalized. Despite similar trends, the other angles were not normalized because the linear trends were not statistically significant.

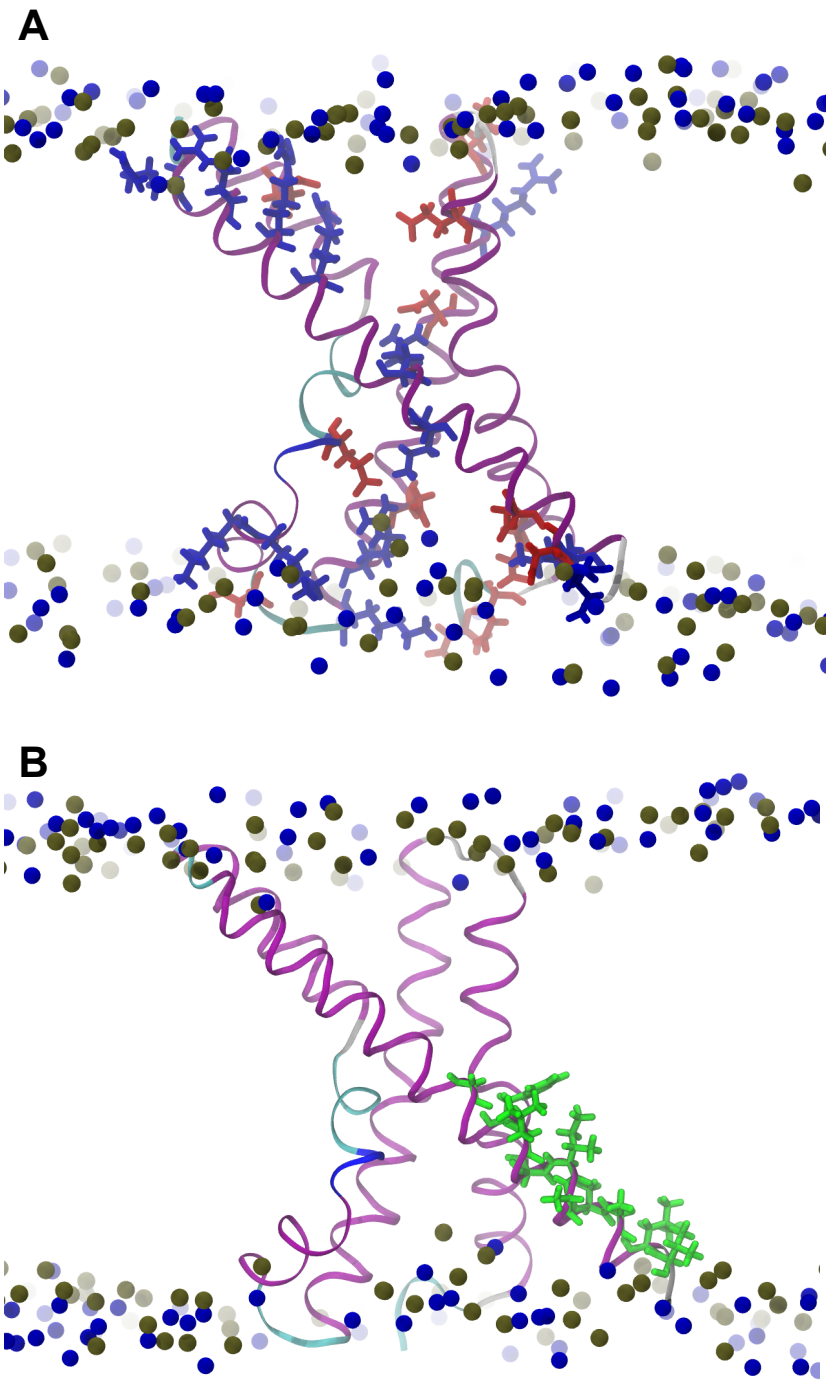


Figure 3.23. Charged and hydrophobic residues in KvAP.
A) The extracellular facing part of the KvAP VSD is positively charged. B) The hydrophobic residues of the S4 helix C-terminus are oriented towards the middle of the bilayer.
The molecular representations of PDB deposited structure of KvAP are shown in ribbons. The basic and acidic residues are colored in blue and red, respectively in the upper panel, and the hydrophobic residues of the S4 C-terminus are colored in green in the lower panel. The nitrogen and phosphorus atoms of the phospholipids are represented as blue and tan spheres, respectively.

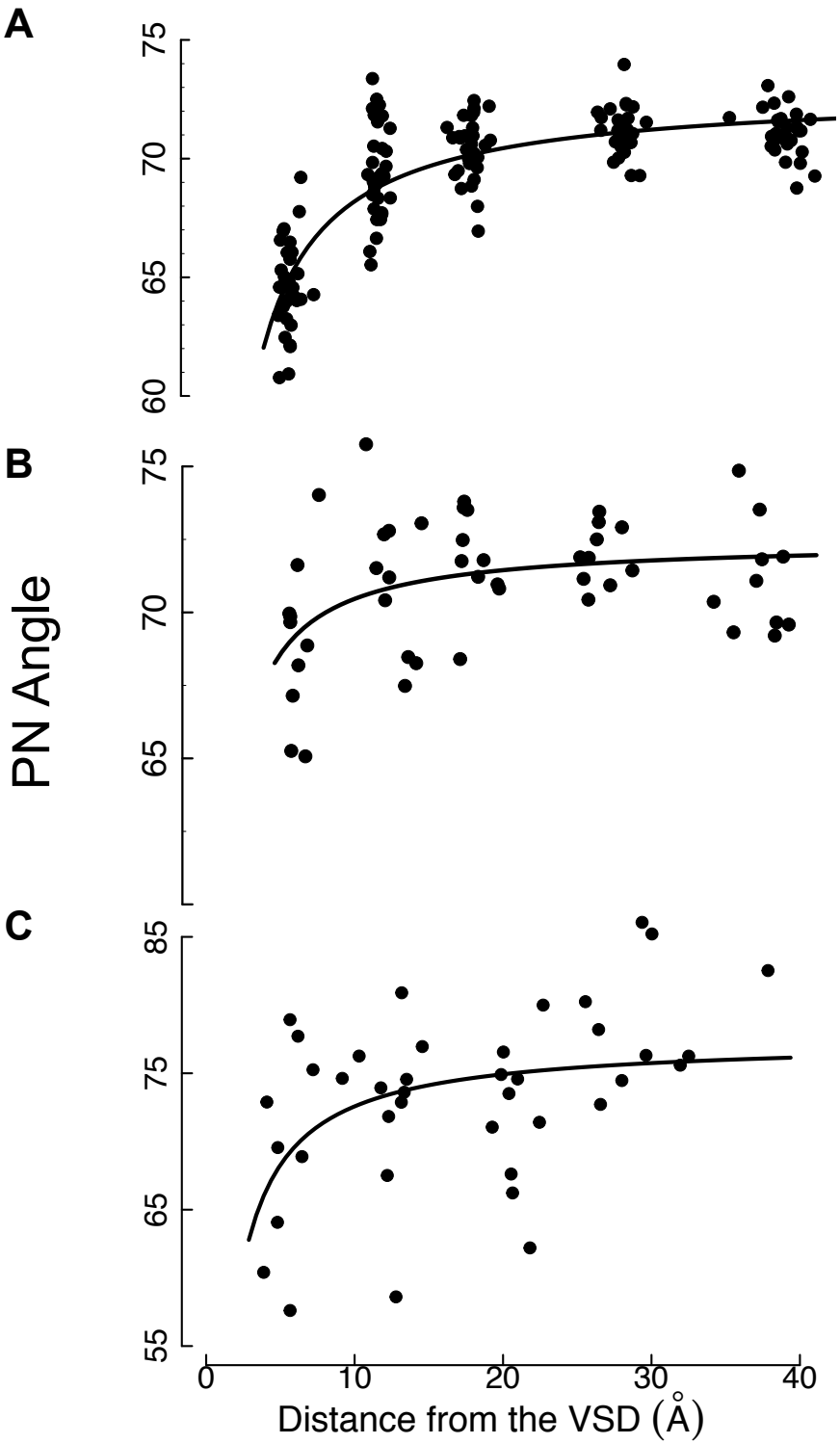


Figure 3.24. The angle of the lipid P-N vectors relative to the normal of the bilayer decreases near the VSD.

Each point represents the average value of the lipid molecules at a given average distance to the toxin during the last 20 ns of independent simulations. The smooth curves correspond to hyperbolic functions with asymptotes ≈ 70 degrees for POPC and ≈ 76 degrees for POPS. (A: extra-cellular leaflet, 100 POPC molecules, $p<0.001$, $n = 32$ simulations; B: intra-cellular leaflet, 50 POPC molecules, $p < 0.05$, $n = 10$; C: 50 POPS molecules, $p < 0.05$, $n = 10$). Note that, in POPC, N represents the choline group, while in POPS, it represents the amine (see Figure 1.1)

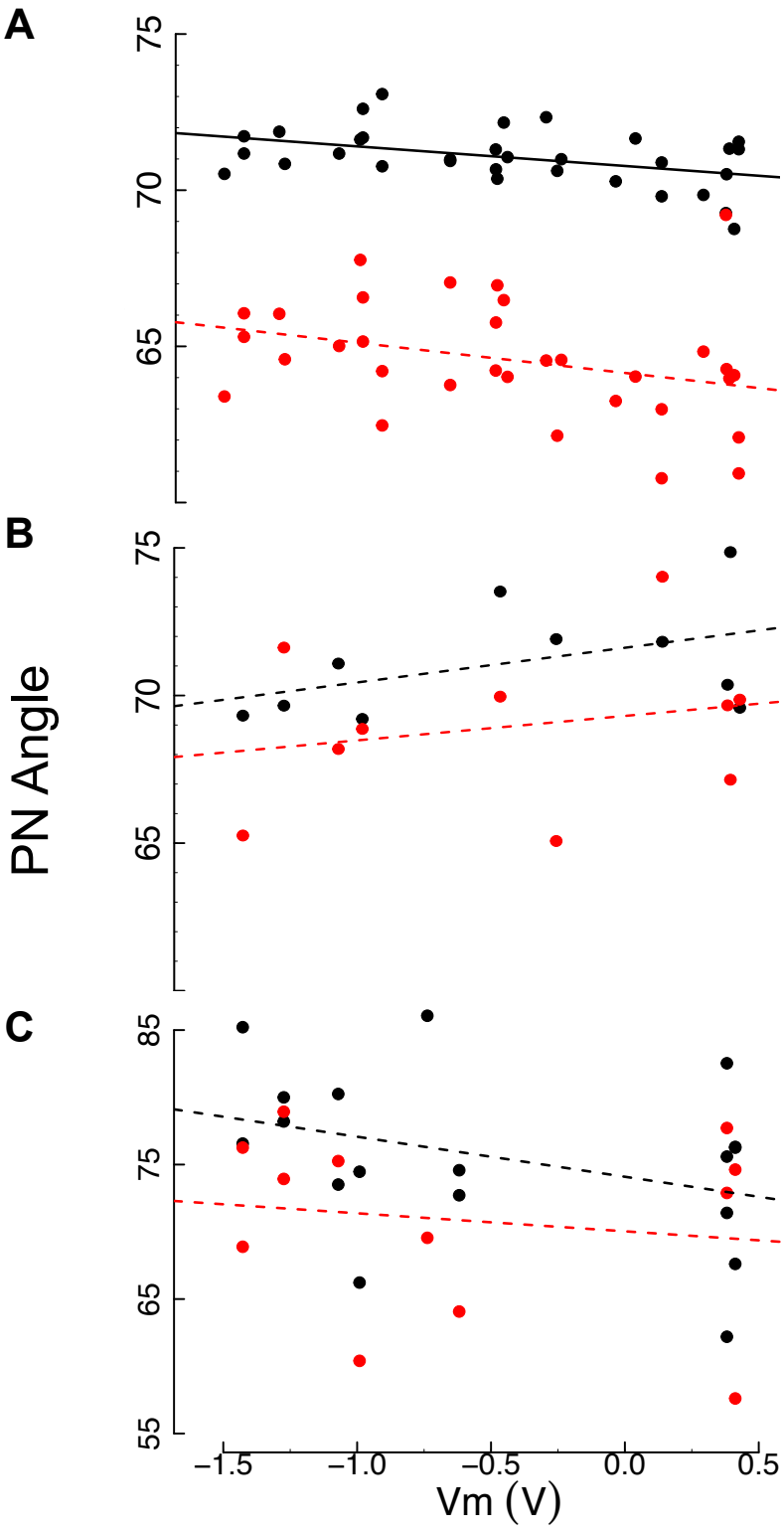


Figure 3.25. The P-N vectors respond to the electric field.
Each point corresponds to the average value of the bulk (black) or near to the VSD (red) lipid molecules under a given applied trans-membrane potential during the last 20 ns of one of 43 (A: extra-cellular leaflet with 100 POPC molecules) or 10 (B: intra-cellular leaflet, with 50 POPC molecules; C: 50 POPS molecules) independent simulations. The solid lines highlight linear regressions for which $p < 0.05$.

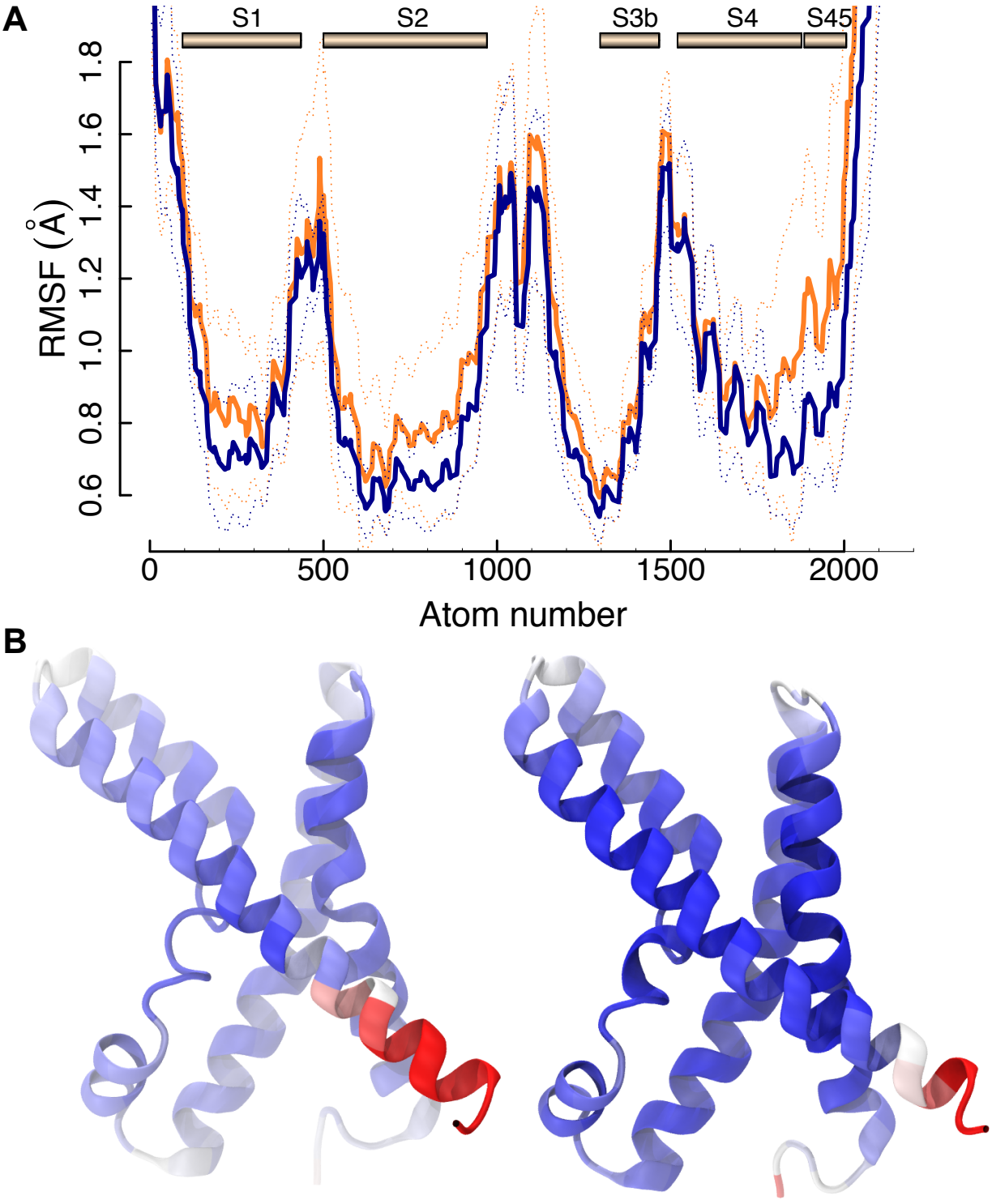


Figure 3.26. Larger RMS fluctuations and B-factors under polarized potential.
A) The average (solid lines) and standard error (dotted lines) of the RMSF of 20 VSD exposed to a depolarized potential (blue) and polarized potential (orange) are shown. The helices and "pseudohelical segments" according to the NMR structural determination are shown as cylinders (*1*). B) The representations of a VSD simulated under a polarized potential (left) and a VSD simulated under a depolarized potential (right) are colored according to their B-factors (calculated during 200 ns each, where atoms with values below 20 $Å^2$ and above 100 $Å^2$ are shown in dark blue and red, respectively).

3.3.2.4 A novel membrane potential induced conformational change of the voltage-sensor domain

In the last years, there have been tremendous progresses relative to the description of the VSD conformational changes in response to the membrane potential (*1*-*6*). Several mechanisms by which the positively charged residues of S4 may exchange interactions between the extracellular and intracellular sides of the transmembrane electric field have been proposed. A translation mostly along the principal axis of S4 of up to 15 Å has been deduced from avidin accessibility to biotin reagents of different length (*1*). Electron paramagnetic resonance (EPR) spectroscopy measurements suggest a shorter S4 translation, accompanied by its rotation (*1*). According to these models, some residues may have to exchange their interactions from an apolar to a more polar environment. Since the crystal structure of KvAP is assumed to represent the activated state, a translation of 5 or 15 Å would lead some of the hydrophobic residues of the S4 C-terminus to be exposed to the polar head groups or to the intracellular cytoplasm (Figure 3.23). Models by which the S4 helix would adopt α to 3_{10} transitions have been proposed. These imply more flexibility, and it was calculated that a small reorientation of the S4 helix would suffice to induce mechanical constraints on the pore.

The MD simulations performed in this work revealed an unexpected conformational change of the S4 helix of KvAP in response to the membrane potential. The root mean-square fluctuations (RMSF) of the VSD during the whole trajectories of 20 systems exposed to a polarized membrane potential and 20 systems exposed to a depolarized potential were compared. The fluctuations of the loops were similar and, as expected, larger than the fluctuations of the helices for all the applied potential. The structure deposited in the PDB, solved by X-ray crystallography, is assumed to capture the protein in a conformation corresponding to a depolarized potential. Interestingly, the helices displayed larger fluctuations in the systems simulated under polarized membrane potential than in the systems held at a potential close to zero (Figure 3.26A), in line with the idea that these structures should adopt a resting state under longer simulations. Since these large fluctuations occurred notably at the level of the second half of the S4 helix, this part of the protein was further analyzed. B-factors are proportional to the square of the RMSF. Thus, the comparison of B-factors calculated from two simulations conducted under different membrane potentials also illustrates this idea. A typical representation of the B-factors computed over a 200 ns simulation at about 0.4 V display values around 10-20 Å2 from residue 120 to residue 141, which can be compared to a simulation performed under a polarized potential of around -1.0 V, where the B-factors of residues 136 to 141 increase to about 50-100 Å2 (Figure 3.26B). In other words, whereas this segment remained rather rigid under depolarized membrane potential, it displayed much larger fluctuations when a negative membrane potential is applied.

A second observation was related to the transport of electric charges between the extra- and intracellular compartments. The charges held by Arg residues of the S4 helix are called gating charges. It is assumed that they change their position with respect to the membrane electric field upon polarization or depolarization. The current generated by this movement is

accordingly called gating current. In the simulations, the membrane potential remained constant during the whole simulation time in most of the cases. However, in a few simulations under a polarized potential, an important drop with respect to the initial membrane potential occurred (Figure 3.27A). In other words, one or two positive charges changed compartment in the direction of the electric field. This gating charge transport was due to the breaking of the central Arg133-Asp62 salt bridge in the middle of the bilayer, as observed earlier (*1*). An interesting time series, in which a gating current occurred, is shown in Figure 3.27B. The breaking of the salt bridge coincided with a gating charge transport, which reduced the membrane potential by approximately 300 mV within 100 ns. As shown in Figure 3.27C, the charge transport is due mainly to the reorientation of the Arg133 towards the intracellular water compartment. At t = 200 ns, the simulated membrane potential was switched back to +0.5 V. This depolarization was almost immediately followed by the re-formation of the salt bridge.

The breaking of the salt bridge and the gating charge transport were further accompanied by the formation of a kink at residue Gly134 in the middle of S4. The formation of this kink may allow a translation of the positively charged residues of S4 without exposing the hydrophobic residues of the C-terminus to a polar environment. The S4 movement implicated by the simulations could reconcile the ideas of a translation of S4 with the one of a global secondary structure modification.

There is experimental evidence in support of the formation of a kink in S4. Interestingly, the available experimental structures of the KvAP voltage-sensor domain show some discrepancies concerning the exact definition of this segment. According to the structure solved by crystallography in 2003, the whole segment between residues 116 and 148 constitutes an α-helix (*1*). Yet, a study on this VSD by electron paramagnetic resonance indicated that this helix might be constituted of two helices separated by a hinge somewhere in its middle (*1*, *2*). An NMR determination of the KvAP VSD in micelles has been reported in 2009 (*124*). The periodicity of amide proton secondary chemical shifts typically reflects the helicity of a segment. The loss of this periodicity at the level of Gly134 was therefore interpreted as a hint that the S4 helix is constituted of two helices connected by a hinge constituted of Ile131, Ser132 and Arg133. In addition, according to other features of the NMR results, the authors defined the S4 segment as a rigid helix, termed S4, between Phe116 and Arg133, continued by a pseudohelical segment, S45, between Gly134 and Ala142. However, the solution structure determined by Shenkarev et al. (*1*) in 2010 has not been deposited in the RSCP Protein Data Bank (*1*). However, the analysis of deposited structures reveals additional hints regarding a kink in S4 of KvAP. Strikingly, among the 20 converged conformations of the solution NMR solved by Butterwick and MacKinnon (*1*)(PDB code 2KYH), three of them display also a kink in the middle of S4, very similar to the one observed in the simulations performed under polarized conditions in this work (Figure 3.27E). In addition, the Arg133-Asp62 salt bridge is broken in these three structures, whereas it is intact in

the other 17 structures. A better understanding of the S4 helix behavior under different conditions would require investigations of the full-length channel Butterwick and MacKinnon (*1*) reported that, while the X-ray structure of the isolated VSD suggests an alpha helical structure from residues 117 to 147, S4 is expected to break when connected to the ion conducting pore. Accordingly, in the full-length deposited structure (PDB code 2AOL), solved by X-ray crystallography, S4 is indeed broken. However, this break occurs at residue Ser139, whereas the EPR investigations, the solution structure of Shenkarev and colleagues and the MD simulations suggest that the break occurs at Gly134.

Finally, a sequence alignment of voltage-gated ion channels published in 2008 (*1*), showed some conservation of Gly and Pro at the position corresponding to Gly134 in the middle of S4. Precisely, a Gly residue is found in Hv1 and a Pro residue in Cav2.1, Cav3.1 and Nav1.2. These results indicate that the conformational change observed in KvAP may apply to other voltage-gated ion channels. It would be of major interest to find out, whether these channels also form a kink in S4, and, if yes, if this is related to the membrane potential. The formation of the kink in S4, since it modifies the orientation of the coupling between S4 and S5, may consequently affect the strength of this coupling.

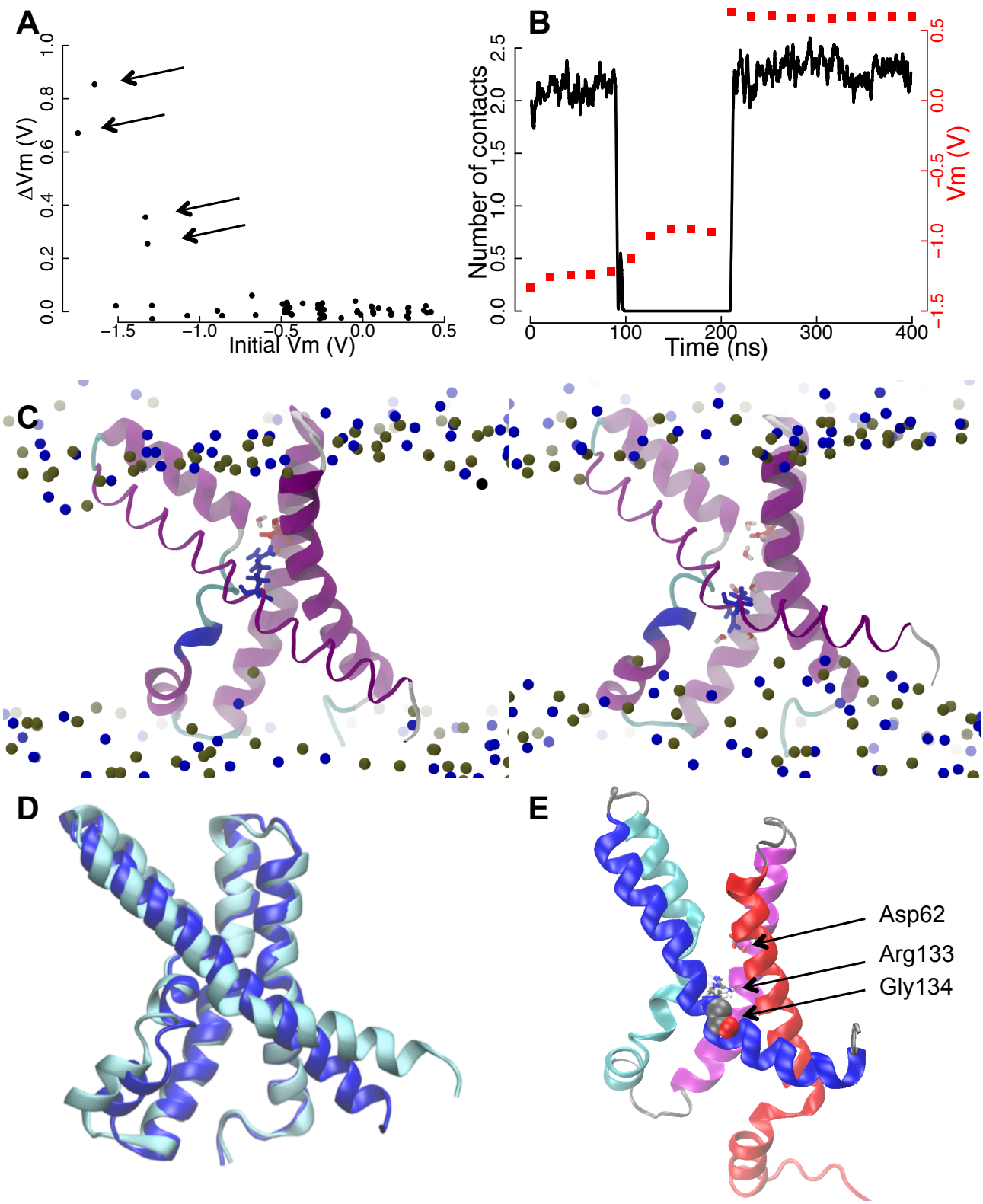


Figure 3.27, previous page: Discovery of a novel conformational change: the S4 helix forms a kink under polarized potential

A) Each point represents the absolute difference between the membrane voltage at the beginning of the simulations, and after 200 ns, as a function of the initial Vm (n = 65 simulations). The arrows identify four gating charge transport events. B) Time course of the number of contacts between the side chains of Arg133 and Asp62 (black line) and of the membrane potential (red dots). At t=200 ns, the potential was switched from polarized to depolarized. C) Breaking of the central salt bridge between Arg133 and Asp62. The left panel shows the intact salt bridge at the beginning of the simulations. The right panel shows the structure after the breaking of the salt bridge. The water molecules in the first hydration shell of the side chain of Arg133 are shown as red (oxygen) and white (hydrogen) spheres. D) Molecular representation of the KvAP simulated under a polarized membrane potential showing the structure at the beginning of the simulation (blue) and after the breaking of the Arg133-Asp-62 salt bridge (cyan). E) Second converged structure of the NMR solution ensemble (PDB code 2KYH). The helices are colored as follows: S1: red, S2: magenta, S3a and S3b: cyan, S4: blue. The heavy atoms of Gly134 are shown as spheres.

All these experimental findings support the view that S4 of KvAP forms a kink under some conditions. The MD simulation analyses indicate that this kink occurs under a polarized membrane potential, is associated with the breaking of the central salt bridge between Arg133 and Asp 62, and the whole leads to gating charge transport.

3.3.2.5 Conclusion

The replication of 134 bilayers, each of them containing the same VSD, and exposed to different Vm values between -1.7 and 0.5 Volts has led to fascinating new findings. A modulation of the interaction between a protein and the bilayer by the membrane potential has been described here for the first time. In the simulations, the VSD induced a significant decrease of the acyl-chain order parameters under a depolarized potential. When the system was exposed to a polarized membrane potential, the decrease of the order parameters was significantly enhanced. Based on the conformational changes of the VSD observed during these 200 ns long simulations, three possible origins of the phenomenon were proposed for further investigations.

The orientation of the head groups was affected by the VSD and by the membrane potential. Their orientation followed the direction of the electric field in both leaflets, and the amplitude of the reorientation was comparable to other studies (*1*). On the other hand, the angle formed by the head groups and the membrane normal was found to decrease near the VSD. It would be interesting to test whether, similarly to the analyses of the toxin interactions with the phosphate groups, the formation of hydrogen bonds between the Arg side chains and the phosphodiesters may be at the origin of the strong reorientation of the head groups.

Previous NMR and EPR investigations predicted a possible kink in the middle of the helix S4. The MD simulations revealed that this kink is related to polarized membrane potentials, is associated with a gating charge transport, and additionally might be associated with the modulation of the VSD induced disordering effect by Vm.

This kink formation allows S4 to translate toward the intracellular compartment without exposing hydrophobic side chains to a polar environment. In other words, these results indicate that the observed kink is formed when the VSD undergoes the transition to the resting state.

4 General conclusion

Fascinating discoveries about the conformational propensities of short peptides, the driving force of folding and a novel conformational change of the voltage-sensor domain as a response to the membrane potential have been revealed by the molecular dynamics simulations performed during this work. Additionally, lipid-mediated mechanisms by which spider toxins may modify the gating of voltage-sensor domains, without direct interaction with the target proteins, were studied. According to the results, mechanisms involving the reshaping of the acyl chain global structure or the reorientation of the head groups might be ruled out for further investigations. An alternative mechanism, by which the toxin and the voltage-sensor domain would compete for the phosphodiester groups, could not be demonstrated. In line with these results, the question whether spider toxins must partition into the membrane in order to modify the gating of the voltage-gated ion channel has been investigated recently through surface plasmon resonance, fluorescence spectroscopy and molecular dynamics. The results involving two gating modifiers led the authors to conclude, that “membrane interaction is not a prerequisite for modification of channel gating”(*125*).

The first project of this thesis investigated how single amino acid substitutions in a short peptide affected its conformational propensities. Whereas NMR experiments showed that the substitution of X by an aromatic residue in the sequence EGAAXAASS might induce the formation of a kink, the MD simulations revealed that this short peptide formed a helix or a turn. Further investigations suggested that bulky side chains impede the hydration of neighboring amide and carbonyl groups. This reduced access to water consecutively favors the formation of intramolecular hydrogen bonds and folding. This folding nucleation may be decisive for folding, since it has a relatively high entropic cost, whereas the following elongation of an helix can be assumed to occur cooperatively.

The exact role of water in protein conformational changes is still a matter of debate. In section 2.3, it was proposed to use a cross-correlation function to investigate the succession of events occurring during the folding und unfolding of a β-hairpin. The analyses suggest that the hydration fluctuations precede the peptide conformational changes by a few hundreds of picoseconds.

It was concluded by calling for an extended use of a combination of unbiased classical MD simulations, NMR investigations and the statistical tools introduced in this thesis to elucidate the conformational changes occurring in proteins or other biomolecules. A proper description of the formation of a one-turn helix could then be progressively extended by investigations of longer peptides, as exposed in the proposed project in section 2.4. This combination of methods could efficiently be used to study other conformational changes for which the mechanisms are difficult

to capture experimentally. The exact mechanisms, by which electrolytes like TMAO or urea favor the folded or unfolded states, respectively, are accessible by this approach.

Several experimental discoveries led us to ask whether gating modifiers might indirectly affect voltage-gated ion channels. Recently, a large number of studies have emphasized the important function of the diversity among membrane lipids. Lipids are not only scaffolds, but were shown to have fundamental and functional roles (*1*). Lately, a study involving measurements of voltage gated ether à gogo (EAG) potassium channels in diverse lipids demonstrated the role of phosphatidylinositol 4,5-bisphosphate (PIP_2) as a modulator of EAG channels (*1*). Voltage-sensor domains have been shown to be sensitive to modification of the membrane environment, since removal of the phosphodiester groups renders the Kv2.1 channel irresponsive to changes of the membrane potential. The contrary is observed upon removal of the choline head group of sphingomyelin by sphingomyeliase. The result is a membrane enriched in negatively charged phospholipids, in which the VSD were shown to activate at a more negative potential (*1*). Former experimental evidence had shown that electrically charged molecules can have a large effect on the structure of the phospholipids (*1*). The finding that spider toxins use a membrane access mechanism to modify the gating of their target (*1*) may have contributed to several investigations of their partitioning in the membrane (*1*) and their orientation within the bilayer (*1*). Spider toxins are electrically charged molecules and the exact mechanism of gating modification remains elusive. In the light of these facts, our motivation was to ask whether the mechanism of spider toxins might involve perturbations of the membrane structure. One could hypothesize that alterations of the lipid head groups could modify the global properties of the membrane, which in turn would affect the VSD. However, experimental evidence demonstrated that the spider toxins can be highly specific, targeting a distinct segment of an ion channel (*1*). For completeness, two different toxins for which the specificity for a given channel is well documented were chosen. The choice was based on the idea that any differences between their interactions with the membrane could be regarded as hint to explain their specificity, while similar effects would have the potential to describe common mechanisms of actions.

The MD simulations revealed significant modifications of the bilayer structure, but the range of perturbations of the two toxins was very similar and could not be related to any conformational change of the VSD. This similarity suggests that a putative lipid-mediated mechanism of gating modifiers is not related to reshaping of the acyl chain global structure or to the reorientation of the phospholipid head groups.

The double bilayers containing voltage-sensor domains exposed to varying membrane voltages were constructed with the aim to investigate the spider toxin mechanisms. However, the analysis of their trajectories revealed a behavior of the VSD S4 helix that has, to our knowledge, never been described. We observed a kink in the middle of the S4 helix that was predicted by EPR (*1*) and NMR (*1*) investigations. The novelty of the finding presented in this thesis, is that

this kink was related to the polarized membrane potential, and was associated with the breaking of the central salt bridge between Arg133 and Asp62. This conformational change allows a displacement of S4 without exposing hydrophobic residues to the polar environment of the lipid head groups or the solvent.

The investigation of sequence alignments suggests that the proton channel Hv1 also has a Gly residue in the middle of the corresponding helix, whereas a few VSD were found, in which a Pro residue is inserted at this position.

These findings urge for experimental and computational investigations of KvAP and Hv1 under varying membrane potential values, aimed at describing the occurrence of a kink in the middle of S4 as a conformational change towards the resting state.

References

1. P. E. Wright, H. J. Dyson, Linking folding and binding. *Current Opinion in Structural Biology*. **19**, 31–38 (2009).

2. T. Kiefhaber, A. Bachmann, K. S. Jensen, Dynamics and mechanisms of coupled protein folding and binding reactions. *Current Opinion in Structural Biology*. **22**, 21–29 (2012).

3. C. B. Anfinsen, Principles that govern the folding of protein chains. *Science*. **181**, 223–230 (1973).

4. C. Levinthal, *proteins_levinthal_1969* (Mossbauer spectroscopy in Biological Systems Proceedings, 1969), vol. 67.

5. K. Neupane *et al.*, Direct observation of transition paths during the folding of proteins and nucleic acids. *Science*. **352**, 239–242 (2016).

6. A. K. Dunker *et al.*, Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling*. **19**, 26–59 (2001).

7. L. Pauling, A theory of the structure and process of formation of antibodies. *J. Am. Chem. Soc.* **62**, 2643–2657 (1940).

8. A. Sigalov, D. Aivazian, L. Stern, Homooligomerization of the Cytoplasmic Domain of the T Cell Receptor ζ Chain and of Other Proteins Containing the Immunoreceptor Tyrosine-Based Activation Motif †. *Biochemistry*. **43**, 2049–2061 (2004).

9. P. Tompa, M. Fuxreiter, Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends in Biochemical Sciences*. **33**, 2–8 (2008).

10. A. K. Dunker *et al.*, The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics*. **9**, S1 (2008).

11. B. Alberts *et al.*, Eds., Molecular Biology of the Cell (2002).

12. O. Quehenberger *et al.*, Lipidomics reveals a remarkable diversity of lipids in human plasma. *The Journal of Lipid Research*. **51**, 3299–3305 (2010).

13. L. Ginsberg, S. RafiqUE, J. H. Xuereb, S. I. Rapoport, N. L. Gershfeld, Disease and Anatomic Specificity of Ethanolamine Plasmalogen Deficiency in Alzheimers-Disease Brain. *Brain Res.* **698**, 223–226 (1995).

14. M. Kosicek, S. Hecimovic, Phospholipids and Alzheimer's Disease: Alterations, Mechanisms and Potential Biomarkers. *IJMS*. **14**, 1310–1322 (2013).

15. B. Fadeel, D. Xue, The ins and outs of phospholipid asymmetry in the plasma membrane: roles in health and disease. *Critical Reviews in Biochemistry and Molecular Biology*. **44**, 264–277 (2009).

16. S. Manno, Y. Takakuwa, (2002).

17. B. D. Smith, T. N. Lambert, Molecular ferries: membrane carriers that promote phospholipid flip-flop and chloride transport. *Chem. Commun.*, 2261 (2003).

18. A. W. Partridge, R. A. Melnyk, C. M. Deber, Polar Residues in Membrane Domains of

Proteins: Molecular Basis for Helix−Helix Association in a Mutant CFTR Transmembrane Segment †. *Biochemistry*. **41**, 3647–3653 (2002).

19. G. Nagel *et al.*, Channelrhodopsin-1: a light-gated proton channel in green algae. *Science*. **296**, 2395–2398 (2002).

20. A. Anishkin, S. H. Loukin, J. Teng, C. Kung, Feeling the hidden mechanical forces in lipid bilayer is an original sense. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7898–7905 (2014).

21. C. D. Pivetti *et al.*, Two Families of Mechanosensitive Channel Proteins. *Microbiology and Molecular Biology Reviews*. **67**, 66–85 (2003).

22. W. A. Catterall, Structure and function of voltage-gated sodium channels at atomic resolution. *Experimental Physiology*. **99**, 35–51 (2013).

23. J. G. McGivern, Advantages of voltage-gated ion channels as drug targets. *Expert Opinion on Therapeutic Targets*. **11**, 265–271 (2007).

24. H. Wulff, N. A. Castle, L. A. Pardo, Voltage-gated potassium channels as therapeutic targets. *Nat Rev Drug Discovery*. **8**, 982–1001 (2009).

25. K. J. Swartz, Sensing voltage across lipid membranes. *Nature*. **456**, 891–897 (2008).

26. M. C. Trudeau, J. W. Warmke, B. Ganetzky, G. A. Robertson, Herg, a Human Inward Rectifier in the Voltage-Gated Potassium Channel Family. *Science*. **269**, 92–95 (1995).

27. Q. Wang *et al.*, Positional cloning of a novel potassium channel gene: KVLQT1 mutations cause cardiac arrhythmias. *Nature Genetics*. **12**, 17–23 (1996).

28. Q. Li, S. Wanderling, P. Sompornpisut, E. Perozo, Structural basis of lipid-driven conformational transitions in the KvAP voltage-sensing domain. *Nat. Struct. Mol. Biol.* **21**, 160–166 (2014).

29. B. L. Tempel, D. M. Papazian, P. J. Schwartz, L. Y. Jan, Sequence of a Probable Potassium Channel Component Encoded at Shaker Locus of Drosophila. *Science*. **237**, 770–775 (1987).

30. F. Bezanilla, The voltage sensor in voltage-dependent ion channels. *Physiological Reviews*. **80**, 555–592 (2000).

31. Y. Jiang *et al.*, X-ray structure of a voltage-dependent K+ channel. *Nature*. **423**, 33–41 (2003).

32. Y. Murata, H. Iwasaki, M. Sasaki, K. Inaba, Y. Okamura, Phosphoinositide phosphatase activity coupled to an intrinsic voltage sensor. *Nature*. **435**, 1239–1243 (2005).

33. Z. Lu, A. M. Klem, Y. Ramu, Ion conduction pore is conserved among potassium channels. *Nature*. **413**, 809–813 (2001).

34. R. C. Thomas, R. W. Meech, Hydrogen ion currents and intracellular pH in depolarized voltage-clamped snail neurones. *Nature*. **299**, 826–828 (1982).

35. M. Sasaki, M. Takagi, Y. Okamura, A voltage sensor-domain protein is a voltage-gated proton channel. *Science*. **312**, 589–592 (2006).

36. I. S. Ramsey, M. M. Moran, J. A. Chong, D. E. Clapham, A voltage-gated proton-selective channel lacking the pore domain. *Nature*. **440**, 1213–1216 (2006).

37. P. Jurkiewicz, L. Cwiklik, A. Vojtíšková, P. Jungwirth, M. Hof, Biochimica et Biophysica Acta. *BBA - Biomembranes*. **1818**, 609–616 (2012).

38. J. V. Raimondo, R. J. Burman, A. A. Katz, C. J. Akerman, Ion dynamics during seizures. *Front. Cell. Neurosci.* **9**, 11521 (2015).

39. A. A. Alabi, M. I. Bahamonde, H. J. Jung, J. I. Kim, K. J. Swartz, Portability of paddle motif function and pharmacology in voltage sensors. *Nature*. **450**, 370–375 (2007).

40. N. J. Saez *et al.*, Spider-Venom Peptides as Therapeutics. *Toxins*. **2**, 2851–2871 (2010).

41. J. Kalia *et al.*, From Foe to Friend: Using Animal Toxins to Investigate Ion Channel Function. *Journal of Molecular Biology*, 1–18 (2014).

42. V. Chi *et al.*, Development of a sea anemone toxin as an immunomodulator for therapy of autoimmune diseases. *Toxicon*. **59**, 529–546 (2012).

43. R. MacKinnon, Determination of the subunit stoichiometry of a voltage-activated potassium channel. *Nature*. **350**, 232–235 (1991).

44. K. Gupta *et al.*, Tarantula toxins use common surfaces for interacting with Kv and ASIC ion channels. *eLife*. **4** (2015), doi:10.7554/eLife.06774.

45. A. A. Vassilevski, S. A. Kozlov, E. V. Grishin, Molecular diversity of spider venom. *Biochemistry Moscow*. **74**, 1505–1534 (2010).

46. R. MacKinnon, C. MILLER, Mutant Potassium Channels with Altered Binding of Charybdotoxin, a Pore-Blocking Peptide Inhibitor. *Science*. **245**, 1382–1385 (1989).

47. H. H. Jung *et al.*, Structure and Orientation of a Voltage-Sensor Toxin in Lipid Membranes. *Biophysj*. **99**, 638–646 (2010).

48. V. Herzig *et al.*, ArachnoServer 2.0, an updated online resource for spider toxin sequences and structures. *Nucleic acids research*. **39**, D653–D657 (2010).

49. P. G. Scherer, J. Seelig, Electric charge effects on phospholipid headgroups. Phosphatidylcholine in mixtures with cationic and anionic amphiphiles. *Biochemistry*. **28**, 7720–7728 (1989).

50. J. Seelig, P. M. Macdonald, P. G. Scherer, Phospholipid head groups as sensors of electric charge in membranes. *Biochemistry*. **26**, 7535–7541 (1987).

51. B. Bechinger, J. Seelig, Interaction of electric dipoles with phospholipid head groups. A 2H and 31P NMR study of phloretin and phloretin analogues in phosphatidylcholine membranes. *Biochemistry*. **30**, 3923–3929 (1991).

52. A. Laganowsky *et al.*, Membrane proteins bind lipids selectively to modulate their structure and function. *Nature*. **510**, 172–175 (2014).

53. B. Han *et al.*, Human EAG channels are directly modulated by PIP. *Nature Publishing Group*, 1–13 (2016).

54. Y. Xu, Y. Ramu, Z. Lu, Removal of phospho-head groups of membrane lipids immobilizes voltage sensors of K+ channels. *Nature*. **451**, 826–829 (2008).

55. D. Schmidt, Q.-X. Jiang, R. MacKinnon, Phospholipids and the origin of cationic gating charges in voltage sensors. *Nature*. **444**, 775–779 (2006).

56. S.-Y. Lee, R. MacKinnon, A membrane-access mechanism of ion channel inhibition by voltage sensor toxins from spider venom. *Nature*. **430**, 232–235 (2004).

57. S. A. Dames *et al.*, Residual Dipolar Couplings in Short Peptides Reveal Systematic Conformational Preferences of Individual Amino Acids. *J. Am. Chem. Soc.* **128**, 13508–13514 (2006).

58. M. Chaplin, Do we underestimate the importance of water in cell biology? *Nat Rev Mol Cell Biol*, 1–6 (2006).

59. M. S. Cheung, A. E. Garcia, J. N. Onuchic, Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 685–690 (2002).

60. S. S. Cho, G. Reddy, J. E. Straub, D. Thirumalai, Entropic stabilization of proteins by TMAO. *J. Phys. Chem. B*. **115**, 13401–13407 (2011).

61. O. F. Lange *et al.*, Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*. **320**, 1471–1475 (2008).

62. K. A. Dill, S. B. Ozkan, M. S. Shell, T. R. Weikl, The protein folding problem. *Annu. Rev. Biophys.* **37**, 289 (2008).

63. P. Radivojac *et al.*, Intrinsic disorder and functional proteomics. *Biophysical Journal*. **92**, 1439–1456 (2007).

64. T. Chouard, Structural biology: Breaking the protein rules. *Nature*. **471** (2011), pp. 151–153.

65. K. A. Dill, J. L. MacCallum, The protein-folding problem, 50 years on. *Science*. **338**, 1042–1046 (2012).

66. C. Seife, What is the universe made of. *Science*. **309**, 78–78 (2005).

67. H. M. Berman, J. Westbrook, Z. Feng, The protein data bank. *Nucleic acids research*. **28**, 235–242 (2000).

68. H. J. Dyson, P. E. Wright, Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*. **6**, 197–208 (2005).

69. A. Bartesaghi *et al.*, 2.2 angstrom resolution cryo-EM structure of beta-galactosidase in complex with a cell-permeant inhibitor. *Science*. **348**, 1147–1151 (2015).

70. K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, How fast-folding proteins fold. *Science*. **334**, 517–520 (2011).

71. M. Vendruscolo, E. Paci, C. M. Dobson, M. Karplus, Three key residues form a critical contact network in a protein folding transition state. *Nature* (2001).

72. The UniProt Consortium, UniProt: a hub for protein information. *Nucleic acids research*. **43**, D204–D212 (2015).

73. P. Tompa, Intrinsically unstructured proteins. *Trends in Biochemical Sciences*. **27**, 527–533 (2002).

74. O. Bignucolo, H. T. A. Leung, S. Grzesiek, S. Bernèche, Backbone Hydration Determines the Folding Signature of Amino Acid Residues. *J. Am. Chem. Soc.* **137**, 4300–4303 (2015).

75. M. Zweckstetter, NMR: prediction of molecular alignment from structure using the PALES software. *Nat Protoc*. **3**, 679–690 (2008).

76. Y. Shen, A. Bax, Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J Biomol NMR*. **38**, 289–302 (2007).

77. H. T. A. Leung *et al.*, A Rigorous and Efficient Method To Reweight Very Large Conformational Ensembles Using Average Experimental Data and To Determine Their Relative Information Content. *J. Chem. Theory Comput.* **12**, 383–394 (2016).

78. M. R. Shirts, V. S. Pande, Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *J. Chem. Phys.* **122**, 134508 (2005).

79. G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten, P. G. Wolynes, Water in protein structure prediction. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 3352–3357 (2004).

80. P. H. Yancey, Organic osmolytes as compatible, metabolic and counteracting cytoprotectants in high osmolarity and other stresses. *Journal of Experimental Biology*. **208**, 2819–2830 (2005).

81. S. Honda, K. Yamasaki, Y. Sawada, H. Morii, 10 residue folded peptide designed by segment statistics. *Structure* (2004), doi:10.1016/j.str.2004.05.022.

82. W. Xu, T. Lai, Y. Yang, Y. Mu, Reversible folding simulation by hybrid Hamiltonian replica exchange. *J. Chem. Phys.* **128**, 175105 (2008).

83. D. Matthes, B. L. de Groot, Secondary Structure Propensities in Peptide Folding Simulations: A Systematic Comparison of Molecular Mechanics Interaction Schemes. *Biophysical Journal*. **97**, 599–608 (2009).

84. A. Suenaga *et al.*, Folding Dynamics of 10-Residue β-Hairpin Peptide Chignolin. *Chem. Asian J.* **2**, 591–598 (2007).

85. D. Satoh, K. Shimizu, S. Nakamura, T. Terada, Folding free-energy landscape of a 10-residue mini-protein, chignolin. *FEBS Letters*. **580**, 3422–3426 (2006).

86. R. Harada, A. Kitao, Exploring the Folding Free Energy Landscape of a β-Hairpin Miniprotein, Chignolin, Using Multiscale Free Energy Landscape Calculation Method. *J. Phys. Chem. B*. **115**, 8806–8812 (2011).

87. W. Jorgensen, J. Chandrasekhar, J. Madura, R. Impey, M. Klein, Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **79**, 926–935 (1983).

88. D. Van Der Spoel *et al.*, GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **26**, 1701–1718 (2005).

89. R. B. Best *et al.*, Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ 1and χ 2Dihedral Angles. *J. Chem. Theory Comput.* **8**, 3257–3273 (2012).

90. W. L. DeLano, PyMOL. *DeLano Scientific* (2002).

91. T. O. Street, D. W. Bolen, G. D. Rose, A molecular mechanism for osmolyte-induced protein stability. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 13997–14002 (2006).

92. S. Honda *et al.*, Crystal Structure of a Ten-Amino Acid Protein. *J. Am. Chem. Soc.* **130**, 15327–15331 (2008).

93. L. Wang, H. R. Eghbalnia, J. L. Markley, Nearest-neighbor effects on backbone alpha and beta carbon chemical shifts in proteins. *J Biomol NMR*. **39**, 247–257 (2007).

94. S. I. S. McDonough, R. A. R. Lampe, R. A. R. Keith, B. P. B. Bean, Voltage-dependent inhibition of N- and P-type calcium channels by the peptide toxin omega-grammotoxin-SIA. *Molecular Pharmacology*. **52**, 1095–1104 (1997).

95. V. Ruta, R. MacKinnon, Localization of the Voltage-Sensor Toxin Receptor on KvAP †. *Biochemistry*. **43**, 10071–10079 (2004).

96. E. Redaelli *et al.*, Target Promiscuity and Heterogeneous Effects of Tarantula Venom Peptides Affecting Na+ and K+ Ion Channels. *Journal of Biological Chemistry*. **285**, 4130–4142 (2010).

97. J. A. Freites, D. J. Tobias, Voltage Sensing in Membranes: From Macroscopic Currents to Molecular Motions. *J Membrane Biol*. **248**, 419–430 (2015).

98. C. T. Armstrong, P. E. Mason, J. L. R. Anderson, C. E. Dempsey, Arginine side chain interactions and the role of arginine as a gating charge carrier in voltage sensitive ion channels. *Nature Publishing Group*, 1–10 (2016).

99. S.-I. Ozawa *et al.*, Structural basis for the inhibition of voltage-dependent K(+) channel by gating modifier toxin. *Nature Publishing Group*. **5**, 14226 (2015).

100. M. Mihailescu *et al.*, Structural interactions of a voltage sensor toxin with lipid membranes. *Proceedings of the National Academy of Sciences*. **111**, E5463–E5470 (2014).

101. H. Takahashi *et al.*, Solution structure of hanatoxin1, a gating modifier of voltage-dependent K+ channels: common surface features of gating modifier toxins. *Journal of Molecular Biology*. **297**, 771–780 (2000).

102. H. J. Jung *et al.*, Solution Structure and Lipid Membrane Partitioning of VSTx1, an Inhibitor of the KvAP Potassium Channel †,‡. *Biochemistry*. **44**, 6015–6023 (2005).

103. S. Jo, J. B. Lim, J. B. Klauda, W. Im, CHARMM-GUI Membrane Builder for Mixed Bilayers and Its Application to Yeast Membranes. *Biophysj*. **97**, 50–58 (2009).

104. A. P. Demchenko, S. O. Yesylevskyy, Nanoscopic description of biomembrane electrostatics: results of molecular dynamics simulations and fluorescence probing. *Chemistry and Physics of Lipids*. **160**, 63–84 (2009).

105. A. A. Gurtovenko, I. Vattulainen, Membrane Potential and Electrostatics of Phospholipid Bilayers with Asymmetric Transmembrane Distribution of Anionic Lipids. *J. Phys. Chem. B*. **112**, 4629–4634 (2008).

106. A. A. Gurtovenko, I. Vattulainen, Calculation of the electrostatic potential of lipid bilayers from molecular dynamics simulations: Methodological issues. *J. Chem. Phys.* **130**, 215107 (2009).

107. C. Kutzner, H. Grubmüller, B. L. de Groot, U. Zachariae, Computational Electrophysiology: The Molecular Dynamics of Ion ChannelPermeation and Selectivity in Atomistic Detail. *Biophysj*. **101**, 809–817 (2011).

108. M. Ø. Jensen *et al.*, Mechanism of Voltage Gating in Potassium Channels. *Science*. **336**, 229–233 (2012).

109. H. J. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, J. R. Haak, Molecular

dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684 (1984).

110. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).

111. T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089 (1993).

112. R. C. Team, R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2012 (2013).

113. M. Milescu *et al.*, Tarantula Toxins Interact with Voltage Sensors within Lipid Membranes. *The Journal of General Physiology*. **130**, 497–511 (2007).

114. J. M. Wang, Molecular Surface of Tarantula Toxins Interacting with Voltage Sensors in Kv Channels. *The Journal of General Physiology*. **123**, 455–467 (2004).

115. L. R. Phillips *et al.*, Voltage-sensor activation with a tarantula toxin as cargo. *Nature*. **436**, 857–860 (2005).

116. K. J. Swartz, Tarantula toxins interacting with voltage sensors in potassium channels. *Toxicon* (2007), doi:10.1016/j.toxicon.2006.09.024.

117. L. S. Vermeer, B. L. de Groot, V. Réat, A. Milon, J. Czaplicki, Acyl chain order parameter profiles in phospholipid bilayers: computation from molecular dynamics simulations and comparison with 2H NMR experiments. *Eur Biophys J*. **36**, 919–931 (2007).

118. V. Miguel *et al.*, Biochimica et Biophysica Acta. *BBA - Biomembranes*. **1858**, 38–46 (2016).

119. S. K. Kandasamy, R. G. Larson, Effect of salt on the interactions of antimicrobial peptides with zwitterionic lipid bilayers. *Biochimica et Biophysica Acta (BBA) - Biomembranes*. **1758**, 1274–1284 (2006).

120. P. C. Dave, E. K. Tiburu, K. Damodaran, G. A. Lorigan, Investigating structural changes in the lipid bilayer upon insertion of the transmembrane domain of the membrane-bound protein phospholamban utilizing 31P and 2H Solid-State NMR Spectroscopy. *Biophysical Journal* (2004), doi:10.1016/S0006-3495(04)74224-1.

121. C. Song *et al.*, Structure and Dynamics of the Human Antimicrobial Peptide Dermcidin Oligomer: It is an Ion Channel. *Biophysj*. **102**, 471A–471A (2012).

122. J. Lind, J. Nordin, L. Mäler, Lipid dynamics in fast-tumbling bicelles with varying bilayer thickness: effect of model transmembrane peptides. *Biochimica et Biophysica Acta (BBA)- …* (2008), doi:10.1016/j.bbamem.2008.07.010.

123. J. A. Killian, F. Borle, B. de Kruijff, J. Seelig, 1986 - J. Seelig - Comparative 2H- and 3tP-NMR study on the properties of palmitoyllysophosphatidylcholinein bilayers with gramicidin, cholesterol and dipalmitoylphosphatidylcholine. *Biochim. Biophys. Acta*. **854**, 133–142 (1986).

124. Z. O. Shenkarev *et al.*, NMR structural and dynamical investigation of the isolated voltage-sensing domain of the potassium channel KvAP: implications for voltage gating. *J. Am. Chem. Soc.* **132**, 5630–5637 (2010).

125. E. Deplazes *et al.*, Biochimica et Biophysica Acta. *BBA - Biomembranes*. **1858**, 872–882 (2016).