Protein-water interactions studied by molecular dynamics simulations

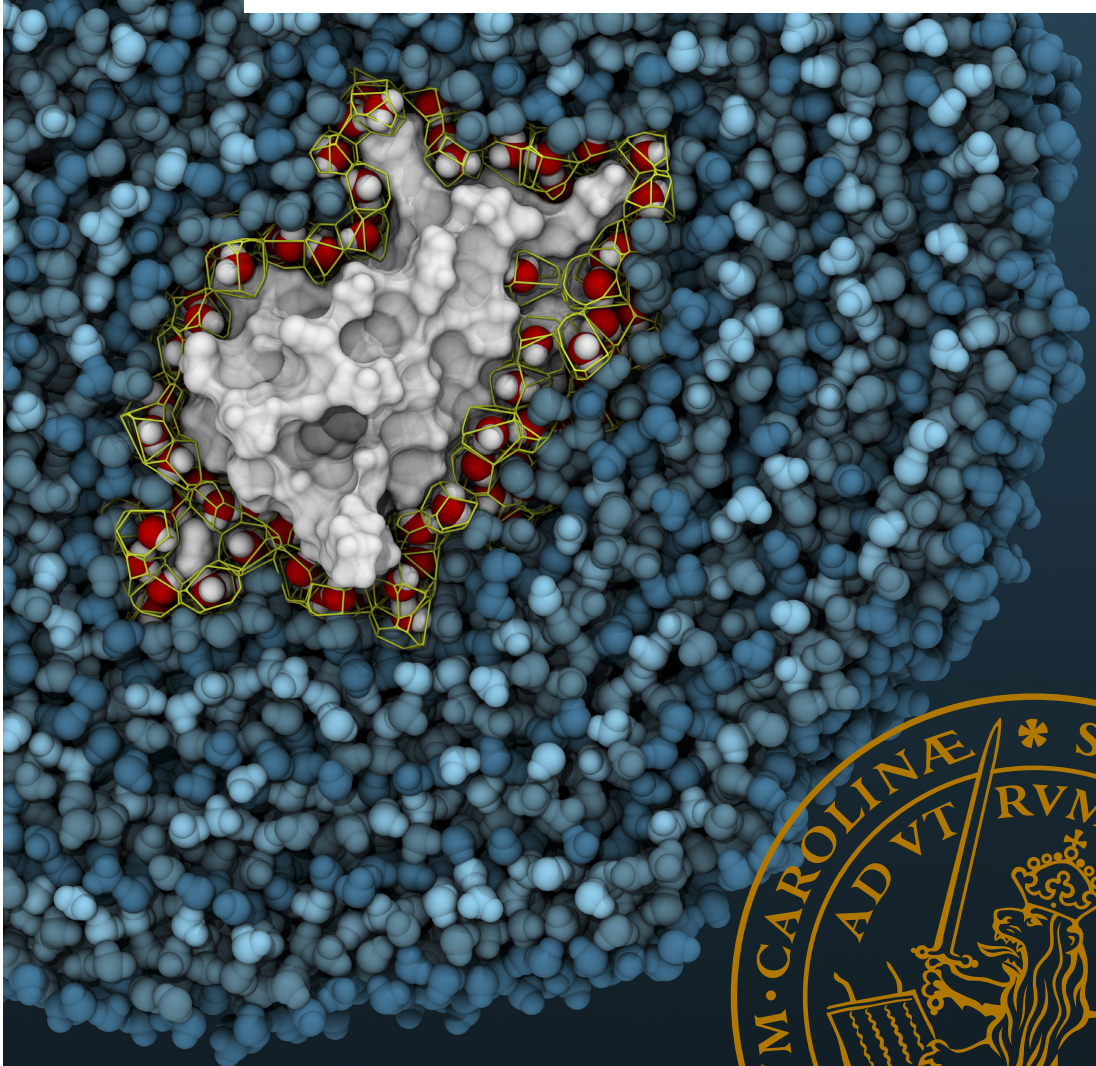
Persson, Filip

2018

Document Version:
Publisher's PDF, also known as Version of record

Protein-water interactions studied by molecular dynamics simulations

FILIP PERSSON | DIVISION OF BIOPHYSICAL CHEMISTRY | LUND UNIVERSITY



Protein-water interactions
studied by molecular dynamics simulations

by Filip Persson

LUND
UNIVERSITY

Thesis for the degree of Doctor of Philosophy
Thesis advisor: Prof. Bertil Halle
Faculty opponent: Prof. Kresten Lindorff-Larsen

To be presented, with the permission of the Faculty of Engineering (LTH) of Lund University, for public criticism in KC:G lecture hall at the Center for Chemistry and Chemical Engineering on Thursday, the 22th of March 2018 at 10:15.

Organization	Document name
LUND UNIVERSITY Division of Biophysical Chemistry Box 118 SE–221 00 LUND Sweden	**DOCTORAL DISSERTATION** Date of disputation 2018-03-22 Sponsoring organization
Author(s) Filip Persson	

Title and subtitle

Protein-water interactions studied by molecular dynamics simulations

Abstract

Most proteins have evolved to function optimally in aqueous environments, and the interactions between protein and water therefore play a fundamental role in the stability, dynamics, and function of proteins. Although we understand many details of water, we understand much less about the protein-water interface. In this thesis we use molecular dynamics (MD) simulations to cast light on many structural and dynamical properties of protein hydration for which a detailed picture is lacking.

We show that the 1 ms MD simulation of the bovine pancreatic trypsin inhibitor (BPTI) by Shaw *et al.* (Science 2010, 330, 341) reproduces the mean survival times from magnetic relaxation dispersion (MRD) experiments by computing the relevant survival correlation function that is probed by these experiments. The simulation validates several assumptions in the model used to interpret MRD data, and reveals a possible mechanism for the water-exchange; water molecules gain access to the internal sites by a transient aqueduct mechanism, migrating as single-file water chains through transient tunnels or pores. The same simulation was also used to reveal a possible mechanism for hydrogen exchange of backbone amides, involving short-lived locally distorted conformations of the protein whereby the amide is presolvated by two water molecules before the catalyst can approach the amide through a water wire.

We perform MD simulations of several small globular proteins in dilute aqueous solution to spatially resolve protein hydration. Defining mono-molecular thick hydration shells as a metric from the protein surface, we compute structural and dynamical properties of water in these shells and show that the protein-induced water perturbation is short ranged, essentially only affecting water molecules in the first hydration shell, thus validating the model used to interpret MRD data. Compared to the bulk, the first shell is 6 % more dense and 25-30 % less compressible. The shell-averaged rotation of water molecules in the first hydration shell is retarded by a factor 4-5 compared to bulk, and the contributions to this retardation can be resolved based on a universal confinement index. The dynamical heterogeneity in the first shell is a result of water molecules rotating by different mechanisms on a spectrum between two extremes: a collective bulk-like mechanism and a protein-coupled mechanism where water molecules in confined sites are orientationally restricted and require an exchange event.

Key words

Protein hydration, Water, Dynamics, Density, Compressibility, Hydration shell, MD simulation, Amide hydrogen exchange, Internal water, NMR, MRD

Classification system and/or index terms (if any)

Supplementary bibliographical information	Language
	English
ISSN and key title	ISBN
	978-91-7422-573-0 (print) 978-91-7422-574-7 (pdf)
Recipient's notes	Number of pages: 408 / Price
	Security classification

Protein-water interactions
studied by molecular dynamics simulations

by Filip Persson

LUND
UNIVERSITY

Cover illustration front: A snapshot from a simulation of BPTI solvated with almost 30,000 water molecules. The solvent excluded surface of BPTI is shown in white and water molecules in van der Waals representation. Water molecules in the first hydration shell are depicted in red and white togheter with their associated (additivley weighted) Voronoi cells (yellow). The following 12 hydration shells are depicted with three reoccurring colors of blue.

Cover illustration back: Exchange event of internal water molecules in BPTI from a snapshot of an ultra-long MD simulation.

Funding information: The thesis work was financially supported by the Swedish Research Council.

Faculty of Engineering (LTH), Division of Biophysical Chemistry

ISBN: 978-91-7422-573-0 (print)
ISBN: 978-91-7422-574-7 (pdf)

Printed in Sweden by Media-Tryck, Lund University, Lund 2018

MILJÖMÄRKT 3041 0903

ISO 14001:2004 CERTIFICATION Intertek

MADE IN SWEDEN

Media-Tryck is an environmentally certified and ISO 14001 certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

May all beings be happy

Preface

This is it. This is the main station on an almost life-long journey to understand biology at its core. Ever since I first saw the classical human body poster in kindergarten, around the same time my father showed the viscera of a lab rat to me and my brother, I have been fascinated by the stupendous complexity inherent in the machinery of life. For most people, the same fascination (and horror) does never become so tangible as when a baby is born or when our bodies cease to function normally. My quest to understand biology has taken me around the life sciences on a path I never imagined. Starting from basic chemistry, to cell biology, to physiology and pathology, just to realize the detailed explanations about life processes that I sought was never answered in a satisfactory way. In despair, I equipped myself with technological skills in bio-engineering to at least exploit my current knowledge in an industrial setting. To my surprise, the mathematics and physical-chemistry I rather unwillingly acquired at the time, turned out to provide the necessary framework to address and answer the driving forces governing life, down to the protein level. Well, it continued down to the atomic level. As Richard Feynman pointed out in his *Lectures on Physics* [1]:

> ...if we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms.

The only experimental technique (although purist might disagree) that allows this level of detail is by means of computer simulation. Suddenly I found myself in a challenging field, that almost consumed me, that provided the tools to address questions on a level I had never imagined. This thesis is a contribution in the quest to understand the machinery of life from the necessary view point of the water molecule.

On a stalled train to Stockholm
January 7th 2018

Contents

List of publications

This thesis is based on the following publications, referred to by their Roman numerals:

I **Transient access to the protein interior: simulation versus NMR**

Filip Persson and Bertil Halle
J. Am. Chem. Soc., **2013**, 135(23), pp 8735-8748

II **Analysis of protein dynamics simulations by a stochastic point process approach**

Bertil Halle and **Filip Persson**
J. Chem. Theory Comput., **2013**, 9(6), pp 2838–2848

III **How amide hydrogens exchange in native proteins**

Filip Persson and Bertil Halle
Proc. Natl. Acad. Sci. U S A., **2015**, 112(33), pp 10383–10388

IV **The geometry of protein hydration**

Filip Persson, Pär Södehjelm and Bertil Halle
Manuscript

V **Compressibility of the protein-water interface**

Filip Persson and Bertil Halle
Manuscript

VI **How proteins modify water dynamics**

Filip Persson, Pär Södehjelm and Bertil Halle
Manuscript

Acknowledgements

Doing a PhD is a long term commitment, and a difficult one. You are stuck with your research project and scientific problems whose solution requires pursuing dead-end paths and fail time and again. Yet, you continue banging your head against the wall. Because when a piece in the puzzle finally falls into place, you may have uncovered a tiny fraction of the universe, and you feel connected to it. To prevent you from perishing in the process, however, you depend heavily on the love and help of many important persons.

First of all I want to thank my supervisor, **Bertil**, who injected confidence and curiosity in a somewhat lost student, now almost seven years ago, when I started working on your challenging master project that later morphed into my PhD research project. Although the road has been bumpy, the journey has been inspiring. Your aptitude in seeking up and tackle down interesting problems in science is impressive - I have learned a lot.

Thanks also to my co-supervisor **Pär**; you provided the vital three-body dynamics for the project to succeed and you helped me with everything MD-related. A big round of thanks to the other seniors at CMPS: **Kristofer** for your passion in teaching and interesting discussions about science and pedagogics; **Bengt** for introducing me to the field; **Mikael** for the protein NMR introduction; **Sara** for the positivity and encouragement; **Ingemar** for the laughs and one or two questions about programming. **Tom** for the tech and training discussions. A special thanks also to **Marie** at Teoretisk Kemi for providing temporary refuge and support. To **Anders** and **Joachim** at LUNARC for all the help with MATLAB DCS, various compilation issues, everything GPU-related and increasing my storage-quota when needed.

Thanks to all PhD students and Post-Docs I had the pleasure to share my PhD with. **Johan Q** for all the help during my first year at BPC. **Risto** for the fun, positivity and soccer nights. **Erik**, **Carl** and **Mikael** for the BPC spirit. To the BPC tennis team: **Zhiwei** for the wisdom in physics and the panicky nights in Gothenburg solving problems in statistical mechanics; **Bhakat** for the tennis coaching and all the laughter and creativity; **Olof** for always lifting my spirits no mater how grumpy I am, it would have been difficult to survive this without you. **Karin** for the candor and the champagne. **Shuji** for the memorable beers at the Les Houches School of Physics. **Uli** for making sure we never missed a coffee break, the updates on South Park episodes and the scientific discussions. **Gleb** for the laughs and start-up creativity; **Michal** ”like a boss” for the lunches. **Sven** for the coffee-timing and latest summaries of the news outlets. **Johan W** for all the fun and interesting discussions as well as designing my training plan for S:t Hans Extreme. **Stefan** for the beer club, oysters and life hacks. **Tanja**, **Mattias** and **Kristine** for guidance in the roller coaster experience that is raising a toddler. **Angus** for the laughs and scientific curiosity and to **Samuel** for

the cheer-leading during the writing of my thesis.

Thanks to **Liqing** at the Hancock lab at Children's Hospital of Philadelphia for the internship and the fascinating immunology research you introduced me to. Thanks also to **Carl-Magnus** and **Lars Erik** at MentLife for making the gap to the life outside academia less intimidating. Thanks to everyone at the Northwest Vipassna Center, Onalaska WA, for the 10 days of noble silence that changed my life.

Tack till Rauhrackelgänget 🍻: **Josef** för att du alltid varit där i vått och torrt; **Daniel** för all humor; **Fredrik** för gästfriheten i Seattle.

Tack till BMC-gänget ⚠️: **Rasmus** för att du stod ut med mina ”föreläsningar” under tenta-pluggen; **Johan** för festerna och insläpp på medicinska föreningen; **Daniel** för ärligheten, squashen, festerna och filosofilektionerna; **Olof** för de oförglömliga tenta-pluggen och timmarna med Pro Evolution Soccer.

Ett sort tack till min fina familj. **Mamma** för den oändliga hjälpen och kampen mot suboptimala myndigheter och andra institutioner, oavsett storlek 😸💪. **Pappa** för all uppmuntran av mina olika intressen under åren och all lek; datorerna, elektronikbyggsatserna, metanolRCbilen, kamerorna, fyrverkeripjäserna...😀. **Sebastian** för allt du lärt mig vare sig det handlat om att spränga leksaker i luften eller programmera. **Caroline** för värmen, träningen och all hjälp.

Till sist tack till **Jessica**, för att du svajpade höger, all stöttning och kärlek. Det här hade varit omöjligt utan dig ❤️. Tack till **Hjalmar** för att du påminner mig om vad som är viktigt.

Popular summary in English

Around 4 billion years ago, our dry and scorching hot planet endured an incessant bombardment of dirty snowballs from outer space. The water that these meteorites carried eventually formed vast oceans as the planet cooled, and within these oceans, life emerged a few hundred million years later. These lifeforms used complex biomolecules, such as proteins, to self-organize and catalyze chemical reactions. From that moment on, all lifeforms on our planet have been dependent on liquid water to thrive.

Reflect on the stupendous timespan that these proteins have had to adapt to and exploit the properties of the ever-present water molecules in their surroundings. You will not be surprised when I tell you that most proteins embed water molecules as a building scaffold in their structure, that water force the protein to hide water-hating building blocks, or that water is an active participant in protein-catalyzed chemical reactions. It is with these proteins that the drug prescribed by your doctor interacts. If the drug is a good one, you will hopefully feel better as the drug molecule takes control over its target protein. If it is bad, we have to come up with something better. But this requires that we know exactly how proteins function, and we therefore have to bring water into the picture as it is not a passive bystander.

The interactions between water molecules and proteins is known as protein hydration, and involves all water molecules that have different properties compared to the bulk water. We say that these water molecules are perturbed by the protein. For decades, the magnitude and the spatial range of the protein-induced water perturbation have been a matter of debate, depending on the interpretation of various experiments. Some claim that water molecules are significantly affected far away from the protein, whereas most evidence point to a short-ranged perturbation. Ideally, we would like to have a microscope allowing individual water molecules to be monitored, but no experimental technique available can do this for us. The next best thing at our hand is therefore a computational microscope, made out of supercomputers, sophisticated software, and mathematical functions to describe the chemistry. The computational microscope will simulate and record the behavior of the protein under experimental conditions, giving us a movie showing the motion of water molecules and the protein.

In this thesis we have used molecular dynamics (MD) simulations as our computational microscope to map out and measure the protein-water perturbation. By assigning water molecules into shells we obtain a convenient handle to describe distances from the protein surface. Each shell is one water molecule thick and the first shell contains all water molecules in contact with the protein surface. For each shell we study several water properties, and many can be compared to results obtained from experiments. For instance, we have looked at how tightly packed water molecules are in each shell and how they fluctuate. We have also studied the rotation of water molecules to determine how long time it takes before a water molecule has lost its

positional memory. Virtually all properties that we look at are only changed in the first shell compared to bulk water. This verifies assumptions used in the analysis of experimental data, and it casts doubt over the claims by some research groups that the protein perturbs water even in the eighth shell.

We also used data from a "super-long" protein-water MD simulation to uncover how internal water molecules exchange with the surrounding bulk via water-filled tunnels and pores that form as the protein spontaneously change its structure. This finding led us to further investigate how different parts of the protein are transiently open and exposed to the surrounding water molecules. By analyzing the open state, we have postulated a mechanism for how water and protein hydrogen atoms swap places, a question that has remained unanswered for more than 60 years.

Populärvetenskaplig sammanfattning på svenska

För circa 4 miljarder år sedan blev vår torra och stek-heta planet bombarderad av smustiga snö-bollar från yttre rymden. Vattnet som dessa meteoriter bar på bildade så småningom stora hav när jorden avsvalnade, och några hundra miljoner år senare, uppstår liv i dessa hav. De enkla livsformerna använde stora biomolekyler, såsom proteiner, för att organisera och föröka sig. Liv på jorden har ända sedan dess varit beroende av flytande vatten för att frodas.

Reflektera över vilken otrolig tidsrymd som proteiner har haft för att anpassa sig till och uttnyttja egenskaperna hos de ständigt närvarande vatten-molekylerna i dess omgivning. Du kommer inte bli förvånad när jag säger att nästan alla proteiner har vatten-molekyler inbyggda i sin struktur, att vattnet tvingar proteinet att gömma undan vatten-skygga byggelement, eller att vatten är en aktiv del i protein-katalyserade kemiska reaktioner. Det är med dessa proteiner som medicininen utskriven av din läkare samspelar med. Om det är en bra medicin kommer du känna dig bättre när läkemedelsmolekylen tar över kontrollen över dess mål-protein. Om det är en dålig medicin däremot, måste vi göra den bättre. Men det kräver att vi föstår hur proteiner fungerar, och därför måste vi ta med vattnet i vår förståelse eftersom det inte är en passiv åskådare.

Interaktionerna mellan vattenmolekyler och proteiner kallas för proteinhydratisering, och inkluderar alla vattenmolekyler vars egenskaper skiljer sig från rent vatten. Vi säger att dessa vattenmolekyler är störda av proteinet. I flera årtionden har man dividerat över hur mycket och över vilken räckvidd som vattnet störs av proteinet. Grunden för denna oenighet är att experimentella resultat kan tolkas på flera sätt. Vissa menar att vattenmolekyler påverkas över väldigt långa avstånd från proteinytan, medan de flesta andra menar att det bara är vattenmolekylerna precis i närheten av proteinet som påverkas. Om vi hade fått önska skulle vi vilja ha ett mikroskop där vi kan studera enskilda vattenmolekylers beteeende när de närmar sig proteinytan. Tyvärr kan ingen experimentell teknik idag göra detta för oss. Det näst bästa vi har tillgång till är mikroskop bestående av super-datorer, avancerad mjukvara och matematiska modeller för att beskriva kemi. Detta datormikroskop simulerar hur rikiga proteiner beter sig i olika vatten-miljöer, och det vi får ut i slutändan är en film som visar rörelserna hos vattenmolekylerna och proteinet.

I den här avhandlingen har vi använt molekyldynamik-simuleringar (MD) som vårt datormikroskop för att kartlägga och mäta proteinets påverkan på omgivande vattenmolekyler. Genom att fördela alla vattenmolekyler i skal runt proteinet får vi ett behändligt avståndsmått till proteinytan. Varje skal är en vattenmolekyl tjockt, och det första skalet är de vattenmolekyler som är i kontakt med proteinet. För varje skal tittar vi på flera egenskaper hos vattnet och många av dem går att jämföra med experiment. Vi har bland annat undersökt hur tätt vattenmolekylerna packas i varje skal och hur detta varierar över tid. Vi tittar även på vattnets rotation och bestämmer

hur lång tid det tar för vatten i de olika skalen att tappa sitt positionsminne. För praktiskt taget alla egenskaper som vi undersöker ser vi att det bara är vatten i det första skalet som skiljer sig från rent vatten. Detta beräftar flera antaganden i olika experiment, men det kastar också stort tvivel på de forskargrupper som hävdar att proteinet påverkar vatten upp till åttonde skalet.

Vi har även använt en superlång MD-simulering för att identifera hur vattenmolekyler inbäddade i proteinet byter plats med det omgivande bulkvattnet genom kortlivade tunnlar som uppstår då proteinet spontant ändrar sin struktur. Den här observationen gjorde oss nyfikna på ett annat fenomen, nämligen hur delar av proteinet tillfälligt öppnas upp och exponeras för det omgivande vattnet. Genom att analysera det öppna tillståndet kunde vi beskriva en möjlig mekanism för hur väteatomer i delar av proteinet byter plats med väteatomer hos vattnet. Detta har varit ett mysterium i över 60 år.

We wish to pursue the truth no matter where it leads — but to find the truth, we need imagination and scepticism both. We will not be afraid to speculate, but we will be careful to distinguish speculation from fact. The cosmos is full beyond measure of elegant truths; of exquisite interrelationships; of the awesome machinery of nature. The surface of the Earth is the shore of the cosmic ocean. On this shore we've learned most of what we know. Recently we've waded a little way out, maybe ankle deep, and the water seems inviting. Some part of our being knows this is where we came from. We long to return. And we can. Because the cosmos is also within us. We're made of star-stuff. We are a way for the cosmos to know itself.

— Carl Sagan[1]

[1]Episode 1 in the TV series *Cosmos: A Personal Voyage* (1980)

Chapter 1

Introduction

Follow the water[1]

When Carl Sagan said to his viewers "we are made out of star-stuff", he meant it literally; the atoms in our body are traceable to the stars that cooked the light atoms hydrogen and helium into heavier ones. Among them carbon, oxygen, nitrogen and other ingredients fundamental for life. The enriched guts of the stars were scattered all across the galaxy as they became unstable in their later years and finally exploded, forming gas clouds that later condensed to solar systems with orbiting planets, Earth one amongst them some 4.5 billion years ago [3]. A little bit later, bombardment of ice-carrying meteorites may have brought water to Earth's hot surface that eventually formed oceans as the planet cooled. We do not know exactly when or how, but some 4 billions years ago life emerged in these oceans [4]. Ever since then, life on Earth cannot be sustained without liquid water.

The large biomolecules, such as proteins, comprising life's machinery have consequently had a "very long" time to adapt and exploit the conditions set by the physical and chemical properties of liquid water. If we want to understand how proteins perform their function, their stability, structure and dynamics must be viewed against this aqueous backdrop. Although we have detailed knowledge on bulk water's structure and dynamics, we understand much less about how water behaves near the protein surface. What is the spatial range over which the structure and dynamics of a water molecule deviates from bulk water due to the presence of a protein? How does this perturbation vary with distance and what is the nature of its coupling? These questions have become increasingly contentious in the scientific community, especially with the ever-increasing sophistication of experimental tools for which less sophisticated physical models may be used to extract meaningful information about protein hydration. Because no experimental technique can unambiguously determine the

[1] NASA's mantra in the search for life in outer space [2].

number of water molecules affected by a protein, the best we can do at the moment to resolve these questions is via computer simulations.

In this thesis we have used molecular dynamics simulations to characterize the interactions between protein and water. Chronologically, we have worked our way from the protein interior, via exchange of internal water molecules (paper [I-II]), to the exterior via protein conformational changes transiently exposing buried backbone amide hydrogens to water (paper [III]). As we continued farther away from the protein, we were motivated to properly define hydration shells as a metric for water-distance to the protein surface (paper [IV]). Having defined robust hydration shells, we could then analyze several structural (paper [IV] and [V]) and dynamical (paper [VI]) properties for water molecules in each shell, some of which can be compared directly to results obtained from experiments.

1.1 Why do simulations?

Much of our current understanding about the molecular properties of proteins has come through experiments, accompanied by models representing a simplified picture of the observable that is being measured. In order to provide an understanding that "makes sense", the model has to trade-off accuracy for simplicity. An example of a model frequently adopted is the two-state model for protein configurations as used in amide hydrogen exchange (section 2.2) for instance. But a simple model will not give us detailed information about all molecular properties in a complex system such as a protein in aqueous solution. The more difficult and interesting our questions are, the more desirable it becomes to have detailed and exact (in some sense) data about our system. This is where computer simulations come into the picture. Here, the model is detailed instead, but much more accurate on the other hand. The models in themselves do not necessarily provide any interesting information, but when plugged into powerful computers they will provide vast amount of data that allows in principle any property to be "measured". Whereas the subtle details about molecular motions, and the fast time-scale over which they occur, are difficult to probe experimentally, they represent no obstacle to a simulator.

The simulation provides a path from the microscopic details of the system to the macroscopic properties observed in experiments. If the model in the simulation is good, the results can be compared to experiments and provide insights to the experimentalist which can simplify the (often) complicated interpretation of experimental data. Because of this bridging role, connecting models and experimental results, and the way simulations are carried out, simulation techniques are rightfully called "computer experiments".

Chapter 2

Protein hydration

In this chapter we will cover a selection of the many aspects of protein hydration that are addressed in this thesis. It includes the structure and dynamics of water inside and outside of the protein surface - the hydration shell. The connection between water inside of proteins, so called internal water molecules, and the outside bulk is related to the process of amide hydrogen exchange that will also be covered.

Before continuing, we interject the definition of hydration which is ambiguous. The term mainly refers to (1) the total interaction of a solute with its aqueous solvent environment; and (2) the perturbation of the properties of water as a result of the interaction with the solute [5]. The second definition is more restrictive and will be used here. At some distance from the protein surface, the aqueous environment should display properties of bulk-like water, i.e. pure water without the protein. The problem at hand when understanding protein hydration is to understand to what extent water near the protein is different from the bulk.

2.1 Internal water molecules

Native proteins fold spontaneously from the polypeptide chain to adopt a tertiary structure that is necessary for function. The principal driving force for this folding is the hydrophobic effect [6–8]; apolar side-chains are driven away from entropically unfavourable contacts with water. During the folding process, water molecules may be incorporated into the structure to achieve minimal frustration in the folding energy landscape, balancing the (free-energy) optimization problem of maximizing the number of hydrogen bonds and, at the same time, the packing density [9]. Thus, these internal water molecules provide favourable hydrogen bonds to be formed with otherwise unsatisfied polar atoms while maintaining optimal packing [10, 11]. In this way, internal water molecules heal packing defects that would otherwise form empty cavities. In addition, they also provide ways for catalytic or binding processes to occur [12, 13]. Internal water molecules should therefore be regarded as an integral part

of the protein, and they are conserved to the same extent as amino acid sequence [14]. Figure 2.1 shows the protein systems studied in this thesis, with internal water molecules depicted.

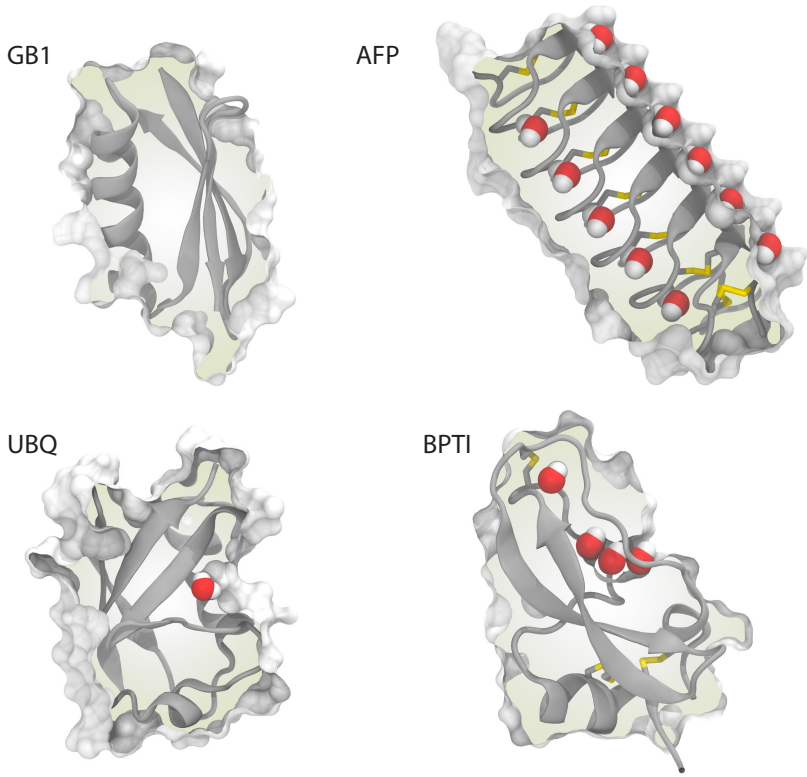


Figure 2.1: Crystal structures of the four proteins studied in this thesis, showing the outline of the solvent accessible surface (white), the secondary structure (gray), disulfide bridges (yellow) and internal water molecules. Missing residues or hydrogen atoms have been added. **GB1** the immunoglobulin-binding domain B1 of protein G from *Streptococcus sp.* (PGB1 [[15]]) contains no internal water molecules. **AFP** the insect antifreeze protein from *Tenebrio molitor* (1EZG [[16]]) with five internal water molecules together with waters on the ice-binding surface. **UBQ** mammalian ubiquitin (1UBQ [[17]], residues R74, G75, G76 removed) contains one internal water molecule close to the protein surface. **BPTI** bovine pancreatic trypsin inhibitor (5PTI [[18]]) contains four internal water molecules of which three form a hydrogen-bonded water chain.

Internal water molecules are very frequent in globular proteins. A statistical survey of high-resolution (r<1.5 Å) crystal structures found internal water molecules in 90 % of the 261 examined proteins[1] [19]. The number of internal water molecules between proteins is very variable. It correlates with protein size but not with the fold type, although fewer internal water molecules are observed for proteins containing many helical secondary structures[11, 19]. Instead, internal water molecules tend to be in regions with residues in loop conformations. Following O. Carugo, internal water

[1]The proteins had a length of (mean±std) 217 ± 6 residues

molecules can be classified as ”lake-like” or ”bay-like”[19]. Lake-like water molecules are completely isolated from the bulk solvent, whereas bay-like water molecules are connected to the bulk through a surface water molecule. On average, there are 2.4 lake-like and 2.8 bay-like water molecules per 100 amino acid residues. Lake-like water molecules are never found to be deeply buried in the protein; the minimum distance between the protein surface and a water molecule in a lake-like cluster is 2.7 Å on average, suggesting that internal water molecules are just beneath the protein surface. As might be expected, comparing crystallographic B-factors shows that lake-like water molecules are as rigid as protein atoms, and that bay-like water molecules are slightly less rigid.

Since the protein is not static but samples many conformational sub-states, these internal water molecules will occasionally undergo exchange with the external bulk water. The average life time of internal water molecules have been measured by magnetic relaxation dispersion (MRD, section 2.3.2). Depending on the hydration site, analysis of MRD data shows that internal water molecules exchange with external ones on a time scale ranging from tens of nanoseconds to hundreds of microseconds [20–22]. Thus, by probing the exchange rate of internal water molecules, which is a rare and transient event on the molecular time scale, one obtains information about the underlying protein dynamics. However, the exchange mechanism between internal hydration sites and bulk solvent is unknown, but large-scale conformational fluctuations are thought to be necessary[20]. In paper [I] we do a detailed characterization of internal-water exchange in BPTI using an ultra-long MD simulation.

2.2 Hydrogen exchange in proteins

Even though proteins have a rather high packing density, they undergo fluctuations that expose the most deeply buried parts of the polypeptide chain to the external solvent. This was first suggested more than 60 years ago by Hvidt and Linderstrøm-Lang, who demonstrated that all backbone amide hydrogens in insulin exchanged with the surrounding water hydrogens [23]. It has now become clear that all backbone amide hydrogens in proteins eventually undergo exchange, with half-times ranging from seconds to years. By monitoring amide hydrogen exchange, we can therefore obtain information about the structure, flexibility and, in favourable cases, the dynamics of proteins.

Hydrogen atoms covalently bonded to protein O, N and S atoms are labile and will undergo a hydrogen exchange reaction (HX) when exposed to solvent. Because one hydrogen atom is replaced by another one, the reaction is monitored in D_2O so that they can be distinguished in the exchange process

$$\mathrm{P{-}H + DOD \longrightarrow P{-}D + HOD} \tag{2.1}$$

where D is the exchanging deuterium atom and P is a protein O, N or S atom. An NMR spectrometer tuned to hydrogen will not "see" deuterium (due to different spin numbers) and the signal from the P-H atom will therefore gradually disappear in a HX experiment[1]. The difference in mass between the hydrogen atom and deuterium atom also allows the exchange process to be measured by mass-spectrometry (MS) experiments.

Hydrogen exchange is catalyzed by both acids and bases, including the autoprotolysis products of water, the hydronoium ion H_3O^+ and the hydroxide ion OH^-. In a buffer-free aqueous solution, the pH-dependence of hydrogen exchange rate k_{ex} is the sum of contributions from acid-, base-, and a pH-independent water catalysis according to

$$k_{ex} = k_w + k_a[H_3O^+] + k_b[OH^-] = k_w + k_a 10^{-\mathrm{pH}} + k_b 10^{\mathrm{pH}-\mathrm{p}K_w} \qquad (2.2)$$

where the second order rate constants k_a and k_b are the acid- and base catalysed rates respectively, k_w is the rate constant for water catalysis, and $K_w = [H_3O^+][OH^-]$ is the ionization constant for water with $pK_w = 14.00$ at 25 °C[25]. The rate constants k_a and k_b have been determined for model compounds where the labile hydrogen atom is fully solvent-exposed. For instance, the exchange rate for the amide hydrogen in poly-D,L-alanine at 25°C is plotted in Fig 2.2 with $k_a = 42\ M^{-1}min^{-1}$ and $k_b = 1.1 \cdot 10^{10}\ M^{-1}min^{-1}$ [26]. As can be seen, the minimum of the pH-dependent curve (pH_{min}) around pH 3 is the result of the much more effective base catalysis. The pH-independent (water-catalysed) exchange is only significant in experimental exchange rates measured at pH near pH_{min}. The position of pH_{min} varies considerably due to the inductive and steric blocking effects imposed by the neighbouring sidechains. This effect has been quantified in a set of correction factors [26] to the rate constants in Eq 2.2, allowing the exchange rate to be predicted for any structureless peptide sequence. Figure 2.2 shows the exchange rate profiles for two unstructured dipeptides as predicted by Eq 2.2 with correction factors to rate constants for PDLA [26].

In the native state, the measured exchange rate of protein amide hydrogens is lower than for solvent-exposed peptides since most of the backbone peptide groups are buried inside the protein. Nevertheless, even the deeply buried amides are exposed to solvent as the protein undergoes conformational changes. Because of this transient exposure, the analysis of HX experiments is based on the following kinetic scheme (the Linderstrøm-Lang model) [27]

$$\mathrm{N{-}H_{(C)}} \underset{k_{cl}}{\overset{k_{op}}{\rightleftharpoons}} \mathrm{N{-}H_{(O)}} \xrightarrow{k_{int}} \mathrm{N{-}D} \qquad (2.3)$$

[1]HX is typically measured using 1D ^{1}H NMR or 2D NMR such as $^1H^{15}N$ HSQC [24].

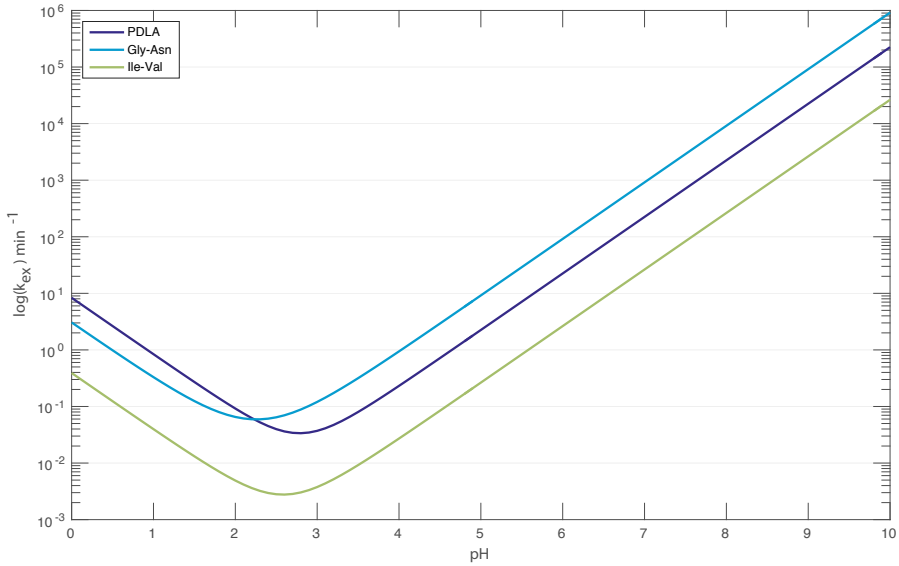


Figure 2.2: Hydrogen exchange rate profiles ($\log(k_{ex})$ for two structureless dipeptides as described by Eq 2.2 using rate constants for poly-DL-Alanine (PDLA) at 25 $^{\circ}C$ with correction factors from reference [26].

where each amide can exist in a closed (C) state, where exchange cannot occur, or in an open (O) state, where exchange can occur at the intrinsic rate k_{int}. The general rate equation describing this process is given by [27]

$$k_{ex} = \frac{k_{op} + k_{cl} + k_{int} - \left[(k_{op} + k_{cl} + k_{int})^2 - 4k_{op}k_{int}\right]^{1/2}}{2} \tag{2.4}$$

In the two state model above, the protein fluctuates between the C and O state with equilibrium constant K_{op}. At equilibrium, we have the detailed balance condition

$$k_{op} f_C = k_{cl} f_O \tag{2.5}$$

where f_O and f_C are the fractional equilibrium populations of the two states. The equilibrium constant K_{op} can then be written

$$K_{op} = \frac{k_{op}}{k_{cl}} = \frac{f_O}{f_C} = \frac{\tau_O}{\tau_C} \tag{2.6}$$

where we have introduced the mean life times in the two states

$$k_{op} = \frac{1}{\tau_C} \tag{2.7}$$

$$k_{cl} = \frac{1}{\tau_O} \tag{2.8}$$

For a buried amide, solvent exposure will be a rare event and the population of the closed state (f_C) can therefore be assumed to be much larger than the population for the open state (f_O). In view of detail balance (Eq 2.5), $f_C \gg f_O$ so that $k_{cl} \gg k_{op}$. In Eq 2.4, we can then assume $(k_{op}k_{int}) \ll (k_{op} + k_{cl} + k_{int})^2$ to obtain the simplified equation

$$k_{ex} = \frac{k_{op}k_{int}}{k_{op} + k_{cl} + k_{int}} = \frac{k_{op}k_{int}}{k_{cl} + k_{int}} \tag{2.9}$$

where $k_{cl} \gg k_{op}$ was invoked in the last step. Equation 2.9 can be further simplified in two limiting cases, known as the EX1 and EX2 limit.

2.2.1 The EX2 limit

Under non-perturbing conditions, the protein structure can be regarded as stable such that $k_{cl} > k_{op}$. For instance, if we assume the mean life time of the open state, τ_O, to be less than 1 μs [1], we have $k_{cl} > 10^6$ s^{-1}, which means that k_{cl} is much faster than k_{int} even at high pH (see Fig 2.2). Under these conditions, opening and re-closing of the open state occurs many times before a successful exchange can occur, and Eq 2.9 reduces to the EX2 limit.

$$k_{ex} = \frac{k_{op}}{k_{cl}}k_{int} = \frac{f_O}{f_C}k_{int} \tag{2.10}$$

where the last step follows from Eq 2.5. The vast majority of HX experiments are performed under conditions where the EX2 limit applies, and consequently do not provide any information about the conformational dynamics underlying the exchange. In order to make practical use of Eq 2.10, we further have to assume that the intrinsic exchange rate can be approximated with the exchange rate from model peptides as described by Eq 2.2. This allow us to express a protection factor κ defined as

$$\kappa \equiv \frac{k_{int}}{k_{ex}} \tag{2.11}$$

Thus, the protection factor on a buried backbone amide reports on how much the exchange rate is slowed down compared to a solvent exposed peptide. In view of Eq 2.6 and 2.7, the protection factor can also be expressed

$$\kappa = \frac{f_C}{f_O} = \frac{1}{K_{op}} = \frac{\tau_C}{\tau_O} \tag{2.12}$$

Since protection factors scale with the inverse of K_{op}, they also provide information about the free energy change , ΔG_{op}, associated with the opening process

[1] The mean life time of the unfolded state from MD simulations of several fast folding proteins [28].

$$\Delta G_{op} = -k_B T \ln\left(\frac{f_O}{f_C}\right) = k_B T \ln \kappa \tag{2.13}$$

The free energy of the opening process can be compared with the free energy of global unfolding ΔG_{UF} from denaturation experiments. Indeed, the amides in peptide groups deep in the apolar core typically exchange by global unfolding as suggested by $\Delta G_{op} \approx \Delta G_{UF}$ [29]. Hydrogen exchange in the more peripheral amides seems to require only local unfolding based on denaturation sensitivity. However, the structural features of the locally unfolded (open) state has been a matter of debate for decades, as well as the mean life-time of the open state. Two models have been proposed for how the exchange catalyst, in most cases the hydroxide ion, access the protein interior for amide hydrogens that do not exchange in the unfolded state. In the "penetration model" [30], the catalyst enters the protein via transiently formed channels and cavities. Speculations on how these channels arise include redistribution of interior hydrogen bonds [31] or from random association of pre-existing cavities [32]. In the "local unfolding model" [33, 34] on the other hand, structural elements, like the α-helix, transiently unfolds into the bulk solvent where exchange can occur [35]. It is assumed that the main barrier to exchange is provided by hydrogen bonds to amide hydrogens. In this model, correlated exchange behaviour has been suggested since adjacent amide hydrogens in the unfolded region are predicted to exchange at roughly the same rate.

Given that the nature of the open state is not known, it is difficult to escape the fact that the analysis of hydrogen exchange in the EX2 limit fully depends on the assumption that exchange in the open state is equivalent to that of solvent-exposed model peptides. In paper [III], we try to characterize the hydrogen exchange mechanism using an ultra-long MD simulation.

2.2.2 The EX1 limit

Provided that the protein is not degraded, it is possible to reach the EX1 limit at very high pH. Here, $k_{cl} \ll k_{int}$ so exchange occurs immediately when the amide hydrogen atom is in the open state. In this limit, Eq 2.9 simplifies to

$$k_{ex} = k_{op} = \frac{1}{\tau_C} \tag{2.14}$$

and measured exchange rates thus report on the dynamics of the fluctuations underlying the exchange. The distinction between the EX1 and EX2 limits is determined by the pH-dependence of k_{ex}. Whereas exchange in the EX1 limit is essentially pH-independent, exchange in the EX2 limit depends on pH the same way as k_{ex} for model peptides shown in Fig 2.2 [24].

2.3 The hydration shell

The first experimental studies of protein hydration was performed by adding water incrementally to dry protein powders. The process was continued until a level of hydration was reached in which the experimental quantity did not change with further addition of water. This was termed the hydration end point, and the hydration shell was defined as the amount of water covering the protein on average at the endpoint. For many of the properties studied (such as the heat capacity and enzyme activity), the hydration level end point was at around 0.3-0.4 g water per gram protein. This was interpreted as a hydration shell corresponding to a monolayer of water molecules where each water on average cover 20 $Å^2$ on the protein surface[36]. However, the hydration level will depend trivially on the protein size, making the translation to the number of water molecules on the protein surface questionable.

Although the term hydration shell originally referred to the water molecules in contact with the protein [37], the term has become ambiguous with a qualitative and a quantitative interpretation [38]. Qualitatively, the hydration shell is the one-molecule thick layer of water molecules that fully wrap the protein. Contrary to experiments, this qualitative picture can be realized (more or less) in molecular simulations by applying a set of geometric conditions to assign water molecules to the shell. A common method to define the hydration shell from a simulation-generated configuration is to include all water oxygen atoms within a given maximum distance from the closest protein atom. Typically, a uniform 3.5 Å distance-cutoff to heavy protein atoms is used so that any water oxygen within the cutoff is assigned to the shell [38–41]. Another method is based on topological neighbours based on Voronoi-tessellation (see section 4.2) where all heavy atoms are assigned a polyhedron, so that any point inside of it is closest to that particular atom; all water polyhedra that share a face with protein polyhedra are defined to be in the first shell [42–45]. By the same token, successive hydration shells can be defined by both methods and the spatial range of the protein-induced water perturbation can be studied in each shell.There is no consensus on how to define these hydration shells and in paper [IV] we do a thorough comparison between the most widely used methods.

Quantitatively, the hydration shell comprise all water molecules with properties different from bulk water. However, this perturbative view of the hydration shell is non-trivial as it can depend on the particular property being probed, and thus on the experimental technique. Indeed, the magnitude and the spatial range of the perturbation - the thickness of the hydration shell - is controversial as no experimental technique can unambiguously provide the number of water molecules that are perturbed by the protein.

We will not attempt to review all experimental techniques used to study protein hydration, which can be found elsewhere [5, 38]. Instead, we will outline the current understanding of the protein-induced water perturbation, and its contrasting views.

We will do one small exception, however, concerning magnetic relaxation dispersion (MRD) experiments, since parts of this thesis have been motivated by the need to quantitatively test the approximations in the model used to interpret MRD data.

2.3.1 Structure

The structure of water in the hydration shell has been studied by X-ray and neutron diffraction which provide — in most cases — generic information about (time-averaged) positional correlations [5]. The electron density maps from X-ray diffraction provide spatial information on the heavier atoms such as oxygen, nitrogen and carbon, while neutron diffraction allows hydrogen atoms to be detected. The position of individual water molecules can be derived from diffraction data on protein crystals, provided that they are ordered to yield maxima in the electron density map. Small angle X-ray and neutron scattering (SAXS and SANS) are also used to study hydrated proteins which provide information on the radial pair distribution function (section 4.1) of atom pairs.

Diffraction studies have shown that the highly corrugated protein surface, with its heterogeneity in polar, non-polar, and charged groups, results in different local hydration geometries [38]. From a Voronoi volume analysis (see section 4.2) of protein crystals, it has been suggested that the water density at the protein surface is ∼20% higher compared to bulk [46], with higher water densities in concave regions. Scattering experiments on hydrated proteins have also shown a mean density-excess of 10-15 % in the hydration shell [47], and a complementary MD simulation has confirmed this [48]. Yet, other MD simulations have suggested a modest density increase between 1-3 % for proteins [42, 43] and polypeptides [49, 50]. This discrepancy is scrutinized in paper [IV].

2.3.2 Dynamics

The range of the perturbation has been studied by NMR on simple model systems, showing that only water molecules in contact with the solute surface have dynamics significantly different from bulk water [5]. This has also been suggested from MD simulations [45, 51], although the decay length of the short-range perturbation has not been characterized in great detail. This is one of the objectives in paper [VI]. In contrast, measurements from terahertz (THz) spectroscopy[1] suggest that the protein significantly perturbs water up to distances of 20 Å — corresponding to 7-8 monolayers of water — from the protein surface [52]. For an insect antifreeze protein, even longer perturbations was claimed [53]. In both cases, the evidence for long-range perturbation was argued to be supported by an MD simulation showing perturbed

[1] THz spectroscopy probes the collective hydrogen-bond distortions via absorbance in the far infrared frequency range.

hydrogen-bond dynamics and rotational relaxation up to a distance of 7 Å (i.e. 2-3 monolayers) [52].

MRD

While most evidence points to a perturbation range involving the first 1-2 monolayers, the magnitude of the perturbation is also debated. The most convincing evidence on its magnitude comes from magnetic relaxation dispersion experiments (MRD), which is one of the few methods that selectively probes the dynamics of water molecules in dilute protein solutions. In MRD, the longitudinal spin relaxation R_1 rate of the quadrupole water nuclei ^{2}H and/or ^{17}O in isotope enriched water is measured as a function of the resonance frequency ω determined by the applied magnetic field. Typically, measurements of R_1 on an aqueous protein solution spans several frequency decades to generate a dispersion profile. The dispersion profile, $R_1(\omega)$, shows the excess relaxation rate compared to bulk due to slower rotational water dynamics in the hydration shell and internal water molecules. Figure 2.3 depicts a typical dispersion profile measured for a dilute protein solution, and the relaxation rate is described by

$$R_1(\omega) = R_1^{bulk} + 0.2j(\omega) + 0.8j(2\omega) \tag{2.15}$$

Molecular level information is extracted from the frequency-dependent spectral density function $j(\omega)$; it is the Fourier transform of the rotational time correlation function (section 4.3) describing how fast a water molecule looses its orientational memory (section 4.3.3). The spectral density function describing $R_1(\omega)$ has the form

$$j(\omega) = \alpha + \beta\frac{\tau_\beta}{1 + (\omega\tau_\beta)^2} \tag{2.16}$$

where τ_β is the rotational correlation time (section 4.3.2). The parameters α and β describe the dynamics of two types of water in the hydration shell that exchange rapidly with the surrounding bulk water molecules. The constant α is the contribution to R_1 from water molecules rotating on a time scale faster than 1 nanosecond, but slower than the picosecond rotational correlation time τ_0 in bulk at room-temperature. The effect is seen as a frequency independent increase of the relaxation rate above the bulk value, R_1^{bulk}. The nanosecond-limit is set by the experimentally accessible timescale ($\sim$ 100 MHz), and the limit serves as an operational definition for slow and fast water molecules; those rotating slower or faster, respectively, than 1 nanosecond. If we know the number of water molecules that are perturbed by the protein, N_{hyd}, it is possible to extract the mean rotational correlation time $\langle\tau_{hyd}\rangle$ of those waters. In MRD it is assumed that only water molecules in contact with the protein are affected (the primary hydration shell), so that N_{hyd} can be estimated simply by dividing A_s,

the solvent-accessible surface area [54] (SASA)[1] of the protein, by the mean SASA that a water molecule occupies on the protein surface,a_H; $N_{hyd} = A_s/a_H$ in short. Computing SASAs is a standard tool in many molecular software packages, and many of them use the numerical algorithm by Shrake and Rupley [55]. In this thesis we have computed SASAs using the analytical algorithms implemented in MSMS [56] and getArea [57].

The second contribution to R_1 is from a few slow water molecules with rotational correlation times longer than 1 nanosecond. These are typically internal water molecules (section 2.1) or waters residing in deep pockets on the protein surface, where the rotation is highly restricted until the water is exchanged with external water molecules due to a protein conformational change. The slow water molecules produce the observable frequency dependence in the dispersion profile, and their contribution to R_1 is described by the β parameter. From the MRD profile, it is possible to determine the number of slow water molecules, and how rotationally restricted they are via an order parameter.

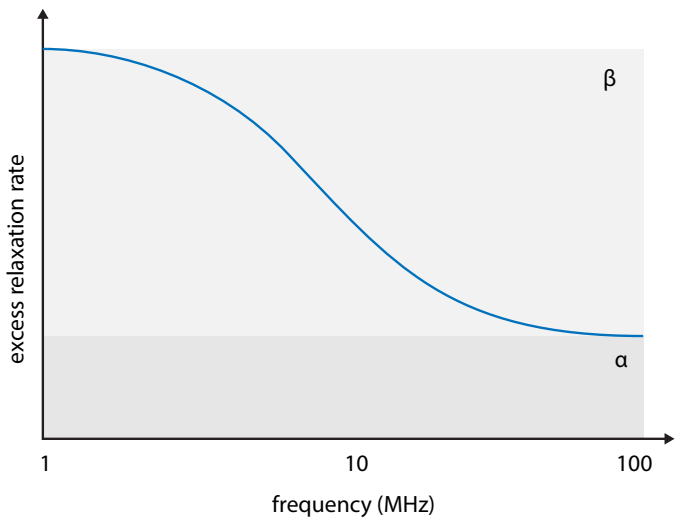


Figure 2.3: Schematic dispersion profile from magnetic relaxation dispersion (MRD) experiments. The excess relaxation rate relative to bulk is a sum of two contributions α and β, containing dynamical information about fast and slow water molecules, respectively, in the hydration shell.

MRD measurements on dilute protein solutions have established that water rotation in the primary protein hydration shell is only moderately perturbed compared to bulk water. Using $a_H = 15$ Å ^{2} and measurements on 11 proteins (fitted using Eq 2.15-2.16), gave a retardation factor $\langle \tau_{hyd} \rangle / \tau_0 = 5.4 \pm 0.6$ [5]. This is stronger than the retardation factors around 1-2 seen for small organic molecules and peptides [58–60].

[1]The SASA is the locus of points traced out by a water-like probe sphere as it rolls over the protein's vdW surface. A probe radius of 1.4 Å is typically used.

The main determinant for the degree of slowed dynamics appears to be the topography of the protein, resulting in various local geometries, such as pockets and grooves, that may interfere with the cooperative motions underlying water rotation and translation [5, 38]. For the most mobile half of water molecules, retardation factors around 2 have been estimated from MRD [61]. The origin of this dynamical heterogeneity is investigated in paper [VI].

Chapter 3

Molecular dynamics

Nature and Nature's laws lay hid in night; God said, Let Newton be! and all was light.
— Alexander Pope[1]

Molecular dynamics (MD) refers to the solution of Newton's laws of motion to propagate a set of molecules over time. In other words, we use the same laws of classical mechanics that were first postulated to study the motion of planets, stars, and other celestial objects. Although the actual behavior of microscopic systems is described correctly by quantum mechanics, this classical approach turns out to be a surprisingly good approximation at the molecular level [2]. In this chapter we cover the basic (and non-rigorous) foundation of molecular dynamics simulation and discuss some of the practical aspects involved in setting up a protein MD simulation. For a more rigorous description of MD there are many good books, and *Understanding molecular simulations* [64] by Frenkel & Smit is a good starting point.

3.0.1 Equations of motion

MD simulations are largely based on Newton's second law, stating that bodies accelerate under the action of an external force according to

$$\mathbf{F}_i = m\mathbf{a}_i = m\ddot{\mathbf{r}}_i \tag{3.1}$$

where $\mathbf{F}_i$ is the force on atom i with (Cartesian) position vector $\mathbf{r}_i$, m and $\mathbf{a}_i$ is its

[1] Epitaph indented for Sir Isaak Newton, Westminister Abbey (1730) [62].

[2] This simple classical treatment is justified within the Born-Oppenheimer approximation [63] — only nuclear positions have to be considered. Also, quantum effects can mostly be ignored in condensed systems with heavier atoms. For an ideal gas, the classical limit applies when the thermal de Broglie wavelength is much smaller than the inter-particle distance.

mass and acceleration respectively. Here we have adopted Newton's notation for differentiation, so that $\ddot{\mathbf{r}}_i$ above is defined as $d^2 r_i / dt^2$.

When working with a complex dynamic system, it is more convenient to use a reformulation of classical mechanics known as Hamiltonian mechanics [65]. Hamilton's equations of motion can be obtained from a generating function known as the Hamiltonian. The Hamiltonian $\mathcal{H}$ is usually the internal energy E of the system. For a system of N particles, the Hamiltonian may be written as the sum of kinetic ($\mathcal{K}(\mathbf{p})$) and potential ($\mathcal{V}(\mathbf{q})$) energy functions as [66]

$$\mathcal{H}(\mathbf{p}, \mathbf{q}) = \mathcal{K}(\mathbf{p}) + \mathcal{V}(\mathbf{q}) = \ = \frac{1}{2m} \sum_{i=1}^{N} \mathbf{p}_i \cdot \mathbf{p}_i + \mathcal{V}(\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_N) \tag{3.2}$$

where $\mathbf{q}_i$ is the position of atom i and $\mathbf{p}_i$ is the momentum of the atom. The coordinates $\mathbf{q}_i$ and $\mathbf{p}_i$ are generalized. This means we do not necessarily have to use a Cartesian coordinate system, which is sometimes useful when treating molecules as rigid bodies for instance. By differentiating $\mathcal{H}$ we obtain Hamilton's equations of motion:

$$\dot{\mathbf{q}}_i = \frac{\partial \mathcal{H}}{\partial \mathbf{p}_i} = \frac{\mathbf{p}_i}{m} \tag{3.3a}$$

$$\dot{\mathbf{p}}_i = \frac{\partial \mathcal{H}}{\partial \mathbf{q}_i} = \mathbf{F}_i \tag{3.3b}$$

In general, Hamilton's equations can be very complicated, but for simple liquids where the Cartesian coordinate system can be used, they become rather simple. In this case, Newton's second law can be recovered by eliminating $\mathbf{p}_i$ above, verifying that no new physics is introduced in this formalism.

3.0.2 Conservation laws

If $\mathcal{H}$ is both invariant under translation and rotation about an axis (by a judicious choice of generalized coordinates), it can be shown that the total linear and angular momentum are conserved [66]. In practice the angular momentum is actually not conserved in most MD simulations. This is because we have to use different box geometries (see section 3.3) for our system that break the symmetry required for the conservation to apply. However, the most important conservation law to mention is the conservation of energy. If $\mathcal{H}$ does not depend on time (explicitly), we may write the total time derivative of $\mathcal{H}$ as

$$\frac{d\mathcal{H}}{dt} = \sum_{i=1}^{N} \left[\frac{\mathbf{p}_i \cdot \dot{\mathbf{p}}_i}{m} + \frac{\partial \mathcal{V}(\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_N)}{\partial \mathbf{q}_i} \cdot \dot{\mathbf{q}}_i \right] \tag{3.4}$$
$$= \sum_{i=1}^{N} \left[\frac{\mathbf{p}_i \cdot \dot{\mathbf{p}}_i}{m} - \mathbf{F}_i \cdot \mathbf{q}_i \right] = 0$$

showing that $\mathcal{H}$ is a constant of motion and thus that Hamilton's equations conserve energy under these conditions.

3.0.3 The arrow of time

An interesting point to make regarding the equations of motions is that they are time symmetric [65]. This means that they are invariant under the transformation of time $t \to -t$. Thus, if we change all the sign of the velocities or momenta, the molecules in our simulation will retrace their motional course. This time-insensitivity of the equations of motion obviously seems to contradict the second (statistical) law of thermodynamics; how come that entropy (almost always) increases as time goes forward even though our equations do not distinguish between past and future? This inconsistency is known as Loschmidt's paradox.

3.1 Statistical ensembles

Since most simulations are carried out to facilitate the interpretation of experiments, we have to set up our simulations in a way that mimics the relevant experimental conditions. This implies we should solve the equations of motion for a stupendous number of degrees of freedom - 1 mole consists of $6.022 \cdot 10^{23}$! This gap is bridged by the realizations of statistical mechanics that macroscopic properties are not heavily dependent on the exact motions of every particle in the macroscopic system, but rather on averages of the microscopic details. This is the basis of the ensemble concept used in statistical mechanics [65]. An ensemble is a collection of systems that have different microscopic configurations (states) but share a set of common macroscopical (thermodynamic) properties (such as the total energy, volume and number of particles). The particles in each system of the ensemble evolve from different initial conditions so that every system is a unique microscopic state at any point in time. If the ensemble is in equilibrium, the experimental (macroscopic) observables can be obtained by averaging over all the microscopic states at any instant.

If we define an ensemble having constant number of particles N and volume V, and apply Newton's or Hamilton's equation, we are automatically in the so called microcanonical (NVE) ensemble in which the energy E of the system is constant (as described by Eq 3.4 above). However, our experiments are performed under constant

temperature and not constant energy. But temperature is problematic here since it is an ensemble average; we cannot know beforehand what the temperature will be given a value of the energy. In MD simulations we do not have ensembles *per se*, but we do obtain different microscopic states over time. Thermodynamic properties can then be obtained as time averages. If the simulation is long enough, then the time average is equal to the ensemble-average according to the ergodic hypothesis. To obtain the temperature from an MD simulation for instance, we average the kinetic energy $\mathcal{K}$ as

$$\frac{3}{2}Nk_BT = \langle \mathcal{K}(\mathbf{p}) \rangle = \Big\langle \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m} \Big\rangle \tag{3.5}$$

where k_B is the Boltzmann constant and the brackets denote averaging over time. This is the equipartition principle, stating that an average energy of $k_BT/2$ per each degree of translational freedom. An instantaneous kinetic temperature in the simulation can be defined as $\mathcal{T} = 2\mathcal{K}(\mathbf{p})/3Nk_B$.

3.1.1 Constant-temperature MD

There are various methods to achieve constant temperature during a simulation, and they are known as thermostats. All of them involve a tampering of the equations of motion, and they can be classified as either deterministic or stochastic thermostats [66] depending on the nature of the tampering. When studying dynamical properties we therefore have to be cautious which thermostats to use, otherwise there is a risk of obtaining spurious dynamics of the system. In this thesis, the MD simulations have used the stochastic Langevin thermostat and the deterministic Berendsen thermostat [67], and we will briefly outline how they work.

In both thermostats the system is coupled to a heat bath, an infinite energy reservoir, with a certain temperature. Energy is allowed to flow to the system and back to the reservoir. In the Berendsen thermostat [67] (also known as the weak-coupling thermostat), the system is coupled to the heat bath during the simulation and the momenta are scaled as

$$\mathbf{p}_i' = \mathbf{p}_i \sqrt{1 + \frac{\Delta t}{\tau}\Big(\frac{T}{\mathcal{T}} - 1\Big)} \tag{3.6}$$

where T is the desired (thermodynamic) temperature and $\mathcal{T}$ is the instantaneous kinetic temperature. The coupling-strength to the heat bath is set by the time constant τ in the scaling factor $\Delta t/\tau$. Although this method ensures a reasonable total kinetic energy for the desired temperature, it is not guaranteed that the temperature will be equal in all parts of the system, despite particle collisions tending to even out the temperature distribution [68, 69]. Hence, the distribution is non-canonical and sensibly depends on the scaling factor [66]. Despite this, the Berendsen thermostat is widely used as it efficiently relaxes the system to the desired temperature.

In the Langevin thermostat the equations of motion are modified by using Langevin dynamics as [66]

$$\dot{\mathbf{p}}_i = \mathbf{F}_i - \gamma_i \mathbf{p}_i + \mathbf{f}_i(t) \tag{3.7}$$

where γ is a friction constant (sometimes called a collision frequency constant) and $\mathbf{f}_i(t)$ is a random force that is uncorrelated with the dynamical variables (i.e. positions and momenta). During the simulation, all particle then receive a random force (this is the stochastic element) and have their momenta lowered by γ, which corresponds to collisions with imaginary heat-bath particles. The value of the random force is not set arbitrarily, but is connected to γ in a way that preserves the fluctuation-dissipation theorem [1]. This is the reason why the Langevin thermostat ensures the proper sampling in the canonical ensemble. However, when implementing the thermostat in a simulation one has to rely on pseudo-random numbers which may introduce synchronization artifacts if the seed to the random number generator is not updated [71].

3.1.2 Constant-pressure MD

In addition to constant temperature, most experiments in chemistry are carried out under constant pressure and it is desirable to replicate these conditions in our simulations as well. This is achieved by carrying out the simulation in the isothermal-isobaric (NPT) ensemble.

The pressure in an MD simulation can be obtained from the virial theorem. Depending on the functional form of the potential $\mathcal{V}(\mathbf{q})$, discussed in section 3.2.2, the virial pressure for pairwise interacting particles is given by [64]

$$P = \frac{k_B T N}{V} + \frac{1}{V}\left\langle \sum_{i=1}^{N}\sum_{j>i}^{N} \mathbf{f}_{ij}(\mathbf{r}_{ij}) \cdot \mathbf{r}_{ij} \right\rangle \tag{3.8}$$

where $\mathbf{f}_{ij}$ is the force excerted by particle i on j, at a distance $\mathbf{r}_{ij}$, and the brackets denote the time average. To adjust the pressure during the simulation we inevitably have to adjust the volume of the system, and methods to do this are called barostats. As for the thermostats, there exist both deterministic and stochastic barostats in which the system is coupled to a pressure "bath" at the desired pressure P_0.

The Berendsen barostat [67] is a commonly used (deterministic) pressure-control method that rescales the pressure in the same way as the Berendsen thermostat. During the simulation the system volume is rescaled by a factor μ. If our system is within a cubic box (and isotropic), then

[1] At equilibrium, the magnitudes of thermal fluctuations are related to how fast the system approaches equilibrium from a small perturbation. The theorem relates many transport properties, such as diffusion and viscosity, from these fluctuations [70]

$$\mu = 1 - \frac{\kappa_T \Delta t}{3\tau_P}(P_0 - P) \tag{3.9}$$

where P is the instantaneous pressure, κ_T is the isothermal compressibility and τ_P is a time constant that governs the coupling strength to the bath. However, this barostat does not sample rigorously from the NPT ensemble and will not generate the correct volume fluctuations [66], which is important when computing the isobaric compressibility κ_P as done in see paper [V].

To get the proper sampling from the NPT ensemble, one can implement the Monte-Carlo barostat that change the system volume by a stochastic process. During the simulation, this barostat will do trial moves where the current volume V_{old} is changed to a new volume V_{new}, with system coordinates rescaled accordingly. The probability of accepting this move, χ_{acc}, is [66]

$$\chi_{acc} = \min\Big[1,\ \exp(-\delta H/k_B T)\Big] \tag{3.10}$$

with

$$\delta H = \delta \mathcal{V} + P(V_{\text{new}} - V_{\text{old}}) - \ln\left(\left(V_{\text{new}}/V_{\text{old}}\right)^{N/k_B T}\right) \tag{3.11}$$

where δH is the energy change between the new (trial) and old (initial) system configuration. If the system energy is lower in the new configuration, the move is immediately accepted. If not, a random number ξ is generated uniformly on $[0, 1]$ and compared with $\exp(-\delta H/k_B T)$. If ξ is lower than $\exp(-\delta H/k_B T)$, the move is accepted. This procedure is the Metropolis Monte Carlo method [72], where the name 'Monte Carlo' refers to the gambling part of the method, i.e. the heavy use of random numbers.

3.2 Practical implementation

Having outlined the basic molecular mechanics involved in an MD simulation, we will now cover the practical aspects involved in setting up a protein MD simulation.

3.2.1 Numerical methods

The simplest form of the equations of motion such as Eq (3.1) cannot be solved analytically, and we therefore have to resort to numerical methods. A method that works surprisingly well is the finite-difference method, which can be outlined as follows: Given the positions and velocities at time t, we seek to find the positions and velocities at a later time $t + \Delta t$. The time interval Δt is called the time step in our simulation since we solve the equations of motion one step at a time; after many steps we have approximated the trajectory of each particle in the system, $\mathbf{r}_i(t)$.

If we do a Taylor expansion about $\mathbf{r}(t)$ in Newton's second law, Eq (3.1), we obtain [66]

$$\begin{aligned}\mathbf{r}_i(t+\Delta t) &= \mathbf{r}_i(t) + \Delta t\,\dot{\mathbf{r}}_i(t) + \tfrac{1}{2}\ddot{\mathbf{r}}_i(t)\Delta t^2 + \ldots \\ \mathbf{r}_i(t-\Delta t) &= \mathbf{r}_{\mathbf{i}}(t) - \Delta t\,\dot{\mathbf{r}}_i(t) + \tfrac{1}{2}\ddot{\mathbf{r}}_i(t)\Delta t^2 + \ldots\end{aligned} \tag{3.12}$$

Addition of these expansions yields the Verlet algorithm [73]

$$\mathbf{r}_i(t+\Delta t) = 2\mathbf{r}_i(t) - \mathbf{r}_i(t-\Delta t) + \Delta t^2\ddot{\mathbf{r}}_i + \mathcal{O}(\Delta t^2) \tag{3.13}$$

where the error is of order Δt^2. Although this method has very good energy-conserving properties, two problems are associated with it. First, since we subtract two large quantities from another, $2\mathbf{r}_i(t)$ and $\mathbf{r}_i(t-\Delta t)$, to obtain a small one, numerical imprecisions will arise. Secondly, the momenta are not present in this equation and must therefore be computed in order to estimate the kinetic energy (and hence the total energy). They may be obtained by another finite-difference as [66]

$$\dot{\mathbf{r}}_i(t) = \frac{\mathbf{r}_i(t+\Delta t) - \mathbf{r}_i(t-\Delta t)}{2\Delta t} \tag{3.14}$$

One improved version of the basic Verlet algorithm, that mitigates the numerical imprecisions, is the so called half-step 'leapfrog' algorithm [66]:

$$\begin{aligned}\dot{\mathbf{r}}_i(t+\tfrac{1}{2}\Delta t) &= \dot{\mathbf{r}}_i(t-\tfrac{1}{2}\Delta t) + \ddot{\mathbf{r}}_i(t)\Delta t \\ \mathbf{r}_i(t+\Delta t) &= \mathbf{r}_i(t) + \dot{\mathbf{r}}_i(t+\tfrac{1}{2}\Delta t)\Delta t\end{aligned} \tag{3.15}$$

The algorithm consists of half-advancing the velocities at time $t+\frac{1}{2}\Delta t$, from a time point $t-\frac{1}{2}\Delta t$, thus leaping over the coordinates at time t. The velocities at $t+\frac{1}{2}\Delta t$ are then used to determine the new positions at $t+\Delta t$, ahead of the velocities, at which point the accelerations are determined as well.

The Lyapnov instability

Because we use approximate methods, the trajectory will not follow the true trajectory, $\mathbf{r}(t)_{true}$, indefinitely. Even if we had exact methods, we are using finite precision arithmetic on our computers which will introduce errors (although tiny). Two systems with particles at identical positions, but with a tiny difference $\delta\mathbf{p}$, in momenta, will diverge from one another exponentially with time according to the Lyaponov divergence [64]

$$\Delta\mathbf{r}(t) \sim (\delta\mathbf{p})exp(\lambda t) \tag{3.16}$$

where $\Delta\mathbf{r}(t)$ is the distance (in phase space) between the trajectories, and λ is the so called Lyaponov exponent. This means that it is difficult to reproduce a simulation unless the systems are identical down to the very bit-level of precision. Any small difference in the prediction of position and momenta at each time step will result in large deviations at longer times. Arguably, this realization seems to render the whole idea of undertaking MD simulations pointless. But as pointed out in section 3.1, it suffice to obtain information about the statistical properties over a long time rather than predicting the true trajectory, which is necessary for satellites in space on the other hand. This is similar to the experimental reality where measurements may be taken over periods of time, and where each measurement represents a population average of the measured property.

3.2.2 The force field

Up until now, we have left out any details about the functional form of the potential function $\mathcal{V}(\mathbf{r})$, which contains all the interesting information about the molecular interactions. The mathematical form of the potential is called a force-field. Given that the major computational cost lies in computing the forces one needs to have a force field that is simple, yet can produce results consistent with experimental data of the same system. An optimal trade-off between these conflicting requirements have resulted in semi-empirical force fields, which are based on a combination of experimental results and quantum-mechanical calculations.

There are many different force fields to model proteins ansd several carry names associated with the MD-software for which they were originally developed. The most commonly used families of force-fields are AMBER [74], CHARMM [75], OPLS [76], and GROMOS [77]. These force-fields are similar, but differ in the set of functions used and the associated parameters. However, they are all built on two fundamental approximations: 1) electrons are not modeled explicitly, hence we cannot describe chemical processes where bonds are broken (e.g. enzyme catalysis); 2) the total potential energy is given by the sum of interactions between pairs of atoms, so many-body contributions have to be effectively included in the parameters.

The contributions to the potential are divided into two groups referred, somewhat inaccurately, as "bonded" and "non-bonded" energy terms. The bonded energy terms include interactions reflecting deformations of the local geometry, whereas the non-bonded energy terms describe interactions between atoms separated by more than two or three bonds, or atoms belonging to different molecules. In this thesis, we have used the AMBER family of force-fields with the functional form [74]

$$\mathcal{V}(\mathbf{r}) = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} k_\phi[1 + cos(n\phi - \phi_0)]$$
$$+ \sum_i \sum_{j>i} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_i \sum_{j>i} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{3.17}$$

where the first three terms are the bonded interactions and the remaining two terms are the non-bonded interactions. The functional form of each sum is depicted in Fig 3.1.

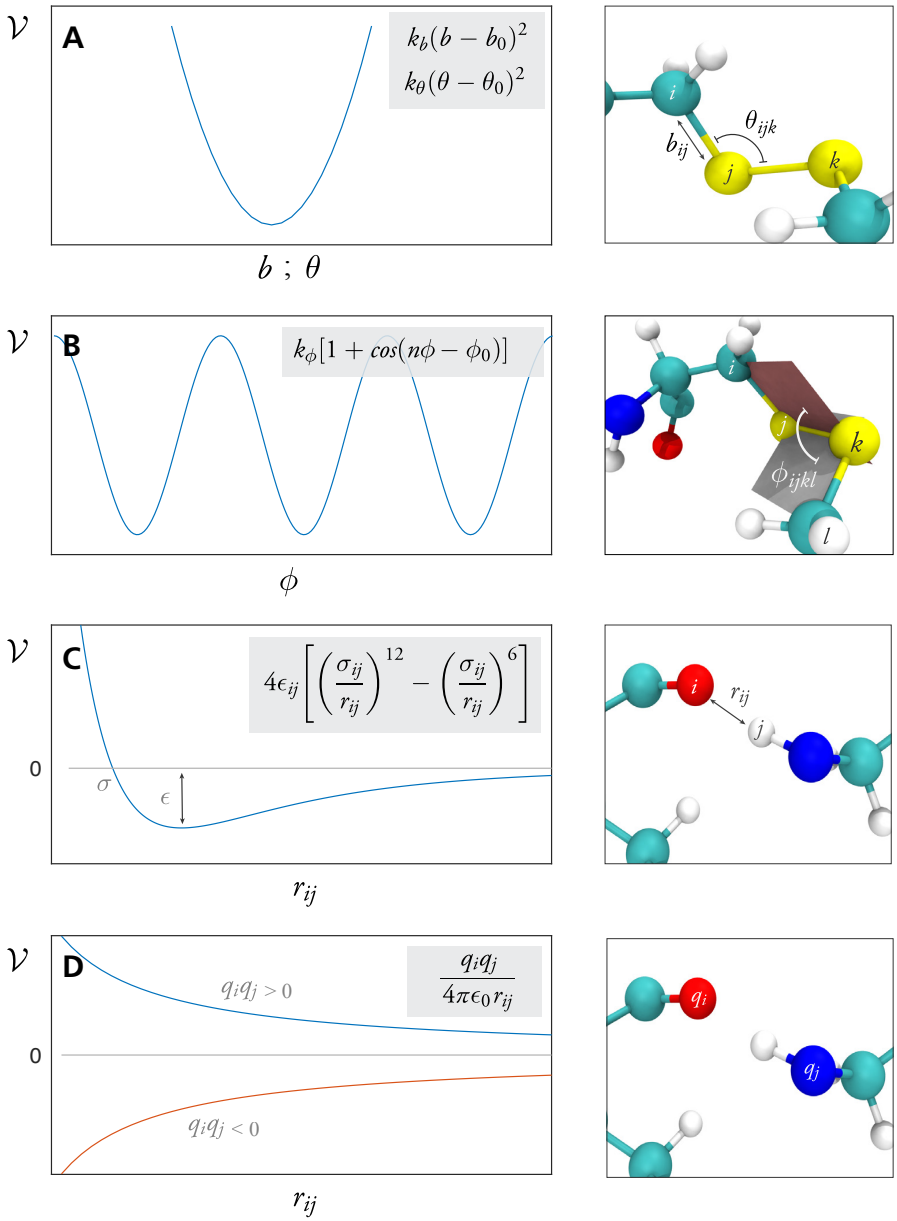


Figure 3.1: The interaction terms in a typical protein force-field. The bonded interactions describes local deformations in geometry with energy contributions from stretching (A), bending (A) and rotations (torsion) of bonds (B). Nonbonded interactions describes interactions of atoms separated by more than three bonds, with energy contributions from short-range dispersion (Lennard-Jones potential) (C) and long-range Coulomb interactions.

Bonded terms

The "bonded" interactions in Eq (3.17) includes energy contributions from bond stretching, angle bending and rotations about a bond, so called dihedral (torsion) rotations. The bond term is a sum over all *ij* pairs of atoms connected and describes the energy of deforming a bond length b (by stretch vibrations) from its ideal value b_0. The angle term is a sum over groups of three consecutively bonded atoms *ijk* and describes the energy of bending a bond (by bend vibrations) with angle θ from its ideal angle θ_0. Both the bond and angle term are modelled as Hookean springs with force constant k_b and k_θ respectively. The thirds sum in Eq (3.17) runs over all groups of four consecutive atoms *ijkl* and describe the energy change when the dihedral angle ϕ is rotated from 0 to 2π. The parameter k_ϕ is a force constant, while the parameters n is an integer (commonly set to 3) and ϕ_0 is the phase shift respectively.

Non-bonded terms

The "non-bonded" terms in Eq (3.17) consists of summing Lennard-Jones [78] (LJ) and Coulomb interactions. Both summations exclude atoms separated by one or two bonds (so called 1-2 and 1-3 pairs). Atoms connected by three atoms (1-4 pair) are partly described by the torsion term, so their LJ-interaction is scaled down. The LJ potential describes the exchange-repulsion and dispersion attraction acting on all atoms, while the Coulomb potential describes the electrostatic interaction between two point charges. Together, the LJ- and Coulomb interactions will also describe hydrogen bonds, which are electrostatic interactions, so we do not have to define them explicitly.

The mixed parameters (ϵ_{ij}, σ_{ij}) are obtained by different combination rules, typically as $\sigma_{ij} = 1/2(\sigma_{ii} + \sigma_{jj})$ and $\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}}$ (the Lorentz-Berthelot rules) [69]. While the decay of the dispersion term, r^{-6}, is physically motivated — describing the distance-dependence for the interactions of two freely rotating dipoles — the repulsive term, r^{-12} is less so. It is simply a rapidly decaying function that can conveniently be obtained by squaring the dispersion term. The true decay of the repulsion is better described by an exponential term as done in the Buckingham potential [79].

Since the LJ-parameters depends on the electronic distribution in the atoms, which is in turn affected by bond types, we need different parameters for the same element. In the AMBER ff14SB protein force-field [80] used in this thesis, parameters for 10 different carbon atoms are used, of which seven are unique for Histidine and Tryptophan residues.

The last term in Eq 3.17 is the Coulomb potential, where q_i and q_j are the partial charges on two atoms separated by the distance r_{ij}, and ϵ_0 is the relative permittivity in vacuum. The partial charges are positioned at the atomic centers, and their values are determined by QM-calculations such that they approximate the electrostatic potential from the continuous electron distribution [69].

Parametrisation

The process of assigning parameters to the functional form of the force field is called parametrisation. This involves fitting the functional forms to experimental results or quantum-mechanical (QM) calculations. The parameters b and θ_0 are typically obtained from crystal structures of small organic molecules, and the force constants k_b and k_θ from gas-phase vibrational spectra of similar molecules. The torsional parameters k_ϕ and ϕ_0 are often obtained by fitting to a QM-calculated potential for small organic molecules [69]. The LJ-parameters ϵ_{ii} and σ_{ii} are usually obtained by fitting to thermodynamic data on liquids, such as density and heat of vaporization [69, 74, 77].

3.3 Defining the system

We would like to carry out our simulations under bulk-like conditions in order to mimic the macroscopic sample probed by experiments. Because proteins are strongly affected by the surrounding solvent, we need to include the effect of the bulk solvent environment on our system. As mentioned in section 3.1, this implies we should account for on the order of 10^{23} molecules. Common computational resources limit this number to a couple of million atoms [81] with standard protein force-fields, and one therefore have to resort to approximations to mimic bulk-like behaviour.

There are two approaches we can employ to simulate the solvent environment. The first approach is to model the solvent implicitly as a continuum dielectric medium (using the generalized Born model for instance). However, this approximation is only accurate if the length scale of interest is large compared to the size of a water molecule [82]. The second approach, which is more rigorous, is to include a small number of solvent molecules explicitly in a simulation cell. Here we will only focus on the latter approach since all simulations in this thesis have been carried out using explicit solvent.

With explicit solvent, a typical MD simulation of a globular protein in solution would include several layers of water molecules inside a simulation cell. If it is a membrane protein, phospholipids would have been included as well. As an example, a simulation cell with a 15 kDa protein would consist of around 2000 protein atoms and around 5000 water molecules together with ions to neutralize the protein net charge. Clearly, the behaviour at the free boundaries of the simulation cell has to be specified, otherwise we would simulate an evaporation process in which the molecules diffuse throughout space. At the boundary, the fraction of molecules is proportional to $N^{-1/3}$ if the geometry of the simulation cell is a cubic box [64]. In the example above, some 30 % of the atoms will then be found at the surface, and their properties would be different from those in the bulk.

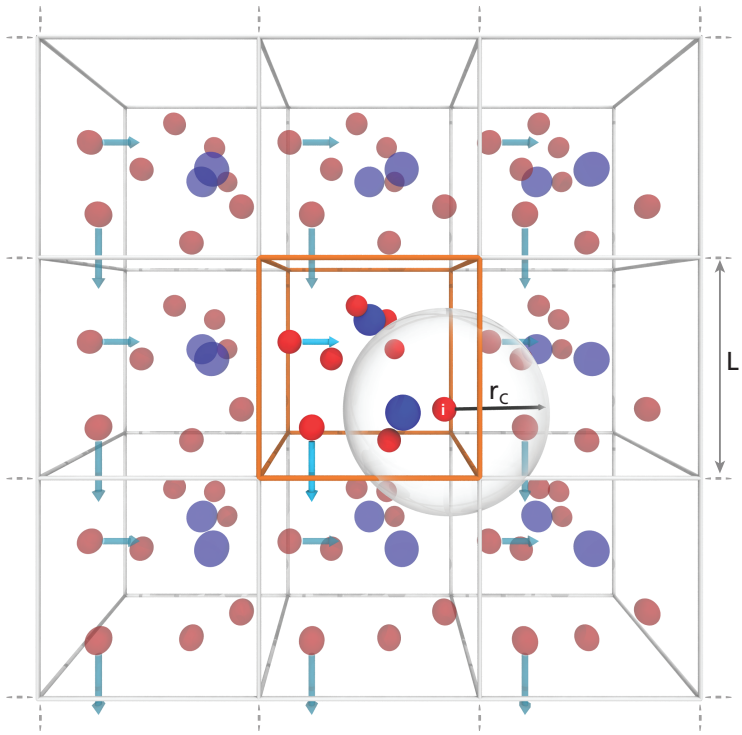


Figure 3.2: A three dimensional periodic system with two types of particles. The cubic primtitive cell (orange box), with side length L, is tiled in every dimension to create an infinite lattice with periodic images of the original cell (white boxes). When a molecule leaves the primitive cell, its periodic image moves in the same direction, so that one of its image enters on the opposite side. The transparent sphere with radius $r_C < L/2$ shows the cut-off distance for interactions to that particle, allowing interactions with particles in the closest neighbouring images.

Periodic boundary conditions

A solution to achieve a bulk-like system is to apply periodic boundary conditions [66]. The original simulation cell (often called the primitive cell) is then tiled throughout space to form an infinite grid of boxes (a lattice) called image cells, illustrated in Fig 3.2. The image cells have no degrees of freedom; as a molecule moves in the original cell, its periodic image in the neighbouring cells moves in the same direction. When the molecule leaves the original cell, it reappears on the opposite side via its periodic image. Thus, there are no walls in the primitive cell and no molecules on the surface. Practically, we do not follow the coordinates of all molecules in the images (they are infinitely many) but only their coordinates in the primitive cell. Other geometries than cubic systems can be used with periodic boundary conditions, including the rhombic dodecahedron and the truncated octahedron. Since these shapes are more spherical, they can significantly reduce the number of solvent molecules needed in the system.

When applying periodic boundaries we have to consider the range of the inter-

molecular potential. If the range is longer than the side length L of the simulation cell, then a particle will interact with itself (via neighbouring images) and we will impose spurious correlated motions. To prevent this we define a cut-off, r_C, for the interaction range, taken as $r_C < L$, so that only the interaction of a particle i with the nearest periodic image of particle j has to be considered. This is the nearest image convention. If the intermolecular potential is not zero for $r \geq r_C$, then the potential will have a discontinuity at r_C and we will introduce a systematic error in the total energy of the system. The total energy will then not be conserved. For short range interactions, we can correct for this by adding a small tail contribution to the potential. However, for the long-ranged[1] electrostatic interactions, with r^{-1} distance dependence, the tail correction will diverge [64]. Instead, lattice methods have to be used that sum interactions with all periodic images. One such method is the Ewald sum[83, 84], which is an effective technique to perform summation over all periodic images [66]. This technique is usually optimized for computational performance by assigning all the charges to a fine regular mesh which allows the fast Fourier transform algorithms to be used. This is implemented in the particle-particle/particle mesh (PPPM) algorithm [85]. A version of PPPM frequently used is the particle-mesh Ewald (PME) algorithm [86]. An important notion to make is that Ewald-summation techniques require the system to be electroneutral. This can be achieved by adding counter ions to the system or rescale the charges. For non-neutral systems, a uniform background charge is applied in the Ewald algorithms to effectively neutralize the system. However, for non-homogeneous systems such as protein in water, this may result in significant artifacts [87].

Water models

Most of the computational overhead will be spent on simulating the water molecules. Because of this, water molecules are often modelled as rigid bodies in most force fields so that the three internal degrees of freedom, involving bending and stretching of intramolecular bonds, are excluded. Thus, only the nonbonded interactions are included in the force field.

The simple rigid water models are grouped based on the number of interaction points, called sites, included in the model. In this thesis we have used the 3-site models TIP3P [88] and SPC/E [89], and the 4-site model TIP4P-Ew [90] which has been developed to be used with the Ewald summation technique. All models use a single LJ site for the oxygen atom and three partial charges, which is illustrated in Fig 3.3. The 3-site models have a negative partial charge at the centre of the oxygen atom and a positive partial charge at the centre of each hydrogen atom, whereas the 4-site models have the negative charge placed on a dummy atom (M) along the bisectris of the α_{HOH} angle. The geometries for TIP3P and TIP4P-Ew are taken from the

[1] Usually defined as a force decaying slower than r^{-d}, where d is the dimensionality of the system.

Table 3.1: Parameters for water models used in the thesis.

Model	r_{OH} (Å)	α_{HOH} (°)	σ (Å)	ϵ (kJ/mol)	$q(H)$	$q(O)$	$q(M)$	r_{OM} (Å)
TIP3P [88]	0.957	104.52	3.151	0.636	0.417	−0.834	0	0
SPC/E [89]	1.000	109.47	3.166	0.650	0.424	−0.848	0	0
TIP4P-Ew [90]	0.957	104.52	3.164	0.1627	0.524	0	−1.048	0.125

experimental gas-phase geometry of water monomers [90] whereas SPC/E adopts an HOH-angle permitting tetrahedrical bonding patterns. Parameters for these water models are given in table 3.1.

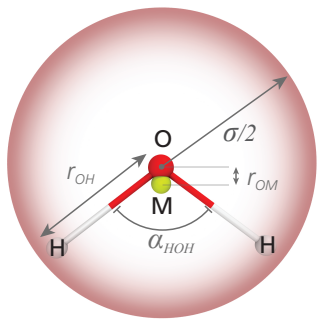


Figure 3.3: Parameters for the rigid 3- and 4-site water models.

The charges in all models are defined so that an effective liquid-phase dipole moment μ of 2.3 Debye is achieved. The current estimate for the dipole moment for liquid water is 3 Debye [94, 95]. The 3-site models accurately predict the densities ρ at fixed pressure, but TIP3P overestimates the dynamics by a factor two for the self-diffusion coefficient D [96]. The 4-site model, TIP4P-Ew, reproduces many of the qualitative features of the water phase-diagram [90], including a density maximum at around $1°C$. Dynamical and structural properties, such as the water-oxygen radial distribution function (see section 4.1) have also been shown to be in very good agreement with experiment. Table 3.2 summarizes some calculated physical properties for the water models. In paper [VI], we further benchmark these water models with respect to experimentally determined rotational correlation times.

Starting up

If we had unlimited compute resources, we could take any configuration of the unfolded protein and let it fold spontaneously during the simulation. This is not yet a viable option unless we are simulating a small and fast-folding protein [28]. Instead, the initial configuration of the protein is taken from crystal structures.

Table 3.2: Calculated and experimental properties for the water models at 298 K and 1 atm.

Model	μ [D]	ϵ	ρ [kg/m^3]	D [10^{-9}m^2/s]	C_p [J/(mol K)]	α_V [10^{-5}K^{-1}]	κ_T [GPa^{-1}]
TIP3P [a]	2.35	82	1000.2	5.19	83.6	92	0.65
SPC/E [b]	2.35	71.1	999.5	2.46	86.3	49.1	0.47
TIP4P-Ew [c]	2.32	63.9	995.4	2.4	80.3	32	0.49
Expt. [d]	1.86*	78.4	997.5	2.3	75.3	25.6	0.46

References [a][[91]], [b][[92]], [c][[90]] and [d][[93]]. *gas phase.

Before the real simulation can start, the system usually has to be prepared as follows. First, missing atoms (typically hydrogen) in the protein structure are added. Solvent molecules are placed in the simulation box, without creating substantial overlap in the configuration. Any steric clashes or covalent strain will lead to large forces which is a potential issue when solving the equations of motion. This is prevented by performing an energy minimization of the system to the nearest local minimum in the energy landscape. The steepest descend method [66] will relax the system by setting the velocities at the start of each step to zero, which allows the system to evolve "downhill" in the direction of the forces. Physically, this corresponds to a rapid cooling of the protein to 0 K.

The initial velocities of all molecules have to be specified in order to propagate the system by the equations of motion. Often, this is done by randomly assigning each molecule velocity components drawn from a Maxwell-Boltzmann distribution [64] at the desired temperature.

Most protein simulations are carried out at room-temperature to mimic common experimental conditions, and the system must then be heated from 0 K to around 300 K. The heating step may be performed over a short time interval by including restraints in the protein force field. Since the barostats can cause instabilities at low temperature, the heating step is often done in the NVT ensemble. The system is then allowed to equilibrate in the NPT ensemble until the density has converged. If all goes well, the system is then setup to begin the production simulation used for the main analysis.

The length of the time step

In an MD simulation, the force calculations are by far the most time consuming part of the simulation as they have to be determined at every time step. Consequently, we do not want Δt be too short, but not too long either; Eqns 3.13-3.15 are truncated Taylor series and only accurate approximations if the time step Δt is sufficiently small. If the time step is too long the total energy of the system will not be conserved, which manifests as an energy drift during the simulation. To guarantee energy conservation, the time step is often taken to be an order of magnitude less than the fastest motions

in the system. For proteins in solution the fastest motion is the vibrational motions of bonds to hydrogen atoms. For example, the main band in the vibrational spectra of liquid water at 298 K occurs at $\sim$ 3400 cm^{-1} [97], which corresponds to a period of 10 femtoseconds (1 fs $= 10^{-15}$ s). Thus, the time step in a the simulation has to be 0.5-1 fs to ensure that the simulation is stable (no energy drifts over time) if this type of motion is of interest in the simulation. However, these bond stretching vibrations are of minimal interest when studying protein function and dynamics, which involve molecular motions on a much longer time scale. By constraining the bonds to hydrogens to a fixed length one can increase the time step to 2 fs. This involves adding constraints where the equations of motion are solved while imposing the constraint. Commonly used algorithms for applying constraints are SHAKE [98], RATTLE [99] and LINCS [100].

Chapter 4

Analysis of MD simulations

Data! Data! Data! I can't make bricks without clay!
— Arthur Conan Doyle [1]

Once the dust has settled and the production simulation is finished we are left with a trajectory of the system saved to a storage device. Because successive time steps in the simulation are correlated, the trajectory only contains snapshots (frames) of the simulation sampled at time intervals adequate for the subsequent analysis [2]. Each frame usually contains coordinates for the atoms, but it can also contain velocities or forces. As the frames are time ordered, the trajectory is a movie of the evolution of the system. From the trajectory we can extract various types of information reporting on the structural and dynamical properties of the molecules in the system, some of which can be compared to experiment and others that cannot be experimentally determined. Here we will focus on structural and dynamical correlation functions, including rotational correlation functions of relevance for NMR. We will also discuss the nontrivial task of decomposing molecular volumes from an MD-generated configuration.

4.1 Radial distribution function

The structure in a liquid (such as water) can be probed by the radial distribution function (or pair correlation function) $g(r)$. We will leave out its formal definition in statistical mechanics and only consider its operational definition, i.e. how it is determined from a simulation. If we have a simulation of a liquid with N particles in a periodic cell with volume V, then the average (uniform) particle density is $\rho(\mathbf{r}) = \rho = N/V$.

[1] The Adventure of the Copper Beeches, page 322.

[2] In order to determine an oscillatory motion for instance, the sampling interval has to be at most $1/(2B)$ seconds for a motional frequency of B Hertz. This is the so called Nyqvist-Shannon sampling theorem.

We will drop the vector notation since the liquid is assumed to be homogeneous. The time-averaged density in a shell at radius r from a reference particle can be defined as $\rho g(r)$, where $g(r)$ is the probability of finding a particle at a separation r from the reference particle. However, the radial distribution function should not be regarded as a density *per se*; it does not represent a packing density that can be obtained physically. This point is elaborated in detail in paper [IV].

To calculate $g(r)$ from a simulation we do the following. First, we choose a reference particle i with position vector $\mathbf{r}_i$. With particle i as the center, we define spherical shells of radius r and thickness Δr. We count the $n_i(r, \Delta r)$ number of particles in the shell. Each particle j in the shell is separated from i at a distance $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$, so that $r - \Delta r \leq r_{ij} < r$. We then divide by the volume of the shell. Repeating the procedure for N reference particles, and taking the average we obtain the pair density $\rho^{(2)}(r)$

$$\rho^{(2)} = \frac{1}{N} \sum_{i}^{N} \frac{n_i(r, \Delta r)}{4\pi r^2 \Delta r} \tag{4.1}$$

By normalizing with the average particle density, $\rho = N/V$, we obtain

$$g(r) = \frac{V}{4\pi r^2 \Delta r N^2} \sum_{i}^{N} n_i(r, \Delta r) \tag{4.2}$$

Provided that we choose sufficiently thin shells, determined by the bin size for the r_{ij} particle distances, Eq 4.2 is the estimation of the true radial distribution function (RDF). Figure 4.1 illustrates the procedure for the RDF calculation, and how the local environment of a reference particle is reflected in the RDF — notice the short-range order in $g(r)$ which is a hallmark for liquids. The first and second maximum in $g(r)$ describes the coordination shell of the nearest neighbors and the second nearest neighbours respectively. Because of the disorder in the liquid, the peaks become less pronounced with increasing distance from the reference particle, and the distribution eventually approach the homogeneous density limit. Thus, the density of a shell at a large distance should approach N/V. Inserting this into Eq 4.2, we get

$$g(r) = \frac{V}{N^2} \sum_{i}^{N} \frac{n_i(r, \Delta r)}{4\pi r^2 \Delta r} \approx \frac{V}{N^2} \sum_{i}^{N} \frac{N}{V} = 1 \tag{4.3}$$

This shows that at a distance much larger than some characteristic correlation length ξ, the RDF goes to unity, i.e.

$$g(r \gg \xi) = 1 \tag{4.4}$$

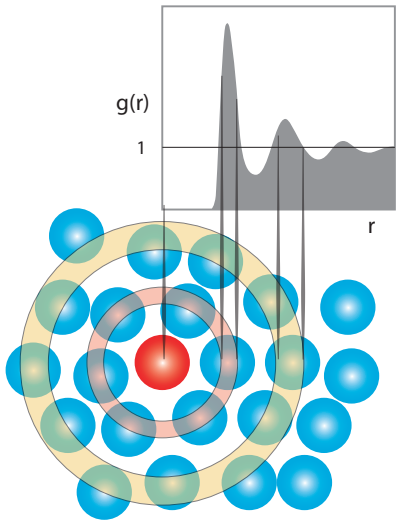


Figure 4.1: The radial distribution function $g(r)$ for a fluid of hard spheres. The first coordination shell (red area) is associated with the first peak in $g(r)$. The second shell (yellow area) has a broader second peak in $g(r)$ due to the more loose coordination of the spheres. As the distance increases, the correlation to the reference sphere is lost and $g(r)$ approach unity (the homogeneous limit).

4.1.1 Coordination numbers

The coordination number is the average number of particles n_c within a distance r_c from the reference particle. Figure 4.1 shows that n_c is related to the integral of the radial distribution function, and it is obtained as [65]

$$n_c = 4\pi\rho \int_0^{r_c} dr\, r^2 g(r) \tag{4.5}$$

To define the average number of particles in the first coordination shell we have to define an appropriate integration range. Although not unique, the position $r_c = r_{min}$ of the first minimum is typically used to define the first shell. A more general "running" coordination number can be calculated as

$$n_c(r) = 4\pi\rho \int_0^{r} d\,r'\, r'^2 g(r') \tag{4.6}$$

which defines the average number of particles coordinating a reference particle out to a distance r. From the MD trajectory we calculate the coordination number within distance $r_M = M\Delta r$ as

$$n_c(r_M) = \sum_{m=1}^{M} \frac{N}{V} 4\pi r_m^2 \Delta r\, g(r_m) \tag{4.7}$$

where r_m is the distance for (shell) bin m.

4.1.2 Experimental determination

The radial distribution function can be determined experimentally by diffraction techniques using X-rays and neutrons which have wavelengths comparable to the interatomic distances in liquids. In these experiments, the intensity $I(\theta)$ is measured with respect to the scattering angle θ from the incoming beam. The scattered intensity is described by an atomic form factor $f(k)$ and a structure factor $S(k)$ according to [65]

$$I(\theta) = f(k)NS(k) \tag{4.8}$$

where k is the magnitude of the scattering vector $\mathbf{k}$. While the atomic form-factor $f(k)$ is unique for the atomic species (and depends on instrumental details), the structure factor $S(k)$ contains the positional information of the atoms. The radial distribution function is essentially the Fourier transform of $S(k)$.

4.2 Voronoi diagrams

In MD simulations we can obtain other structural details that are not easily probed by experiment. Because we have the spatial coordinates in our simulation, we can compute a variety of geometrical quantities, such as the volume of the protein or analyzing shape complementary for a ligand bound to an active site. In this thesis we have analyzed the volumes of the protein and solvent molecules by a method known as Voronoi decomposition (also known as the Voronoi diagram/tessellation). Before we continue, it is pertinent to point out that volume is not an intrinsic molecular property (such as mass) as it depends on the environment. A molecule in a liquid will have different volumes depending on the nature of the interactions with the surrounding molecules. We should therefore distinguish between thermodynamic volumes, such as partial volumes obtained from experiments, and geometric volumes obtained from a simulation as described here.

Unlike for a macroscopic object, a dividing surface cannot be uniquely and precisely defined for a molecule. The volume for a molecule will thus depend on where we place the dividing surface. A common method is to partition the space by creating a Voronoi diagram, named after Georgy Voronoi who extended the method to higher dimensions in 1907 [101][1]. If we represent the atoms as points, that we call sites, then each site is assigned a region in space called a Voronoi cell. Each point in space is assigned to the nearest site, the Voronoi assignment, and the subsequent subdivision is the Voronoi diagram [103]. Operationally, the Voronoi cell for a site is created by defining a plane between each pair of surrounding sites, such that it is halfway and perpendicular to the line connecting them. The faces thus define a polygon in 2D, which is illustrated in Fig 4.2 A, and a closed polyhedron in 3D. The Voronoi cell has

[1] Voronoi diagrams had studied earlier by Johann G. Dirichlet in 1850 [102]

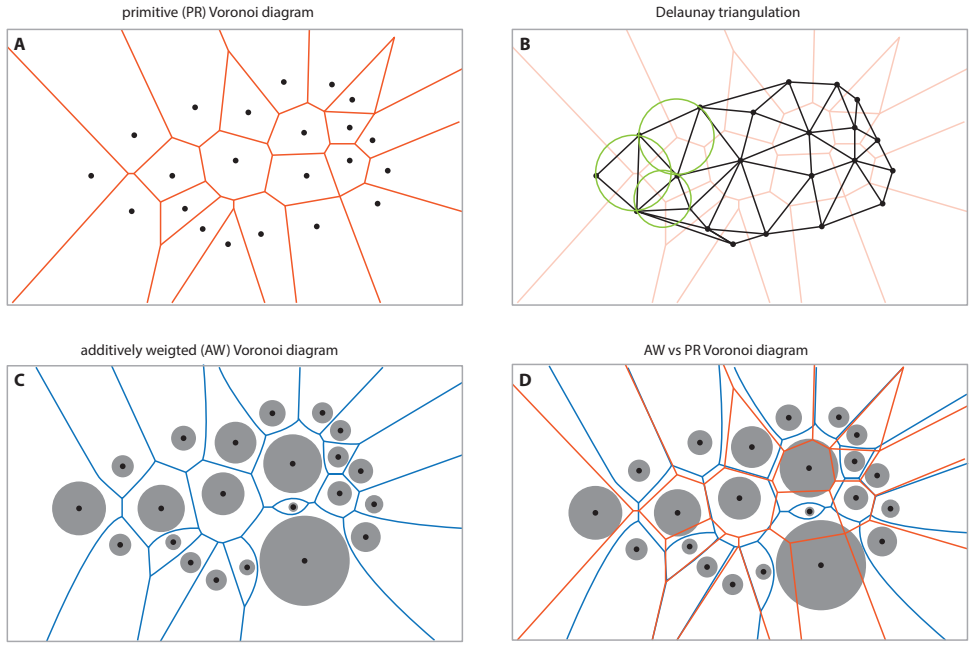


Figure 4.2: The Voronoi decomposition. **A-B** illustrates the relationship between the primitive (PR) Voronoi diagram and its dual graph, the Delaunay triangulation. Each triangle in B is created such that no point is inside the circumscribed circle (green) associated with that triangle. **C** The additievly weighted (AW) Voronoi diagram for a set of disks. In contrast to the planar surfaces in (A), the dividing plane is put halfway between the radii of the disks which creates curved surfaces. **D** Superposition of the PR and the AW Voronoi diagram.

the property that every point inside the cell is closer to the included site than to any other site. This fact has made Voronoi diagrams a versatile geometric structure with applications ranging from social geography to astronomy [103].

Mathematically, the Voronoi diagram is the dual graph of the Delaunay triangulation, which is usually determined as the first step in computing the Voronoi diagram. In other words, if we have the Delaunay triangulation for a set of points we automatically have the Voronoi diagram for the same set of points. For points in the plane, the Delaunay triangulation creates triangles such that no point in the set is inside the circumscribed circle for any triangle. This is illustrated in Fig 4.2 B. There are many algorithms available to compute the Delaunay and Voronoi tesselations from a set of points in 2, 3 or higher dimensions, but the most common is the Quickhull algorithm [104] implemented in the Qhull C++ library. In this thesis we have used both Qhull and the delaunayTriangulation class implemented in MATLAB.

When applying the Voronoi decomposition to molecular systems, the dividing plane is placed between the atomic centers and therefore does not take into account any difference in atomic size. This will be a rather unphysical approach for allocating space to atoms of different species. A variant of the Voronoi diagram called additively weighted (AW) Voronoi diagram, will instead put the dividing plane between

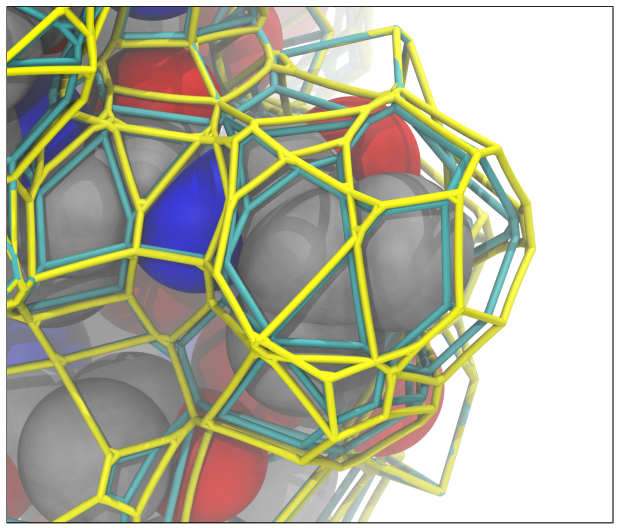


Figure 4.3: The Voronoi diagram for the protein-water interface using AW and PR tesselation. The edges in the AW diagram are shown in yellow, and the edges belonging to the PR diagram are shown in cyan. Oxygen (red), nitrogen (blue) and carbon (grey) atoms are shown with standard van der Waals radii. The tessellations were done on heavy atoms.

the surface of the atoms, typically defined by their van der Waals (vdW) radius. As a consequence, the surfaces of the Voronoi cells will be curved instead of planar. In this thesis we have used the Voronota algorithm [105] which has been developed specifically to obtain AW tesselations from vdW radii in molecular structures. For a system with the same atomic species, i.e. with the same atomic radii, the AW Voronoi diagram is equal to the PR Voronoi diagram. Figure 4.2 C shows the AW Voronoi diagram for disks in the plane, and the difference to the PR diagram is shown in Fig 4.2 D. It is not obvious how this difference translates to a Voronoi tesselation of a more complex system, such as a protein surrounded by water molecules. Figure 4.3 compares the AW and PR the diagram defining the protein-water Voronoi faces for heavy atoms. As can be seen, the differences are associated with faces belonging to carbon atoms that have a larger vdW radius than nitrogen and oxygen atoms. Although the difference between AW and PR looks small, which would render the AW tesselation unnecessarily complicated, the difference is crucial for the correct analysis of density an volume fluctuations in protein hydration (see paper [IV]).

The volume of a Voronoi cell can be computed easily by defining a point inside the cell, the centroid for instance, and triangulate to obtain irregular tetrahedrons and add their volumes. The volume of an irregular tetrahedron can be obtained via the lengths of the edges as [106]

$$288V^2 = \begin{vmatrix} 0 & \mathrm{d}_{12}^2 & \mathrm{d}_{13}^2 & \mathrm{d}_{14}^2 & 1 \\ \mathrm{d}_{12}^2 & 0 & \mathrm{d}_{23}^2 & \mathrm{d}_{24}^2 & 1 \\ \mathrm{d}_{13}^2 & \mathrm{d}_{23}^2 & 0 & \mathrm{d}_{34}^2 & 1 \\ \mathrm{d}_{14}^2 & \mathrm{d}_{24}^2 & \mathrm{d}_{34}^2 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{vmatrix} \tag{4.9}$$

where d_{12} is the length of the edge connecting vertex 1 and 2, and similarly for the other pair of the four vertices. For the Voronoi cells generated by the AW Voronoi diagram, the calculation of the volume is much more complicated. The cell has to be triangulated to a fine mesh to define small sub volumes that are added up, but numerically adding many small values of finite precision will lead to rounding errors. Thus, volumes from PR diagram are preferable over the AW diagram whenever the decompositions involve the same type of atoms.

4.3 Time-correlation functions

Many stochastic processes can be characterized by their time correlation function (TCF) which measures statistical correlations across time signals. In MD simulations, TCFs are often computed since their time integral and Fourier transforms are more or less directly accessible to experiments, such has measurements from NMR or infrared spectroscopy (IR).

Statistical dependence between two different quantities A and B is often measured by the correlation coefficient as

$$c_{AB} = \frac{\mathrm{cov}(A, B)}{\sigma_A \sigma_B} \tag{4.10}$$

where *cov* is the covariance of the variables and σ their standard deviations. The correlation coefficient is normalized so that a high degree of correlation will give c_{AB} equal to 1 (or -1), whereas a value of 0 indicates no correlation. If we have sampled M values of the quantities above, we obtain the correlation coefficient as

$$c_{AB} = \frac{\frac{1}{M}\sum_{i=1}^{M}\big(A_i - \langle A\rangle\big)\big(B_i - \langle B\rangle\big)}{\left[\left(\frac{1}{M}\sum_{i=1}^{M}\left(A_i - \langle A\rangle\right)^2\right)\left(\frac{1}{M}\sum_{i=1}^{M}\left(B_i - \langle B\rangle\right)^2\right)\right]^{1/2}} \tag{4.11}$$

where the brackets denote the average of the A_i and B_i values over the data set.

By evaluating the variables at two different time points, the correlation coefficient is extended to a time correlation function (TCF) $c_{AB}(t)$. The non-normalized correlation function is defined as

$$C(\tau)_{AB} = \big\langle A(\tau)B(0)\big\rangle \tag{4.12}$$

where the brackets denote the expectation value in an equilibrium ensemble. If the quantities A and B are different the TCF is referred to as a cross-correlation. If they are the same, i.e. $B = A$, the TCF $C_{AA}(\tau)$ is referred to as an autocorrelation function. The autocorrelation function can be viewed as a measure of the "memory" of the system, i.e. to what extent the system has a memory of its previous values. In other words, $C_{AA}(\tau)$ shows how long time it takes for A to loose its memory.

4.3.1 Time symmetry

Equation 3.4 showed that Hamilton's equations conserve energy and that the Hamiltonian $\mathcal{H}$ is a constant of time. This fact will also result in the equilibrium ensemble being invariant in time, which has implications for the time correlation functions [65]. Perhaps the most important property is that time correlations are independent of the reference time t such that

$$\big\langle A(t+\tau)B(t)\big\rangle = \big\langle A(\tau)B(0)\big\rangle \tag{4.13}$$

This leads to interesting symmetry properties. For instance, the autocorrelation function is symmetric in time

$$\big\langle A(0)A(-\tau)\big\rangle = \big\langle A(0)A(\tau)\big\rangle \tag{4.14}$$

Thus, the time correlation function can be obtained from an average of a trajectory in the equilibrium ensemble. If our MD simulation is sufficiently long, the statistical correlation function (Eq 4.12) can be estimated from the trajectory by averaging over reference times t

$$C_{AB}(\tau) = \frac{1}{\Delta t - \tau}\int_0^{\Delta t-\tau} dt\, A(t+\tau)B(t) \tag{4.15}$$

where Δt is the length of the trajectory, and τ is the time period in which correlations are followed. The time available for averaging is $\Delta t - \tau$, which means our trajectory has to be long enough relative to τ in order to obtain sufficient statistical accuracy. Given that we sample the MD simulation at discrete time intervals δt, so that τ is a multiple m of δt, the TCF is estimated as

$$C_{AB}(\tau) \approx \frac{1}{M-m}\sum_{n=1}^{M-m} A(t_{n+m})B(t_n) \tag{4.16}$$

where M is the total number of time steps in the trajectory.

4.3.2 Correlation times

In MD simulations of systems at equilibrium we are often computing quantities that fluctuate in time. Although the data will look noisy, the time dependence is stationary. The time scale of the fluctuations are often characterized by computing the correlation time (or relaxation time) τ_{corr} which can be estimated from experiments. The correlation time is characteristic for the quantity investigated, and reports on how long time it takes for the quantities $A(\tau)$ and $B(0)$ to loose their correlation. After a time $\tau \gg \tau_{corr}$, A and B will be statistically independent so that

$$C_{AB}(\tau \gg \tau_{corr}) = \langle A \rangle \langle B \rangle \tag{4.17}$$

As for the correlation coefficient (Eq 4.10), correlation functions are typically normalized. For instance, the autocorrelation function $C_{AA}(\tau)$ is normalized as

$$\begin{aligned} c_{AA}(\tau) &= \frac{\langle \delta A(t) \delta A(0) \rangle}{\langle \delta A(0)^2 \rangle} \\ \delta A(t) &= A(t) - \langle A \rangle \end{aligned} \tag{4.18}$$

where the brackets denote the average. Thus, $c_{AA}(0) = 1$ and $c_{AA}(\tau \to \infty) = 0$. If the decay time is exponential, i.e. $c_{AA}(\tau) = exp(-\tau/\tau_{corr})$, we obtain the correlation time as

$$\tau_{corr} = \int_0^\infty d\tau \, c_{AA}(\tau) \tag{4.19}$$

For more complicated correlation functions, Eq 4.19 may still be used as a measure of the relaxation time, but then it is more pertinent to call it an integral correlation time.

4.3.3 The spectrum

Apart from reporting on the correlation time, auto correlation functions may contain a wealth of information about the underlying dynamics in the system. To extract this information in a way that is more easy to interpret, the TCF is decomposed into a frequency spectrum that can be compared to experiment. The spectrum is obtained by taking the Fourier transform, defined as

$$f(\omega) = \int_{-\infty}^\infty d\tau \, e^{-i\omega\tau} c_{AA}(\tau) \tag{4.20}$$

Evaluating the correlation at negative times is possible due to the time symmetry property in Eq 4.14, so that

$$c(-\tau)_{AA} = c(\tau)_{AA} \tag{4.21}$$

By the same token, we can obtain the spectrum as the cosine transform along the positive time axis only

$$f(\omega) = 2\int_0^\infty d\tau \cos(\omega\tau) c_{AA}(\tau) \tag{4.22}$$

and the spectrum $f(\omega)$ will in turn be symmetric in the frequency ω

$$f(-\omega) = f(\omega) \tag{4.23}$$

We can recover $c_{AA}(\tau)$ using the Fourier inversion theorem

$$c_{AA}(\tau) = \frac{1}{2\pi}\int_0^\infty d\omega\, e^{-i\omega\tau} f(\omega) \tag{4.24}$$

Examples

The Fourier transform of the velocity autocorrelation for hydrogen atoms in liquid water can be used to assign its vibrational spectra obtained from IR measurements. For a N-atom system, the velocity autocorrelation function is defined as

$$c_{vv}(\tau) = \frac{1}{N}\sum_{i=1}^{N} \frac{\langle \mathbf{v}_i(t)\cdot\mathbf{v}_i(0)\rangle}{\langle \mathbf{v}_i(0)\cdot\mathbf{v}_i(0)\rangle} \tag{4.25}$$

where $\mathbf{v}_i(t)$ is the velocity of atom i at time t, and the brackets denote the statistical average. From a simulation, the average is thus taken both over time origins and over all the atoms; $c_{vv}(\tau)$ is thus a single-particle TCF. The velocity autocorrelation contains information about the vibrational motion in the system at short times, whereas its long time behavior is related to the diffusive motion.

If we consider the angular velocity of a molecule instead, we get the angular velocity correlation function (TCF)

$$c_{\omega\omega}(\tau) = \frac{1}{N}\sum_{i=1}^{N} \frac{\langle \omega_i(t)\cdot\omega_i(0)\rangle}{\langle \omega_i(0)\cdot\omega_i(0)\rangle} \tag{4.26}$$

which is the angular velocity analog of the linear velocity correlation function in Eq 4.25. Thus, $c_{\omega\omega}(\tau)$ indicates the degree to which the angular velocity of a molecule at time t is related to its angular velocity at time 0.

Molecular rotations can be probed by several experimental techniques to yield the rotational correlation time τ_R^L of different ranks L. The rank depends on the type of

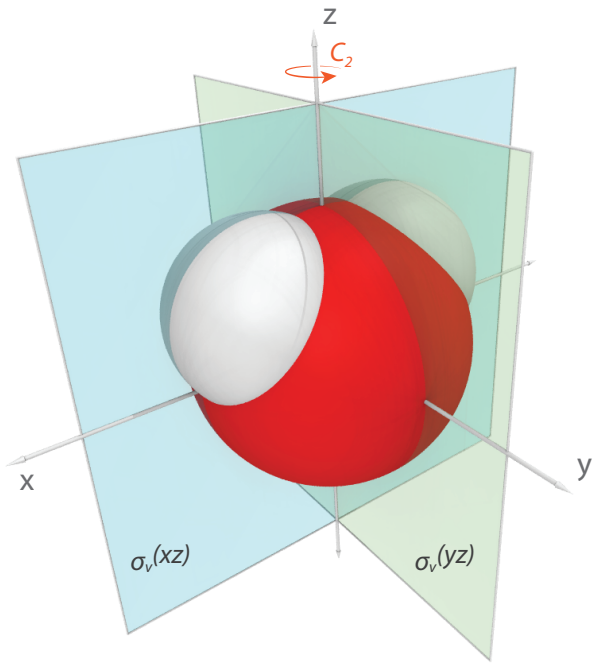


Figure 4.4: The water molecule in a body fixed coordinate system used in the calculation of rotational TCF. The two symmetry planes, $\sigma_v(xz)$ and $\sigma_v(yz)$, defines the principal axes (C_2, z) which is along the dipole direction.

interaction (tensor) that is involved in the relaxation process of the technique. For instance, rotational TCFs associated with NMR relaxation measurements are computed via the rank 1 and rank 2 Legendre polynomials

$$c_{uu}^{(1)}(\tau) = \frac{\langle \mathbf{u}(\tau)\cdot\mathbf{u}(0)\rangle}{\langle \mathbf{u}(0)\cdot\mathbf{u}(0)\rangle} = \langle \mathbf{u}(\tau)\cdot\mathbf{u}(0)\rangle = \langle \cos\theta(\tau)\rangle \tag{4.27}$$

$$c_{uu}^{(2)}(\tau) = \left\langle \frac{3}{2}\left[\mathbf{u}(\tau)\cdot\mathbf{u}(0)\right]^2 - \frac{1}{2}\right\rangle \tag{4.28}$$

of the reorientation for a body fixed axes $\mathbf{u}$ that has rotated an angle $\theta(\tau)$ in the time interval τ. Figure 4.4 shows the body-fixed coordinate system for the water molecule used in the calculation of rotational TCFs in paper [VI]. The Fourier transform of $c_{uu}^{(2)}$ is probed by ^{2}H MRD (section 2.3.2).

Chapter 5

Summary of thesis work

In this chapter we will overview our main findings in the different papers. For the initiated reader, the terse abstracts for each paper will serve equally as well.

5.1 Paper I&II

Many proteins rely on brief visits to highly excited conformations in order to perform their function, such as conformationally gated ligand binding and release or solvent access to internal cavities. In magnetic relaxation dispersion experiments, water molecules buried in such cavities can be used to probe the underlaying transient protein motions that govern their exchange. For the bovine pancreatic trypsin inhibitor (BPTI), the four crystallographically identified internal water molecules in the interloop region [18] exchange on a timescale ranging from tens of nanoseconds to hundreds of microseconds [20, 21]. However, it is experimentally challenging to characterize these transient states, and little is therefore known about the exchange mechanism. In order to extract exchange kinetics, three assumptions are used in the exchange-mediated orientational randomization (EMOR) model [107, 108] used to interpret MRD data: 1) water exchange is instantaneous; 2) once a water leaves the protein it has the same probability to return as any other water; and 3) the probability that a water has not exchanged - its survival probability - can be described by the stochastic Poisson process. To validate these approximation requires the full atomic detail provided by MD simulations. But the long simulation time required to sample enough statistics on water exchange was hopelessly beyond reach, until 2010 when the first all-atom millisecond long MD simulation was published by D.E. Shaw *et al.* [109] for BPTI solvated in water. The previously longest MD simulation published just 3 years earlier was only 10 microseconds long, and this milestone simulation was made possible thanks to the development of Anton, a super-computer optimized for running MD simulations [110].

In paper [I] we used the 1 ms MD simulation to test the validity of the MRD model, characterize the exchange mechanism and benchmark the forcefield. In order to do so we had to address many non-trivial computational issues in the process. First we had to identify persistent hydration sites occupied by long-lived water molecules in a protein structure that undergoes large conformational changes during the simulation. This will result in dynamical disorder [111] where the water exchange rate depends on the conformational state of the protein. The simulation does not reproduce the equilibrium conformations of the C14-C38 disulfide bond [112] determined from NMR measurements [113, 114]; the dominant M1 state has a population of 95 % whereas it is 25 % in the simulation. We therefore defined conformational states based on a cluster analysis of the rotamer states sampled by the disulfide bond, and analyzed sub-trajectories for each conformational state. Since, the exchange event was negligibly short compared to the time between exchanges, we could describe the exchange process as a stationary point-process [115, 116] and characterize the dynamical disorder. The information in such a process is completely contained in the residence correlation function (RCF), $Q_R(\tau)$, which is the probability to observe a water molecule occupying a hydration site longer than a time τ. This general framework, presented in paper [II], can be used to extract the essential dynamical characteristics for any reoccurring transient event observed in an MD simulation.

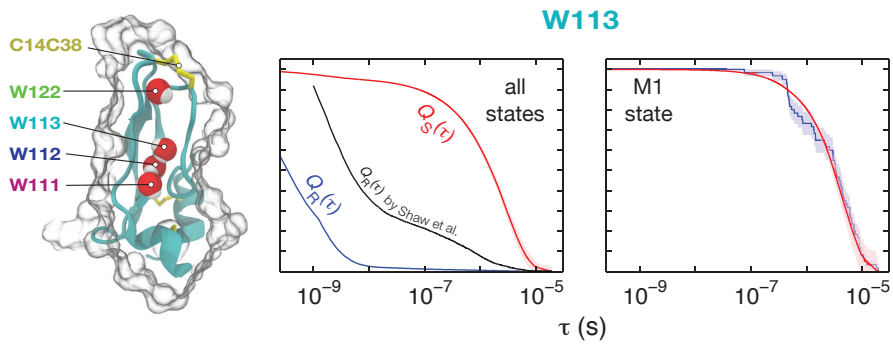


Figure 5.1: Internal hydration sites in BPTI and the disulfide bonds (yellow). Residence $Q_R(\tau)$ and survival $Q_S(\tau)$ correlation functions for W113 in all states and the experimentally dominant M1 state. If residence times are independent and exponentially distributed, we have a Poisson process for which $Q_R(\tau) = Q_S(\tau) = exp(-\tau/\tau_S)$.

Because experimental measurements are not synchronized with the exchange event, the RCF cannot be used to compare simulation with results from MRD. Instead, the proper correlation function is the survival correlation function (SCF), $Q_S(\tau)$. This is the probability that a water molecule residing in a hydration site at a randomly chosen time point does not leave the site in the subsequent time interval τ. Measurements on immobilized proteins by MRD yields essentially the Fourier transform of the SCF, and the MRD-derived survival times (ST), τ_S is the integral of the SCF. In paper [II] we present an algorithm for computing the SCF that is several order of magnitudes

faster than the original algorithm by Impey *et al* [117], and we give a detailed error analysis of both the statistical error and the binning error for the SCF.

For SCFs computed from the M1 state, the simulation reproduces the experimental STs remarkably well for the four internal hydration sites. For the sites W111-W113, the activation energy discrepancy is 1.5 k_BT, and for the W122 site, adjacent to the C14-C38 disulphide bond, the discrepancy is 3 k_BT. However, the forcefield underestimates the C_2 flip barrier in the sites by as much as 6 k_BT. The essentially exponential SCFs validates the Poisson approximation in the EMOR model. The fast exchange events, absence of site correlations, and the very low return probability for recently exchanged water molecule also validates the EMOR model.

The exchange mechanism for water exchange of the three deepest sites (W111, W112 and W122) revealed, in all cases, a short-lived ($\leq$ 5 ns) transition state where the buried site is accessed via a single-file water chain migrating through a transient tunnel or pore. We call this the aqueduct mechanism which was observed to have two variants. The water chain either involved the adjacent sites W111-W113, or one or more new tunnels or pores. The latter variant was dominated for W122 in the M2 state of the C14-C38 disulfide bond and resulted in a higher water content in the interloop region.

5.2 Paper III

The transient solvent exposure of the protein interior observed in Paper [I] spurred us to see if the 1 ms simulation of BPTI [109] could cast light on another mechanism that has remained elusive for more than 60 years. Like internal water molecules, amide hydrogen exchange (HX) requires transient solvent access to the interior sites, but here the site must also be accessed by a catalytic ion (usually OH^-). Under native conditions, few amides exchange by global unfolding so exchange must involve subglobal structural changes for the majority of amides. For the past decades there has been much debate about the nature of these subglobal fluctuations and their frequency, duration, amplitude and cooperativity. Measuring HX rates is usually done by NMR in the EX2 limit of the standard HX model [27] where the measured HX rate k_{HX} reports on the protection factor κ of the amide

$$\kappa = \frac{k_{int}}{k_{HX}} = \frac{f_C}{f_O} \tag{5.1}$$

where f_O and f_C are the fractional populations of the open (O) exchange competent state and the closed (C) exchange incompetent state. It is implicitly assumed that k_{int} is the same as the HX rate from structureless peptides. Protection factors are often determined to gain information about protein structure and flexibility. But because the exchange mechanism is unclear, the interpretation of these experiments are of qualitative value at best.

We therefore set out to see if we could use the ultra-long simulation to reproduce experimental PFs and thereby gain insight into the HX mechanism. Because no bonds are broken in an MD simulation, the O state cannot be identified directly. Instead, we had to postulate a structural criterion that must be satisfied for HX to occur. Our O-state criterion requires an amide hydrogen N-H to have at least two water oxygens within $R_{HO} = 2.6$ Å. This criterion will almost always guarantee that any intramolecular hydrogen bonds are broken, which is often used to correlate experimental PFs [118–120]. Out of 53 amides, 41 accessed the O state in the subtrajectory where the C14C38 disulfide bond is in the experimentally dominant M1 configuration. As might be expected, weakly protected amides required a smaller structural adjustment to become exchange competent. However, the rigidity as measured by crystallographic B-factors did not show any such correlation, in contrast to a previous suggestion [121].

We determined PFs using the O-state definition and compared them against reliable experimental PFs on BPTI [122–124] at the simulation temperature 300 K. For the 30 PFs available for comparison, the computed PFs agreed well with experiment (see Fig 5.2). Except for three amides, the simulation-based O/C free energy difference agrees to better than 2.5 k_BT. Expressed as a signed average, $\beta\langle\Delta G_{sim} - \Delta G_{exp}\rangle = 0.44\ k_BT$. Our O-state definition is also supported by the observation that none of the eight amides in the beta-sheet core access the O-state. These amides exchange by global unfolding and should not be expected to access the O-state in the analysed native-state trajectory.

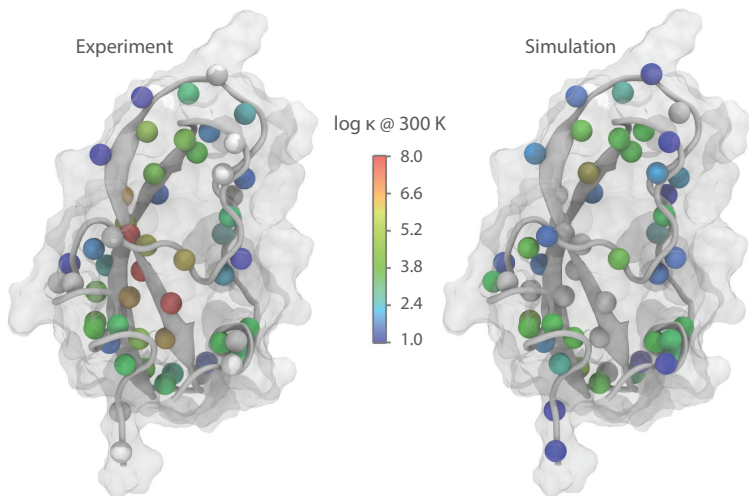


Figure 5.2: Backbone amide hydrogen atoms in BPTI, color-coded according to the HX protection factor (κ). (Left) Experimentally determined protection factors from reference [[122–124]] for 41 backbone amide hydrogens (temperature corrected). (Right) Protection factors determined from MD simulation where 41 backbone amide hydrogens accessed the postulated open state.

Most of the visits to the O-state were only a single frame, so the mean residence life time (MRT) of the O-state must be shorter given that only a fraction are recorded in the 0.25 ns sampling resolution. We could correct for this systematic binning error by modelling the $C \leftrightarrows O$ fluctuations as an alternating Poisson process. The approach was validated by the MRT distribution for the C-state which is close to exponential. While the MRTs for the C-state ranged from 1 ns to 2 μs, the MRTs for the O-state varied by only a factor of three, with mean and standard deviation 81 ± 18 ps. The O-state is thus highly unstable, so the large variation in PF must almost entirely be due to the variation of MRTs in the C-state.

Two competing models for the HX mechanism have been proposed where the amide either gets exposed to solvent by "solvent penetration" or by "local fluctuations" [30, 32, 34]. We find that these imprecisely defined models are not mutually exclusive; a few amides gain access to solvent by the aqueduct mechanism seen for internal water exchange in Paper [I], but most amides gain access to solvent by more local structural distortions. However, because no protons are actually exchanged in the simulation, the exchange-competent state cannot unambiguously be established. Nevertheless, we believe that the proton transfer occurs via a Grotthuss-type (proton jumping) structural diffusion in which the amide has to be "pre-solvated" by two water molecules before the catalyst can approach the amide through a water wire.

5.3 Paper IV

The protein hydration shell is a well established concept but with different operational definitions. In its simplest interpretation, the protein hydration shell consists of the water molecules in contact with the protein surface, but no consensus exists on how to identify these water molecules from molecular simulations. We were motivated to examine methods to construct hydration shells as they provide a simple and robust metric in the analysis of the spatial range of the protein-induced water perturbation; the hydration shell index represents the number of water monolayers that may separate a perturbed water molecule from the protein surface.

Two different types of methods, based on spatial or topological proximity between protein and water oxygen atoms, have been used to define the hydration shells. We compared these methods on how well they produce shells that are one water molecule thick and that fully covers the protein surface or the inner hydration shell. Our analysis is based on molecular dynamics simulations of four globular proteins in dilute aqueous solution, with three different water models. For all systems, the best method to construct the first shell is a 5 Å water-carbon cutoff (CC) which almost completely covers the protein surface, whereas a popular method, using a ~ 3.5 Å cutoff to any protein heavy atom [39, 40, 125, 126], only covers half of the surface. The topological method based on Voronoi-tessellation is being used more and more to define the first and higher order shells [42–44, 49–51, 127, 128]. We find that this method produce

too thick shells (4 Å) owing to the fact that the majority of topological neighbors are 3.5-6 Å away from the water molecule. Using a 4 Å water-water cutoff (WC) we obtained higher order shells with 95 % coverage and a shell thickness of 2.8-2.85 Å that closely matches the first maximum in the bulk water oxygen-oxygen radial distribution function (RDF) [129–131]. Figure 5.3 shows nine hydration shells constructed using the CC/WC and the VN definitions.

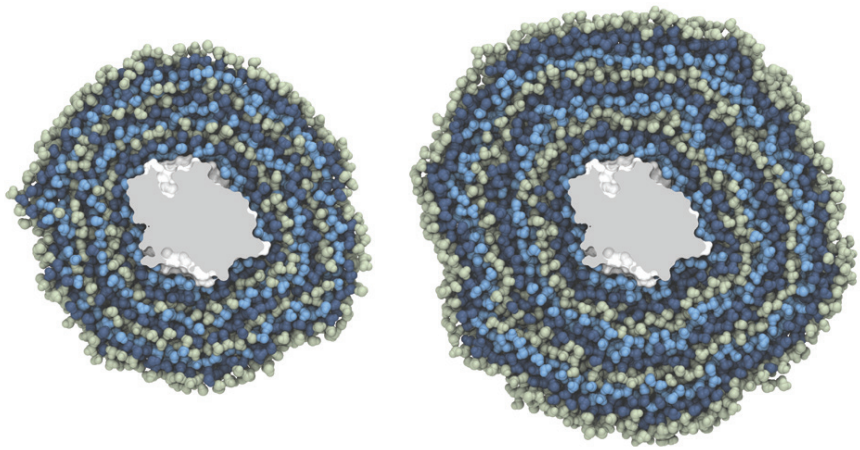


Figure 5.3: The first nine hydration shells of Ubiquitin, constructed from the first frame of the production simulation using the WC/CC (left) and VN (right) shell definitions.

Having defined monolayer hydration shells we then analyzed different, essentially, geometric properties for the protein-water interface. At the molecular scale, geometric volumes cannot be unambiguously defined; any estimate of the hydration water density relies on a suitable choice of a protein-water dividing surface. If the dividing surface is put at equal distance between any atom pairs, i.e. the ordinary Voronoi cell, we obtained a density increase of 1 % compared to bulk. But a more physically realistic volume decomposition is to put the dividing surface at equal distance, at any point, from the vdW surfaces of the neighboring atoms. This decomposition is called additively weighted Voronoi tessellation [105, 132] and yields a 6 % density increase compared to bulk. The experimentally measured first-shell density is estimated to 10 % from scattering experiments [47]. Because these experiments cannot disentangle the hydration shell thickness and the density, we argue that the difference to our results can be explained by 1) the too small estimate on the amount of water in the shell, taken from studies of non-freezing water [133, 134], which typically corresponds to only half of the first shell; and 2) the assumed 22 % effective excess density of this shell based on measurements of unit cell volumes of hydrated and anhydrous crystals of (mostly) inorganic salts [133–135].

Including the higher shells, the protein-induced relative density perturbation is short-ranged and highly invariant for all systems as shown in Fig 5.4b; it is reduced

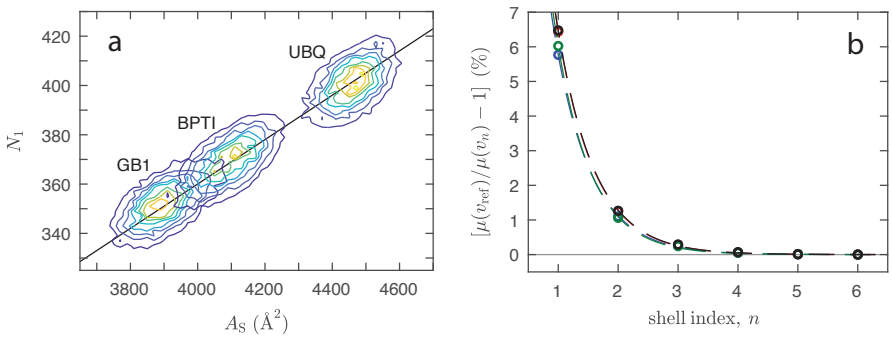


Figure 5.4: **[a]** Contours of the bivariate distribution $f(A_S, N_1)$ where A_S is the SASA and N_1 is the number of water molecules in the first hydration shell. The line corresponds to IWA= 11.11 Å^2. **[b]** Percent difference from bulk water of the mean water density in the n:th hydration shell for the four proteins studied. The dashed curves resulted from exponential fits.

five-fold in going from one shell to the next higher shell. This correspond to a decay "length" of 0.6 shells, the length at which the perturbation has decayed to a fraction of $1/e$ of its value in the first shell. Using an average shell thickness of 2.8 Å, this gives a physical decay length of 1.70 Å.

Several experimental techniques, such as small-angle scattering and magnetic relaxation dispersion (MRD), rely on an estimate of the number of water molecules in contact with the protein surface. If one knows the mean interfacial water-area (IWA), i.e. the average amount of the protein solvent-accessible surface area (SASA) occupied by a water molecule on the protein surface, the number of water molecules in the first shell is simply approximated as SASA/IWA. We determined the IWA to 11.1 Å^2 and conjecture that this value applies to most globular single-domain proteins (Fig 5.4a).

We also characterized the neighborhood of individual water molecules by computing local coordination numbers resolved on shells and subsets thereof. The coordination numbers were defined as the number of "ligand" (L) atoms within a cutoff distance from a water oxygen atom (W), with L = C, N, O and W atoms. The cutoff distances were defined from the first minimum in the W-L RDFs. The distribution of the polar coordination number (L = N+O+W) differs very little among the four proteins. Remarkably, this distribution differs very little from the bulk-water distribution. The mean polar coordination number is 4.26, a mere 1 % below the bulk water, and is the net result of a near cancellation of a 4 % excess (relative bulk) in the polar subset and a 5.4 % deficit in the nonpolar subset.

5.4 Paper V

Having studied molecular volumes in paper [IV], we also investigated the fluctuations of protein and hydration shell volumes, which is related to the compressibility for a system. The isothermal compressibility describes the stability of a protein against pressure denaturation [136–139] and its functional relevant [140, 141] mechanical properties [142] and volume fluctuations [143]. Because the work of cavity formation is (inversely) related to compressibility [144], the compressibility of the hydration shell is linked to hydrophobic effects at the protein-water interface [145].

For a closed system with volume $\langle V \rangle$, with fixed particle numbers $\{N\}$, the isothermal compressibility κ describes how the volume responds to an isothermal pressure change, defined as

$$\kappa \equiv \frac{1}{\langle V \rangle} \left(\frac{\partial \langle V \rangle}{\partial p} \right)_{\{N\},T} \tag{5.2}$$

The isothermal compressibility can be obtained from molecular dynamics (MD) simulations performed at different pressures, or it can be obtained from the isothermal volume fluctuation $\delta V = V - \langle V \rangle$ of a closed system in an NpT MD simulation,

$$\kappa = \frac{\langle (\delta V)^2 \rangle}{k_B T \langle V \rangle} \tag{5.3}$$

We were motivated to determine κ since experiments, which measures the partial protein compressibility, cannot disentangle the contributions to κ from the protein and the hydration shell. To compute κ for the hydration shells in our NpT simulations of small globular proteins, we had to solve many non-trivial problems, some that had been recognized before but never analyzed or discussed. First, as shown in paper [IV], the density is higher near the protein than further away, but the compressibility definitions above are not defined for this inhomogeneous solvent. Secondly, the hydration shell is an open system, containing a fluctuating number of water molecules, but Eqns 5.2 and 5.3 are only valid for a closed system; one obtains a pseudo compressibility $\tilde{\kappa}_n$ that will differ greatly from the true (intrinsic) compressibility $\hat{\kappa}_n$ for shell n, because it includes a (negative) contribution η_n from molecular exchange between regions of different density. Thus, $\tilde{\kappa}_n = \hat{\kappa}_n + \eta_n$. Third, because geometry and interactions cannot be rigorously disentangled at the molecular level [146] a protein-water dividing surface must be imposed and this choice affects the compressibility. Here we use the realistic dividing surface obtained by so-called additively weighted Voronoi tessellation [105, 132], but so far only primitive Voronoi tessellation has been used to compute compressibilities [127, 146–149]

Because compressibility is a collective property, reflecting coupled volume fluctuations among several water molecules, $\hat{\kappa}_n$ can be decomposed into self $\hat{\kappa}_n^{self}$ and

cross $\hat{\kappa}_n^{cross}$ contributions from waters in the hydration shells. The total cross contribution, $\hat{\kappa}_n^{cross}$, are made up of rapidly decreasing contributions from nearby shells (or the protein), with negligible contribution from shells beyond $n \pm 4$. They account for between 50 to 60 % of the total $\tilde{\kappa}_n$, and many authors have ignored these contributions [42, 125, 150]

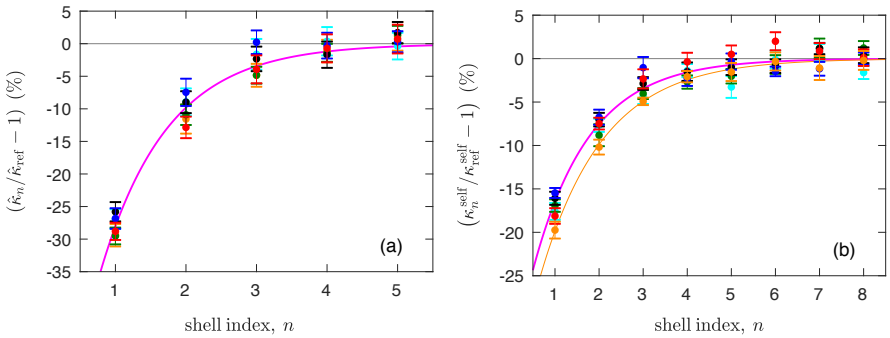


Figure 5.5: **[a]**Relative variation of the intrinsic compressibility $\tilde{\kappa}_n$ with shell index n. **[b]** Relative variation of the intrinsic self compressibility $\tilde{\kappa}_n^{self}$ with shell index n. The magenta curve is a joint exponential fit to all six data sets.

We find that the intrinsic compressibility $\hat{\kappa}$ is 25-30 % lower in the first hydration shell compared to bulk. This difference is larger than the static properties determined in paper [IV], but it is largely a trivial effect of the non-local character of $\hat{\kappa}$: the proximity to a more rigid material (the protein) suppresses volume fluctuations in the first shell, thereby reducing the self and correlated contributions. Figure 5.5a shows the intrinsic compressibilities for shells 1-5 for the studied systems, with an exponential fit yielding a decay length of 0.95 shells. Figure 5.5b shows the intrinsic self compressibility for shells 1-8, with an exponential fit yielding a decay length of 1.4 shells.

Finally, we show how to compute the experimentally measured partial protein compressibility $\bar{\kappa}_P$ from simulations. For our systems there is a negative hydration contribution to $\bar{\kappa}_P$, and it is of similar magnitude to the intrinsic partial protein compressibility, so that $\bar{\kappa}_P$ is close to zero. Although no experimental data is available for our small proteins, surface-to-volume scaling suggest that the negative hydration contribution should be more important for small proteins [151]. We therefore regard our results as being consistent with the available experimental database [152–154].

5.5 Paper VI

The dynamics at the protein-water interface is important in many biological processes, and water motions at or near the protein surface has been characterized by several experimental techniques. The most compelling experimental evidence comes from ^{17}O magnetic relaxation dispersion (MRD) experiments which selectively probes the

motions of single water molecules. Apart from providing information about internal water molecules (paper [I]), this technique provides the average rotational correlation time τ_R of the primary hydration shell, which is often expressed as the rotational perturbation factor (RPF) $\xi_R = \tau_R/\tau_R^{bulk}$. MRD measurements of a large number of native proteins has found that $\xi_R \approx$ 3-5 at room temperature [20–22, 59, 61, 155–167]. For the past decade, RPFs from molecular dynamics (MD) simulations have achieved semi-quantitative agreement with those measured by MRD [45, 168–171], but these studies suffer from several shortcomings to allow a rigorous comparison to MRD data. On the other hand, what has been demonstrated in these simulations is a strong dynamical heterogeneity within the hydration shell around proteins [51, 168–171], showing rotational correlation times spanning three orders of magnitude. Yet, the precise distribution of correlation times is not clear, nor is the molecular mechanisms that give rise to these wide distributions.

Using MD simulations, we therefore set out to do a comprehensive analysis of protein hydration dynamics in order to add missing pieces to this puzzle as well as performing the most rigorous comparison with MRD results to date. Our analysis is based on simulations of four globular proteins, with three different water models, in dilute aqueous solution at room temperature. As a spatial metric, we assigned water molecules to monolayer hydration shells (as established in paper [IV]) and subsets thereof. We compute three different rotational time correlation functions: two uniaxial TCFs of rank 1 (U1) and 2 (U2) describing the rotation of a water-molecule fixed vector and one biaxial TCF of rank 2 (B2) describing the rotation of a water-molecule fixed tensor. The B2 TCF must be computed in order to compare with ^{17}O MRD results (which is rarely done [45, 166]). Because MRD essentially gives the integral rotational correlation time (IRCT) (at zero frequency), we computed the TCFs up to 1 ns which is a much wider range of delay time than what is customary. In most previous MD studies, the IRCT has been extracted by fitting the TCF to an exponential at short times (typically less than 10 ps [169, 170], which will lead to an underestimation of the RPF (typically by a factor of 2 as shown in Fig 5.6) as well as missing the information about confined water molecules.

We determined RPFs for polar and nonpolar subsets of the first hydration shell since water dynamics has been suggested to depend strongly on site polarity. Water molecules within polar subsets were subdivided if the site involved charged or neutral protein atoms. RPFs increased in the order nonpolar < positive < neutral < negative, ranging from $\sim$ 2 (nonpolar) to 7-11 (negative). The slowest dynamics at negatively charged sites have been found before [40, 169], but some authors have claimed that rotation is slowest at positively charged sites [172] or even at nonpolar sites [173] - in stark contrast to our results. However, the correlation on site polarity is merely a correlation and may instead depend on the surface topography (which in turn may be correlated to polarity). Slower water dynamics have been noted in several MD simulations, with water in concave sites, pockets or clefts being more perturbed than

exposed, convex sites [39, 39, 40, 169–171, 174–176]. But no quantitative correlation has been established. Guided by these observations, we assigned each water molecule a confinement index z_C, defined as the number of carbon atoms within 5.0 Å of the water oxygen atom.

This simple definition turned out to capture the essence of water confinement and reveal several key insights about water perturbation: With increasing confinement index z_C, the RPF $\xi(z_C)$ increases exponentially for $z_C < 10$, whereas the number of water molecules with confinement index z_C, $N_1(z_C)$, decreases exponentially. For the most confined sites, the RPF increases more strongly and with more protein specificity. Among the three TCFs, the B2 TCF is the most sensitive probe for water confinement; for every additional carbon atom the (B2) RPF increase with 27 %.

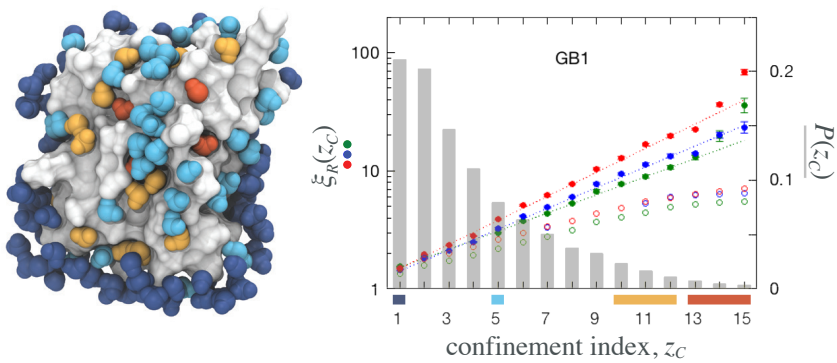


Figure 5.6: **[Left]** Water molecules in the first shell of GB1, color-coded according to their confinement index (only a fraction of the $z_C = 1$ subset is shown). **[Right]** The rotational perturbation factor $\xi(z_C)$ based on IRCTs (filled circles) and $\xi(z_C)$ based on exponential fitting to the TCF in 2-10 ps interval (open symbols) versus confinement index z_C. Dotted lines resulted from exponential fits for $z_C \leq 10$. TCF type: U1 (green) and U2 (blue) and B2 (red). The bars shows the fraction $P(z_C)$ of first shell water molecules with a given z_C.

The confinement index also correlates with the number of neighbouring polar atoms. Although the number of neighbouring water molecules decrease with increasing z_C, the number of polar protein atoms increase with z_C. Thus, our confinement index measures the extent of the protein-water contact regardless of whether it involves polar or nonpolar protein atoms.

Our discovery of a universal and exponential dependence of the RPF on confinement index indicates that water molecules in the hydration shell rotates by different mechanism on a spectrum of two extremes. At the lower end, the water molecules with $z_C = 1$ at nonpolar (non) sites coordinate almost the same number of water molecules as in bulk and therefore rotate by a bulk-like mechanism, with a cooperative motion of several water molecules. This is supported by the TCF rank dependence, $\tau_R^{non}(U1)/\tau_R^{non}(U2) = 2.55$, which is the same as in bulk. For the most confined water molecules at the high end of the spectrum, orientation is restricted and rotation cannot occur by concerted motions as in the bulk. Rotation therefore requires an

exchange event, whereby another water molecule enters the confined site and the original one now can rotate with little or no retardation: this is the exchange-mediated orientational randomization (EMOR) mechanism. For water molecules rotating by the EMOR mechanism, the asymptotic decay time should be the same for all three TCFs, on the time scale of the mean survival time. This is indeed what we see for the most confined water molecules, as shown in Fig 5.7a.

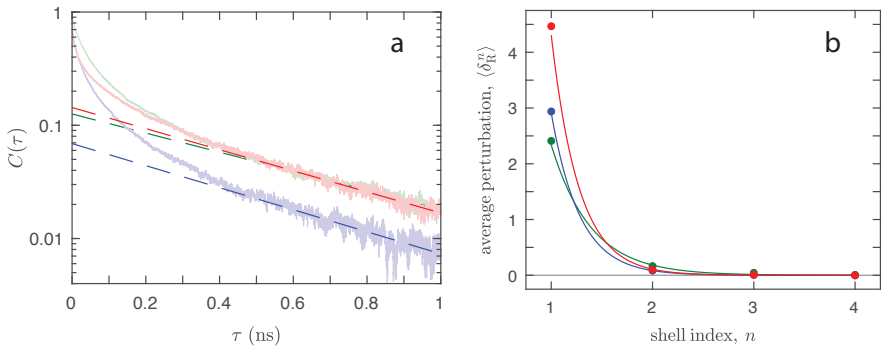


Figure 5.7: **[a]** The three TCFs for the most confined ($z_C = 15$) water molecules in the first hydration shell of Ubiquitin. Exponential fits (dashed line) in the interval 0.5-1.0 ns. **[b]** Excess rotational perturbation factor, $\delta_R^n = \xi_R^n - 1$, derived from the three TCFs for water molecules in the n:th hydration shell, and averaged over the four proteins. TCF type: U1 (green) and U2 (blue) and B2 (red).

By computing the B2 TCF for all water molecules in the systems we could benchmark the simulation force-field with model-free MRD results. Like previous studies we obtain a semi-quantitative agreement between simulation and MRD, supporting the simulation data and the conclusions drawn from it. However, our rigorous analysis shows that the simulation overestimates the MRD-derived (generalized excess) RPF by 25-30 % for three of the four proteins. The same discrepancy is seen for the other water models, and we therefore attribute the difference between simulation and experiment to the protein force-field; because the RPF is heavily influenced by a small number of highly confined sites, it depends sensitively on the protein water-interactions which might not be described correctly by the protein force-field.

Finally, we address the contentious issue of the spatial range of the protein-induced perturbation on water dynamics by computing RPFs for each monolayer-thick hydration shell. The perturbation is short-ranged as shown in Fig 5.7b; on going from one shell to the next higher one, the perturbation is reduced by an order of magnitude. This corresponds to an exponential decay-length of 0.4 or 0.3 shells for the uniaxial and biaxial (B2) TCFs respectively. Translated to a decay length, with an average shell-thickness of 2.8 Å (paper [IV]), this yields 1.1 and 0.8 (B2) Å.

However, the only long range perturbation that we observe is a weak alignment of the water molecules by the electric field of the protein, which decays as R^{-3} for the electroneutral proteins studied here. Such a weak alignment hardly affects the local water dynamics, but it introduces a persistent orientational correlation. Complete

randomization of a water molecule's orientation then requires diffusion around the protein, which is manifested in the TCFs as two distinct time-scales: picosecond water rotation brings the TCF down to a small plateau value, whereupon nanosecond water diffusion completes the decay towards zero. The weak long-time tail associated with this isotropic averaging of the local electric field could be observed for the U1 TCF up to the sixth shell, but it has already decayed to 1 % of its initial value in the second shell. The effect of the second and higher shells contribution to the total perturbation measured by ^{17} O MRD is only 3 %, verifying that the (generalized excess) RPF can, to a very good approximation, be assigned to water molecules in the first shell.

References

[1] Feynman, R. P.; Leighton, R. B.; Sands, M. *The Feynman Lectures on Physics, boxed set: The New Millennium Edition*; Basic Books, 2011.

[2] Carroll, M. *Drifting on Alien Winds: Exploring the Skies and Weather of Other Worlds*; Springer, 2011; p 125.

[3] Dalrymple, G. B. *Geol. Soc. London, Spec. Publ.* **2001**, *190*, 205–221.

[4] Dodd, M. S.; Papineau, D.; Grenne, T.; Slack, J. F.; Rittner, M.; Pirajno, F.; O'Neil, J.; Little, C. T. *Nature* **2017**, *543*, 60–64.

[5] Halle, B. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **2004**, *359*, 1207–1223; discussion 1223–1224, 1323–1328.

[6] Kauzmann, W. *Advances in Protein Chemistry* **1959**, *14*, 1–63.

[7] Dill, K. A. *Biochemistry* **1990**, *29*, 7133–7155.

[8] Baldwin, R. L. *Proceedings of the National Academy of Sciences* **2014**, *111*, 13052–13056.

[9] Richards, F. M. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151–176.

[10] Williams, M. A.; Goodfellow, J. M.; Thornton, J. M. *Protein Sci.* **1994**, *3*, 1224–1235.

[11] Park, S.; Saven, J. G. *Proteins Struct. Funct. Genet.* **2005**, *60*, 450–463.

[12] Meyer, E. *Protein Sci.* **1992**, *1*, 1543–1562.

[13] Petrone, P. M.; Garcia, A. E. *J. Mol. Biol.* **2004**, *338*, 419–435.

[14] Baker, E. N. In *Protein-Solvent Interactions*; Gregory, R., Ed.; CRC Press, 1995; Chapter 2, pp 143–189.

[15] Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. *Biochemistry* **1994**, *33*, 4721–4729.

[16] Liou, Y.; Tocilj, A.; Davies, P.; Jia, Z. *Nature* **2000**, *406*, 322–324.

[17] Vijay-Kumar, S.; Bugg, C.; Cook, W. *J.Mol.Biol.* **1987**, *194*, 531–544.

[18] Wlodawer, A.; Walter, J.; Huber, R.; Sjolin, L. *J.Mol.Biol.* **1984**, *180*, 301–329.

[19] Carugo, O. *Amino Acids* **2016**, *48*, 193–202.

[20] Denisov, V. P.; Peters, J.; Hörlein, H. D.; Halle, B. *Nat. Struct. Biol.* **1996**, *3*, 505–509.

[21] Persson, E.; Halle, B. *J. Am. Chem. Soc.* **2008**, *130*, 1774–1787.

[22] Kaieda, S.; Halle, B. *J. Phys. Chem. B* **2013**, *117*, 14676–14687.

[23] Hvidt, A.; Linderstrøm-Lang, K. *Biochimica et Biophysica Acta* **1954**, *14*, 574–575.

[24] Dempsey, C. E. *Prog. Nucl. Magn. Reson. Spectrosc.* **2001**, *39*, 135–170.

[25] Bandura, A. V.; Lvov, S. N. *J. Phys. Chem. Ref. Data* **2006**, *35*, 15–30.

[26] Bai, Y.; Milne, J. S.; Mayne, L.; Englander, S. W. *Proteins Struct. Funct. Bioinforma.* **1993**, *17*, 75–86.

[27] Hvidt, A.; Nielsen, S. O. *Adv. Protein Chem.* **1966**, *21*, 287–386.

[28] Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517–520.

[29] Arrington, C. B.; Robertson, A. D. *Methods Enzymol.* **2000**, *323*, 104–124.

[30] Woodward, C.; Simon, I.; Tuchsen, E. *Mol. Cell. Biochem.* **1982**, *48*, 135–160.

[31] Nakanishi, M.; Tsuboi, M.; Ikegami, A. *J. Mol. Biol.* **1973**, *75*, 673–682.

[32] Richards, F. M. *Carlsberg Res. Commun.* **1979**, *44*, 47–63.

[33] Englander, S. W. *Ann. N. Y. Acad. Sci.* **1975**, *244*, 10–27.

[34] Kossiakoff, A. A. *Nature* **1982**, *296*, 713–721.

[35] Miller, D. W.; Dill, K. A. *Protein Sci.* **1995**, *4*, 1860–1873.

[36] Rupley, J. A.; Careri, G. In *Adv. Protein Chem.*; C.B. Anfinsen John T. Edsall and David S. Eisenberg, F. M. R., Ed.; Academic Press, 1991; Vol. 41; pp 37–172.

[37] Kuntz, I. D.; Kauzmann, W. *Adv. Protein Chem.* **1974**, *28*, 239–345.

[38] Laage, D.; Elsaesser, T.; Hynes, J. T. *Chem. Rev.* **2017**, *117*, 10694–10725.

[39] Henchman, R. H.; McCammon, J. A. *Protein Sci* **2002**, *11*, 2080–2090.

[40] Schröder, C.; Rudas, T.; Boresch, S.; Steinhauser, O. *J. Chem. Phys.* **2006**, *124*, 234907–1–234907–8.

[41] Beck, D. A. C.; Alonso, D. O. V.; Daggett, V. *Biophys. Chem.* **2003**, *100*, 221–237.

[42] Gerstein, M.; Tsai, J.; Levitt, M. *J. Mol. Biol.* **1995**, *249*, 955–966.

[43] Neumayr, G.; Rudas, T.; Steinhauser, O. *J. Chem. Phys.* **2010**, *133*, 84108.

[44] Voloshin, V. P.; Medvedev, N. N.; Smolin, N.; Geiger, A.; Winter, R. *J. Phys. Chem. A* **2015**, *119*, 1881–1890.

[45] Braun, D.; Schmollngruber, M.; Steinhauser, O. *Phys. Chem. Chem. Phys.* **2016**, *18*, 24620–24630.

[46] Gerstein, M.; Chothia, C. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 10167–10172.

[47] Svergun, D. I.; Richard, S.; Koch, M. H. J.; Sayers, Z.; Kuprin, S.; Zaccai, G. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 2267–2272.

[48] Merzel, F.; Smith, J. C. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 5378–5383.

[49] Voloshin, V. P.; Medvedev, N. N.; Andrews, M. N.; Burri, R. R.; Winter, R.; Geiger, A. *J. Phys. Chem. B* **2011**, *115*, 14217–14228.

[50] Voloshin, V. P.; Kim, A. V.; Medvedev, N. N.; Winter, R.; Geiger, A. *Biophys. Chem.* **2014**, *192*, 1–9.

[51] Abseher, R.; Schreiber, H.; Steinhauser, O. *Proteins Struct. Funct. Genet.* **1996**, *25*, 366–378.

[52] Ebbinghaus, S.; Kim, S. J.; Heyden, M.; Yu, X.; Heugen, U.; Gruebele, M.; Leitner, D. M.; Havenith, M. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 20749–20752.

[53] Meister, K.; Ebbinghaus, S.; Xu, Y.; Duman, J. G.; DeVries, A.; Gruebele, M.; Leitner, D. M.; Havenith, M. *Proc. Natl. Acad. Sci.* **2013**, *110*, 1617–1622.

[54] Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379–IN4.

[55] Shrake, A.; Rupley, J. A. *J. Mol. Biol.* **1973**, *79*, 351–371.

[56] Sanner, M. F.; Olson, A. J.; Spehner, J. C. *Biopolymers* **1996**, *38*, 305–320.

[57] Fraczkiewicz, R.; Braun, W. *J. Comput. Chem.* **1998**, *19*, 319–333.

[58] Halle, B. In *Hydration processes in biology*; M.-C.), B.-F., Ed.; IOS Press, 1999; Chapter 10, p 233–249.

[59] Modig, K.; Liepinsh, E.; Otting, G.; Halle, B. *J. Am. Chem. Soc.* **2004**, *126*, 102–114.

[60] Qvist, J.; Persson, E.; Mattea, C.; Halle, B. *Faraday Discuss.* **2009**, *141*, 131–144.

[61] Mattea, C.; Qvist, J.; Halle, B. *Biophysical journal* **2008**, *95*, 2951–63.

[62] Pope, A.; Lisle Bowles, W. *The Works of Alexander Pope*; J. Johnson, 1806; p 447.

[63] Born, M.; Oppenheimer, R. *Ann. Phys.* **1927**, *389*, 457–484.

[64] Frenkel, D.; Smit, B. *Understanding Molecular Simulation, Second Edition: From Algorithms to Applications (Computational Science Series, Vol 1)*; Academic Press, 2001.

[65] Tuckerman, M. E. *Statistical Mechanics: Theory and Molecular Simulation (Oxford Graduate Texts)*; Oxford University Press, 2010.

[66] Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press, 2017.

[67] Berendsen, H. J.; Postma, J. P.; Van Gunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

[68] Harvey, S. C.; Tan, R. K.-Z.; Cheatham, T. E. *J. Comput. Chem.* **1998**, *19*, 726–740.

[69] Leach, A. *Molecular Modelling: Principles and Applications (2nd Edition)*; Pearson, 2001.

[70] Dill, K. A.; Bromberg, S. *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience, 2nd Edition*; Garland Science, 2010.

[71] Sindhikara, D. J.; Kim, S.; Voter, A. F.; Roitberg, A. E. *J. Chem. Theory Comput.* **2009**, *5*, 1624–1631.

[72] Metropolis, N.; Ulam, S. *J. Am. Stat. Assoc.* **1949**, *44*, 335–341.

[73] Verlet, L. *Phys. Rev.* **1967**, *159*, 98–103.

[74] 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. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

[75] MacKerell, A. D. et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

[76] W.J. Jorgensen and J. Tirado-Rives., *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.

[77] Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656–1676.

[78] Jones, J. E. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **1924**, *106*, 463–477.

[79] Buckingham, R. A. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **1938**, *168*, 264–283.

[80] Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

[81] Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. *J. Chem. Theory Comput.* **2013**, *9*, 3878–3888.

[82] Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–52.

[83] Ewald, P. P. *Ann. Phys.* **1921**, *369*, 253–287.

[84] Madelung, E. *Z. Phys.* **1918**, *19*, 524–533.

[85] Eastwood, J. W.; Hockney, R. W.; Lawrence, D. N. *Comput. Phys. Commun.* **1980**, *19*, 215–261.

[86] Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

[87] Hub, J. S.; De Groot, B. L.; Grubmüller, H.; Groenhof, G. *Journal of Chemical Theory and Computation* **2014**, *10*, 381–390.

[88] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*.

[89] Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269–6271.

[90] Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 234505–234511.

[91] Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 234505–6085.

[92] Fennell, C. J.; Li, L.; Dill, K. A. *J. Phys. Chem. B* **2012**, *116*, 6936–6944.

[93] Lide, D. R., Ed. *CRC handbook of chemistry and physics : a ready-reference book of chemical and physical data*, 84th ed.; CRC Press, 2003.

[94] Silvestrelli, P. L.; Parrinello, M. *Phys. Rev. Lett.* **1999**, *82*, 3308–3311.

[95] Gubskaya, A. V.; Kusalik, P. G. *J. Chem. Phys.* **2002**, *117*, 5290–5302.

[96] Mark, P.; Nilsson, L. *J. Phys. Chem. A* **2001**, *105*, 9954–9960.

[97] Kraemer, D.; Cowan, M. L.; Paarmann, A.; Huse, N.; Nibbering, E. T. J.; Elsaesser, T.; Dwayne Miller, R. J. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 437–42.

[98] Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. *J. Comput. Phys.* **1977**, *23*, 327–341.

[99] Andersen, H. C. *J. Comput. Phys.* **1983**, *52*, 24–34.

[100] Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

[101] Voronoi, G. *J. Reine Angew. Math.* **1907**, *133*, 97–178.

[102] Dirichlet, G. L. *J. Reine Angew. Math.* **1850**, *40*, 209–227.

[103] de Berg, M.; Cheong, O.; van Kreveld, M.; Overmars, M. *Computational Geometry: Algorithms and Applications*; Springer, 2008.

[104] Barber, C. B.; Dobkin, D. P.; Huhdanpaa, H. *ACM Trans. Math. Softw.* **1996**, *22*, 469–483.

[105] Olechnovič, K.; Venclovas, Č. *J. Comput. Chem.* **2014**, *35*, 672–681.

[106] Uspensky, J. V. *Theory of equations*; McGraw-Hill Book Co: New York, 1948.

[107] Halle, B. *Prog. Nucl. Magn. Reson. Spectrosc.* **1996**, *28*, 137–159.

[108] Nilsson, T.; Halle, B. *J. Chem. Phys.* **2012**, *137*, 054503.

[109] Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341–346.

[110] Shaw, D. E. et al. Millisecond-scale molecular dynamics simulations on Anton. Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis - SC '09. 2009; p 1.

[111] Zwanzig, R. *Accounts of Chemical Research* **1990**, *23*, 148–152.

[112] Xue, Y.; Ward, J. M.; Yuwen, T.; Podkorytov, I. S.; Skrynnikov, N. R. *J. Am. Chem. Soc.* **2012**, *134*, 2555–2562.

[113] Otting, G.; Liepinsh, E.; Wüthrich, K. *Biochemistry* **1993**, *32*, 3571–3582.

[114] Grey, M. J.; Wang, C.; Palmer, A. G. *J. Am. Chem. Soc.* **2003**, *125*, 14324–14335.

[115] Cox, D.; Miller, H. *The Theory of Stochastic Processes (Science Paperbacks)*; Chapman and Hall/CRC, 1977.

[116] Cox, D.; Isham, V. *Point Processes (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*; Chapman and Hall/CRC, 1980.

[117] Impey, R. W.; Madden, P. A.; McDonald, I. R. *J. Phys. Chem.* **1983**, *87*, 5071–5083.

[118] Milne, J. S.; Mayne, L.; Roder, H.; Wand, A. J.; Englander, S. W. *Protein Sci.* **1998**, *7*, 739–45.

[119] García, A. E.; Hummer, G. *Proteins Struct. Funct. Genet.* **1999**, *36*, 175–191.

[120] Skinner, J. J.; Lim, W. K.; Bédard, S.; Black, B. E.; Englander, S. W. *Protein Sci.* **2012**, *21*, 996–1005.

[121] Bahar, I.; Wallqvist, A.; Covell, D. G.; Jernigan, R. L. *Biochemistry* **1998**, *37*, 1067–1075.

[122] Battiste, J. L.; Li, R.; Woodward, C. *Biochemistry* **2002**, *41*, 2237–2245.

[123] Kim, K.-S.; Fuchs, J. A.; Woodward, C. K. *Biochemistry* **1993**, *32*, 9600–9608.

[124] Tüchsen, E.; Woodward, C. *Biochemistry* **1987**, *26*, 1918–1925.

[125] Dadarlat, V. M.; Post, C. B. *Biophys. J.* **2006**, *91*, 4544–4554.

[126] Nutt, D. R.; Smith, J. C. *J. Am. Chem. Soc.* **2008**, *130*, 13066–13073.

[127] Marchi, M. *J. Phys. Chem. B* **2003**, *107*, 6598–6602.

[128] Voloshin, V. P.; Medvedev, N. N.; Smolin, N.; Geiger, A.; Winter, R. *Phys. Chem. Chem. Phys.* **2015**, *17*, 8499–8508.

[129] Soper, A. K. *Chem. Phys.* **2000**, *258*, 121–137.

[130] Clark, G. N.; Cappa, C. D.; Smith, J. D.; Saykally, R. J.; Head-Gordon, T. *Mol. Phys.* **2010**, *108*, 1415–1433.

[131] Skinner, L. B.; Huang, C.; Schlesinger, D.; Pettersson, L. G.; Nilsson, A.; Benmore, C. J. *J. Chem. Phys.* **2013**, *138*, 074506.

[132] Okabe, A.; Boots, B.; Sugihara, K.; Chiu, S. N. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*; Wiley, 2000.

[133] Hubbard, S. R.; Hodgson, K. O.; Doniach, S. *The Journal of biological chemistry* **1988**, *263*, 4151–8.

[134] Perkins, S. J. *Biophysical Chemistry* **2001**, *93*, 129–139.

[135] Perkins, S. J. *European Journal of Biochemistry* **1986**, *157*, 169–180.

[136] Sasahara, K.; Sakurai, M.; Nitta, K. *Journal of Molecular Biology* **1999**, *291*, 693–701.

[137] Seemann, H.; Winter, R.; Royer, C. A. *Journal of Molecular Biology* **2001**, *307*, 1091–1102.

[138] Scharnagl, C.; Reif, M.; Friedrich, J. **2005**, *1749*, 187–213.

[139] Meersman, F.; Dobson, C. M.; Heremans, K. *Chemical Society Reviews* **2006**, *35*, 908.

[140] Eden, D.; Matthew, J. B.; Rosa, J. J.; Richards, F. M. *Proceedings of the National Academy of Sciences of the United States of America* **1982**, *79*, 815–819.

[141] Gekko, K. *Biochim. Biophys. Acta* **2002**, *1595*, 382–6.

[142] Kharakoz, D. P. *Biophysical Journal* **2000**, *79*, 511–525.

[143] Cooper, A. *Proceedings of the National Academy of Sciences of the United States of America* **1976**, *73*, 2740–2741.

[144] Chandler, D. *Nature* **2005**, *437*, 640–647.

[145] Jamadagni, S. N.; Godawat, R.; Garde, S. *Annual Review of Chemical and Biomolecular Engineering* **2011**, *2*, 147–171.

[146] Paci, E.; Velikson, B. *Biopolymers* **1997**, *41*, 785–797.

[147] Paci, E.; Marchi, M. *Proceedings of the National Academy of Sciences of the United States of America* **1996**, *93*, 11609–11614.

[148] Scharnagl, C.; Reif, M.; Friedrich, J. *Biophysical Journal* **2005**, *89*, 64–75.

[149] Schnell, C.; Reif, M.; Scharnagl, C.; Friedrich, J. *Physical chemistry chemical physics : PCCP* **2005**, *7*, 2217–2224.

[150] Dadarlat, V. M.; Post, C. B. *The Journal of Physical Chemistry B* **2001**, *105*, 715–724.

[151] Kharakoz, D. P.; Sarvazyan, A. P. *Biopolymers* **1993**, *33*, 11–26.

[152] Gekko, K.; Noguchi, H. *Journal of Physical Chemistry* **1979**, *83*, 2706–2714.

[153] Gekko, K.; Hasegawa, Y. *Biochemistry* **1986**, *25*, 6563–6571.

[154] Chalikian, T. V.; Totrov, M.; Abagyan, R.; Breslauer, K. J. *Journal of Molecular Biology* **1996**, *260*, 588–603.

[155] Denisov, V. P.; Halle, B. *J. Mol. Biol.* **1995**, *245*, 682–697.

[156] Denisov, V. P.; Halle, B. *J. Am. Chem. Soc.* **1995**, *117*, 8456–8465.

[157] Wiesner, S.; Kurian, E.; Prendergast, F. G.; Halle, B. *J. Mol. Biol.* **1999**, *286*, 233–246.

[158] Denisov, V. P.; Jonsson, B. H.; Halle, B. *Nat. Struct. Biol.* **1999**, *6*, 253–260.

[159] Langhorst, U.; Loris, R.; Denisov, V. P.; Doumen, J.; Roose, P.; Maes, D.; Halle, B.; Steyaert, J. *Protein Sci.* **1999**, *8*, 722–30.

[160] Modig, K.; Rademacher, M.; Lücke, C.; Halle, B. *J. Mol. Biol.* **2003**, *332*, 965–977.

[161] Modig, K.; Kurian, E.; Prendergast, F. G.; Halle, B. *Protein Sci.* **2003**, *12*, 2768–2781.

[162] Denisov, V. P.; Schlessman, J. L.; Garcia-Moreno, E. B.; Halle, B. *Biophys. J.* **2004**, *87*, 3982–3994.

[163] Denisov, V. P.; Peters, J.; Hörlein, H. D.; Halle, B. *Biochemistry* **2004**, *43*, 12020–12027.

[164] Qvist, J.; Davidovic, M.; Hamelberg, D.; Halle, B. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 6296–6301.

[165] Modig, K.; Qvist, J.; Marshall, C. B.; Davies, P. L.; Halle, B. *Physical Chemistry Chemical Physics* **2010**, *12*, 10189.

[166] Qvist, J.; Ortega, G.; Tadeo, X.; Millet, O.; Halle, B. *J. Phys. Chem. B* **2012**, *116*, 3436–3444.

[167] Kaieda, S.; Halle, B. *J. Phys. Chem. B* **2015**, *119*, 7957–7967.

[168] Marchi, M.; Sterpone, F.; Ceccarelli, M. *J. Am. Chem. Soc.* **2002**, *124*, 6787–6791.

[169] Sterpone, F.; Stirnemann, G.; Laage, D. *J. Am. Chem. Soc.* **2012**, *134*, 4116–4119.

[170] Fogarty, A. C.; Laage, D. *J. Phys. Chem. B* **2014**, *118*, 7715–7729.

[171] Duboué-Dijon, E.; Laage, D. *J. Chem. Phys.* **2014**, *141*, 22D529.

[172] Bandyopadhyay, S.; Chakraborty, S.; Bagchi, B. *J. Am. Chem. Soc.* **2005**, *127*, 16660–16667.

[173] Brunne, R. M.; Liepinsh, E.; Otting, G.; Wüthrich, K.; Van Gunsteren, W. F. *J. Mol. Biol.* **1993**, *231*, 1040–1048.

[174] Makarov, V. A.; Andrews, B. K.; Smith, P. E.; Pettitt, B. M. *Biophys. J.* **2000**, *79*, 2966–2974.

[175] Luise, A.; Falconi, M.; Desideri, A. *Proteins Struct. Funct. Genet.* **2000**, *39*, 56–67.

[176] Hua, L.; Huang, X.; Zhou, R.; Berne, B. J. *J. Phys. Chem. B* **2006**, *110*, 3704–3711.

Chapter 6

Scientific publications

Author contributions

Co-authors are abbreviated as follows: Bertil Halle (BH); Pär Söderhjelm (PS).

Paper I: Transient access to the protein interior: simulation versus NMR

BH designed the project and I computed all the primary data from the simulation. I and BH performed the data analysis. I took part in writing the paper.

Paper II: Analysis of protein dynamics simulations by a stochastic point process approach

BH derived the theoretical framework and did the analysis from the primary data computed by me. I took part in writing the paper.

Paper III: How amide hydrogens exchange in native proteins

I computed all the primary data and performed the data analysis together with BH. I took part in writing the paper.

Paper IV: The geometry of protein hydration

I performed all the simulations and computed all the primary data. The analysis was performed by me and BH. PS provided valuable advice and mentoring. I took part in writing the paper.

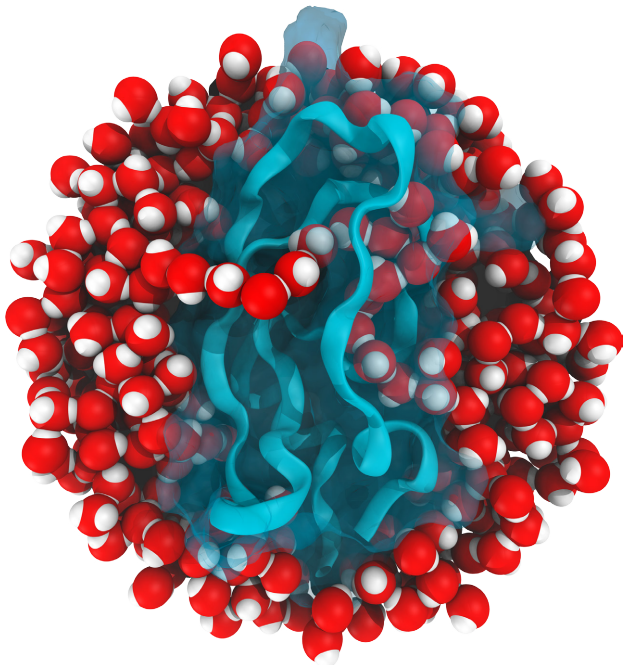
Paper V: Compressibility of the protein-water interface

I performed all the simulations and computed all the primary data. The analysis was performed by me and BH. I took part in writing the paper.

Paper VI: How proteins modify water dynamics

I performed all the simulations and computed all the primary data. The analysis was performed by me and BH. PS provided valuable advice and mentoring. I took part in writing the paper.

Printed by Media-Tryck, Lund 2018 NORDIC SWAN ECOLABEL 3041 0903



Most proteins have evolved to function optimally in aqueous environments, and the interactions between protein and water therefore play an essential role in the stability, dynamics and function of proteins. Although we understand many details of water, we understand much less about the protein-water interface. In this thesis we use molecular dynamics (MD) simulations to cast light on many structural and dynamical properties of protein hydration for which a detailed picture is lacking, such as the exchange mechanism of internal water molecules captured in the MD-snapshot above, or the spatial range of the protein-induced water perturbation depicted on the front cover.

LUND UNIVERSITY

ISBN 978-91-7422-573-0 (print)

Biophysical Chemistry
Department of Chemistry
Faculty of Engineering (LTH)
Lund University

9 789174 225730