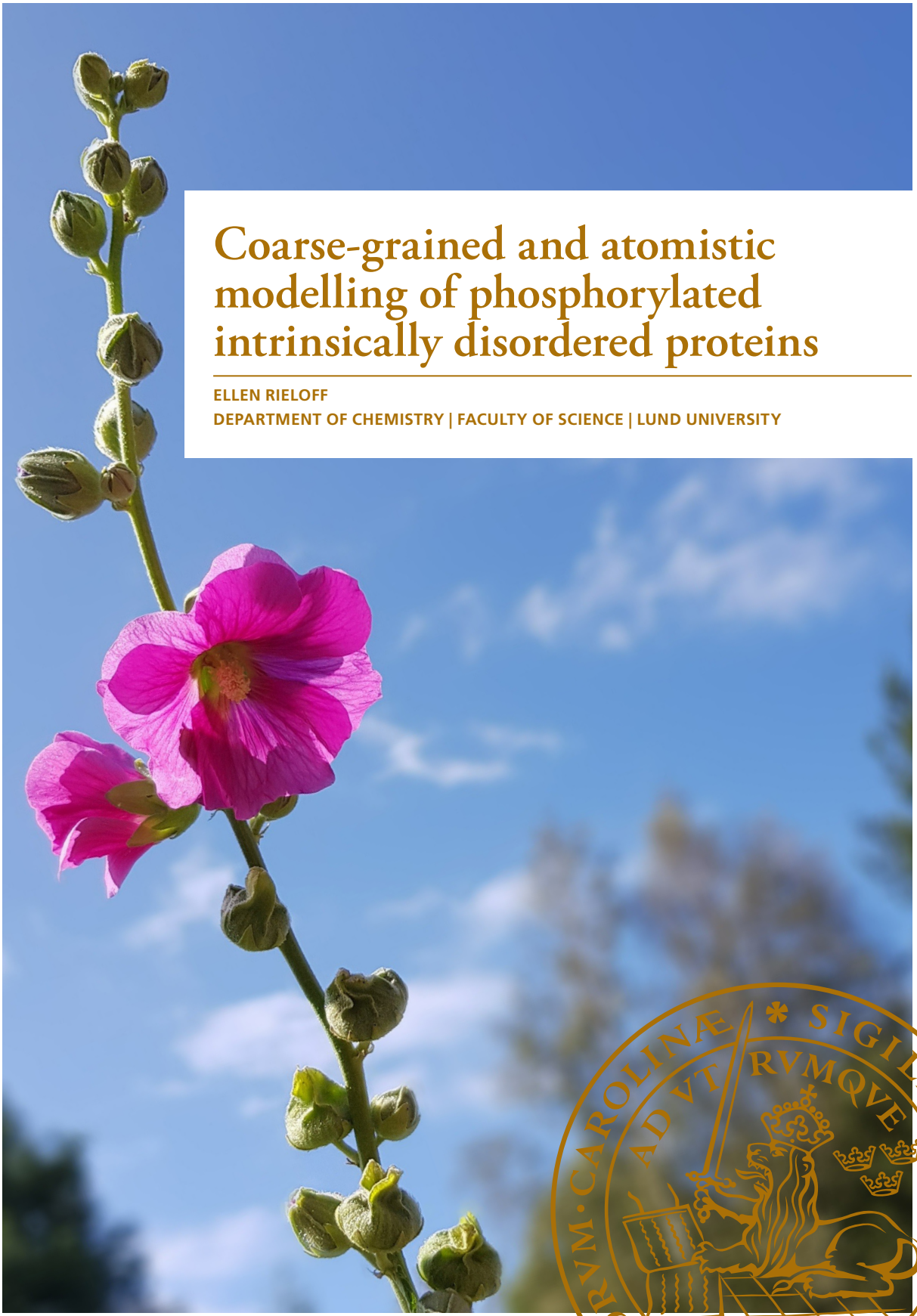
Coarse-grained and atomistic modelling of phosphorylated intrinsically disordered proteins

ELLEN RIELOFF
DEPARTMENT OF CHEMISTRY | FACULTY OF SCIENCE | LUND UNIVERSITY



Coarse-grained and atomistic modelling of phosphorylated intrinsically disordered proteins

by Ellen Rieloff

LUND
UNIVERSITY

DOCTORAL THESIS

by due permission of the Faculty of Science of Lund University, Sweden. To be defended on Friday, the 29th of October 2021 at 13:00 in lecture hall A at Kemicentrum.

Faculty opponent
Assoc. Prof. Elena Papaleo
Technical University of Denmark, Lyngby, Denmark.

Organization	Document name
LUND UNIVERSITY Department of Chemistry	**DOCTORAL DISSERTATION**
	Date of disputation 2021-10-29
	Sponsoring organization
Author(s) Ellen Rieloff	

Title and subtitle

Coarse-grained and atomistic modelling of phosphorylated intrinsically disordered proteins

Abstract

Intrinsically disordered proteins (IDPs) are involved in many biological processes such as signalling, regulation and recognition. One of the main questions regarding IDPs is how sequence, structure and function are related. Phosphorylation, a type of post-translational modification prevalent in intrinsically disordered proteins and regions, is an example of how modifications at the sequence level can induce changes in structure and thereby influence function. The lack of well-defined tertiary structure in IDPs makes them better described by an ensemble of conformations than a single structure. Furthermore, it causes them to be more difficult to study than conventional proteins, so a combined approach of experimental and simulation techniques are often advantageous. However, simulations rely on appropriate models. In this thesis, the conformational ensembles of IDPs, especially the saliva protein statherin, have been investigated using both simulations with different models and the experimental techniques small-angle X-ray scattering and circular dichroism spectroscopy. The aims have been to contribute to the collection of available tools for studying IDPs, by investigating models, and to explore the link between sequence and structure of IDPs, with special focus on phosphorylation. It was shown that a coarse-grained "one bead per residue model" can be used to describe several different IDPs and provide an understanding of how protein length, charge distribution and salt concentration affects IDPs. Furthermore, by including a hydrophobic interaction the model could qualitatively describe the self-association of statherin and provide insight on the balance of interactions and entropy governing the process. The model was however shown to overestimate the compactness of longer and more phosphorylated IDPs. Turning to atomistic simulations, it was revealed that the conformational ensembles of phosphorylated IDPs are highly influenced by salt bridges forming between phosphorylated residues and arginine/lysine/C-terminus, such that over-stabilised salt bridges cause larger compaction than observed in experiments. Another force field could however detect phosphorylation-induced changes in global compaction and secondary structure and relate them to interactions between specific residues, illustrating the potential ability of simulations to provide insight into phosphorylation.

Key words

intrinsically disordered proteins, phosphorylation, simulations, Monte Carlo, molecular dynamics, coarse-graining, atomistic, statherin, small-angle X-ray scattering, circular dichroism

Classification system and/or index terms (if any)

Supplementary bibliographical information	Language English
ISSN and key title	ISBN 978-91-7422-828-1 (print) 978-91-7422-829-8 (pdf)
Recipient's notes / Number of pages 274	Price
Security classification	

Coarse-grained and atomistic modelling of phosphorylated intrinsically disordered proteins

by Ellen Rieloff

LUND
UNIVERSITY

A doctoral thesis at a university in Sweden takes either the form of a single, cohesive research study (monograph) or a summary of research papers (compilation thesis), which the doctoral student has written alone or together with one or several other author(s).

In the latter case the thesis consists of two parts. An introductory text puts the research work into context and summarises the main points of the papers. Then, the research publications themselves are reproduced, together with a description of the individual contributions of the authors. The research papers may either have been already published or are manuscripts at various stages (in press, submitted, or in draft).

Front cover: Photo by Ellen Rieloff.

Parts of this thesis has been published before in:
Rieloff, Ellen, *Assessing self-association of intrinsically disordered proteins by coarse-grained simulations and SAXS* (2019)

Faculty of Science, Department of Chemistry

ISBN: 978-91-7422-828-1 (print)
ISBN: 978-91-7422-829-8 (pdf)

Printed in Sweden by Media-Tryck, Lund University, Lund 2021

Media-Tryck is an environmentally certified and ISO 14001:2015 certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN

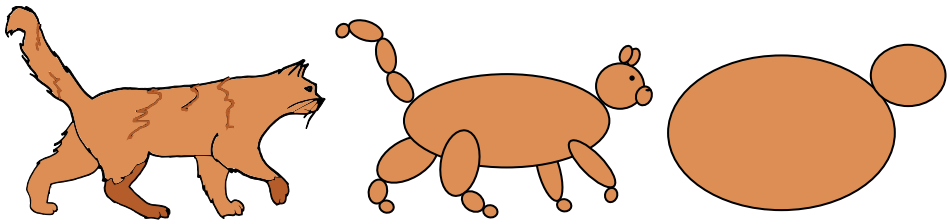
Till Ludvig
(Hoppas du gillar katten)

Contents

Populärvetenskaplig sammanfattning på svenska

Proteiner är en livsnödvändig komponent i våra kroppar. Dels är de viktiga byggstenar eftersom de ingår i kroppens alla vävnader, muskler och benstomme, men de har också andra kritiska uppgifter, såsom att transportera näringsämnen och syre samt försvara oss mot virus och bakterier. Länge trodde man att proteiner behövde en fix struktur för att vara funktionella, och att dess struktur avgjorde funktionen. Detta ifrågasattes dock, när det konstaterades att en betydande del av alla proteiner faktiskt saknar väldefinierad struktur, men ändå är funktionella. Dessa kallas för oordnade proteiner och utmärker sig genom att vara flexibla och byta konformation ofta. Oordnade proteiner är involverade i många biologiska processer där deras brist på väldefinierad struktur faktiskt kan vara en fördel. Till exempel kan de lättare interagera med flera olika partners eftersom de är anpassningsbara, och därmed fungera bra för att reglera processer. När saker går snett med de oordnade proteinerna kan det dock uppstå sjukdomar. Alzheimers, Parkinsons, och vissa typer av cancer är alla exempel på sjukdomar som involverar oordnade proteiner. I vår saliv finns det också flertalet oordnade proteiner som hjälper till med att skydda tandemaljen och slemhinnor, samt att bekämpa virus, bakterier och svamp. Proteinet jag har jobbat mest med heter statherin och har som främsta funktion att binda kalciumsalter i saliven, så det finns lättillgängligt när emaljen måste byggas upp, men inte i så stora mängder att det bildas utfällningar. Genom att förstå hur oordnade proteiner fungerar kan vi förstå sjukdomsförlopp, hitta botemedel och hämta inspiration för utveckling av läkemedel.

Proteiner är uppbyggda som långa kedjor av aminosyror med olika karaktär. Det finns ca 20 olika aminosyror som naturligt ingår i proteiner, och beroende på vilka som ingår och i vilken ordning dessa är uppradade i proteinet, det vill säga vilken sekvens proteinet har, så får proteinet olika struktur och beteende. En av de största frågorna när det kommer till oordnade proteiner är hur den här relationen mellan sekvens, struktur och funktion faktiskt ser ut. För att få svar på det, måste vi studera många olika oordnade proteiner. Det är dock ganska svårt att bestämma struktur av oordnade proteiner, just eftersom de växlar mellan olika konformationer hela tiden och således vara utsträckta i ena stunden och mer kompakta i nästa stund. I de flesta experimentella tekniker som går att tillämpa på oordnade proteiner mäter man på jättemånga proteinmolekyler samtidigt och får ut ett medelvärde över tid. Man kan likna det vid att försöka få en bild av hur människor ser ut genom att ta ett långtidsexponerat foto på ett dansgolv, där de dansande människorna är proteinerna. Fotot kommer mest visa suddiga skuggor. Ett sätt att få en bättre bild av vad som försiggår är genom att använda sig av datorsimuleringar, vilket kan visa exakt hur varje protein ser ut i varje ögonblick, samtidigt som man kan beräkna medelvärden motsvarande den experimentella datan. För att kunna göra simuleringar behövs dock en modell. Modeller kan byggas upp på olika sätt, vilket illustreras i Figur 1. Ju mer detaljer som är med i modellen, desto mer detaljerad information kan fås ut, men det blir både svårare att tolka och mer krävande att simulera, i termer av datorresurser och tidsåtgång.



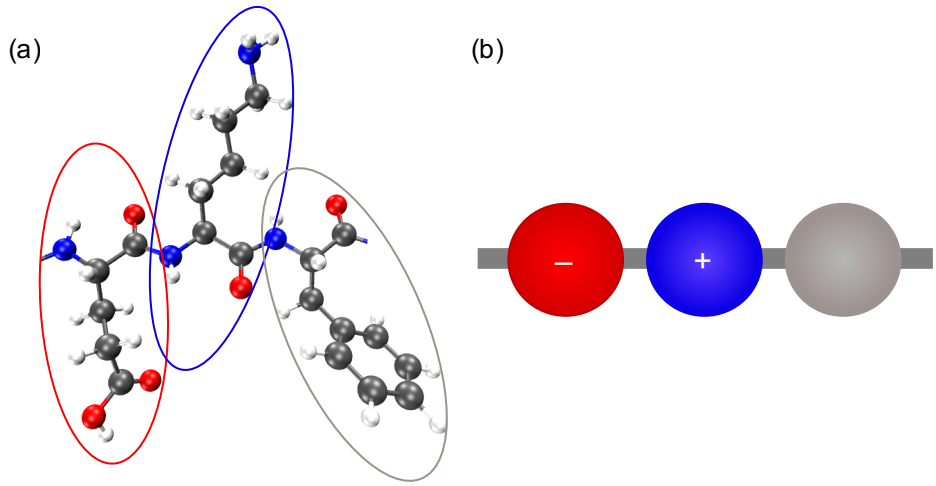
Figur 1: Olika modeller av en katt. Den till vänster är mest detaljerad. Modellerna till höger är grovkorniga och den längst till höger är mest grovkornig.

Beroende på vad vi har för forskningsfråga behöver vi därför ha olika modeller. För att fortsätta på exemplet med katten i Figur 1, så kan det vara viktigt att ha med svansen i en studie av hur katter kommunicerar. Om vi istället vill ta reda på hur många katter som får plats i ett rum räcker det dock med att se varje katt som en boll, vars storlek bestäms av hur stor katten är och hur mycket utrymme den vill ha. Men bara för att en modell innehåller mer detaljer betyder det inte att den ger bättre resultat. För att vara säkra på att modellerna stämmer och ger rätt resultat måste vi således ändå ha experimentella data att jämföra med.

I den här avhandlingen har jag främst haft två mål. Det första har varit att undersöka och vidareutveckla modeller för att beskriva oordnade proteiner, så att vi får fler verktyg för att studera denna typ av proteiner. Det andra har varit att undersöka sambandet mellan sekvens och struktur, framför allt hur fosforylering av proteiner påverkar strukturen. Fosforylering är en typ av reversibel ändring som kan göras på vissa aminosyror i ett protein, och som medför att aminosyran bland annat blir negativt laddad och får annan storlek. För att gå tillbaka till exemplet med katten, så kan vi likna det vid att sätta på katten en strumpa. Det kan påverka hur katten rör sig, och ha olika effekt beroende på vilken tass vi sätter den på, samt hur många tassar som får strumpor.

I mitt arbete har jag använt mig av två olika typer av modeller. Den första typen är en grovkornig modell, som beskriver ett protein som ett pärlhalsband. Varje pärla motsvarar en aminosyra, och har fått en laddning motsvarande den av aminosyran. Den andra typen är atomistisk, vilket innebär att alla atomer i alla aminosyror är representerade, så den är mycket mer detaljerad än den grovkorniga modellen, vilket visas i Figur 2. Den grovkorniga modellen visade sig kunna beskriva flertalet oordnade proteiner och ge en ökad förståelse för vad som kontrollerar proteinets struktur, det vill säga vilka konformationer det helst antar. En lite modifierad version av modellen kunde dessutom beskriva självassociering av statherin, det vill säga processen där flera proteinmolekyler går samman och bildar större kluster. Tillsammans med experimentella data kunde modellen användas för att avkoda vilka interaktioner som är viktiga i statherins självassociering. Den grovkorniga modellen visade sig dock överdriva hur kompakta proteiner som fosforylerats på många ställen är.

För att bättre förstå hur fosforylering påverkar proteiner behövdes en mer detaljerad modell



Figur 2: En bit av ett protein i en a) atomistisk modell och b) grovkornig modell. De färgade ovalerna visar vilka atomer som bakas samman till en pärla i den grovkorniga modellen.

än den grovkorniga, så därför använde jag två olika atomistiska modeller för att studera fosforylerade oordnade proteiner. Dessa modeller gav väldigt olika resultat, vilket visar vikten av att alltid jämföra med experiment. Den ena modellen visade sig kraftigt överskatta hur starka interaktionerna mellan fosforylerade och positivt laddade aminosyror är, vilket gjorde att proteinerna blev mer kompakta än vad experimentella metoder visade. Den andra modellen kunde kvalitativt fånga effekter av fosforylering som påvisats experimentellt och ge en detaljerad bild av vilka aminosyror som spelade roll och på vilket sätt. Detta visade att atomistiska simuleringar kan användas för att ge ökad förståelse av sambandet mellan sekvens och struktur, men att det är väldigt viktigt att fortsätta förbättra modeller.

List of publications

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

I **Utilizing Coarse-Grained Modeling and Monte Carlo Simulations to Evaluate the Conformational Ensemble of Intrinsically Disordered Proteins and Regions**

C. Cragnell, **E. Rieloff**, M. Skepö
Journal of Molecular Biology, 2018, 430, 2478–2492.

II **Assessing the Intricate Balance of Intermolecular Interactions upon Self-association of Intrinsically Disordered Proteins**

E. Rieloff, M. D. Tully, M. Skepö
Journal of Molecular Biology, 2019, 431, 511–523.

III **Phosphorylation of a Disordered Peptide – Structural Effects and Force Field Inconsistencies**

E. Rieloff, M. Skepö
Journal of Chemical Theory and Computation, 2020, 16, 1924–1935.

IV **Molecular Dynamics Simulations of Phosphorylated Intrinsically Disordered Proteins: A Force Field Comparison**

E. Rieloff, M. Skepö
International Journal of Molecular Sciences (in press), 2021.

V **The Effect of Multisite Phosphorylation on the Conformational Properties of Intrinsically Disordered Proteins**

E. Rieloff, M. Skepö
Manuscript (submitted).

Publications not included in this thesis:

Determining R_g of IDPs from SAXS Data

E. Rieloff, M. Skepö
In: Kragelund B., Skriver K. (eds), Intrinsically Disordered Proteins. Methods in Molecular Biology, vol 2141. Humana, New York, NY

Author contributions

Paper I: Utilizing Coarse-Grained Modeling and Monte Carlo Simulations to Evaluate the Conformational Ensemble of Intrinsically Disordered Proteins and Regions

I performed the experiments and part of the simulations and analysis, took part in discussions and contributed to the writing of the paper.

Paper II: Assessing the Intricate Balance of Intermolecular Interactions upon Self-association of Intrinsically Disordered Proteins

I planned the study together with my supervisor, performed the experiments and simulations and implemented cluster moves and analyses. I analysed the data with input from the co-authors, and wrote the manuscript with support from the co-authors.

Paper III: Phosphorylation of a Disordered Peptide – Structural Effects and Force Field Inconsistencies

I planned the study together with my supervisor, performed the simulations, prepared the experimental samples, performed the circular dichroism spectroscopy experiments and analysed all the data. I wrote the manuscript with support from my supervisor and was responsible for the submission and revision process.

Paper IV: Molecular Dynamics Simulations of Phosphorylated Intrinsically Disordered Proteins: A Force Field Comparison

I planned the study together with my supervisor and performed the simulations and data analysis. I wrote the manuscript with support from my supervisor.

Paper V: The Effect of Multisite Phosphorylation on the Conformational Properties of Intrinsically Disordered Proteins

I planned the study together with my supervisor and performed all the experiments, simulations and data analysis. I wrote the manuscript with support from my supervisor.

List of abbreviations

A99	Amber ff99SB-ILDN + TIP4P-D
C36	CHARMM36m
CD	circular dichroism
CMC	critical micelle concentration
FCR	fraction of charged residues
FRET	fluorescence resonance energy transfer
IDP	intrinsically disordered protein
NCPR	net charge per residue
NMR	nuclear magnetic resonance
PBC	periodic boundary conditions
PCA	principal component analysis
PME	particle mesh Ewald
PTM	post-translational modification
R_g	radius of gyration
R_{ee}	end-to-end distance
SAXS	small-angle X-ray scattering

Acknowledgements

First I want to thank my supervisor *Marie* for all the support and guidance you have given me throughout the years. I also want to express my appreciation to all former and current group members and colleagues at the division, for forming a friendly environment, and providing good discussions and fun times at ”fika”. A special thanks to *Stephanie* and *Maria*, for all we have done together during these years. I am also thankful to *Carolina* for teaching me about experimental work with proteins and SAXS, and to *Mona*, *Eric*, and *Amanda* for reading and commenting on this thesis. Furthermore, I want to thank *my family* and my friend *Emil* for support. I feel endless gratitude towards *Max* for always being by my side and supporting me in all kinds of ways. Lastly, a huge thanks to *Ludvig*, for bringing me so much joy and showing me what is truly important in life.

Chapter 1

Introduction

For a long time, the structure–function paradigm dominated the view on proteins. According to this paradigm, protein function is critically dependent on a well-defined and folded three-dimensional structure, determined by sequence [1]. However, since the late 1990s, the field of intrinsically disordered proteins (IDPs) has rapidly evolved [2] and challenged this view. Despite being unfolded at physiological conditions, IDPs have proved to have important functions in our bodies [2–5] and are today recognised as an integral part of protein science. One of the main questions in this field is how sequence, structure, and function are related. Post-translational modification (PTM), such as phosphorylation, is a great example of how function can be regulated by modifications at the sequence level inducing structural changes.

Since IDPs lack well-defined structure they have proven more challenging to study experimentally than conventional proteins. Thus, computer simulations have emerged as a useful complement, to aid in the interpretation of experimental data and to access detailed information on the molecular level. Simulations are also useful for making predictions and investigations at conditions unattainable by experimental methods. However, to obtain successful results from computer simulations, accurate models are required. To this day, there is no model available that can describe everything, hence there is a wide range of specialised models. Simulations are also limited by the computational time and resources it takes to simulate a system, so different types of models are required for different research problems.

To evaluate models an important part is comparison with experimental data, hence, experiments and computer simulations are closely linked, and also in this thesis. The aims of this thesis have been: i) to contribute to the collection of possible tools to use for studying IDPs, by evaluation and further development of suitable models, and ii) to investigate

the link between sequence and structure by studying conformational properties of IDPs in solution, with focus on phosphorylated IDPs.

Chapter 2

Background

This chapter describes IDPs and their biological relevance. The main part of my research has been focused around the saliva protein statherin, so it and its natural environment are given more focus.

2.1 Proteins

Proteins are biological macromolecules essential for life, as they provide a wide range of functions within organisms. Proteins are essentially polypeptides, since they are constructed as chains of amino acid residues connected by peptide bonds. Traditionally, the term protein is applied to long polypeptides consisting of 50 residues or more [6], while those shorter than that are referred to as polypeptides, or just peptides. Although there are many different amino acids, only roughly 20 are incorporated biosynthetically into proteins. These are referred to as *proteinogenic* amino acids. They all share the same basic structure, shown in Figure 2.1, consisting of an amino group ($-NH_2$), a carboxyl group ($-COOH$) and a side

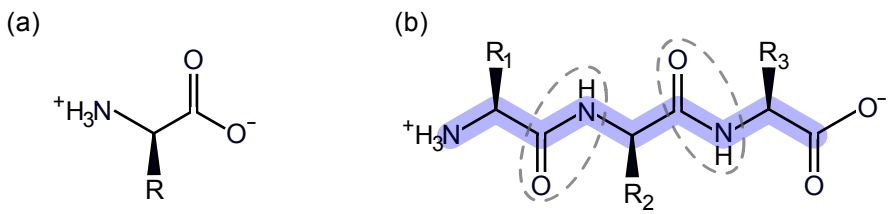


Figure 2.1: General structure of a) an amino acid and b) a tripeptide at pH 7, where R represents side groups. The backbone is highlighted in blue and the peptide bonds are shown within dashed ovals.

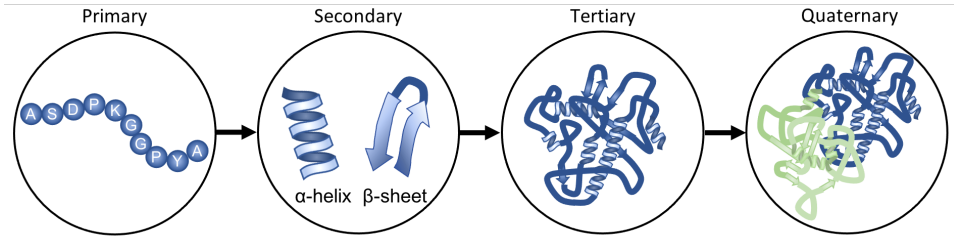


Figure 2.2: Illustration of the different levels of protein structure.

group ($-$R). At pH 7, which roughly corresponds to physiological pH, the amino group is protonised ($-NH_3^+$) and the carboxyl group deprotonized ($-COO^-$), making the amino acid *zwitterionic*. Depending on the characteristics of the side group, the amino acids can be classified as polar, hydrophobic, positively charged, or negatively charged.

The structure of a protein can be described at four different levels, as illustrated in Figure 2.2. The *primary structure* is the sequence of amino acid residues. Local parts of the chain can arrange into regular structures, referred to as *secondary structure*. The most common types of secondary structure are α-helix and β-sheet, which both form as a result of hydrogen bonds between protein backbone atoms [6]. 3_{10}- and π-helix are similar to α-helix, but differ in the hydrogen bond pattern, causing the pitch of the helix to be different. Turn is another rather common secondary structural element, which corresponds to a short segment in which the direction of the polypeptide chain is reversed. Another interesting type of secondary structure is the left-handed polyproline type II helix (PPII), which is a rather extended helix that actually lacks internal hydrogen bonds. Instead, it can be identified by the values of the backbone dihedral angles [7].

The protein can also fold into a well-defined three-dimensional shape, referred to as the *tertiary structure*. The major driving force behind folding is the hydrophobic interaction, trying to hide hydrophobic residues from the surrounding water [8]. In addition, a protein can consist of several different protein chains, each having a three-dimensional structure and making up a subunit of the complete protein. The arrangement of the subunits is called the *quaternary structure*.

2.2 Intrinsically disordered proteins

IDPs are characterized by a lack of well-defined tertiary structure under physiological conditions, which means that they are much more flexible than other proteins and interchange rapidly between many different conformations. Often can protein disorder be recognised already in the primary sequence. IDPs typically have a low sequence complexity and are

generally enriched in charged and polar amino acids, with a low content of bulky hydrophobic amino acids [9, 10].

When IDPs and intrinsically disordered regions first were discovered, they were regarded as non-functional and of no importance, due to the belief that protein function was strongly coupled to the three-dimensional structure. Since then, it has been shown that intrinsic disorder is actually wide-spread in nature. At least 10% of eukaryotic proteins are intrinsically disordered, while even more proteins contain long disordered regions [11–14]. In addition, it has been established that IDPs are involved in many important biological processes, such as regulation, signalling, and recognition, where intrinsic disorder can actually be crucial for the function [3–5, 13, 15–17]. Some advantages of disorder are that it enables interactions of high specificity coupled with low affinity, multiple binding partners, faster association/disassociation rates, and larger interaction surfaces [4]. Furthermore, many IDPs have been shown to have folding induced upon binding to interaction partners [2, 4, 18]. Due to the immense biological functions of IDPs, there is no surprise that they are also associated with pathological conditions, for example Alzheimer's disease, Parkinson's disease, diabetes, and several types of cancer [19, 20].

2.2.1 Classification of IDPs

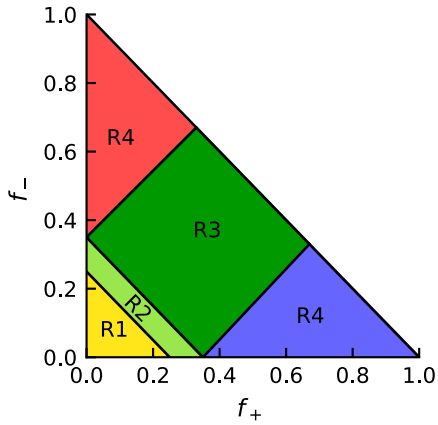
IDPs are a rather heterogeneous group, including less or more compact proteins with different degrees of secondary and tertiary structure [21, 22]. The amino acid composition and charge distribution have been shown to be important for the conformational properties of IDPs, such that they can be used to define conformational classes. From the fraction of positively and negatively charged residues, f_+ and f_-, the fraction of charged residues (FCR) and net charge per residue (NCPR) are defined according to

$$\mathrm{FCR} = f_+ + f_- \tag{2.1}$$

$$\mathrm{NCPR} = |f_+ - f_-|. \tag{2.2}$$

Based on these quantities, Das et al. have introduced a diagram-of-state with four different conformational classes called R1–R4 [23], shown in Figure 2.3. The R1 class consists of globules, while the R3 class are made up by coils and hairpins. The R2 class is an intermediate region, such that IDPs in this class usually adopt both coil and more globule-like conformations. The IDPs in the R4 class are either strongly positively or negatively charged, and behave as semi-flexible rods or coils.

Polymers consisting of positively or negatively charged subunits are called *polyelectrolytes*, while polymers containing subunits of mixed charges are called *polyampholytes*. They can be either weak or strong, depending on their FCR. Applying this terminology to IDPs, weak polyampholytes and polyelectrolytes are found in the R1 class, strong polyampholytes in the



Class	FCR	NCPR	Conformation
R1	<0.25	< 0.25	
R2	0.25–0.35	≤0.35	
R3	>0.35	≤0.35	
R4	>0.35	>0.35	

Figure 2.3: Diagram-of-states showing conformational classes of IDPs based on the fraction of positively (f_+) and negatively (f_-) charged residues, fraction of charged residues (FCR), and net charge per residue (NCPR), as introduced by Das et al. [23]. R1: globules, R2: mix of globules and coils, R3: coils or hairpins, R4: semi-flexible rods or coils.

R3 class, and strong polyelectrolytes in the R4 class. This classification scheme to predict the conformational class of an IDP is valid for IPDs consisting of at least 30 residues, having low hydrophobicity and low proline content. A high proline content is expected to give more extended conformations than the diagram-of-states predicts.

For the IDPs in the R3 class, the distribution of charges throughout the sequence also determines what conformations are adopted. The distribution of charges can be described using the parameter κ, loosely described as a parameter accounting for charge mixing. κ adopts a value between zero and one, where the maximum value corresponds to the sequence with the largest possible segregation of opposite charges for the given composition. IDPs having a low κ are expected to behave more as self-avoiding random walks, while IDPs with a high κ are more likely to adopt hair-pin like conformations. κ can also be useful for predicting the influence of salt concentration, since IDPs with high κ usually show larger conformational changes upon changes in ionic strength [24].

2.3 Phosphorylation

A common regulatory strategy employed by cells is PTM, in which a protein is chemically modified after synthesis by for example the addition of a modifying group. One of the most abundant PTM is phosphorylation, in which a phosphoryl group is attached to a residue, most commonly serine or threonine. Phosphorylation is a reversible process, and especially prevalent among IDPs and disordered regions [4, 25, 26]. As seen in Figure 2.4,

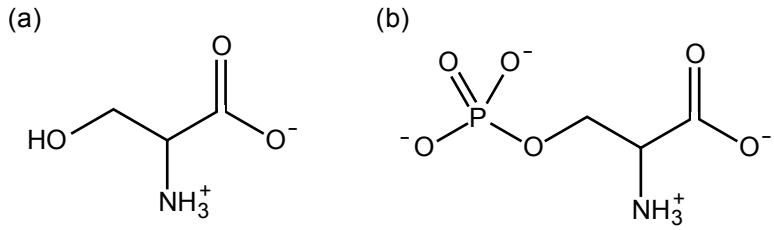


Figure 2.4: The structure of a) serine and b) phosphoserine at physiological pH.

phosphorylation increases the bulkiness of the residue and introduces two additional negative charges at physiological pH, which can greatly influence the electrostatic interactions within a protein or with a binding partner. It has been established that phosphorylation can induce changes in both overall conformation and secondary structure, as well as affect the dynamics and interactions with binding partners [27]. As a consequence, abnormal phosphorylation can be pathological; for example, Alzheimer's disease is associated with hyperphosphorylation of the neuroprotein tau [28]. In the disordered milk proteins caseins and saliva protein statherin, phosphorylated residues are of direct importance for the functionality, by enabling sequestration of calcium [29] and increasing binding to the tooth surface [30, 31].

2.4 Saliva

Saliva is a complex fluid of great importance to our oral health, even though it consists of 99.5% water. The rest involves inorganic components such as sodium, potassium, calcium, and chloride, and organic components such as proteins, lipids, and carbohydrates. Saliva aids speaking and swallowing through lubrication of the oral tissues, helps with digestion, provides protection for the teeth, and is a first line of defence against bacteria, viruses, and fungii [32]. Many of the protective functions of saliva are attributed to proteins, as presented in Figure 2.5. Note that several of these proteins are in fact intrinsically disordered and multi-functional. Many of the proteins are part of the acquired enamel pellicle, which is a thin protein-rich film that forms on the tooth surface. The pellicle protects against acid degradation, provides lubrication that protects the teeth from abrasion and attrition, and also serves as a layer to which bacteria can adhere [33, 34].

The composition, and hence the ionic strength and pH of saliva, varies with a lot of different factors, for example time of day and food intake. The saliva production can also be affected by diseases and medication [33].

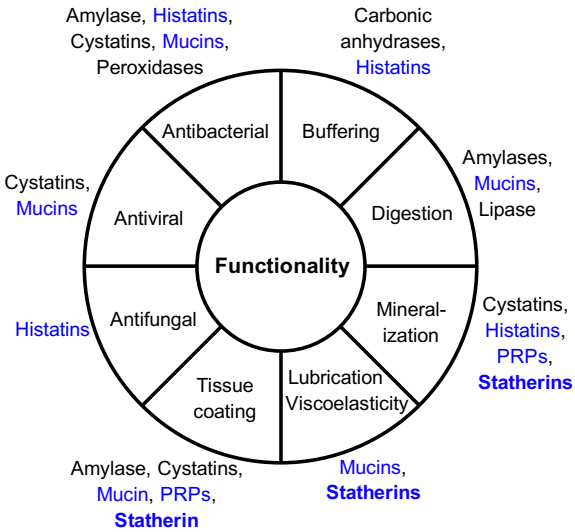


Figure 2.5: Proteins responsible for functionality of saliva, where intrinsically disordered proteins are marked in blue. The figure is adapted from Levine [35].

2.5 Statherin

Statherin is one of the intrinsically disordered salivary proteins that is part of the aquired enamel pellicle. The main function of statherin is to prevent spontaneous precipitation of calcium phosphate salts in saliva, in order to maintain a supersaturated environment [36, 37], which helps with remineralisation after dental erosion [38]. In addition, statherin has also been shown to have lubricative properties [39] and promote adhesion of certain bacteria that are associated with cemental caries and gum disease [40–42].

Statherin is a rather small protein, only 43 amino acids long with a molecular weight of 5.38 kDa, which makes it suitable for modelling. It has a distinct charge distribution, evident in the primary sequence in Figure 2.6, where nine out of ten charged residues are located among the first 13 residues in the N-terminal part. This N-terminal part, including the acidic motif with two phosphorylated serines, has been shown to be of extra importance for the ability of statherin to adsorb to the tooth enamel and prevent crystal growth [30]. Overall, the hydrophobicity is rather low (based on the hydropathy values in the Kyte-Doolittle scale [44]), which is typical for IDPs. However, region 15–43 is rich in prolines and glutamines, which allow for weak association to many other proteins [45], and contain seven tyrosines, whose aromatic side-chains have been established to be of importance for liquid-liquid phase separation [46, 47]. Statherin self-associates upon increased protein concentration [48], such that several protein chains merge to a larger complex. Self-association is further described in the following section.

+DSSEEKFLRRIGRFGYGYGPYQPVPEQPLYPQPYQPQYQQYTF−

Figure 2.6: The primary sequence of Statherin [43]. Amino acids that have a negatively charged side chain at pH 8 are marked in red, and those with a positively charged side chain are marked in blue. The phosphorylated serines (marked in dark red) have a charge of -2e each at pH 8.

2.6 Self-association

Self-association is the spontaneous formation of larger structures from smaller constituents. A typical example of self-association is the micelle formation of surfactants. Surfactants usually consist of a hydrophobic tail and a polar head-group, which means that they are *amphiphilic*. Driven by the hydrophobic interaction (see section 3.9) the surfactants arrange into spherical structures called micelles, hiding the hydrophobic tails in the interior, as shown in Figure 2.7. This only happens above a certain surfactant concentration, named the *critical micelle concentration* (CMC).

Self-association is governed by intermolecular interactions, such as van der Waals interactions, hydrogen bonding, hydrophobic interaction, and screened electrostatic interactions, which are further described in chapter 3. Since these interactions are generally weak, at least compared to covalent bonds, the self-association process is highly affected by solution conditions such as pH and ionic strength. Both the interactions between and within self-assembled structures are affected by changes in the solution conditions, therefore the size and shape of the self-assembled complexes can be modified [49].

Large molecules such as amphiphilic block-copolymers can also form micelles, however, due to their much larger size and sometimes more pronounced amphiphilic nature, the behaviour can differ from surfactants. Proteins can also self-associate, which the intrinsically disordered milk protein β-casein is a good example of. The C-terminal part of β-casein contains many hydrophobic residues, while the N-terminal part has several phosphorylated residues that contributes to a net charge, giving the protein chain an amphiphilic structure. Many studies, only a few mentioned here, have been devoted to the β-casein micelle form-

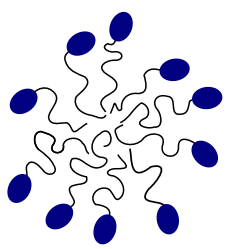


Figure 2.7: A schematic illustration of a micelle formed of surfactants having polar head-groups and hydrophobic tails.

ation and have shown that the micelle size and shape, as well as CMC are sensitive to the solution conditions such as temperature, pH and protein concentration [50–54].

Chapter 3

Intermolecular interactions

Studying proteins from a chemical point of view, we distinguish between two classes of interactions: i) covalent bonds that keep the atoms together in molecules, and ii) non-covalent intermolecular interactions. Although the term *intermolecular* literarily translates to existing or occurring *between* molecules, the interactions also act between different parts of molecules. The intermolecular interactions are generally weak compared to covalent bonds, but are highly important as they account for how proteins behave, for example how they fold and bind to other molecules. The intermolecular interactions that will be described in this chapter can be classified as short-ranged or long-ranged, depending on their distance dependence. The van der Waals interaction, having a $1/r^6$-dependence, is a typical example of a short-ranged interaction, while the Coulomb interaction acting between charged species is considered long-ranged, due to its $1/r$-dependence. The decay of potentials with different distance dependence is shown in Figure 3.1. This chapter is mostly based on the book by Israelachvili [49], which is referred to for a more thorough description.

3.1 Charge–charge interaction

The electrostatic force, F, between two atoms with charges Q_i and Q_j, separated by a distance r, is described by the Coulomb law

$$F(r) = \frac{Q_i Q_j}{4\pi\varepsilon_0\varepsilon_r}\frac{1}{r^2}, \tag{3.1}$$

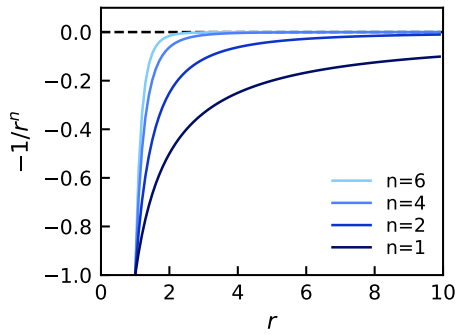


Figure 3.1: Illustration of the decay of potentials with different distance dependence.

where ε_0 is the vacuum permittivity and ε_r is the relative permittivity of the surrounding medium. The interaction free energy, $w(r)$, between the two charges is given by

$$w(r) = \int_0^\infty -F(r)\mathrm{d}r = \frac{Q_i Q_j}{4\pi\varepsilon_0\varepsilon_r}\frac{1}{r}. \tag{3.2}$$

The interaction is long-ranged, but if the charges are surrounded by ions, as in an aqueous salt solution, the interaction is screened, which reduces the range of the interaction. According to the Debye–Hückel theory, a screened Coulomb potential can be expressed as

$$V(r) = \frac{Q_i Q_j}{4\pi\varepsilon_0\varepsilon_r}\frac{1}{r}\exp(-\kappa r), \tag{3.3}$$

where $V(r)$ is the potential energy and κ^{-1} is the Debye length, defined by

$$\kappa^{-1} = \sqrt{\frac{\varepsilon_0\varepsilon_r kT}{2N_A e^2 I}}, \tag{3.4}$$

where k is the Boltzmann constant, T is the temperature, N_A the Avogadro constant, e the elementary charge, and I refers to the ionic strength, defined as

$$I = \frac{1}{2}\sum_{i=1}^{n} c_i Z_i^2. \tag{3.5}$$

Here, n is the number of different ion species, and c_i is the concentration of ion i with charge number Z_i.

3.2 Charge–dipole interaction

Most molecules have no net charge; however, they often possess an electric dipole, caused by an asymmetric distribution of electrons in the molecule. The dipole moment is defined

as

$$\mu = q\mathbf{l}, \tag{3.6}$$

where $\mathbf{l}$ is the distance vector between the two charges $-q$ and $+q$. When a charge and a dipole interact at a distance $r >> l$, the potential energy is given by

$$V(r,\theta) = -\frac{Q\mu\cos\theta}{4\pi\varepsilon_0\varepsilon_\mathrm{r}}\frac{1}{r^2}, \tag{3.7}$$

where the polar angle, θ, is the angle between the distance vector and the dipole (see Figure 3.2a). If the charge is positive, maximum attraction occurs when the dipole points away from the charge ($\theta = 0°$). At large separation or in a medium with high relative permittivity, the angle dependence of the interaction can fall below the thermal energy kT, which allows the dipole to rotate more or less freely. However, conformations allowing for attractive interactions will still be more favourable, so the angle-averaged potential will not be zero. The interaction free energy between a freely rotating dipole and a charge is given by

$$w(r) \approx -\frac{Q^2\mu^2}{6(4\pi\varepsilon_0\varepsilon_\mathrm{r})^2 kT}\frac{1}{r^4} \text{ for } kT > \frac{Q\mu}{4\pi\varepsilon_0\varepsilon_\mathrm{r} r^2}. \tag{3.8}$$

Note that this changes the distance dependence of the potential, making it more short-ranged.

3.3 Dipole–dipole interaction

The interaction energy between two stationary dipoles i and j can be described by the following potential

$$V(r,\theta_i,\theta_j,\phi) = -\frac{\mu_i\mu_j}{4\pi\varepsilon_0\varepsilon_\mathrm{r}}\frac{1}{r^3}(2\cos\theta_i\cos\theta_j - \sin\theta_i\sin\theta_j\cos\phi), \tag{3.9}$$

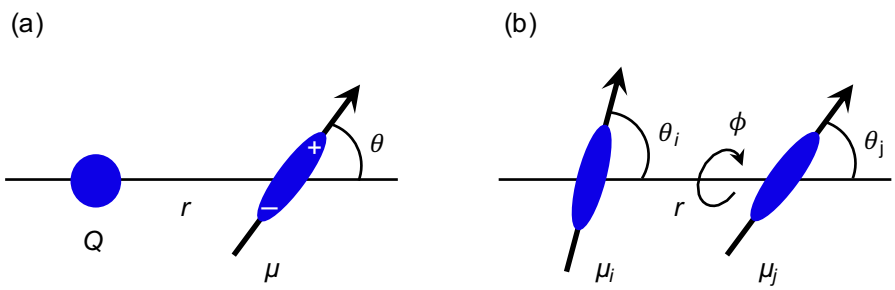


Figure 3.2: Schematic representation of the (a) charge–dipole and (b) dipole–dipole interaction, where r is the distance between the interacting species, θ is the polar angle and ϕ the azimuthal angle.

where ϕ is the azimuthal angle between the dipoles (see Figure 3.2b). Also in this case can the dipoles rotate, so the angle-averaged interaction free energy is

$$w(r) = -\frac{\mu_i^2 \mu_j^2}{3(4\pi\varepsilon_0\varepsilon_r)^2 kT} \frac{1}{r^6} \text{ for } kT > \frac{\mu_i \mu_j}{4\pi\varepsilon_0\varepsilon_r r^3}. \tag{3.10}$$

This interaction is usually referred to as the *Keesom interaction* and is a part of the total van der Waals interaction described in section 3.6.

3.4 Charge–induced dipole interaction

All molecules and atoms, even non-polar ones, are polarised by an external electric field, which means that the electron cloud in the molecule is displaced. Hence, the electric field exhibited by a charge will induce a dipole moment in a non-polar molecule. The potential between the charge and the induced dipole is expressed as

$$V(r) = -\frac{-Q^2\alpha}{2(4\pi\varepsilon_0\varepsilon_r)^2} \frac{1}{r^4}, \tag{3.11}$$

where α is the polarisability of the molecule.

3.5 Dipole–induced dipole interaction

Similarly to the charge–induced dipole interaction, a non-polar molecule can gain an induced dipole moment in the field from a permanent dipole. The interaction is described by the following potential,

$$V(r) = -\frac{\mu^2\alpha}{(4\pi\varepsilon_0\varepsilon_r)^2} \frac{1}{r^6}. \tag{3.12}$$

Notice that this potential is already angle-averaged, since the interaction normally is not strong enough to mutually orient the molecules. This interaction is usually referred to as the *Debye interaction* and is a part of the total van der Waals interaction due to the $1/r^6$-dependence.

3.6 Van der Waals interaction

The total van der Waals interaction includes three different types of interactions, which all have a $1/r^6$-dependence: Keesom, Debye and London (dispersion), of which Keesom

and Debye have been described above (section 3.3 and 3.5). The Keesom interaction is only present between permanent dipoles and the Debye interaction when one of the molecules is a permanent dipole. The last interaction, the *London dispersion interaction* is however present between all types of molecules. It is of quantum mechanical origin, although we can think of it in a simpler manner. For a non-polar atom (or molecule) the time averaged dipole moment is zero, although at any instant it exists a finite dipole moment caused by an uneven electron distribution around the nucleus. This instantaneous dipole generates an electric field that induces a dipole in another nearby atom (or molecule), leading to an attractive interaction.

3.7 Hydrogen bond

In the previous chapter hydrogen bonds where mentioned in the context of protein secondary structure. A hydrogen bond can occur between a highly electronegative atom, such as nitrogen, oxygen or fluorine, and a hydrogen covalently bonded to another such electronegative atom. It is of predominantly electrostatic origin and can be seen as an especially strong dipole–dipole interaction. Unlike normal dipole–dipole interactions it is fairly directional and can be described by a $1/r^2$-dependence, similar to the charge–dipole interaction.

3.8 Exchange repulsion (excluded volume)

At very small interatomic distances, when electron clouds overlap, a strong repulsive interaction of quantum mechanical origin occurs, which limits how close two atoms can come. The repulsion increases steeply with decreased distance and is therefore often modelled with a hard sphere potential which goes directly from zero to infinity, or with a soft core potential of $1/r^{12}$-dependence.

3.9 Hydrophobic interaction

Water is a special solvent due to the possibility to form many hydrogen bonds, which makes the water–water interaction strong. Therefore, the water molecules much rather interact with other water molecules than non-polar molecules. For small non-polar molecules the water can arrange around the non-polar molecule in such a way that no hydrogen bonds are broken. However, this arrangement is more ordered and therefore comes at an entropic cost, which makes it more favourable to separate the non-polar molecules from the water molecules. For large non-polar molecules it is not possible to retain hydrogen bonds, which instead leads to an energy driven separation. Therefore, the cause of separation between

water and non-polar molecules can be both mostly entropic or mostly energetic, however, the net result can always be seen as an effective attraction between non-polar molecules, called a hydrophobic interaction [55].

3.10 Conformational entropy

When a flexible polymer, for example an IDP, approaches a surface or other polymers, restrictions are enforced on the available conformations, which leads to a decrease in conformational entropy. If the restrictions are large enough, the result will be an effective repulsion of entropic origin.

Chapter 4

Statistical thermodynamics

Statistical mechanics provides a connection between macroscopic properties, such as temperature and pressure, and microscopic properties related to the molecules and their interactions. The aim is to provide means to both predict macroscopic phenomenas and understand them on a molecular level. Statistical mechanics applied for explaining thermodynamics is usually referred to as statistical thermodynamics. Here I will provide a brief introduction to the key concepts, while a more in-depth description can be found in for example the book by Hill [56].

A central concept in statistical mechanics is *ensembles*. An ensemble is an imaginary collection of a very large number of systems, each being equal at a thermodynamic (macroscopic) level, but differing on the microscopic level. Ensembles can be classified according to the macroscopic system that they represent, as outlined below.

Microcanonical (NVE) **ensemble**: represents an isolated system in which the number of particles (N), the volume (V) and the energy (E) are constant. Hence, the systems in the ensemble all have the same N, V, and E, and share the same environment, however, they correspond to different microstates.

Canonical (NVT) **ensemble:** corresponds to a closed and isothermal system, by having constant number of particles, volume, and temperature (T).

Grand canonical ensemble (μVT)**:** represents an open isothermal system, in which the chemical potential (μ), the volume, and the temperature are kept constant.

Isothermal-isobaric ensemble (NpT)**:** has constant number of particles, pressure (p), and temperature.

When an experimental measurement is performed, a time average is taken over the observ-

able of interest. If we instead want to calculate the observable from molecular properties, we would need to deal with both a large number of molecules and the requirement to observe them for a sufficiently long time to smear out molecular fluctuations. In practice this would be extremely complicated, however, a different approach is possible due to the *first postulate of statistical mechanics*: a (long) time average of a mechanical variable in a thermodynamic system is equal to the ensemble average of the variable in the limit of an infinitely large ensemble, provided that the ensemble replicate the thermodynamic state and environment. Stated differently, this postulate says that instead of using a time average, we can obtain the same result by performing an ensemble average, given that the ensemble is sufficiently large. This is valid for all ensembles and provides the basis for molecular simulations. There is also a *second postulate of statistical mechanics* which states that for an infinitely large ensemble representing an isolated thermodynamic system, the systems of the ensemble are distributed uniformly over the possible states consistent with the specified values of N, V and E. This postulate is also referred to as the *principle of equal a priori probabilities*, as it says that in the microcanonical ensemble, all microscopic states are equally probable.

In the canonical ensemble, the probability to find the system in a particular energy state E_i is

$$P_i(N, V, T) = \frac{\exp[-E_i(N, V)/kT]}{Q(N, V, T)}, \tag{4.1}$$

where Q is the canonical partition function, given by

$$Q(N, V, T) = \sum_i \exp[-E_i(N, V)/kT], \tag{4.2}$$

where $\exp[-E_i(N, V)/kT]$ is known as the Boltzmann weight. The partition function describes the equilibrium statistical properties of the system and can be used to express the Helmholtz free energy, A, as

$$A = -kT \ln Q. \tag{4.3}$$

The Helmholtz free energy is the characteristic function for the canonical ensemble and can be used to derive other thermodynamic variables, such as the entropy, pressure and total energy.

Here the partition function has been introduced in a quantum mechanical formulation with discrete energy states. However, many simulation methods are based on classical mechanics, in which the microstates are so close in energy that they are approximated as a continuum. In a classical treatment the canonical partition function becomes

$$Q_{\text{class}} = \frac{1}{N!h^{3N}} \int \exp[-H(\mathbf{p}^N, \mathbf{r}^N)/kT] \mathrm{d}\mathbf{p}^N \mathrm{d}\mathbf{r}^N, \tag{4.4}$$

where h is Planck's constant and the integration is performed over all momenta $\mathbf{p}^N$ and all coordinates $\mathbf{r}^N$ for all N particles. $H(\mathbf{p}^N, \mathbf{r}^N)$ is the Hamiltonian of the system, having

one kinetic energy part (dependent on the temperature) and one potential energy part (dependent on the interactions). The kinetic part can be integrated directly, simplifying the partition function to

$$Q_{\text{class}} = \frac{Z_N}{N!\Lambda^{3N}}, \tag{4.5}$$

where

$$Z_N = \int_V \exp[-U_{\text{pot}}(\mathbf{r}^N)/kT]\mathrm{d}\mathbf{r}^N \tag{4.6}$$

is the configurational integral calculated from the potential energy, U_{pot}, and

$$\Lambda = \frac{h}{(2\pi mkT)^{1/2}} \tag{4.7}$$

is the de Broglie wavelength, where m is the mass. If we know the configurational integral, we can calculate the ensemble average of an observable X, according to

$$\langle X(\mathbf{r}^N)\rangle = \frac{\int_V X(\mathbf{r}^N)\exp[-U_{\text{pot}}(\mathbf{r}^N)/kT]\mathrm{d}\mathbf{r}^N}{Z_N}. \tag{4.8}$$

However, solving the integrals is normally a rather challenging problem that requires numerical solution tools, such as the Monte Carlo method that will be discussed in chapter 6.

Chapter 5

Simulation models

A model is a representation of reality and can be constructed with varying degree of detail. When constructing or choosing a model, it is important to consider the properties of interest. The model should include enough detail to be able to accurately describe the properties of interest. Including excessive detail makes the model harder to interpret and increases the computational cost, which can limit the accessible time scale or system size. Hence, different scientific problems requires different models. In this thesis, two different types of models have been used to study IDPs, specifically a coarse-grained model representing each amino acid as a hard sphere, and an atomistic model including all atoms in the system, see Figure 5.1.

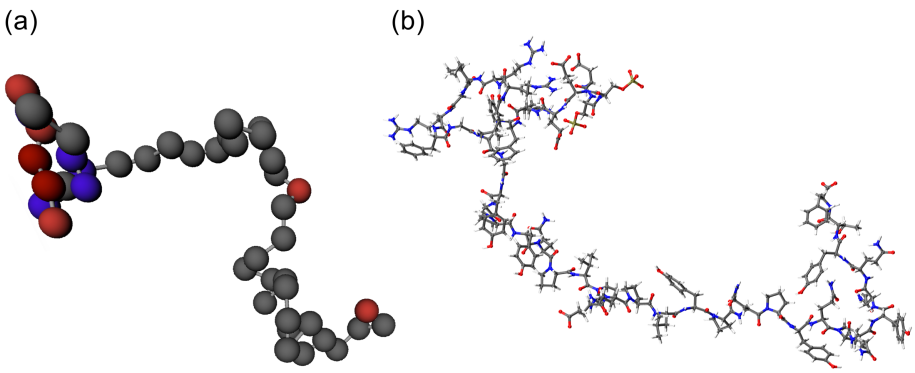


Figure 5.1: Statherin depicted in the different models: a) coarse-grained model, where gray spheres represent neutral residues, blue spheres positively charged residues, red spheres negatively charged residues, and dark red spheres phosphorylated residues, b) atomistic model, where carbon atoms are shown in gray, nitrogen in blue, oxygen in red, hydrogen in white, and phosphorus in tan.

5.1 The coarse-grained model

The coarse-grained model is a bead-necklace model based on the primitive model, in which each amino acid is described as a hard sphere (bead), connected by harmonic bonds. The N- and C-termini are modelled explicitly as charged spheres in each end of the protein chain, so the full length corresponds to the number of amino acids plus two. Each bead has a fixed point charge of $+1e$, 0, $-1e$, or $-2e$, corresponding to the state of the amino acid side chain at the desired pH. The counterions are included explicitly, while the solvent (water) and salt is treated implicitly. The model, as used in Paper I, was parameterised by Cragnell et al. for the saliva IDP histatin 5 [57].

The model contains contributions from excluded volume, electrostatic interactions, and a short-ranged attraction mimicking van der Waals-interactions. The total potential energy is divided into bonded and non-bonded interactions, according to

$$U_{\text{tot}} = U_{\text{bond}} + U_{\text{non-bond}} = U_{\text{bond}} + U_{\text{hs}} + U_{\text{el}} + U_{\text{short}}, \tag{5.1}$$

where U_{hs} is a hard-sphere potential, U_{el} the electrostatic potential, and U_{short} a short-ranged attraction. The non-bonded energy is assumed pairwise additive, according to

$$U_{\text{non-bond}} = \sum_{i<j} u_{ij}(r_{ij}), \tag{5.2}$$

where u_{ij} is the interaction between two particles, $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the center-to-center distance between the two particles, and $\mathbf{r}$ refers to the coordinate vector.

A harmonic bond represents the bonded interaction,

$$U_{\text{bond}} = \sum_{i=1}^{N-1} \frac{k_{\text{bond}}}{2}(r_{i,i+1} - r_0)^2. \tag{5.3}$$

Here, N denotes the number of beads in the protein, k_{bond} is the force constant having a value of 0.4 N/m, and $r_{i,i+1}$ is the center-to-center distance between two connected beads, with the equilibrium separation r_0 = 4.1 Å.

The excluded volume is accounted for by a hard sphere potential,

$$U_{\text{hs}} = \sum_{i<j} u_{ij}^{\text{hs}}(r_{ij}), \tag{5.4}$$

where the summation extends over all beads and ions. Here, u_{ij}^{hs} represents the hard sphere potential between two particles, according to

$$u_{ij}^{\text{hs}}(r_{ij}) = \begin{cases} 0, & r_{ij} \geq R_i + R_j \\ \infty, & r_{ij} < R_i + R_j \end{cases}, \tag{5.5}$$

where R_i and R_j denote the radii of the particles (2 Å). The electrostatic potential energy is given by an extended Debye–Hückel potential,

$$U_{\text{el}} = \sum_{i<j} u_{ij}^{\text{el}}(r_{ij}) = \sum_{i<j} \frac{Z_i Z_j e^2}{4\pi\varepsilon_0\varepsilon_{\text{r}}} \frac{\exp[-\kappa(r_{ij} - (R_i + R_j))]}{(1 + \kappa R_i)(1 + \kappa R_j)} \frac{1}{r_{ij}}. \tag{5.6}$$

Hence, the salt in the system is treated implicitly as a screening of the electrostatic interactions.

The short-ranged attractive interaction is expressed as

$$U_{\text{short}} = -\sum_{i<j} \frac{\varepsilon_{\text{short}}}{r_{ij}^6}, \tag{5.7}$$

where summation extends over all beads. Here, $\varepsilon_{\text{short}}$ reflects an average amino acid polarisability and sets the strength of the attraction. In this model $\varepsilon_{\text{short}}$ is $0.6 \cdot 10^4$ kJ Å/mol, which corresponds to an attraction of $0.6\,kT$ at closest contact.

In Paper II, an additional short-ranged interaction is included in the model, to make the protein chains associate. This mimicks a hydrophobic interaction, which is applied between all neutral amino acids, according to

$$U_{\text{h-phob}} = -\sum_{\text{neutral}} \frac{\varepsilon_{\text{h-phob}}}{r_{ij}^6}, \tag{5.8}$$

where $\varepsilon_{\text{h-phob}}$ is $1.32 \cdot 10^4$ kJ Å/mol. This corresponds to an attraction of $1.32\,kT$ at closest contact. The value of $\varepsilon_{\text{hphob}}$ was set by comparing the average association number with experimental results obtained by small-angle X-ray scattering (SAXS).

5.2 The atomistic model

In the atomistic model, distributed in the GROMACS simulation package [58–62], each atom in the system is included, hence, also solvent molecules and ions are modelled explicitly. The total potential energy consists of bonded and non-bonded interactions, according to

$$U_{\text{tot}} = \underbrace{U_{\text{bond}} + U_{\text{angle}} + U_{\text{d}} + U_{\text{id}}}_{\text{bonded}} + \underbrace{U_{\text{LJ}} + U_{\text{el}}}_{\text{non-bonded}}. \tag{5.9}$$

The bonded potentials act on covalently bonded atoms and each of the interaction potentials are summed over the atoms involved in the interaction. The first bonded term is a harmonic potential representing bond stretching,

$$U_{\text{bond}} = \sum_{b} \frac{1}{2} k_{ij}^{\text{b}} \left(r_{ij} - r_{ij}^0\right)^2, \tag{5.10}$$

where k_{ij}^{b} is a force constant, r_{ij} the distance between two bonded atoms i and j, and r_{ij}^{0} the equilibrium bond length. The second term is the bond angle vibration,

$$U_{\text{angle}} = \sum_{\theta} \frac{1}{2} k_{ij}^{\theta} \left(\theta_{ijk} - \theta_{ijk}^{0} \right)^{2}, \tag{5.11}$$

in which k_{ij}^{θ} is a force constant, and θ_{ijk} the angle between the three atoms *i-j-k*, having the equilibrium angle θ_{ijk}^{0}. The third and fourth term are torsion potentials related to dihedral angles, i.e. angles between two intersecting planes, controlling the rotation of a bond around its own longitudinal axis. Here, the proper dihedral angle is defined according to the IUPAC/IUB convention [63], as the angle ϕ_{ijkl} between the *ijk* and *jkl* planes, with zero corresponding to the cis conformation (atoms i and l on the same side). The proper dihedral angle potential is given by a sinusoidal function with periodicity n and phase ϕ_{s}:

$$U_{\mathrm{d}} = \sum_{\phi} k_{\phi} \left[1 + \cos(n\phi_{ijkl} - \phi_{\mathrm{s}}) \right], \tag{5.12}$$

where k_{ϕ} is a force constant. Unlike for the proper dihedrals, the atoms defining an improper dihedral do not need to be linearly connected. The improper dihedrals are used to keep planar groups (e.g. aromatic rings) planar, and maintain chirality. The improper dihedral angle potential is a harmonic potential,

$$U_{\mathrm{id}} = \sum_{\xi} \frac{1}{2} k_{\xi} \left(\xi_{ijkl} - \xi_{0} \right)^{2}, \tag{5.13}$$

where k_{ξ} is the force constant and ξ_{ijkl} the angle between the planes having an equilibrium dihedral angle ξ_{0}. The bonded interactions are illustrated in Figure 5.2.

Regarding the non-bonded interaction potentials, both are assumed pairwise additive. The Lennard-Jones potential,

$$U_{\mathrm{LJ}} = \sum_{i<j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{5.14}$$

represents steric repulsion and an attractive dispersion interaction. Here, ϵ_{ij} is the depth of the potential well, and σ_{ij} corresponds to the finite distance at which the potential becomes zero. For the force fields used in this work, the Lorentz-Berthelot rules are used to calculate ϵ_{ij} and σ_{ij}, according to

$$\begin{aligned} \epsilon_{ij} &= (\epsilon_{ii}\epsilon_{jj})^{1/2}, \\ \sigma_{ij} &= \frac{\sigma_{ii} + \sigma_{jj}}{2}. \end{aligned} \tag{5.15}$$

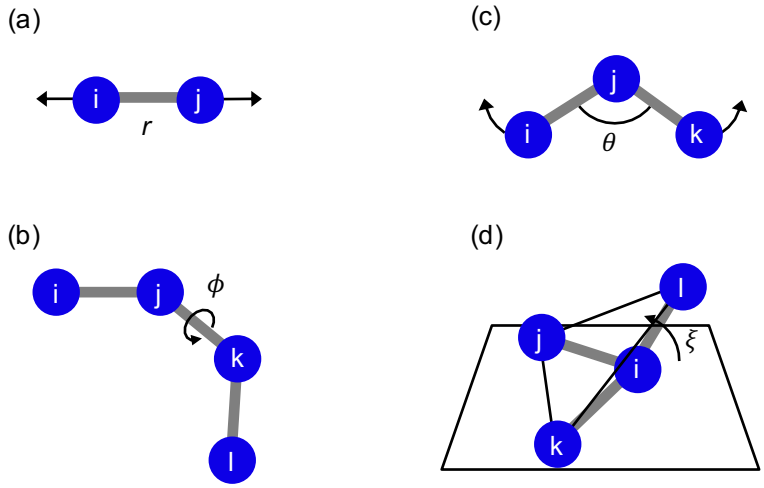


Figure 5.2: Schematic representation of the bonded interactions included in the atomistic model: a) bond stretching, b) bond angle vibration, c) proper dihedral torsion, and d) improper dihedral torsion.

The electrostatic interactions are represented by the Coulomb interaction,

$$U_{\text{el}} = \sum_{i<j} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}}, \tag{5.16}$$

where q_i and q_j are the charges of particle i and j, respectively.

5.2.1 Explicit water models

As previously mentioned, the atomistic simulations include the solvent, i.e. water, explicitly. The reason for this, is that the solvent itself and solvent–biomolecule interactions can have critical influence for biomolecules immersed in solvent. In fact, IDPs have been shown to be especially sensitive to how the water is represented, due to the extended conformations often adopted significantly exposing the protein to solvent [64–66].

There are many different explicit water models available, and due to the large number of water molecules needed to simulate a biomolecular system, the level of complexity of the water model not only influences the accuracy, but also the computational time. Among the most widely used water models today are the rigid point-charge water models with pairwise additive interactions. Due to having a fixed geometry of the water molecule, only non-bonded interactions (Coulomb and Lennard-Jones interactions) are included explicitly, which reduces the required computational effort [67]. The water models can be further dived into classes based on the number of interaction sites they contain. As shown in

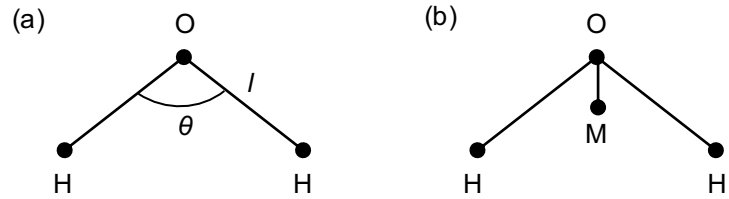


Figure 5.3: Illustration of a a) three-site and b) four-site water model, with the bond length l and bond angle θ. M represents a dummy atom where the oxygen charge is located.

Figure 5.3, three-site models have three sites, one for each atom in the molecule. In four-site models the oxygen charge is displaced to a fourth site M, while the Lennard-Jones term remains on the oxygen. Specific models are defined by their geometry (i.e. bond lengths and angles), Lennard-Jones parameters (σ and ϵ), and charges. The water models that I have used are part of the TIP family, first developed by Jorgensen [68], and are TIP3P [69] with modifications for the CHARMM force field [70, 71] and TIP4P-D [64]. The TI4P-D model uses the same geometry as the preceding TIP4P/2005 model [72], but has increased dispersion interactions (part of the Lennard-Jones interactions), aimed at sampling more extended conformations of IDPs. Another set of three-site models is the SPC family. The key difference between TIP and SPC is the geometry of the water molecule, which in TIP closely approximates experimental values (bond length $l = 0.9572$ Å and bond angle $\theta = 104.52°$), while the SPC water molecule mimics the tetrahedral shape of water molecules in ice ($l = 1$ Å and $\theta = 109.5°$) [67].

5.2.2 Force fields

The potentials described in section 5.2 together with the parameter set (e.g. force constants, equilibrium angles, and charges) constitutes a force field, which provides the foundation of a simulation. Although the dream is to have one force field that can describe all possible types of molecular systems, this is far from reality. Force field parameters are generally obtained from quantum chemical calculations and/or fitting with experimental data for a set of molecules, meaning that different force fields are aimed at different molecular systems. For proteins, the most widely used force fields families are Amber, CHARMM, GROMOS, and OPLS-AA. For a description of similarities and differences between these families, the reader is referred to ref. [73]. When discussing force fields, it is important to point out the relation to water models. Most force fields have been developed to work with a specific water model, and it has been shown that for IDPs even subtle changes in water model can influence the conformational ensemble sampled [74, 75]. Hence, it is important to use a correct combination of force field and water model.

While globular proteins and IDPs can appear indistinguishable at the most basic level; both

being chains of amino acid residues connected by peptide bonds, standard force fields developed for globular proteins have been shown to work poorly for IDPs, by overestimating α-helical and β-strand structure [76–78] and producing overly compact conformations [79, 80]. Therefore, much effort has been put into improvements, resulting in numerous force fields [75, 78, 81–95]. For IDPs, there are mainly two types of improvements that have been relevant. The first is improvement of the propensity of sampling secondary structure, for example by adjustments of backbone dihedral parameters, such as in Amber ff03* and ff99SB* [82], and CHARMM22* [85]. Side-chain torsion potentials have also been improved, resulting in force fields like Amber ff99SB-ILDN [84]. Another approach with the same aim has been the introduction of energetic terms based on backbone dihedral cross-terms, so called grid-based energy correction maps (CMAP), first introduced in the CHARMM22/CMAP (CHARMM27) force field [81]. This force field was still shown to have bias towards α-helical structure, and therefore the CMAP potentials were refined against nuclear magnetic resonance (NMR) data, which together with updated sidechain dihedral parameters resulted in CHARMM36 [86]. Further refinement of CMAP potentials together with updates to Lennard-Jones parameters to correct arginine–glutamate/aspartate/C-terminus salt bridges, were introduced in CHARMM36m [75]. The second type of improvements has been aimed at overcoming collapse by balancing the protein–water and protein–protein interactions, for example by specifically targeting Lennard-Jones parameters between water and protein atoms as in Amber ff03ws [87], or by introducing a new water model [64]. A more profound description of force field development for IDPs can be found in the following reviews: [96–98].

As stated above, force fields generally perform best for systems that have been used in their optimisation. This also extends to the type of properties considered for validation. Hence, different force fields are better at reproducing some properties than others. Therefore, when selecting a force field, it is important to carefully consider the type of system and problem at hand, as well as perform tests and compare to experimental data.

Chapter 6

Simulation methods

Simulations act as a bridge between the microscopic and macroscopic world, and between theory and experiment. Through simulations we can obtain values of observables that can be measured in the lab, based on the interactions described in the model. In this way we can test a model by comparing with experiments, and test theoretical predictions on which the model is built. Given an accurate model, the simulations can also provide information not accessible by experiments.

In this work two different simulation methods have been employed: i) *Monte Carlo* (MC) to simulate the coarse-grained model and ii) *Molecular dynamics* (MD) to simulate the atomistic model. The main difference between MC and MD is that MC calculates ensemble averages based on random sampling, while MD is based on Newton's equations of motion, hence providing time averages. Recalling the first postulate of statistical mechanics stated in chapter 4, provided sufficiently long time and large ensembles, the result is the same.

6.1 Metropolis Monte Carlo simulations

As mentioned in chapter 4, the MC technique can be utilised to compute the ensemble average of an observable, given in Equation 4.8. In the simplest MC technique, often referred to as *random sampling*, this is done by evaluating the observable at a large number of random points in phase space and multiplying the result with the Boltzmann factor. Each point in phase space corresponds to a configuration. However, a lot of the generated configurations would only give a negligible contribution to the average, by having a really small Boltzmann factor. Such configurations are for example the ones in which particles are overlapping, since that results in a very high (or infinite) potential energy.

Metropolis et al. [99] presented a more efficient scheme for evaluating a ratio of integrals for obtaining the ensemble average. In this scheme the sampling is based on the Boltzmann factor, so that the sampling is focused more around configurations with a larger Boltzmann factor. This is a type of *importance sampling* and implies that the number of configurations needed for getting a good result is reduced, which makes the simulations faster. A Metropolis MC algorithm is outlined below [100]:

Metropolis Monte Carlo algorithm

i) Generate a starting configuration.

ii) Calculate the interaction energy within the system, U_{old}.

iii) Choose a particle at random and a type of trial move (see section 6.1.1).

iv) Generate a new configuration by performing the trial move on that particle.

v) Calculate the energy of the new configuration, U_{new}.

vi) Compare the energy of the old and the new configuration to determine if the new configuration is accepted. The probability of acceptance is given by:

$$p_{\text{acc}} = \begin{cases} 1 & \text{if } U_{\text{new}} \leq U_{\text{old}} \\ \exp[-\frac{1}{kT}(U_{\text{new}} - U_{\text{old}})] & \text{if } U_{\text{new}} > U_{\text{old}} \end{cases}.$$

vii) If the new configuration is rejected, restore the old one.

viii) Repeat from step ii.

To perform the MC simulations I have used the simulation package Molsim [101]. After an initial simulation allowing the system to equilibrate, the production run consisted of a single continuous run, divided into macrosteps, on which statistics have been calculated.

6.1.1 Trial moves

Trial moves are applied to generate new configurations of the system, to explore phase space. An advantage with Monte Carlo simulations is that unphysical moves can be used to speed up the exploration. In Paper I, four different types of moves, commonly applied to polymers and proteins modelled as bead-necklaces, were used. In Paper II I also implemented a cluster move, which is advantageous in self-associating systems.

Single particle translation: A single bead in the chain, or an ion, is moved to a new, ran-

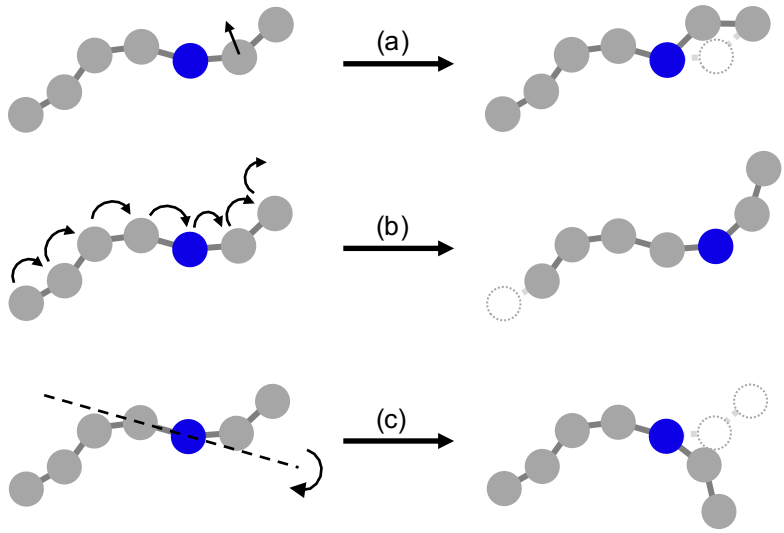


Figure 6.1: Illustration of three types of Monte Carlo moves: a) single particle displacement, b) slithering move, and c) pivot rotation.

domly chosen, position, see Figure 6.1a. The length of the translation is limited by an input parameter defined in the simulation.

Slithering move: In the slithering move, also known as reptation, one of the end beads is displaced to a random position within a bond length. The other beads are moved forward in the chain along the old configuration, as illustrated in Figure 6.1b.

Pivot rotation: One end of the chain is rotated around an axis defined by a randomly selected bond, see Figure 6.1c.

Chain translation: A whole chain is translated. This move does not change the conformation of the chain, only the position in relation to other chains and particles in the system.

Cluster move: A translation of a group of chains. The group includes the chain that the selected particle belongs to and all other chains whose center of mass is less than a predefined distance away. If the number of chains in the cluster changes during the displacement, the move is automatically rejected, as this violates detailed balance[1].

[1] Detailed balance implies that the probability of making a move and reversing it should be the same.

6.2 Molecular dynamics simulations

MD is another technique for computing equilibrium properties of classical many-body systems. In contrary to the MC technique, dynamical information can also be obtained due to the technique following Newton's equations of motion to move the particles. Newton's second law of motion states that for a particle i with constant mass, m_i, the force, $\mathbf{F}_i$ is proportional to the acceleration, $\mathbf{a}_i$, which can be expressed as the second derivative of the position $\mathbf{r}_i$ with respect to time, t:

$$\mathbf{F}_i = m_i \cdot \mathbf{a}_i = m_i \cdot \frac{\partial^2 \mathbf{r}_i}{\partial t^2}. \tag{6.1}$$

Hence, by knowing the forces, new positions and velocities of the particles can be generated by integrating Newton's second law of motion.

To run an MD simulation, starting velocities and positions, as well as the interaction potential are required as input. The forces are computed from the potential $U(\mathbf{r}^N)$, where $\mathbf{r}^N$ represents the complete set of atomic coordinates, according to

$$\mathbf{F}_i = -\frac{\partial U(\mathbf{r}^N)}{\partial \mathbf{r}_i}. \tag{6.2}$$

Since this is a many-body problem, we can only integrate the equations of motion numerically. Of course, the MD program relies on a good algorithm for doing this. I have used a version of the Verlet algorithm, called the *leap-frog algorithm*. In this algorithm, the velocities, $\mathbf{v}$, and the positions, $\mathbf{r}$, are updated at alternating times, as illustrated in Figure 6.2, using the following relations:

$$\mathbf{v}(t + \frac{1}{2}\Delta t) = \mathbf{v}(t - \frac{1}{2}\Delta t) + \frac{\Delta t}{m}\mathbf{F}(t) \tag{6.3}$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \Delta t \cdot \mathbf{v}(t + \frac{1}{2}\Delta t). \tag{6.4}$$

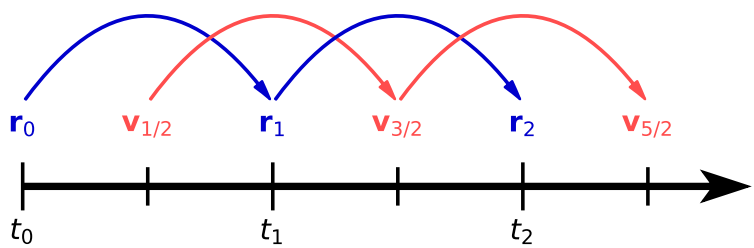


Figure 6.2: Schematic illustration of the leapfrog algorithm. It is called leapfrog due to the positions, $\mathbf{r}$, and velocities, $\mathbf{v}$, leaping over each other like frogs.

This algorithm is time reversible and area preserving, which contributes to its good energy-conserving properties. In addition, the algorithm allows for fairly long time steps, which is desirable since the number of time-consuming force evaluations then can be reduced [100, 102].

By repeatedly calculating the forces, velocities and positions, a trajectory showing how the positions and velocities changes with time is created. In this way, averages of observables can be obtained. A generic MD algorithm is summarized below [103]:

Molecular Dynamics algorithm

i) Initialize system: input the initial conditions (positions and velocities of all atoms in the system, and the potential interaction).

ii) Compute forces.

iii) Update configuration by numerically solving Newton's equations of motion.

iv) Write output.

v) Repeat from step ii.

For the MD simulations I have employed the GROMACS simulation package [58–62]. Each system has been simulated in several replicates, which have been initiated separately from the same structure, to obtain different starting velocities. Before final analysis, the individual replicates have been concatenated to one trajectory.

6.3 Technical details

In a simulation program, there are certain things that can be made to make the simulations more efficient or represent the system that we want. Here, some of those are described.

6.3.1 Periodic boundary conditions

Since this thesis investigates the behaviour of IDPs in solution, the simulations are supposed to represent bulk properties. Simulation systems can however not be as large as what is used in experiments, because that would entail an extremely large number of particles. For example, considering the most dilute samples in Paper II, even a small sample volume such as 0.1 mL contains about 10^{15} protein molecules, which is way too computationally demanding even for a coarse-grained simulation with implicit water. Unfortunately, the

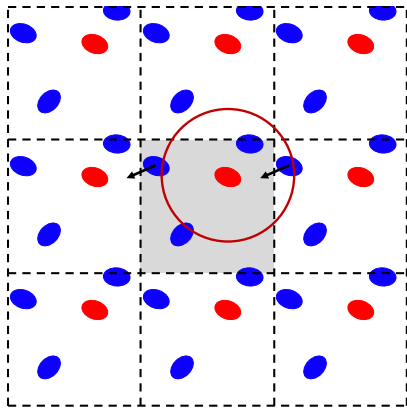


Figure 6.3: A schematic illustration of periodic boundary conditions in two dimensions, where the gray box is replicated in all directions. The arrows represents movement over a border. The red circle represents a spherical cut-off compliant with the minimum image convention for the particle marked in red.

relatively small system size employed in simulations causes a large part of the molecules in the system to be in contact with the walls of the box enclosing the system. Hence, to represent bulk behaviour, we employ periodic boundary conditions (PBC). This means that the simulation box is replicated in all directions to create an infinite lattice, as illustrated in Figure 6.3. In practice, this is achieved by letting a particle that leaves from one side of the box enter again from the opposite side. With this approach there are no walls in the system, hence it resembles the bulk. However, the periodicity of such a system can give rise to artefacts, especially if the simulation box is too small. Therefore, it is good practice to try different box sizes for the system. In the MD simulations, to ensure that the protein is not interacting with one of the periodic images, I have monitored the shortest distance between the protein and its closest periodic image. This distance should not fall below the cut-off applied to the non-bonded interactions. Cut-offs are further described in section 6.3.2.

In the coarse-grained simulations, a cubic box was employed, which is one of the simplest shapes that can be applied. However, in atomistic simulations using explicit solvent, a cubic box is not very efficient, due to the amount of solvent molecules needed to fill the corners of the cube. While a sphere is the most efficient volume, it cannot be combined with PBC. A shape that both has a smaller volume for the same image distance compared to a cube and is applicable for PBC is the rhombic dodecahedron, which has been used in the atomistic simulations.

6.3.2 Truncation

When dealing with an infinite system such as when using PBC, adding all the interactions in the system would lead to an infinite sum, due to the infinite number of particles. So for it to work practically, the interactions need to be truncated. Another reason for using truncation is that it increases the speed of the simulations, by reducing the number of calculations of non-bonded interactions. One approach is to use the minimum image convention, which restricts each molecule to interact only with the closest image of the other molecules. In practice, a spherical cut-off is often used, as illustrated in Figure 6.3. For a cubic box, the cut-off distance should not exceed half the box length, to comply with the minimum image convention. Truncating the interactions is often permissible dealing with short-ranged interactions, as the cut-off can be chosen sufficiently large, such that the interaction potential is zero beyond the cut-off. However, for long-ranged interactions, the contribution from the tail of the potential beyond the cut-off is usually non-negligible. Hence, to avoid errors, another approach is needed.

6.3.3 Long-range force handling

Due to the reasons described above, long-ranged electrostatic interactions are usually handled by the *particle-mesh Ewald* (PME) method [104], which is an improved version of Ewald summation. In Ewald summation the long-ranged interaction is separated into two parts: a short-ranged part treated as a direct sum, and a long-ranged part treated as a summation in reciprocal space. In this way, both parts converge rapidly. However, the computational cost scales as N^2, which makes it unsuitable for large systems. In PME, the reciprocal sum is approximated by a multidimensional piecewise interpolation. The approximate reciprocal energy and forces are expressed as convolutions and can therefore be evaluated using fast Fourier transforms, reducing the order of the algorithm to $N \cdot \ln N$, which makes it substantially faster than the original Ewald summation.

6.3.4 Neighbour lists

By employing cut-offs, the simulation program is sped up since the number of calculations of non-bonded interactions is reduced. However, iterating over all particles to calculate the distance between them, so that it can be determined which particles are within cut-off distance, still takes computational time. In liquids, it is usually the same particles that are in close vicinity over a few simulation steps, since it takes some simulation steps for the particles to move further away. By keeping lists over which particles are close, so-called *neighbour list*, we can avoid doing these calculations in every step. Due to having a "buffer zone" outside the interaction cut-off when creating the neighbour lists, they can be updated

less frequent. For a description of different ways to generate neighbour lists, the reader is referred to ref. [100].

6.3.5 Bond constraints

Another way to reduce the computational cost of MD simulations is by using a longer time step. The size of the time step is constrained by the time scale of the highest frequency motion in the system, which is usually bond vibrations of bonds involving hydrogen, limiting the time step to around 1 fs. Using a longer time step potentially makes the simulations unstable [105]. However, biomolecular simulations usually require simulation times in the order of µs–ms, which has a very high computational cost in terms of resources and/or physical time. By applying constraints on the bonds, such as by the LINCS algorithm [106], the length of the time step can be increased.

6.3.6 Controlling temperature and pressure

Direct use of MD simulations corresponds to the microcanonical (*NVE*) ensemble, since the Verlet-type integrators naturally conserves energy (assuming an appropriate time step). However, other ensembles can be a more convenient choice, for example the isothermal-isobaric (*NpT*) ensemble, having constant pressure and temperature, corresponding to the conditions of many laboratory experiments. The temperature and pressure can be controlled by applying temperature and pressure couplings. While there are several different options available, the *velocity-rescaling thermostat* [107] and the *Parrinello-Rahman barostat* [108] have been used for the MD simulations in this work.

The velocity rescaling thermostat is based on the Berendsen thermostat [109], in which the system is weakly coupled to an external heat bath, fixed at a desired temperature, T_0. The velocities of the particles in the system are rescaled in such a way that the rate of temperature change is proportional to the difference in temperature between the bath and the system:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{T_0 - T}{\tau}. \tag{6.5}$$

Here, τ is a time constant determining how strong the coupling is. A problem with the Berendsen thermostat is that it suppresses the fluctuations of the kinetic energy, meaning that it does not generate a proper canonical ensemble, hence the sampling is incorrect. In the velocity-rescaling thermostat this is corrected by an additional stochastic term that ensures a correct kinetic energy distribution [103]. When applying the Parrinello-Rahman barostat, additional terms involving the box vectors are included in the equations of motion, allowing the volume and shape to fluctuate.

Chapter 7

Simulation analyses

To characterise the simulated protein systems and obtain data that can be compared with experiments, I have performed different analyses, out of which the most important are described below.

7.1 Size and shape

The *radius of gyration*, R_g, is generally used as a measurement of size and is calculated as

$$R_g = \sqrt{\frac{\sum_{i=1}^{n} m_i ||\mathbf{r}_i - \mathbf{r}_{com}||^2}{\sum_{i=1}^{n} m_i}} \tag{7.1}$$

where m_i is the mass of element i, $\mathbf{r}_i$ the position of element i, $\mathbf{r}_{com}$ is the center of mass, and n the total number of elements. In the atomistic simulations the elements are the atoms, while in the coarse-grained simulations they are the beads, with each bead having equal mass.

The *end-to-end distance*, R_{ee}, provides the distance between the N- and C- terminus and is given by

$$R_{ee} = \sqrt{||\mathbf{r}_1 - \mathbf{r}_n||^2}, \tag{7.2}$$

where $\mathbf{r}_1$ and $\mathbf{r}_n$ is the position of the first and last element, respectively.

Defining the shape factor as

$$r_s = \frac{R_{ee}^2}{R_g^2}, \tag{7.3}$$

we obtain a measurement of the shape of the IDP. For a Gaussian chain, r_s is approximately six, while in the rod-like limit it reaches twelve.

7.2 Scattering curves

For a direct comparison between experiments and simulations, scattering curves are measured by SAXS and corresponding curves are calculated in the simulations. The theory behind SAXS can be found in section 8.2. The scattering curves are calculated differently in the coarse-grained and the atomistic simulations, and will be presented separately.

7.2.1 Coarse-grained approach

Each particle (bead) is regarded as a point scatterer. For a system containing N identical scattering objects, the total structure factor is expressed as

$$S(\mathbf{q}) = \left\langle \frac{1}{N} \left| \sum_{j=1}^{N} \exp(i\mathbf{q} \cdot \mathbf{r}_j) \right|^2 \right\rangle, \tag{7.4}$$

where $\mathbf{q}$ is the scattering vector. $S(\mathbf{q})$ can be further decomposed into partial structure factors given by

$$S_{jk}(\mathbf{q}) = \left\langle \frac{1}{(N_j N_k)^{1/2}} \left[\sum_{j=1}^{N} \exp(i\mathbf{q} \cdot \mathbf{r}_j) \right] \left[\sum_{k=1}^{N} \exp(i\mathbf{q} \cdot \mathbf{r}_k) \right] \right\rangle, \tag{7.5}$$

where j and k are particle types. The total and partial structure factors are related through

$$S(\mathbf{q}) = \sum_{j=1}^{N_j} \sum_{k=1}^{N_k} \frac{(N_j N_k)^{1/2}}{N} S_{jk}(\mathbf{q}). \tag{7.6}$$

For identical homogeneous spheres, the scattering intensity can be expressed as a product of the form factor and the structure factor, where the form factor corresponds to intra-particle interference and the structure factor to inter-particle interference. For a point scatterer, the form factor is constant, inferring that the scattering intensity is proportional to the structure factor. Consequently, the calculated structure factor for the point scatterers corresponds to the system's scattering intensity, only lacking a constant scaling factor. If the system is composed of a single protein chain, the calculated scattering profile comes only from intra-chain interference, hence, it is the protein form factor. For comparison with experiments an approximate effective particle form factor needs to be accounted for. This can be solved by dividing both the experimental and calculated scattering profile by their forward scattering, I_0.

7.2.2 Atomistic approach

There are several methods available for calculating solution scattering curves of macromolecules from atomic coordinates, of which the main differences regard the treatment of the solvent. The solvent is of importance because in a SAXS experiment, it is the excess electron density compared to pure solvent that is measured, meaning that the collected pattern corresponds to both the protein and the more dense layer of water molecules surrounding the protein, called the *hydration shell* (or hydration layer).

In this work I have used CRYSOL [110], in which the solvent is treated as a continuous electron density. The hydration shell is a 3 Å thick border layer with a constant excess electron density. The contrast of this hydration shell, i.e how much higher the water density is in this layer compared to the bulk, largely influences the calculated scattering curve. The effect of the contrast is especially evident in the Kratky plot, which provides information about the shape of the macromolecule. Unfortunately, choosing the optimal value of this contrast is not straightforward, as it has been shown to depend on both protein and force field [111]. A more robust way of obtaining the scattering curve is through explicit-solvent methods such as WAXSiS [112], which eliminate free parameters describing the hydration shell. However, it is associated with a higher computational cost.

7.3 Complex analyses

In Paper II, studying the self-association of statherin, several analyses are performed to characterise the result of the self-association, that is, the formed complexes. In these analyses, two chains are regarded as being part of the same complex if the center-to-center distance between a bead in each chain is less than a certain cut-off.

The complex size probability distribution is calculated according to

$$P_n = \frac{n\left\langle N_n^{\text{complex}}\right\rangle}{\sum_n n\left\langle N_n^{\text{complex}}\right\rangle}, \tag{7.7}$$

where $\left\langle N_n^{\text{complex}}\right\rangle$ is the average number of complexes consisting of n chains [113]. Since the number of chains is constant in the simulations, the denominator is equal to the total number of chains in the system. Note that the distribution is weighted by the number of chains in each complex. The average association number is calculated from the complex size probability distribution, as

$$N_{\text{assoc}} = \sum_n nP_n. \tag{7.8}$$

To set the strength of the short-ranged hydrophobic interaction, in addition to comparing the average association number with experimental results, the number of contacts for each chain was monitored along the simulation. The purpose of that was to avoid a too large interaction, which would have prevented chains in complexes from separating. The geometric condition mentioned above was used to determine if two chains were in contact.

The shape of the complexes is determined from the the principal moments of the gyration tensor. For a perfect sphere, all three principal moments are equally large. The gyration tensor is calculated from the x, y and z-coordinates according to

$$S = \frac{1}{N} \begin{pmatrix} \sum_i^N X_i^2 & \sum_i^N X_i Y_i & \sum_i^N X_i Z_i \\ \sum_i^N X_i Y_i & \sum_i^N Y_i^2 & \sum_i^N Y_i Z_i \\ \sum_i^N X_i Z_i & \sum_i^N Y_i Z_i & \sum_i^N Z_i^2 \end{pmatrix}, \tag{7.9}$$

where $X_i = (x_i - x_{\text{com}})$ and similarly for Y and Z, and N is the number of beads in the complex. Through a transformation to a principal axis system such that

$$S = \text{diag}(R_1^2, R_2^2, R_3^2) \tag{7.10}$$

S is diagonalised and $R_1^2 \geq R_2^2 \geq R_3^2$ are the eigenvalues of S, also called the principal moments of the gyration tensor [114]. In the simulations the ensemble averages of the eigenvalues are calculated for each complex size separately. From the principal moments of the gyration tensor, the asphericity, α_s, is calculated according to

$$\alpha_s = \frac{\left(\langle R_1^2 \rangle - \langle R_2^2 \rangle\right)\left(\langle R_2^2 \rangle - \langle R_3^2 \rangle\right)\left(\langle R_3^2 \rangle - \langle R_1^2 \rangle\right)}{2\left(\langle R_1^2 \rangle + \langle R_2^2 \rangle + \langle R_3^2 \rangle\right)^2}. \tag{7.11}$$

The asphericity ranges between 0 and 1, the values for a perfect sphere and a rod, respectively [115].

7.4 Secondary structure

In the coarse-grained model, no information regarding secondary structure of the IDP is available, since that requires finer details. However, from atomistic simulations, secondary structure can be determined. The program DSSP [116] calculates secondary structure based on hydrogen bonding patterns. Hydrogen bonds are defined through an electrostatic interaction energy between C=O and N–H groups, employing a generous cut-off. Secondary structure types that lack hydrogen bonding, such as bends, are determined based on geometric conditions. The secondary structure types defined in DSSP are α-helix, β-bridge,

β-sheet, 3_{10}-helix, π-helix, hydrogen bonded turn, and bend. Residues not fulfilling the criteria for any of the aforementioned types are classified as having irregular structure. In IDPs, PPII structure is also common, which can be identified by DSSP-PPII [7, 117], an extension to the DSSP program. The DSSP-PPII program acts solely on what DSSP has classified as irregular, and uses a definition of PPII based on dihedral angles.

There are many available programs for secondary structure assignment, although DSSP is one of the most used. Another wide-spread program is STRIDE [118], which uses both hydrogen bonding patterns and dihedral angles. In the visualization tool VMD [119], secondary structure is assigned by STRIDE. Although DSSP and STRIDE often are in good agreement for structured proteins, especially in the assignment of α-helix and β-sheet, disagreement is somewhat larger among IDPs, where structural elements are usually shorter and more distorted. Differences are largest among turns, where DSPP and STRIDE use different definitions [120].

Experimentally, we have used circular dichroism (CD) spectroscopy to probe secondary structure. As will be discussed in section 8.3.2, it is challenging to obtain reliable quantitative measurements of secondary structure for IDPs from CD data. However, as an alternative, there are algorithms available that can calculate CD spectra from atomic coordinates [121, 122]. Such an algorithm can therefore be used to calculate the CD spectra from simulations, to compare with experimental data. However, recent studies have suggested that they are currently not reliable for IDPs [123, 124].

7.5 Salt bridges

In proteins, salt bridges can form between oppositely charged amino acid residues. In terms of intermolecular interactions, a salt bridge is a combination of an attractive charge–charge interaction and a hydrogen bond. Phosphorylated residues have the ability to form salt bridges with positively charged residues, and as Papers III–V show, this can greatly influence the conformational ensemble. We analyse salt bridges between phosphorylated and positively charged side groups based on formed hydrogen bonds, defined according to the Wernet-Nilsson criterion [125],

$$r_{\mathrm{DA}} < 3.3\ \text{Å} - 0.00044 \cdot \theta_{\mathrm{HDA}}^2, \tag{7.12}$$

where r_{DA} is the distance between donor and acceptor heavy atoms, and θ_{HDA} is the angle made by the hydrogen, donor, and acceptor atoms, given in degrees, with zero corresponding to a perfectly straight bond.

7.6 Principal component analysis

An important part of characterising IDPs is to get a view of the conformational ensemble. The complete energy landscape contain all information about a molecule and is described by $3N - 6$ internal coordinates, where N is the number of atoms of the system [126]. For most systems, this is a huge number of dimensions, making it impossible to handle. Additionally, the information content of a complete energy landscape is much larger than what we are interested in. Usually the goal is to find a few conformational classes, or arrive at a low-dimensional energy landscape that captures the relevant behaviour of the system in only a small set of coordinates. For this, *principal component analysis* (PCA) can be applied. It is a mathematical method for reducing the dimensionality of data while still retaining most of the variability, i.e. information content. PCA transforms the data from the original set of possibly correlated variables, into a new set of uncorrelated variables called principal components. The principal components are constructed as linear combinations of the original variables, in such a way that the first principal component accounts for as much of the variation of the data as possible. Each succeeding principal component account for as much of the remaining variation as possible, while still being orthogonal to the preceding components [127]. Hence, the information content is largest in the first few components, which makes it possible to scrap the remaining components and still retain a reasonable description of the system.

To construct low-dimensional energy landscapes of the IDPs in atomistic simulations, we follow the Campos and Baptista approach [126], where PCA is applied to the cartesian coordinates of the backbone atoms of the protein, obtained after translational and rotational least square fitting on a reference structure. Due to IDPs lacking an experimental reference structure, the central structure of the simulation, i.e. the conformation that differs least from all the sampled conformations, is used as reference. In mathematical terms, it is the conformation i among N sampled conformations that minimizes the dispersion measure

$$D_i = \left(\frac{1}{N-1} \sum_{j}^{N} \mathrm{RMSD}_{ij}^2 \right)^{1/2}, \tag{7.13}$$

where RMSD_{ij} is the root mean square deviation between backbone conformations i and j. After PCA, the probability density function, $P(\mathbf{r})$, in the representation space is estimated using a Gaussian kernel density estimator. The conditional free energy is then calculated according to

$$E(\mathbf{r}) = -RT \ln \frac{P(\mathbf{r})}{P_{\max}}, \tag{7.14}$$

where $P_{\max}$ is the maximum value of $P(\mathbf{r})$. This corresponds to assigning zero energy to the maximum of the probability density. The resulting energy landscape can be used to

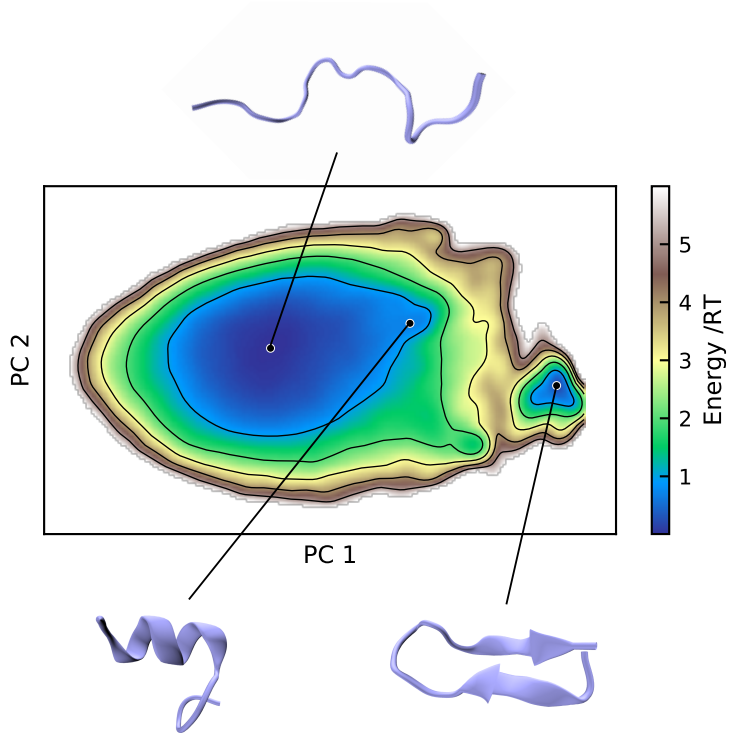


Figure 7.1: Schematic representation of an energy landscape constructed from the first two principal components that can be used to identify conformational classes.

compare different simulations and identifying conformational classes, as exemplified in Figure 7.1. However, for a complete picture of the conformational classes, more than the first two principal components are often required. This is due to that the major groups of conformations not necessarily are arranged in a non-overlapping way in this subspace, despite the first two principal components accounting for most of the variation.

7.7 Quality of sampling

In molecular simulations, there are two main factors causing errors: i) inaccurate models, and ii) insufficient sampling [128]. Hence, to be able to trust the simulation results and accredit discrepancies between simulations and experiments to model inaccuracies, we need to ensure proper sampling. It is important to keep in mind that it is much easier to rule out proper sampling than to prove it. In addition, without previous knowledge of phase space, there is no way to ensure that all important regions have been visited. Hence, focus

needs to be on assuring good quality sampling in the regions visited. Here I will describe the methods used in this work, while a more profound guide can be found in for example these references [128, 129].

To check that basic equilibration has occurred, the time series of single observables can be observed, such as R_g and R_{ee}. For IDPs which exhibit a wide range of interchanging conformations, these observables usually show large fluctuations, however, systematic changes can often still be detected. The quality of sampling of single observables can be assessed by observing correlation and calculating error estimates. For a time-ordered series of values of an observable $f(t)$, the *auto-correlation function* at a time separation t' is given by

$$c_f(t') = \frac{\langle (f(t) - \langle f \rangle)(f(t+t') - \langle f \rangle) \rangle}{\sigma_f^2}, \tag{7.15}$$

where angular brackets denote the arithmetic mean, and σ_f^2 is the variance calculated as

$$\sigma_f^2 = \frac{1}{N-1} \sum_{i=1}^{N} (f_i - \langle f \rangle)^2, \tag{7.16}$$

where N is the number of values sampled. The auto-correlation function starts at one and decays towards zero as the correlation between values diminishes, i.e the simulation looses memory of earlier values. The time it takes for the simulation to loose memory is called the *correlation time*, and is more rigorously defined as

$$\tau = \int_0^\infty c_f(t') \mathrm{d}t'. \tag{7.17}$$

From the correlation time, it is possible to estimate the number of statistically independent values as the total simulated time divided by the correlation time, which can be used as a measurement of the quality of sampling of the observable. As a rule-of-thumb, the number of statistically independent values should be at least around 20 for the sampling of that observable to be considered reliable.

In *block averaging*, the trajectory is divided into M blocks of length n. For each block, the average of the observable, B_i, is calculated, yielding a total of M values. The block size n is gradually increased, and for each block size, the block-averaged standard error is calculated as

$$\mathrm{BSE}(n) = \frac{\sum_{i=1}^{M} (B_i - \langle B \rangle)^2}{M(M-1)}, \tag{7.18}$$

where $\langle B \rangle$ is the total average for the given block size. When the block length is substantially larger than the correlation time, i.e. the blocks are independent of each other, the BSE is a reliable estimator of the true standard error. For very small block sizes, when the

consecutive blocks are highly correlated, BSE greatly underestimates the statistical error. Hence, BSE(n) increases with n until it reaches an asymptote to the true standard error. A converged BSE plot therefore signalizes that the error estimate for that observable has converged.

While the described methods above provides information about the sampling of single observables, it says little about the global sampling quality, i.e. how well the conformational space is sampled. Therefore, best practice is to always run several replicates with different initial conditions to compare.

Chapter 8

Experimental methods

In order to ensure that the simulation models describe the real world, we need to evaluate them against experimental data. Some of the most common techniques for experimental studies of IDPs are SAXS, single-molecule fluorescence resonance energy transfer (smFRET), and NMR, which all provide ensemble averaged data. This chapter focuses on the experimental techniques applied in this work, namely SAXS and CD spectroscopy. First however, I give a description of my protein purification process. In contrary to simulations were we are in complete control over what is included in the simulation box, real-world products purchased are never 100% pure. Therefore, the sample preparation and especially the protein purification is an important step in every experiment. In addition, the last section highlights some things to be aware of when using experimental data as validation.

8.1 Protein purification and determination of concentration

Statherin and the peptide fragments used in this work were purchased as lyophilised powders. The statherin powder contained trifluoroacetate, which lowered the pH, so that small addition of sodium hydroxide was necessary to dissolve the protein in buffer. To remove impurities and other buffer remains, the proteins and peptides were purified by two alternative methods. In the first, the protein solution was rinsed with buffer corresponding to at least 30 times the final sample volume, by centrifugation at a maximum speed of 358g at 8 °C in concentration cells with a 2 kDa cutoff. In the second method, dialysis was performed in room temperature and at 6 °C against a buffer of at least 400 times the sample volume, using 0.5–1 kDa membranes and exchanging the buffer 4 times during 48 h.

In both SAXS and CD experiments, the recorded signal depends on the protein concentration. Hence, for processing and interpreting the data it is important to know the concen-

tration. I have determined the concentration by absorption measurements using a Nanodrop 2000 spectrometer. For statherin, measurements were performed at 280 nm using an extinction coefficient of 8740 $M^{-1}cm^{-1}$. Since the 15 residue long N-terminal fragment of statherin lacks residues with aromatic rings, measurements were instead performed at 214 nm, using an extinction coefficient of 24000 $M^{-1}cm^{-1}$, calculated based on contributions of the peptide bond and the individual amino acids present, according to Kuipers and Gruppen [130]. In Paper III, due to limitations posed by available equipment, the concentration of the statherin fragment samples for SAXS were determined at 257 nm, where phenylalanine absorbs. The extinction coefficient used was 390 $M^{-1}cm^{-1}$, based on the value reported by Mihalyi [131]. However, here the absorption was rather low, so this approach was associated with a larger uncertainty.

8.2 Small-angle X-ray scattering

SAXS is a low-resolution technique commonly used to probe the average size, shape, and structure of particles in the nanometer length scale, typically between 1 and 100 nm. It can be applied to samples in different states such as liquid and solid, but here we focus on solution scattering of biological macromolecules.

8.2.1 Basic principle

In a SAXS experiment, a narrow beam of X-rays is sent through a sample. The X-rays interact with the electrons in the atoms, which causes the atoms to emit spherical scattered waves. The scattered waves interfere, which gives rise to an interference pattern at the detector, from which structural information can be extracted. A schematic set-up of the main parts of a SAXS instrument is found in Figure 8.1.

Scattering can occur with or without the loss of energy, however, it is the elastic scattering,

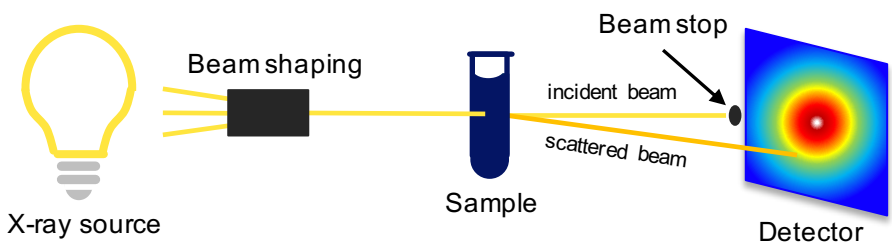


Figure 8.1: A schematic representation of the main components in a SAXS instrument. The beam stop hinders the incident beam from reaching the detector and overshadowing the sample scattering.

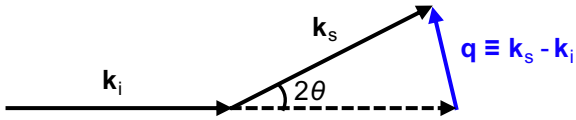


Figure 8.2: A schematic representation of the scattering vector $\mathbf{q}$, defined by the incident wave vector $\mathbf{k}_i$ and the scattered wave vector, $\mathbf{k}_s$.

that occurs without energy loss, that is of importance for SAXS. Both the incident beam and the scattered beam can be considered as planar waves defined by a wave vector, $\mathbf{k}_i$ and $\mathbf{k}_s$, respectively. The momentum transfer, usually referred to as the scattering vector, $\mathbf{q}$, is defined as the difference between the incident and scattered wave vectors, as illustrated in Figure 8.2. The magnitude of the incident wave vector is $||\mathbf{k}_i|| = 2\pi/\lambda$, where λ is the wavelength of the incident beam. Since there is no loss of energy in elastic scattering, $||\mathbf{k}_s|| = ||\mathbf{k}_i||$, hence, the magnitude of $\mathbf{q}$ can be expressed as

$$q = \frac{4\pi}{\lambda} \sin(\theta), \tag{8.1}$$

where 2θ is the angle between the incident and scattered wave vector [132].

Since the X-rays are scattered due to interactions with electrons, the more electrons a sample contains, the stronger the scattering signal is. The difference in electron density throughout the sample is therefore responsible for creating the contrast. Biological macromolecules contain mostly light elements such as hydrogen and carbon, thus the difference in electron density compared to the aqueous solution is small. Hence, the resulting signal is especially weak [132]. Therefore, for biological samples, it can be advantageous to use X-rays produced from a synchrotron, a type of large circular accelerator, instead of a lab source. The synchrotron produces X-rays with much higher brilliance, which means that the exposure time needed for detecting a useful signal is much shorter, often a few seconds compared to hours. However, the risk of radiation damage to the sample is much higher. Therefore, several frames are recorded of each sample, to compare for radiation damage and collect statistics. Also, I have used Tris buffer, which acts as a radical scavenger and therefore reduces radiation damage, in contrary to phosphate buffer which can promote it [133].

8.2.2 The scattering intensity

The detector records the scattering intensity at positions in two dimensions, however, since thermal motion causes the orientation of the particles to be random in respect to the incident beam, the scattering signal is a spherical average and can therefore be reduced to one dimension. The scattering intensity is usually presented as a function of q, to be independent of the wavelength. When performing a SAXS experiment, the scattering of the full

sample is recorded. To obtain the scattering curve of only the solute of interest, in my case the protein, we need to subtract the background. Therefore, the scattering of a matching buffer is also measured. A poorly matched buffer will greatly affect the data, so to ensure a good match, I dialysed all stock solutions overnight. The resulting dialysis buffers were used for background measurements and to dilute the samples into a concentration series.

The scattering intensity contains information on both the single particle (intraparticle interference) and relation between different particles (interparticle interference). Assuming the system consists of identical homogeneous spheres, the scattering intensity can be expressed as

$$I(q) = P(q) \cdot S(q), \tag{8.2}$$

where $P(q)$ is the form factor and $S(q)$ is the structure factor. From the form factor the size and shape of the individual particle can be determined. The structure factor contains information on the distance between particles, which can show if the particles are repelling or attracting each other. Attraction will increase the scattering curve at low q and repulsion will decrease it. In dilute and weakly interacting systems no structure is formed in the solution, meaning that the structure factor is a constant. Hence, at such conditions the form factor can be determined. Different form factors are illustrated in Figure 8.3a.

Note that IDPs adopt many different conformations, so the measured SAXS pattern corresponds to an average over all these conformations. Likewise, when dealing with polydisperse samples containing particles of different sizes, the resulting SAXS curve is an average over the different sizes present.

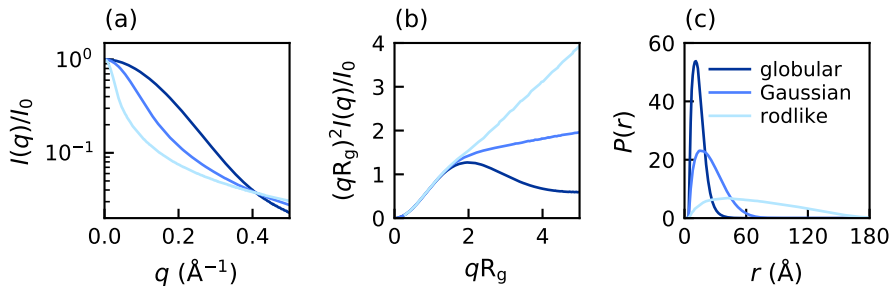


Figure 8.3: Illustration of the differences between a more globular, flexible (Gaussian chain-like) and rodlike protein. a) Form factor, b) dimensionless Kratky plot, and c) pair distance distribution function.

8.2.3 Data analysis

For proteins some standard analyses which do not require any modelling are usually performed. Besides providing information regarding particle shape and size, they also serve as a check of data quality.

The Guinier approximation

The Guinier approximation [134] provides a relation between the scattering curve at low q and the object size given by R_g, according to

$$\ln I(q) = \ln I_0 - (R_g q)^2/3, \tag{8.3}$$

where I_0 is the forward scattering (the scattering signal extrapolated to $q = 0$). Usually $\ln I(q)$ is linear with respect to q^2 at small q, normally in the region $qR_g < 1.3$ for well-folded proteins. For IDPs, this region can be reduced to $qR_g < 0.8$ [135]. Using a too large q-range tends to underestimate the R_g. If the Guinier plot shows an upswing at low q this indicates considerable aggregation in the sample, while a downswing corresponds to intermolecular repulsion. In both cases the data quality is compromised and detailed analysis should be avoided.

The forward scattering is related to the molecular weight by

$$M_w = \frac{I_0 \cdot N_A}{c([\rho_p - \rho_s]\nu_p)} \tag{8.4}$$

where I_0 is given in absolute units (cm^{-1}) and c is the protein concentration. The electron density of the protein, ρ_p, the electron density of the solvent, ρ_s, and the partial specific volume of the protein, ν_p, can all be calculated theoretically. The forward scattering is measured in arbitrary units that differs between detectors, but can be transformed to absolute units, for example by measuring the scattering of water. Normally a difference less than 10% between the measured and the theoretical weight is regarded as good [54, 136]. For self-associating proteins such as statherin, the average association number can be calculated from the measured molecular weight. Note however that for a polydisperse sample, this average is not the number average. The scattering from a sphere can be expressed analytically, from which it can be shown that in the $q \to 0$ limit, $I \propto R^6$, where R is the sphere radius [132]. Hence, large particles contribute more to the average than small particles. This is also the reason why SAXS is so sensitive to aggregates in the sample. To remove possible large aggregates from the samples, I centrifuged all protein stock solutions at approximately 18000g for at least 2 hours, after which the bottom 1/3 of the samples were discarded.

Kratky plot

To assess the flexibility of a protein and differentiate between globular and disordered proteins the Kratky plot is useful. A dimensionless Kratky plot allows for comparison between proteins of different sizes, and is constructed as $(qR_g)^2 I(q)/I_0$ vs qR_g [137]. Figure 8.3b illustrates the different behaviour of a more globular, Gaussian chain-like and rodlike protein. An intrinsically disordered protein usually exhibits a plateau as the Gaussian chain, while the actual slope depends on for example the amount of partial structure.

Pair distance distribution function

The pair distance distribution function, $P(r)$, provides information on shape, since it shows the distribution of pair distances within the protein. It is expressed in real space, compared to the scattering pattern that contains information in inverse space. $I(q)$ and $P(r)$ are related by a Fourier transform, according to [132]

$$P(r) = \frac{1}{2\pi^2} \int_0^\infty I(q) qr \sin(qr) \mathrm{d}q. \tag{8.5}$$

Since $I(q)$ is not known over the full interval $0 \leq q \leq \infty$, $P(r)$ can not be obtained directly, hence an indirect Fourier transformation method [138, 139] is often used. By definition, $P(r)$ is equal to zero at $r = 0$ and $r = D_{\max}$, the maximum distance within the protein. Since proteins do not have hard surfaces, the distribution is expected to approach zero smoothly. Problems of reaching zero or small peaks at larger r values are indicative of aggregation in the sample [140].

The $P(r)$ provides easy differentiation between globular and unfolded proteins, such as IDPs, as illustrated by Figure 8.3c. For a globular protein, the $P(r)$ is a symmetric bell-shaped curve, while for an unfolded protein the $P(r)$ shows an extended tail. If a protein has multiple domains it can be detected in the $P(r)$ as two different peaks.

R_g and I_0 can also be calculated from $P(r)$, by using the equations below [135]

$$R_g^2 = \frac{\int_0^{D_{\max}} r^2 P(r) \mathrm{d}r}{2 \int_0^{D_{\max}} P(r) \mathrm{d}r} \tag{8.6}$$

$$I_0 = 4\pi \int_0^{D_{\max}} P(r) \mathrm{d}r. \tag{8.7}$$

Since the Guinier method only uses a small region of the scattering curve, while $P(r)$ is based on more or less the whole curve, the Guinier method is more susceptible to experimental noise, giving rise to larger uncertainties. Hence, the $P(r)$ method can be more reliable.

However, the Guinier method normally has better reproducibility between users, as it is an easier method to apply. Ideally, the R_g determined from both methods should be in agreement. Note however, that R_g determined from SAXS is not directly comparable to the R_g calculated in simulations using equation 7.1, due to the scattering pattern including contributions from the hydration shell surrounding the protein [111, 141].

8.2.4 Size exclusion chromatography-coupled SAXS

A size exclusion chromatography (SEC) column is used for separating a sample according to size. A SEC column usually contains porous beads that allow small molecules to travel into the bead pores, while large objects only moves in between the beads. Hence, smaller objects travel a longer route and will be eluted later than large objects. A SEC column can therefore be used in-line with SAXS to separate the sample according to size and measure SAXS directly as it is eluted. For polydisperse samples it is therefore possible be to obtain SAXS curves for the different sized objects individually and hence obtain a size distribution. SEC-SAXS is also useful in obtaining the form factor for samples prone to aggregate, since the aggregates and the monomeric protein are eluted at different times.

8.3 Circular dichroism spectroscopy

CD spectroscopy is a highly sensitive but low-resolution technique based on the adsorption of polarised light and provides information on the secondary structure content in proteins.

8.3.1 Basic principle

Light is a type of electromagnetic radiation, which comprises an electric field and a magnetic field. These fields oscillate in perpendicular planes, that also are perpendicular to the direction of propagation. Normally light is unpolarised, which means that it oscillates in all possible directions. In linearly polarised light, the oscillations are restricted to only one direction, as illustrated in Figure 8.4a. In circularly polarised light, the electric vector rotates around the direction of propagation, undergoing a full revolution per wavelength. Clockwise rotation corresponds to right circularly polarised light, and counterclockwise to left circularly polarised light [142].

Linearly polarised light can be viewed as made up by two components of circularly polarised light of equal magnitude and phase, rotating in opposite directions (left and right), as illustrated in Figure 8.4b. If the two components are of different amplitudes, the light will be elliptically polarised, as the electric vector instead will trace an ellipse, see Figure 8.4c.

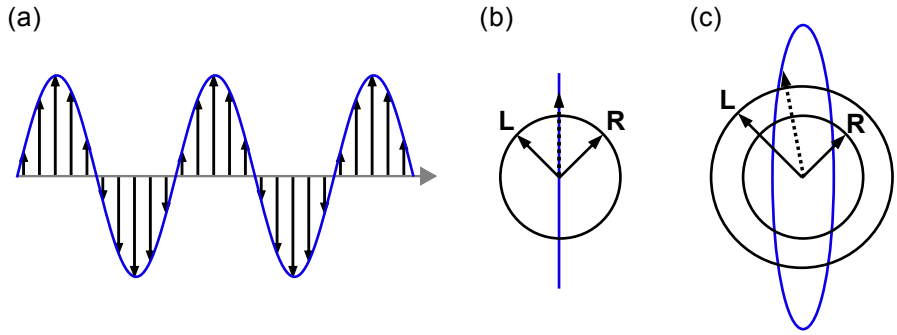


Figure 8.4: a) An illustration of linearly polarised light. The grey arrow corresponds to the direction of propagation and the black arrows represent the electric vector at different points along the propagation. b) Linearly polarized light made up by two components of circularly polarized light L and R rotating in opposite directions. The dashed arrow represents the electric vector and corresponding to the sum of the two components, which is always oriented along the blue line. (c) Different amplitude of the two components causes the electric vector (dashed arrow) to trace an ellipse, outlined in blue.

This is what happens during a CD spectroscopy experiment, as an optically active sample absorbs the left and right circularly polarised light to different extents [143].

An optically active sample contains chromophores, i.e. light-absorbing groups, that are chiral, covalently linked to a chiral centre, or situated in a chiral environment due to the three-dimensional structure of the molecule. In a protein, the chromophores of largest interest are the peptide bond, aromatic amino acid side chains and the disulphide bond. The far UV-region (approximately 170-250 nm) is dominated by peptide bond absorption, and it is in this region different secondary structure give rise to characteristic patterns, see Figure 8.5 [142].

8.3.2 Data analysis

A CD experiment monitors the difference in absorption of left and right circularly polarised light for different wavelengths. To ensure a good signal from the protein, the absorbance of the buffer should be low. Chloride ions strongly absorbs light at wavelengths in the lower end of the UV region of interest [143], and therefore I used sodium fluoride instead of sodium chloride in the CD samples. Also Tris absorbs in this region, so phosphate buffer was used instead. Aggregates and dust particles can create artefacts in the data [143], so all samples were filtered through a 0.22-µm hydrophilic filter before measurement.

Due to historic reasons the spectrum is usually presented in terms of ellipticity, with the unit degrees, and not as a difference in absorbance (ΔA). The ellipticity, θ, is calculated from the major and minor axes of the resulting ellipse and is related to the absorbance by $\theta = 32.98\Delta A$. The magnitude of the CD signal depends on the sample concentration

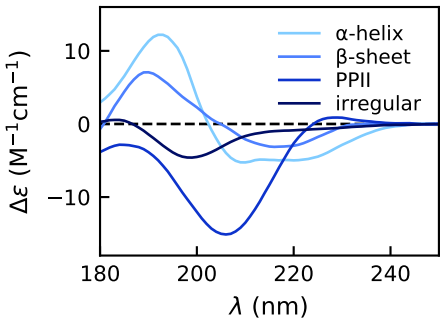


Figure 8.5: CD spectra of proteins with different secondary structure. The spectra are obtained from the Protein Circular Dichroism Data Bank [144] with the following spectrum id: CD0000117000 (α-helix) [145], CD0000118000 (anti-parallel β-sheet) [145], CD0004553000 (PPII) [146], and CD0006124000 (irregular) [147].

and the path length, so to be able to compare different measurements, the signal needs to be normalised. A common approach is to express the signal as the mean residue ellipticity (unit: deg·cm^2·dmol^{-1}), calculated as

$$[\theta]_{\mathrm{MRW}} = \frac{\theta \cdot \mathrm{MRW}}{10 \cdot d \cdot c}, \tag{8.8}$$

where θ is the observed ellipticity (in mdeg), d the path length of the cell (in cm), and c the protein concentration (in mg/mL). The mean residue weight, MRW, is the molecular weight (in Da) divided by the number of peptide bonds [143]. Data in absorption units is often expressed as the molar differential extinction coefficient, $\Delta\varepsilon$ (unit: M^{-1}cm^{-1}), calculated as

$$\Delta\varepsilon = \frac{\Delta A}{C \cdot d}, \tag{8.9}$$

where C is the molar concentration (in M).

By observing the shape of the CD spectrum, it is usually possible to discern the dominating type of secondary structure. Monitoring the shape is a straightforward method for detecting conformational and structural changes upon changes in environment, such as salt concentration or temperature. To obtain a quantified measurement of the secondary structure composition from the CD spectrum, there are several different methods available. They are all based on the approximation that a given protein CD spectrum can be expressed as a linear combination of spectra of different secondary structure components [148]. Hence, a good reference data set is vital to the results. A big reference set is often advantageous to account for some of the structural variability within a secondary structure type. Still, results can vary with both method used and applied reference set. Since irregular structure, sometimes referred to as random coil, is not a defined secondary structure, rather

the lack of other structural elements, its variability is especially large. Hence, structural assessment of IDPs from CD data is particularly challenging. Furthermore, most methods are optimized for globular proteins, meaning the result for short peptides and IDPs can be questionable. It is therefore advantageous to compare the result of different methods and/or basis sets before drawing conclusions, or only use CD spectroscopy as an indicative tool of changes in secondary structure.

8.4 Using experimental data to evaluate simulation models

By using experimental data for investigating whether the simulation models are correct, we assume that the experimental data is representative of the real world. However, even when disregarding errors that can occur in the execution of experiments, as we have seen above, approximations and assumptions are often used in the processing of data. This of course affects the final data and is another possible source of discrepancy between simulations and experiments. It is therefore preferable if the observables measured in experiments can be calculated directly in simulations.

Something else to consider is that the methods described above are rather low in resolution and measure ensemble averages. This implies that it is easier to prove a model incorrect than correct, since for example a given SAXS curve can agree equally well with different ensembles of structures. Hence, best practice is to always use several experimental methods to compare with. Just as SAXS, smFRET provides information on the overall chain dimensions, by probing long-range distances within IDPs. Connecting fluorophores to the N- and C-terminus, R_{ee} can be determined by assuming a shape of the distance distribution based on polymer theory [149]. However, the necessary fluorophores have actually been shown to influence the conformational properties of the IDP, which needs to be corrected for [150]. NMR data in the form of chemical shifts and scalar couplings contain information about local-level phenomena such as secondary structure content, and have also been applied for force field validation [65, 77, 92, 151]. In fact, regarding atomistic simulations, it has been shown that overall chain dimensions and secondary structure content is largely independent of each other, such that experimental data of both types need to be used in proper validation of force fields [152].

Lastly, when comparing results of a simulation model to experimental data, we should be aware of the intended purpose of the model. Quantitative agreement with experimental data is not always required for a model to be useful. In fact, qualitative agreement through trends can be enough, depending on the research question asked.

Chapter 9

The research

This chapter summarises and discusses the papers compiling this thesis. Overall, the research has been focused on investigating models and force fields and explore the conformational ensembles of IDPs. The first two papers explored the coarse-grained "one bead per residue"-model. Paper I investigated the generality of the model in dilute conditions, while Paper II applied the model to the self-association of statherin. In Paper III–V focus was shifted to the role of phosphorylation, which required an atomistic approach to capture changes in secondary structure. Paper III studied the 15 residue long N-terminal fragment of statherin using two different force fields. The force fields were further evaluated in Paper IV for an additional four peptides, and in Paper V the most appropriate force field was used to investigate the conformational effects induced by phosphorylation.

9.1 The generality of the coarse-grained model at dilute conditions

To test the generality of the coarse-grained model, in Paper I MC simulations of a single chain with explicit counterions and implicit salt and water, were performed for the ten different intrinsically disordered proteins or regions summarized in Table 9.1. According to the Das-Pappu plot in Figure 9.1a, this selection of IDPs represent all four conformational classes of IDPs. Hence, although the number of IDPs studied is fairly small, they still provide a good representation.

The R_g determined from simulations were compared to the R_g reported from SAXS measurements at 150 mM. As Figure 9.1b shows, the simulated values were overall in rather good agreement with the experimental values, suggesting that the model can be applied to a range of different IDPs. However, for some sequences the simulated value was distinctly smaller than the experimental value, considering the reported uncertainty, namely

Table 9.1: Length, number of phosphorylated residues (N_{phos}), fraction of charged residues (FCR), net charge per residue (NCPR), proline content (Pro), and hydrophobic content (H-phob) of the IDPs studied in Paper I. The name of the phosphorylated IDPs are printed in red, while yellow represents proline-rich IDPs.

IDP	Length	N_{phos}	FCR	NCPR	Pro (%)	H-phob (%)
histatin 5_{4-15}	12	0	0.42	+0.42	0	17
histatin 5	24	0	0.38	+0.21	0	8
statherin	43	2	0.28	-0.09	16	16
IB5	73	0	0.11	+0.08	40	7
ash1	83	0	0.20	+0.18	15	14
pash1	83	10	0.45	-0.06	14	14
sic1	92	0	0.12	+0.12	16	22
psic1	92	6	0.25	-0.01	16	20
II-1ng	141	0	0.19	+0.11	36	1
RNase E	248	0	0.39	+0.05	6	22

for pAsh1, pSic1, II-1ng, and RNase E. For RNase E it is plausible that the discrepancy was caused by a slight degree of self-association affecting the SAXS data. II-1ng is rich in prolines, which is known to increase stiffness. This effect has not been accounted for in the model, hence a smaller simulated value could be expected. The discrepancies for II-1ng and RNase E were however relatively small, compared to the discrepancies for pAsh1 and pSic1, which are most probably due to their high number of phosphorylated residues, which will be discussed later on.

Further-on, the experimental R_g could be fitted to a power law expression typical for polymers:

$$R_g = \rho_0 N^{\nu}, \tag{9.1}$$

where ρ_0 is a prefactor, N is the number of monomers (i.e amino acid residues), and ν is the Flory exponent, determined to 0.59, which agrees with the value for a self-avoiding random walk (SARW), which is approximately 0.6. This indicates that this selection of IDPs can be approximated as SARWs under the experimental conditions used, namely high ionic strength (150 mM). Therefore, it suggests that the intramolecular interactions are dominated by electrostatic interactions, which are highly screened at 150 mM.

Using a model system without charges, resembling the SARW, it was shown that the range of R_g values sampled increased with chain length, implying a relation between the conformational entropy and chain length. For all chain lengths, the probability distribution of the shape factor was a broad bell-shaped curve ranging between zero and twelve (the rod-like limit) with a maximum value of 15% at six, the value for an ideal chain. This shows that IDPs indeed adopt a wide range of different conformations, so that the conformational ensemble description is necessary.

Since IDPs are generally rather sensitive to environmental changes due to their rather flat conformational landscapes, the effect of ionic strength is of interest. Indeed the number of

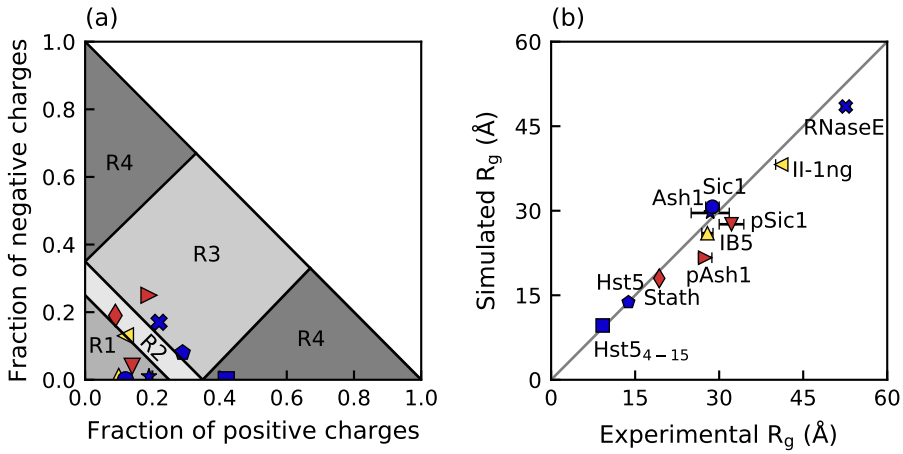


Figure 9.1: a) Classification of the IDPs included in Paper I according to the Das-Pappu plot. The regions are globules (R1), globules and coils (R2), coils/hairpins (R3), and coils/semiflexible rods (R4). Radii of gyration obtained from simulations *versus* the radii of gyration determined from SAXS experiments. In both panels proline-rich IDPs are shown in yellow, phosphorylated in red, and the rest in blue.

charged residues and their distribution throughout the sequence controlled the response to changes in ionic strength. For example, RNase E expanded upon increased ionic strength, in agreement with its classification as a strong polyampholyte, while Ash1 showed polyelectrolytic behaviour, i.e. a contraction. Although it was concluded that the IDPs could be approximated as SARWs at an ionic strength of 150 mM, Figure 9.2a confirms that this is an approximation. For Ash1, full agreement with the distribution of a SARW was reached first at 1000 mM, although the largest change occurred between 10 and 150 mM. In fact, the ionic strength was shown to have a considerable effect on the form factor. The form factor from simulations at both 150 mM and SARW conditions were in agreement with the experimental form factor collected at 150 mM NaCl, see Figure 9.2b,c. The form factor at 10 mM deviated, which implies that using the form factor collected at 150 mM salt to obtain the structure factor at 10 mM salt is indeed an approximation. However, depending on the system this approximation can be valid or contribute to errors.

To summarise, it appears that many IDPs can be described by this coarse-grained model including only steric contributions, electrostatic interactions and an approximate van der Waals interaction. The model is able to provide a basic understanding of the importance of chain length and charge distribution, and predict the outcome of changes in ionic strength. Of course, the model has its limitations. As pointed out above, the R_g of IB5 was slightly underestimated, and the stiffness shown by the Kratky plot as well. Including an angular potential made it possible to accurately represent the shape in accordance with the Kratky plot, however, this instead caused an overestimation of the R_g. To obtain a better repres-

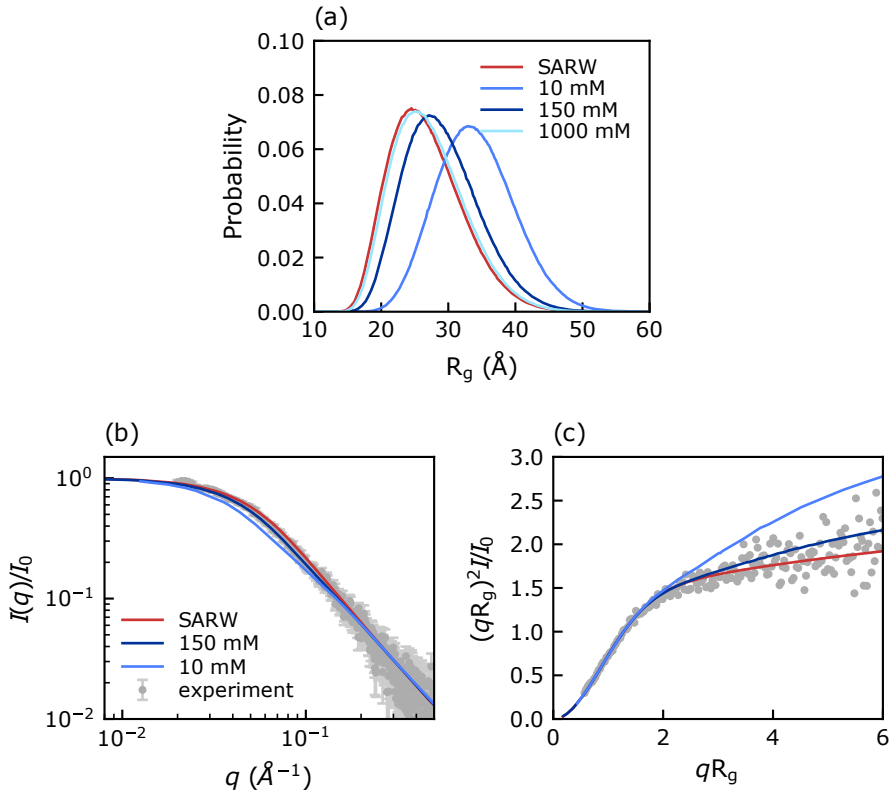


Figure 9.2: a) Probability distribution of the radius of gyration for Ash1. b) Form factor and c) dimensionless Kratky plot of Ash1 at 10 and 150 mM salt, and modelled as a SARW, compared to the experimental form factor collected at 150 mM NaCl, obtained from [153].

entation of both size and shape, a different approach, for example including local stiffness, would be necessary. The phosphorylated IDPs were also shown to be a challenge for the model. Statherin, the shortest and least phosphorylated of the three, showed a matching scattering curve and decent agreement of R_g, but for pSic1 and pAsh1 the model produced more collapsed ensembles than the experimental references. Interestingly, the agreement was much better using a charge of only $-1e$ on the phosphorylated residues. What appears as an overestimation of charges in the model may instead be caused by experimental deficiencies and/or errors and approximations within the model. For example, there can be a natural variation of the number of phosphorylated residues in the experimental sample, as well as traces of multivalent ions binding to some phosphorylated residues, meaning that the simulated and experimental sample might not be the same. Since the model has been parameterised by comparing with the form factor of histatin 5, the fact that the calculated R_g from simulations does not take into account a hydration shell, is not expected to cause discrepancy as long as the hydration shell is rather similar to that of histatin 5.

However, for Ash1/pAsh1 it was recently shown that the SAXS-derived R_g includes a larger hydration shell for the phosphorylated species, which makes it appear larger and therefore partly masks conformational changes induced by phosphorylation [141]. In addition, this model uses fixed charges, and it is possible that $-2e$ is an overestimation of the negative charge, considering the pK_a being approximately six [154] and possible influence from the local environment. As Section 9.3 will show, phosphorylation contributes with more than only charge–charge interactions, and these other factors can influence the conformational ensemble, such that a more detailed description than what this model provides might be necessary for an accurate description of phosphorylated IDPs.

9.2 Self-association of statherin

While Paper I showed that a coarse-grained model can be useful for exploring the conformational ensemble of IDPs at dilute conditions, one of the greatest benefits of a coarse-grained approach is that it enables studies of larger and more complex systems, where the computational load of an atomistic model is too large to be feasible. Hence, in Paper II the aim was to apply the model for understanding the balance between interactions in a self-associating IDP system. The saliva protein statherin was used as a model system, due to its amphiphilic character and relatively short chain length. Using SAXS, it was shown that statherin forms complexes upon increased protein concentration, see Figure 9.3a. The self-association ceased with the addition of 8 M urea, and diminished by increased temperature or lowered ionic strength. Changes in the Kratky plot (Figure 9.3b) and $P(r)$ showed that the formed complexes were more globular than the monomeric protein.

Although the exact mechanism of how urea affects proteins and self-associating systems has

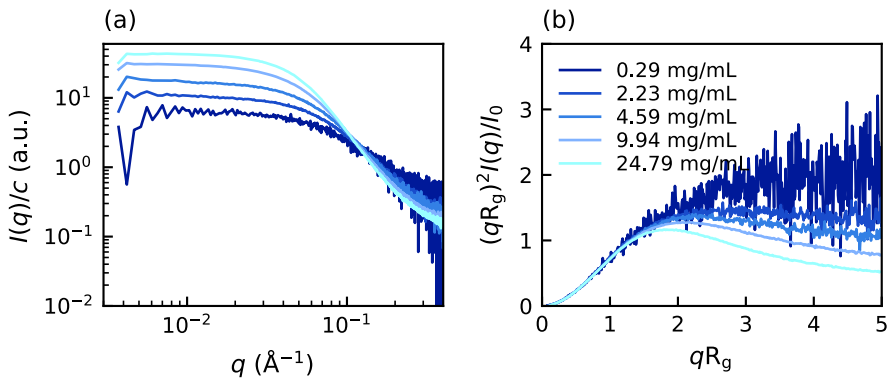


Figure 9.3: a) Scattering intensities and b) dimensionless Kratky plot of increasing concentrations of statherin in 20 mM Tris, 150 mM NaCl, pH 8, and 20 °C. The legend applies to both panels.

been long debated, urea is regarded as being able to weaken hydrophobic interactions in aqueous solution [155, 156]. Thus, that the self-association occurred both at high and low salt concentration and was hindered by urea, was interpreted as it being hydrophobically driven. To induce self-association within the model, an additional short-ranged attractive potential between neutral residues was needed, mimicking a smeared hydrophobic interaction. The strength of this potential was determined by comparing the average association number between simulations and experiments at 150 mM NaCl and 20 °C. The model was then able to capture the trends regarding protein concentration, salt concentration, and temperature. In line with the experimental findings, the complexes were shown to be more globular/spherical than the monomeric protein, see Figure 9.4a. In addition, the simulations also revealed polydispersity, as shown in Figure 9.4b. The reduction of average association number with decreased ionic strength demonstrated that electrostatic repulsion between the chains contributes to limit the growth of complexes. Substituting the phosphorylated residues with non-charged residues within the model gave larger complexes, revealing the electrostatic contribution of the phosphorylated residues. Excluding charges all together pinpointed the contribution of chain entropy in limiting the growth of complexes, which I therefore believe is the dominating factor behind the temperature effect observed in this system.

To conclude, the adjusted model successfully captured the experimentally observed trends and aided in the explanation of the observed effects in terms of a balance between different interactions and entropy. However, some limitations of the model were also encountered. First, upon inclusion of the additional attractive potential, the shape and size of the monomeric protein were no longer in agreement with SAXS data, as shown in Figure 9.5a. It might be possible to counteract this by also including an angular potential, but it would require careful balancing against the short-ranged attraction. Also, at high salt concentrations

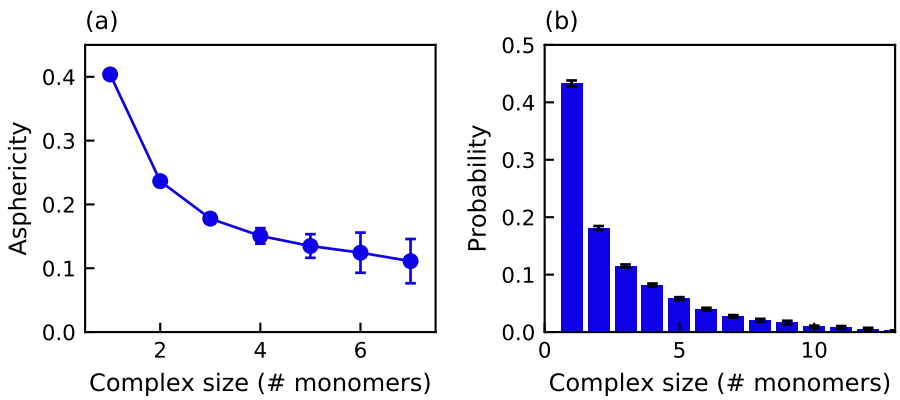


Figure 9.4: a) Asphericity of complexes of different size and b) size distribution in the simulation of 5 mg/mL statherin.

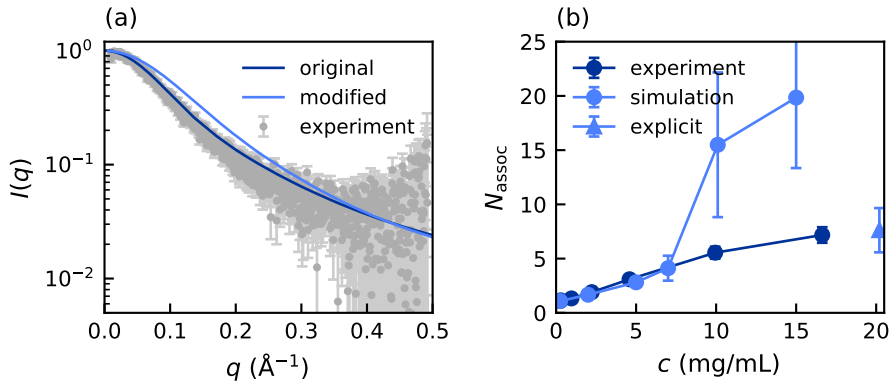


Figure 9.5: a) Form factor of statherin calculated in the original model and after inclusion of an additional short-ranged potential necessary for simulating self-association, compared to the experimentally determined form factor (collected by SEC-SAXS at pH 8 and 20 °C, with 20 mM Tris and 150 mM NaCl). b) Average association number against statherin concentration calculated from SAXS data (experimental) and determined from simulations at an ionic strength of 150 mM and 20 °C. The triangular data point is the result of a simulation using explicit salt.

the model was only applicable at low protein concentrations, as seen in Figure 9.5b. At high protein concentrations all protein chains aggregated into one large complex. This was discovered to depend on the implicit treatment of salt. With explicit salt no such breakdown was observed, which shows that the model performs better with a more accurate description of the electrostatic interactions than the extended Debye-Hückel potential. However, an explicit treatment of salt greatly increases the number of particles in the system and therefore poses larger demands on computational resources and the simulation software.

9.3 An atomistic approach to phosphorylated IDPs

The coarse-grained treatment of phosphorylated IDPs in Paper I suggested that depending on the number of phosphorylated residues and their distribution throughout the sequence, short-ranged attractive electrostatic interactions can have dramatic effects on the conformational ensemble. The discrepancies between simulations and experimental references motivated a more detailed investigation, using an atomistic approach. In addition, phosphorylation has been shown to be a versatile method for controlling protein function, as different IDPs have demonstrated varying conformational and structural response. It is therefore desirable to achieve a better understanding of phosphorylation effects.

Due to the computational expense of all-atom simulations, the 15 residue long N-terminal fragment of statherin, SN15, was chosen instead of the full protein for studying phosphorylation effects in Paper III. I selected two different force fields shown to work well

for short IDPs and which had parameters for phosphorylated residues available: i) Amber ff99SB-ILDN [84] with the TIP4P-D water model [64] and the phosaa10 parameter set for phosphorylated residues [157, 158] (A99), and ii) CHARMM36m [75] with the CHARMM-modified TIP3P water model [71] (C36). Note however that the Amber parameters had been developed for a preceding force field. For experimental reference, SAXS and CD spectroscopy were performed. The force fields were shown to be in good agreement for the non-phosphorylated peptide. R_g, R_{ee} and scattering curves were in excellent agreement, and the scattering curves also matched the experimental curve, see Figure 9.6a,b. On the contrary, for the phosphorylated peptide there were large discrepancies between the force fields

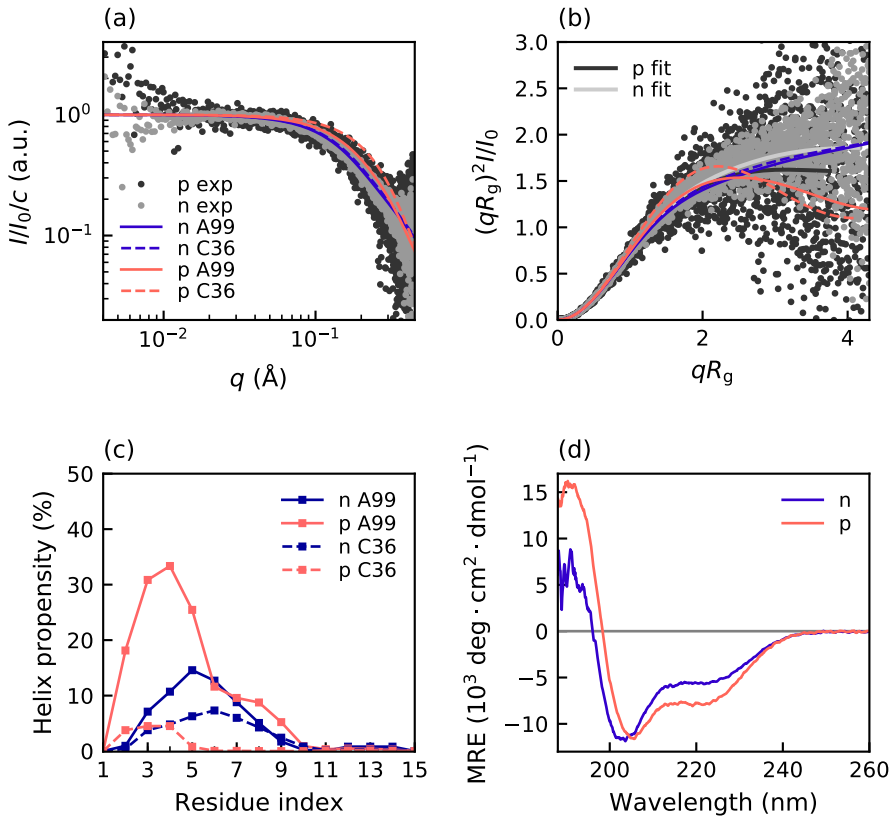


Figure 9.6: a) Form factor and b) dimensionless Kratky plot of non-phosphorylated (n) and phosphorylated (p) SN15 obtained by SAXS at 4 and 1.2 mg/mL, respectively, at 20 °C, 150 mM NaCl, 20 mM Tris, and pH 7.5, and from simulations using AMBER ff99SB-ILDN+TIP4P-D (A99) and CHARMM36m (C36). The lines "fit" correspond to the regularised curves fitted to the experimental SAXS data in the $P(r)$ determination. c) Helix propensity along the sequence for non-phosphorylated and phosphorylated SN15. d) CD spectra of non-phosphorylated and phosphorylated SN15 measured at 20 °C in 20 mM phosphate buffer with 150 mM NaF at pH 7.5, shown as the mean residue ellipticity *versus* wavelength.

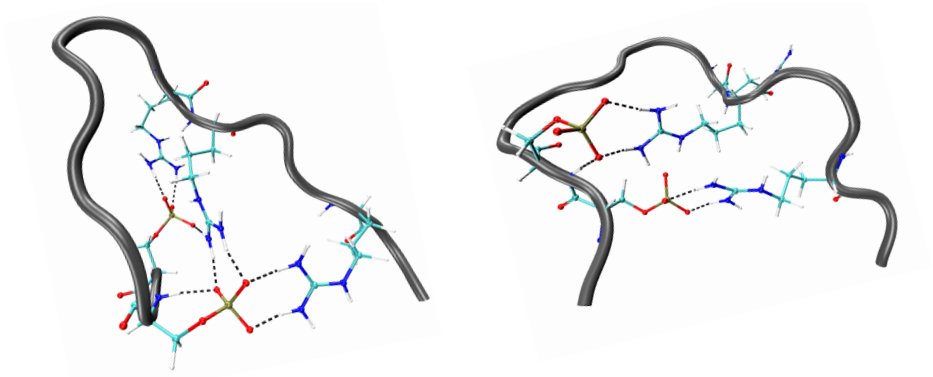


Figure 9.7: Two representative compact conformations of SN15 in CHARMM36m held together by strong salt bridges. All atoms are shown in the positively charged and phosphorylated residues. The black dashed lines represent hydrogen bonds.

regarding overall size, shape and secondary structure. C36 produced much more compact conformations, which were coupled to a higher occurrence of salt bridges between phosphorylated and positively charged residues, see Figure 9.7 for illustrative snapshots. These salt bridges also increased the content of bends in the peptide. The other main difference in secondary structure was the helical content. A substantial increase of α- and 3_{10}-helical content was observed upon phosphorylation in A99, but not in C36, as shown in Figure 9.6c. The differences in CD spectra between non-phoshorylated and phosphorylated SN15 shown in Figure 9.6d, supports an increase of α-helical structure. Both force fields gave a compaction of the peptide upon phosphorylation, however, the R_g determined from SAXS data for the non-phosphorylated and phosphorylated peptide were indistinguishable. Nonetheless, the Kratky plot indicated a small compaction upon phosphorylation, according to Figure 9.6b. Hence, a compaction in accordance with the simulations is plausible, but most probably not as large as in C36. To investigate whether the deficiencies of the force fields were general or specific to SN15, in Paper IV, the study was expanded to an additional four peptides, presented in Table 9.2.

Table 9.2: Full name, number of residues (N_{res}), phosphorylation sites (N_{ph}), positively charged residues (N_+), negatively charged residues (N_-), and net charge of the non-phosphorylated (Z_{no}) and phosphorylated variant (Z_{ph}) of the peptides studied throughout Paper III–V.

Name	Peptide	N_{res}	N_{ph}	N_+	N_-	Z_{no}	Z_{ph}
Tau1	$tau_{173-183}$	11	2	2	0	+2	-2
SN15	$statherin_{1-15}$	15	2	4	3	+1	-3
Tau2	$tau_{225-246}$	22	4	5	0	+5	-3
bCPP	β-$casein_{1-25}$	25	4	2	7	-5	-13
Stath	statherin	43	2	4	4	0	-4

C36 was shown to produce much more compact ensembles than A99 for all the phosphorylated peptides, see Figure 9.8. All peptides showed significantly higher probability of salt bridges in C36 than A99, which was the main reason behind the discrepancy between the force fields. In the 43 residue long statherin, where the phosphorylated and positively charged residues are all located within the first 13 residues, there was also another contribution. The C36 simulation contained more structures with β-strand and β-bridge formation between the middle and C-terminal end, and less structures where the protein was allowed more extended conformations. Additionally, all peptides contained a higher fraction of bends in C36, which in most cases could be linked to the salt bridges. Another noteworthy observation regarding secondary structure was that C36, in contrary to A99, did not sample any helical content at all in the N-terminal region of statherin. Although the N-terminal end of statherin is considered to be mainly irregular in water, helical propensity has been

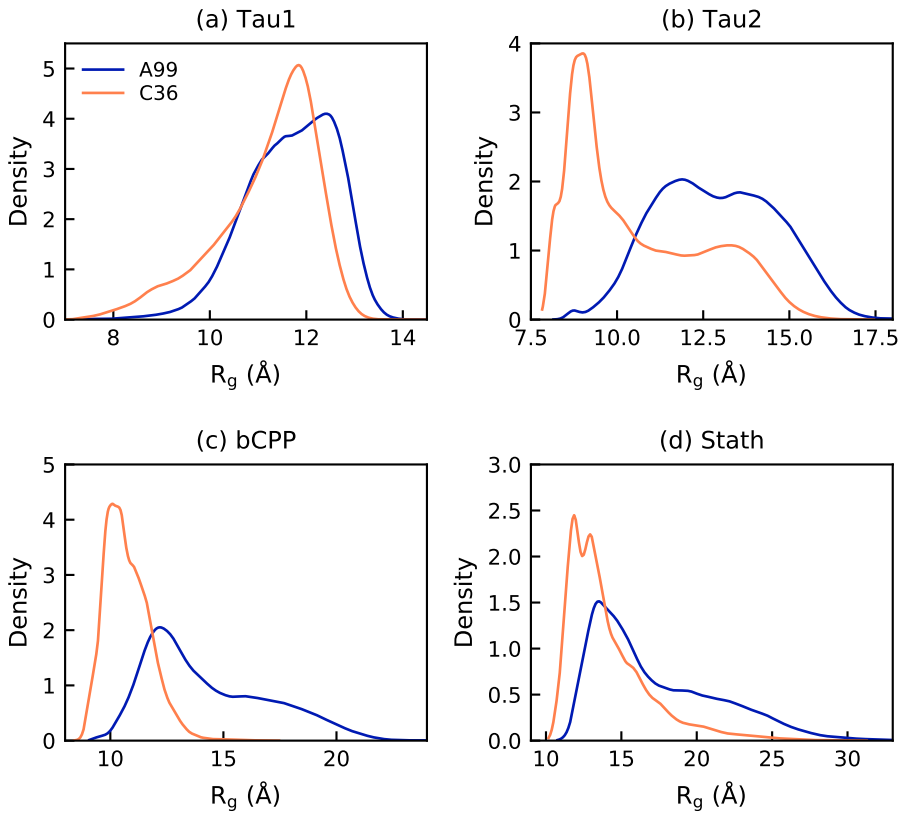


Figure 9.8: Radius of gyration distribution of a) Tau1, b) Tau2, c) bCPP and d) statherin, simulated with AMBER ff99SB-ILDN (A99) and CHARMM36m (C36).

detected in experiments [30, 159].

Noticing the large influence of salt bridges on the conformational ensemble, it was worth considering the influence of screening by addition of salt. These simulations have been performed in a salt-free environment, only with counterions to neutralise the system. So, for bCPP that showed the largest deviations between the force fields, in line with being the most charged peptide with the greatest separation of oppositely charged residues, additional simulations with 150 mM NaCl were performed. Although the probability of several salt bridges were greatly reduced in C36 when adding salt, the conformational ensemble did not change much, as was shown by comparing the R_g distributions (Figure 9.9a). In fact, the most probable conformations were still heavily influenced by salt bridges and the electrostatic interactions involving phosphorylated residues. In A99 only one salt bridge was significantly reduced, and the R_g distributions were highly similar. The calculated scattering curves were also indistinguishable in both force fields, see Figure 9.9b. Hence, the inclusion of 150 mM salt had little to no effect on the conformational ensemble, and the salt bridges were still of importance. It has been indicated that many force fields have a tendency to overestimate salt bridges [85, 141, 160, 161], hence, it is possible that both A99 and C36 overestimate the importance of salt bridges in phosphorylated IDPs. Compared to available experimental data for the shortest peptide Tau1 and the longest IDP statherin, A99 appeared as the better choice for simulating phosphorylated IDPs. However, for a better evaluation of the force fields, more experimental data is needed. Here NMR plays an important role, by being able to detect secondary secondary structure propensity for individual residues and salt bridges by scalar couplings, chemical shifts and NOEs.

In Paper v the A99 force field was used to also simulate the non-phosphorylated variants of

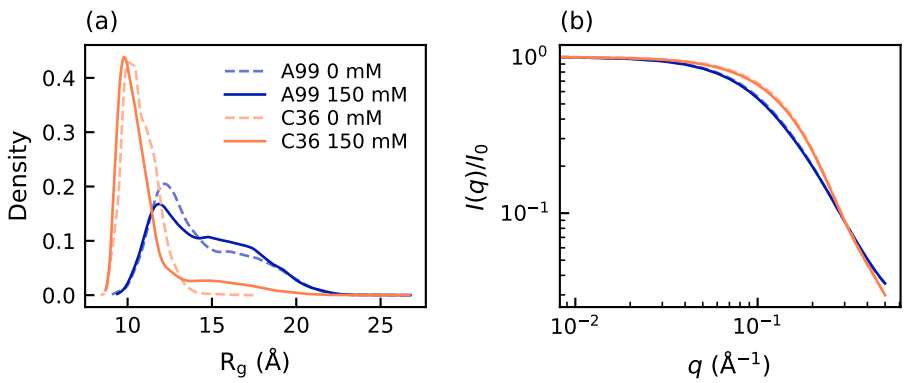


Figure 9.9: a) Radius of gyration distribution and b) calculated form factor of bCPP simulated with Amber ff99SB-ILDN (A99) or CHARMM36m (C36) in the presence of 0 or 150 mM NaCl.

the peptides, to study the conformational and structural effects induced by phosphorylation. To fully observe the electrostatic effects, the simulations were performed without additional salt. However, complementary simulations of bCPP at 150 mM demonstrated that phosphorylation effects still remained at 150 mM, although slightly diminished. Recently it was hypothesised that the global conformational changes could be predicted from the net charge of an IDP in non-phosphorylated state, such that a positively charged IDP contracts, while a neutral or negatively charge IDP expands [141]. Both Tau1 and bCPP were shown to contradict this hypothesis, see Table 9.3. In bCPP the electrostatic attraction between the arginine termini residues and the phosphorylated region drove a contraction of the peptide (see Figure 9.10), despite a local expansion of region E13–E21, containing the phosphorylation sites. Salt bridge formation between arginine/lysine and phosphorylated residues was indeed shown to be a major reason behind compaction upon phosphorylation in SN15, Tau2, and bCPP. Another contributing factor in SN15 and Tau2 was helix formation. These peptides, as well as statherin, which also exhibited increased helix propensity upon phosphorylation, all have a lysine three or four steps away from the phosphorylated residue, a pattern known to stabilise helices through salt bridge formation between the side groups [162].

In statherin, phosphorylation induced a compaction of the first 15 residues, but an overall expansion. The expansion was not caused by electrostatic repulsion, but instead explained by the preference of forming arginine-phosphoserine salt bridges over arginine–tyrosine cation–π-interaction. In non-phosphorylated statherin, arginine–tyrosine interaction caused β-sheet formation, which disappeared upon phosphorylation, when the arginine residues instead became involved in salt bridges with phosphoserine. The disruption of the β-sheet caused a global expansion. Relating back to Paper I, it is worth noticing that these effects are not captured by the coarse-grained model, since it only includes electrostatic effects between charged residues. In fact, the coarse-grained model provides a small decrease in R_g upon phosphorylation, originating from the compaction of the N-terminal region where the phosphorylated residues reside.

To conclude, the studies performed in Paper III-V showed that phosphorylation induces changes in both overall dimensions and structural content, and that salt bridge formation

Table 9.3: Net charge of the non-phosphorylated peptide and mean radius of gyration (R_g) and end-to-end distance (R_{ee}) of the non-phosphorylated (n) and phosphorylated (p) variants.

		R_g (Å)		R_{ee} (Å)	
Peptide	Net charge	n	p	n	p
Tau1	+2	9.3 ± 0.1	9.8 ± 0.1	27.4 ± 0.6	28.9 ± 0.2
SN15	+1	10.0 ± 0.1	9.0 ± 0.1	25.4 ± 0.9	23.0 ± 0.3
Tau2	+5	14.6 ± 0.2	12.9 ± 0.3	38.3 ± 0.9	32.7 ± 1.7
bCPP	-5	15.3 ± 0.3	14.3 ± 0.3	38.0 ± 0.8	30.9 ± 1.5
Stath	0	15.6 ± 0.4	17.3 ± 0.9	33.0 ± 0.4	40.5 ± 1.7

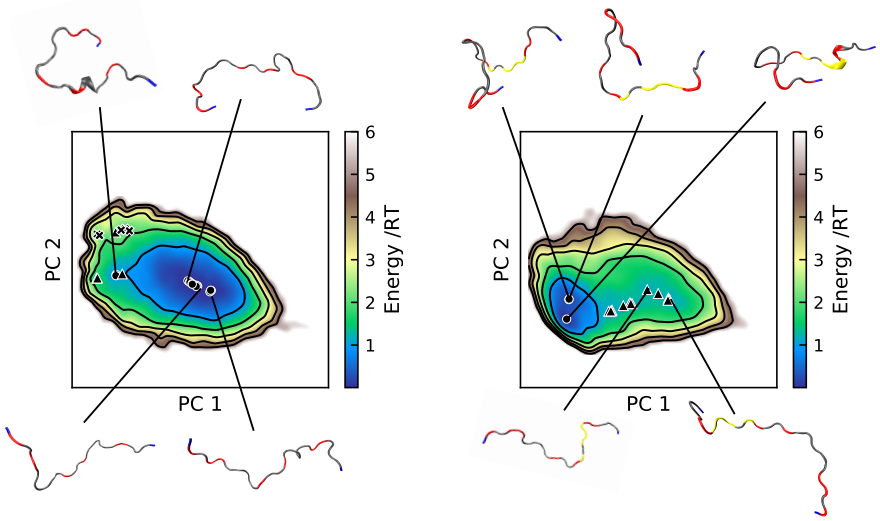


Figure 9.10: Energy landscapes with conformations in selected minima of bCPP for non-phosphorylated (left) and phosphorylated (right) bCPP. The energy landscapes were constructed using the first two components from principal component analysis, using the same basis set for both variants. Hence, they are directly comparable. Contour lines are drawn for integer energy levels in the interval $1 \leq RT \leq 5$ and the minimum of each basin is represented by a marker depending on the energy: ●: $\leq 1RT$, ▲: $\leq 2RT$, ✖: $\leq 3RT$. In the conformations positively charged residues are shown in blue, negatively charged residues in red and phosphorylated residues in yellow.

is an important contributor to this. Vast over-stabilisation of salt bridges was shown to have large effects on the global dimensions, demonstrating the need for revised force fields. Also at 150 mM salt did salt bridges between phosphorylated and positively charged residues influence the conformational ensemble. It was shown that only considering net charge is not enough for predicting the outcome of phosphorylation, and that also non-charged residues can be of importance. Atomistic simulations show great potential in providing deeper knowledge regarding the effect of phosphorylation, however, more experimental studies at both global and local length-scales are required for further revision and validation of force fields.

9.4 Conclusions and outlook

The overall objective of this thesis has been to investigate the conformational ensembles exhibited by IDPs in solution, to explore the dependence on sequence, especially the impact of phosphorylated residues. Due to the conformational polydispersity exhibited by IDPs, it is challenging to extract detailed information from experiments, but combining different experimental and computational techniques has proven to be a fruitful approach. Since a computational approach is dependent on appropriate models, a significant part of the work

has been focused on investigating how models and force fields perform.

One property characterising a great model is it being as simple as possible, but still describing the phenomenon of interest. In this way, it can act as an explanatory tool. The coarse-grained "one bead per residue model" relying on excluded volume, electrostatic interactions and an approximate van der Waals interaction was shown to reproduce R_g for a range of different IDPs under dilute conditions, implying that many IDPs can be thought of as self-avoiding random walks influenced by electrostatic interactions. From this model, a basic understanding of how chain length, charge distribution and salt concentration affects the conformational ensemble can be achieved. Furthermore, with the addition of a hydrophobic interaction, the model was shown to qualitatively describe the self-association process of statherin and provided a deeper understanding of the balance of interactions. This demonstrates that the model is applicable also in larger and more complex systems, where coarse-grained approaches are currently the only feasible option considering the computational expense versus resources. Other adaptations of the original model have also been applied to studies of crowding [123, 163] and zinc-initiated oligomerisation [164], showcasing the potential and adaptability of this model within the field of IDP research. However, all models come with limitations. Here it was shown that the model in current form could not simultaneously provide a good representation of both size and level of stiffness for the proline-rich proteins and that the size of the highly phosphorylated IDPs was underestimated. Since IDPs are a very diverse group of proteins, it is by no means surprising that not all IDPs can be described by this model. For the phosphorylated proteins, better agreement was achieved with a reduced charge of the phosphorylated residues. It is therefore of interest to further explore whether this is due to an overestimation of electrostatic interactions in the model, ill-matching of the experimental conditions or if a fixed charge of $-2e$ is a poor representation of the charge state of phosphorylated residues at physiological pH. Also, in the simulations of self-association, the implicit treatment of salt caused the model to break down at higher protein concentrations. While an explicit treatment of salt provides better results, it comes with a larger computational cost and limits to the accessible system size.

Regarding the effects of phosphorylation, this problem required a more detailed model. Atomistic simulations were shown to detect changes in global compaction and secondary structure, and relate them to interactions between specific residues. Especially salt bridges between phosphorylated and positively charged residues were shown to have major impact on the conformational ensemble, which highlighted the importance of having force fields that accurately estimate the strength of salt bridges. Other force field deficiencies regarding secondary structure were also detected. In the continued strive for understanding the implications of phosphorylation of IDPs, it is therefore important to revise force fields, and to especially consider the strength of salt bridges involving phosphorylated residues. Therefore, the collection of more experimental data suitable for use as benchmarking is also required, which extends beyond the techniques applied in this work. NMR was men-

tioned as an example, which has the advantage that scalar couplings and chemical shifts can be calculated from simulations, which facilitates comparison. The interplay between arginines, tyrosines and phosphorylated residues implied by the atomistic simulations of statherin is of specific interest to explore further. In addition, a systematic investigation varying the number of phosphorylated residues and their position in relation to positively charged residues in a controlled manner is suggested for gaining a better understanding of underlaying factors controlling the outcome of phosphorylation.

While this thesis has been focused on the relation between sequence and structure, an area where much is yet to be explored, the link to function is equally important to consider. Since the functionality often involves interaction with binding partners or surfaces, there is a requirement for computational models to handle such situations. Also in this context can statherin be used as a model protein, as binding to hydroxyapatite has been shown to induce more helix formation in the N-terminal end [165, 166] and expose a bacterial binding site in the C-terminal tail [166, 167].

As a final remark, one of the greatest lessons I have learned during these years of research is that it is not at all straightforward to compare experimental and simulation data and draw correct conclusions from it. Here I see great advantages of having practical experience of both parts, as it provides better comprehension of what can affect the data and what is actually compared.

References

[1] R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, and M. M. Babu, "Classification of intrinsically disordered regions and proteins," *Chem. Rev.*, vol. 114, no. 13, pp. 6589–6631, 2014.

[2] C. J. Oldfield and A. K. Dunker, "Intrinsically disordered proteins and intrinsically disordered protein regions," *Annu. Rev. Biochem.*, vol. 83, no. 1, pp. 553–584, 2014.

[3] P. E. Wright and H. Dyson, "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm," *J. Mol. Biol.*, vol. 293, no. 2, pp. 321 – 331, 1999.

[4] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovićá, "Intrinsic disorder and protein function," *Biochemistry*, vol. 41, no. 21, pp. 6573–6582, 2002.

[5] V. N. Uversky and A. K. Dunker, "Understanding protein non-folding," *Biochim. Biophys. Acta, Proteins Proteomics*, vol. 1804, no. 6, pp. 1231 – 1264, 2010.

[6] J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*. New York, USA: W. H. Freeman and Company, international 7th ed., 2011.

[7] Y. Mansiaux, A. P. Joseph, J.-C. Gelly, and A. G. de Brevern, "Assignment of polyproline ii conformation and analysis of sequence – structure relationship," *PLOS ONE*, vol. 6, pp. 1–15, 03 2011.

[8] K. A. Dill, "Dominant forces in protein folding," *Biochemistry*, vol. 29, no. 31, pp. 7133–7155, 1990.

[9] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker, "Sequence complexity of disordered protein," *Proteins*, vol. 42, no. 1, pp. 38–48, 2001.

[10] S. Vucetic, C. J. Brown, A. K. Dunker, and Z. Obradovic, "Flavors of protein disorder," *Proteins*, vol. 52, no. 4, pp. 573–584, 2003.

[11] A. K. Dunker, P. Romero, Z. Obradovic, E. C. Garner, and C. J. Brown, "Intrinsic protein disorder in complete genomes," *Genome Inform.*, vol. 11, pp. 161–171, 2000.

[12] P. Romero, Z. Obradovic, C. Kissinger, J. Villafranca, E. Garner, S. Guilliot, and A. Dunker, "Thousands of proteins likely to have long disordered regions," *Pac. Symp. Biocomput.*, vol. 3, pp. 437–448, 1998.

[13] J. Ward, J. Sodhi, L. McGuffin, B. Buxton, and D. Jones, "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life," *J. Mol. Biol.*, vol. 337, no. 3, pp. 635–645, 2004.

[14] B. Xue, A. K. Dunker, and V. N. Uversky, "Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life," *J. Biomol. Struct. Dyn.*, vol. 30, no. 2, pp. 137–149, 2012.

[15] H. J. Dyson and P. E. Wright, "Intrinsically unstructured proteins and their functions," *Nat. Rev. Mol. Cell Biol.*, vol. 6, pp. 197–208, 2005.

[16] P. Tompa, "Intrinsically disordered proteins: a 10-year recap," *Trends Biochem. Sci.*, vol. 37, no. 12, pp. 509 – 516, 2012.

[17] J. Liu, J. R. Faeder, and C. J. Camacho, "Toward a quantitative theory of intrinsically disordered proteins and their function," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 106, no. 47, pp. 19819–19823, 2009.

[18] P. E. Wright and H. J. Dyson, "Linking folding and binding," *Curr. Opin. Struct. Biol.*, vol. 19, no. 1, pp. 31–38, 2009.

[19] V. N. Uversky, C. J. Oldfield, and A. K. Dunker, "Intrinsically disordered proteins in human diseases: Introducing the d2 concept," *Annu. Rev. Biophys.*, vol. 37, no. 1, pp. 215–246, 2008.

[20] V. N. Uversky, V. Davé, L. M. Iakoucheva, P. Malaney, S. J. Metallo, R. R. Pathak, and A. C. Joerger, "Pathological unfoldomics of uncontrolled chaos: Intrinsically disordered proteins and human diseases," *Chem. Rev.*, vol. 114, no. 13, pp. 6844–6879, 2014.

[21] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic, "Intrinsically disordered protein," *J. Mol. Graphics Modell.*, vol. 19, no. 1, pp. 26–59, 2001.

[22] V. N. Uversky, "Unusual biophysics of intrinsically disordered proteins," *Biochim. Biophys. Acta, Proteins Proteomics*, vol. 1834, no. 5, pp. 932–951, 2013.

[23] R. K. Das, K. M. Ruff, and R. V. Pappu, "Relating sequence encoded information to form and function of intrinsically disordered proteins," *Curr. Opin. Struct. Biol.*, vol. 32, pp. 102–112, 2015. New constructs and expression of proteins / Sequences and topology.

[24] R. K. Das and R. V. Pappu, "Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 110, no. 33, pp. 13392–13397, 2013.

[25] L. M. Iakoucheva, P. Radivojac, C. J. Brown, T. R. O'Connor, J. G. Sikes, Z. Obradovic, and A. K. Dunker, "The importance of intrinsic disorder for protein phosphorylation," *Nucleic Acids Res.*, vol. 32, pp. 1037–1049, 02 2004.

[26] J. Gao and D. Xu, *Biocomputing 2012*, ch. Correlation Between Posttranslational Modification and Intrinsic Disorder in Protein, pp. 94–103. World Scientific Publishing Co. Pte. Ltd., 2012.

[27] L. N. Johnson and R. J. Lewis, "Structural basis for control by phosphorylation," *Chem. Rev.*, vol. 101, no. 8, pp. 2209–2242, 2001.

[28] C. X. Gong and K. Iqbal, "Hyperphosphorylation of microtubule-associated protein tau: a promising therapeutic target for alzheimer disease," *Curr. Med. Chem.*, vol. 15, no. 23, pp. 2321–2328, 2008.

[29] C. G. De Kruif and C. Holt, *Casein Micelle Structure, Functions and Interactions*, pp. 233–276. Boston, MA: Springer US, 2003.

[30] P. A. Raj, M. Johnsson, M. J. Levine, and G. H. Nancollas, "Salivary statherin. Dependence on sequence, charge, hydrogen bonding potency, and helical conformation for adsorption to hydroxyapatite and inhibition of mineralization.," *J. Biol. Chem.*, vol. 267, no. 9, pp. 5968–76, 1992.

[31] K. Makrodimitris, D. L. Masica, E. T. Kim, and J. J. Gray, "Structure prediction of protein–solid surface interactions reveals a molecular recognition motif of statherin for hydroxyapatite," *J. Am. Chem. Soc.*, vol. 129, no. 44, pp. 13713–13722, 2007.

[32] J. A. Loo, W. Yan, P. Ramachandran, and D. T. Wong, "Comparative human salivary and plasma proteomes," *J. Dent. Res.*, vol. 89, no. 10, pp. 1016–1023, 2010.

[33] M. Edgar, C. Dawes, and D. O'Mullane, eds., *Saliva and Oral Health*. London, UK: British Dental Association, 3rd ed., 2004.

[34] W. Siqueira, W. Custodio, and E. McDonald, "New insights into the composition and functions of the acquired enamel pellicle," *J. Dent. Res.*, vol. 91, no. 12, pp. 1110–1118, 2012.

[35] M. J. Levine, "Development of artificial salivas," *Crit. Rev. Oral Biol. Med.*, vol. 4, no. 3, pp. 279–286, 1993.

[36] E. Moreno and R. Zahradnik, "Demineralization and remineralization of dental enamel," *J. Dent. Res.*, vol. 58, no. 2_suppl, pp. 896–903, 1979.

[37] D. Hay, D. Smith, S. Schluckebier, and E. Moreno, "Basic biological sciences relationship between concentration of human salivary statherin and inhibition of calcium phosphate precipitation in stimulated human parotid saliva," *J. Dent. Res.*, vol. 63, no. 6, pp. 857–863, 1984.

[38] M. A. Buzalaf, A. R. Hannas, and M. T. Kato, "Saliva and dental erosion," *J. Appl. Oral Sci.*, vol. 20, no. 5, pp. 493–502, 2012.

[39] W. H. Douglas, E. S. Reeh, N. Ramasubbu, P. A. Raj, K. K. Bhandary, and M. J. Levine, "Statherin: A major boundary lubricant of human saliva," *Biochem. Biophys. Res. Commun.*, vol. 180, no. 1, pp. 91 – 97, 1991.

[40] R. J. Gibbons and D. I. Hay, "Human salivary acidic proline-rich proteins and statherin promote the attachment of actinomyces viscosus LY7 to apatitic surfaces.," *Infect. Immun.*, vol. 56, no. 2, pp. 439–445, 1988.

[41] A. Amano, K. Kataoka, P. A. Raj, R. J. Genco, and S. Shizukuishi, "Binding sites of salivary statherin for porphyromonas gingivalis recombinant fimbrillin," *Infect. Immun.*, vol. 64, no. 10, pp. 4249–4254, 1996.

[42] H. Nagata, A. Sharma, H. T. Sojar, A. Amano, M. J. Levine, and R. J. Genco, "Role of the carboxyl-terminal region of porphyromonas gingivalis fimbrillin in binding to salivary proteins," *Infect. Immun.*, vol. 65, no. 2, pp. 422–427, 1997.

[43] D. H. Schlesinger and D. I. Hay, "Complete covalent structure of statherin, a tyrosine-rich acidic peptide which inhibits calcium phosphate precipitation from human parotid saliva," *J. Biol. Chem.*, vol. 252, no. 5, pp. 1689–1695, 1977.

[44] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.*, vol. 157, no. 1, pp. 105–132, 1982.

[45] C. Holt, "Unfolded phosphopolypeptides enable soft and hard tissues to coexist in the same organism with relative ease," *Curr. Opin. Struct. Biol.*, vol. 23, no. 3, pp. 420–425, 2013. New contructs and expressions of proteins / Sequences and topology.

[46] Y. Lin, S. L. Currie, and M. K. Rosen, "Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs," *J. Biol. Chem.*, vol. 292, no. 46, pp. 19110–19120, 2017.

[47] C. W. Pak, M. Kosno, A. S. Holehouse, S. B. Padrick, A. Mittal, R. Ali, A. A. Yunus, D. Liu, R. V. Pappu, and M. K. Rosen, "Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein," *Mol. Cell*, vol. 63, no. 1, pp. 72–85, 2016.

[48] E. Rieloff, M. D. Tully, and M. Skepö, "Assessing the intricate balance of intermolecular interactions upon self-association of intrinsically disordered proteins," *J. Mol. Biol.*, vol. 431, no. 3, pp. 511–523, 2019.

[49] J. N. Israelachvili, *Intermolecular and Surface Forces*. Oxford, UK: Academic Press, Elsevier, 3rd ed., 2011.

[50] M. T. A. Evans, M. C. Phillips, and M. N. Jones, "The conformation and aggregation of bovine β-casein a. II. Thermodynamics of thermal association and the effects of changes in polar and apolar interactions on micellization," *Biopolymers*, vol. 18, no. 5, pp. 1123–1140, 1979.

[51] K. Takase, R. Niki, and S. Arima, "A sedimentation equilibrium study of the temperature-dependent association of bovine β-casein," *Biochim. Biophys. Acta, Proteins Proteomics*, vol. 622, no. 1, pp. 1–8, 1980.

[52] J. O'Connell, V. Grinberg, and C. de Kruif, "Association behavior of β-casein," *J. Colloid Interface Sci.*, vol. 258, no. 1, pp. 33–39, 2003.

[53] I. Portnaya, U. Cogan, Y. D. Livney, O. Ramon, K. Shimoni, M. Rosenberg, and D. Danino, "Micellization of bovine β-casein studied by isothermal titration microcalorimetry and cryogenic transmission electron microscopy," *J. Agric. Food Chem.*, vol. 54, no. 15, pp. 5555–5561, 2006.

[54] C. Moitzi, I. Portnaya, O. Glatter, O. Ramon, and D. Danino, "Effect of temperature on self-assembly of bovine β-casein above and below isoelectric pH. Structural analysis by cryogenic-transmission electron microscopy and small-angle x-ray scattering," *Langmuir*, vol. 24, no. 7, pp. 3020–3029, 2008.

[55] D. Chandler, "Hydrophobicity: Two faces of water," *Nature*, vol. 417, no. 491, pp. 493–502, 2002.

[56] T. L. Hill, *An Introduction to Statistical Thermodynamics*. Reading, MA, USA: Addison-Wesley Publishing Company, 2nd ed., 1962.

[57] C. Cragnell, D. Durand, B. Cabane, and M. Skepö, "Coarse-grained modeling of the intrinsically disordered protein histatin 5 in solution: Monte carlo simulations in combination with saxs," *Proteins*, vol. 84, no. 6, pp. 777–791, 2016.

[58] H. Berendsen, D. van der Spoel, and R. van Drunen, "Gromacs: A message-passing parallel molecular dynamics implementation," *Comput. Phys. Commun.*, vol. 91, no. 1, pp. 43–56, 1995.

[59] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation," *J. Chem. Theory Comput.*, vol. 4, no. 3, pp. 435–447, 2008.

[60] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl, "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit," *Bioinformatics*, vol. 29, pp. 845–854, 02 2013.

[61] S. Páll, M. J. Abraham, C. Kutzner, B. Hess, and E. Lindahl, "Tackling exascale software challenges in molecular dynamics simulations with gromacs," in *Solving Software Challenges for Exascale* (S. Markidis and E. Laure, eds.), (Cham), pp. 3–27, Springer International Publishing, 2015.

[62] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1-2, pp. 19–25, 2015.

[63] G. P. Moss, "Basic terminology of stereochemistry (IUPAC recommendations 1996)," *Pure Appl. Chem.*, vol. 68, no. 12, pp. 2193–2222, 1996.

[64] S. Piana, A. G. Donchev, P. Robustelli, and D. E. Shaw, "Water dispersion interactions strongly influence simulated structural properties of disordered protein states," *J. Phys. Chem. B*, vol. 119, no. 16, pp. 5113–5123, 2015.

[65] S. Rauscher, V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. de Groot, and H. Grubmüller, "Structural ensembles of intrinsically disordered proteins depend strongly on force field: A comparison to experiment," *J. Chem. Theory Comput.*, vol. 11, no. 11, pp. 5513–5524, 2015.

[66] J. Henriques and M. Skepö, "Molecular dynamics simulations of intrinsically disordered proteins: On the accuracy of the TIP4P-D water model and the representativeness of protein disorder models," *J. Chem. Theory Comput.*, vol. 12, no. 7, pp. 3407–3415, 2016.

[67] A. V. Onufriev and S. Izadi, "Water models for biomolecular simulations," *WIREs Comput. Mol. Sci.*, vol. 8, no. 2, p. e1347, 2018.

[68] W. L. Jorgensen, "Transferable intermolecular potential functions for water, alcohols, and ethers. application to liquid water," *J. Am. Chem. Soc.*, vol. 103, pp. 335–340, 1981.

[69] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, 1983.

[70] S. R. Durell, B. R. Brooks, and A. Ben-Naim, "Solvent-induced forces between two hydrophilic groups," *J. Phys. Chem.*, vol. 98, pp. 2198–2202, 1994.

[71] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins," *J. Phys. Chem. B*, vol. 102, no. 18, pp. 3586–3616, 1998.

[72] J. L. F. Abascal and C. Vega, "A general purpose model for the condensed phases of water: TIP4P/2005," *J. Chem. Phys.*, vol. 123, no. 23, p. 234505, 2005.

[73] O. Guvench and A. D. MacKerell, *Comparison of Protein Force Fields for Molecular Dynamics Simulations*, pp. 63–88. Totowa, NJ: Humana Press, 2008.

[74] S. Boonstra, P. R. Onck, and E. van der Giessen, "CHARMM TIP3P water model suppresses peptide folding by solvating the unfolded state," *J. Phys. Chem. B*, vol. 120, no. 15, pp. 3692–3698, 2016.

[75] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller, and A. D. MacKerell Jr, "CHARMM36m: an improved force field for folded and intrinsically disordered proteins," *Nat. Methods.*, vol. 14, no. 1, pp. 71–73, 2017.

[76] R. B. Best, N.-V. Buchete, and G. Hummer, "Are current molecular dynamics force fields too helical?," *Biophys. J.*, vol. 95, no. 1, pp. L07–L09, 2008.

[77] W. Wang, W. Ye, C. Jiang, R. Luo, and H.-F. Chen, "New force field on modeling intrinsically disordered proteins," *Chem. Biol. Drug. Des.*, vol. 84, no. 3, pp. 253–269, 2014.

[78] Y. Zhang, H. Liu, S. Yang, R. Luo, and H.-F. Chen, "Well-balanced force field ff03CMAP for folded and disordered proteins," *J. Chem. Theory Comput.*, vol. 15, no. 12, pp. 6769–6780, 2019.

[79] S. Piana, J. L. Klepeis, and D. E. Shaw, "Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular

dynamics simulations," *Curr. Opin. Struct. Biol.*, vol. 24, pp. 98–105, 2014. Folding and binding / Nucleic acids and their protein complexes.

[80] J. Henriques, C. Cragnell, and M. Skepö, "Molecular dynamics simulations of intrinsically disordered proteins: Force field evaluation and comparison with experiment," *J. Chem. Theory Comput.*, vol. 11, no. 7, pp. 3420–3431, 2015.

[81] A. D. Mackerell Jr., M. Feig, and C. L. Brooks III, "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations," *J. Comput. Chem.*, vol. 25, no. 11, pp. 1400–1415, 2004.

[82] R. B. Best and G. Hummer, "Optimized molecular dynamics force fields applied to the helix–coil transition of polypeptides," *J. Phys. Chem. B*, vol. 113, no. 26, pp. 9004–9015, 2009.

[83] R. B. Best and J. Mittal, "Protein simulations with an optimized water model: Cooperative helix formation and temperature-induced unfolded state collapse," *J. Phys. Chem. B*, vol. 114, no. 46, pp. 14916–14923, 2010.

[84] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the amber ff99sb protein force field," *Proteins*, vol. 78, no. 8, pp. 1950–1958, 2010.

[85] S. Piana, K. Lindorff-Larsen, and D. Shaw, "How robust are protein folding simulations with respect to force field parameterization?," *Biophys. J.*, vol. 100, no. 9, pp. L47–L49, 2011.

[86] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell, "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ_1 and χ_2 dihedral angles," *J. Chem. Theory Comput.*, vol. 8, no. 9, pp. 3257–3273, 2012.

[87] R. B. Best, W. Zheng, and J. Mittal, "Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association," *J. Chem. Theory Comput.*, vol. 10, no. 11, pp. 5113–5124, 2014.

[88] F. Jiang, C.-Y. Zhou, and Y.-D. Wu, "Residue-specific force field based on the protein coil library. RSFF1: Modification of OPLS-AA/L," *J. Phys. Chem. B*, vol. 118, no. 25, pp. 6983–6998, 2014.

[89] C.-Y. Zhou, F. Jiang, and Y.-D. Wu, "Residue-specific force field based on protein coil library. rsff2: Modification of amber ff99sb," *J. Phys. Chem. B*, vol. 119, no. 3, pp. 1035–1047, 2015.

[90] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb," *J. Chem. Theory Comput.*, vol. 11, no. 8, pp. 3696–3713, 2015.

[91] D. Song, R. Luo, and H.-F. Chen, "The idp-specific force field ff14idpsff improves the conformer sampling of intrinsically disordered proteins," *J. Chem. Inf. Model.*, vol. 57, no. 5, pp. 1166–1178, 2017.

[92] P. Robustelli, S. Piana, and D. E. Shaw, "Developing a molecular dynamics force field for both folded and disordered protein states," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 115, no. 21, pp. E4758–E4766, 2018.

[93] H. Liu, D. Song, H. Lu, R. Luo, and H.-F. Chen, "Intrinsically disordered protein-specific force field CHARMM36IDPSFF," *Chem. Biol. Drug. Des.*, vol. 92, no. 4, pp. 1722–1735, 2018.

[94] H. Liu, D. Song, Y. Zhang, S. Yang, R. Luo, and H.-F. Chen, "Extensive tests and evaluation of the CHARMM36IDPSFF force field for intrinsically disordered proteins and folded proteins," *Phys. Chem. Chem. Phys.*, vol. 21, pp. 21918–21931, 2019.

[95] S. Yang, H. Liu, Y. Zhang, H. Lu, and H. Chen, "Residue-specific force field improving the sample of intrinsically disordered proteins and folded proteins," *J. Chem. Inf. Model.*, vol. 59, no. 11, pp. 4793–4805, 2019.

[96] J. Mu, H. Liu, J. Zhang, R. Luo, and H.-F. Chen, "Recent force field strategies for intrinsically disordered proteins," *J. Chem. Inf. Model.*, vol. 61, no. 3, pp. 1037–1047, 2021.

[97] J. Huang and A. D. MacKerell, "Force field development and simulations of intrinsically disordered proteins," *Curr. Opin. Struct. Biol.*, vol. 48, pp. 40–48, 2018. Folding and binding in silico, in vitro and in cellula • Proteins: An Evolutionary Perspective.

[98] S.-H. Chong, P. Chatterjee, and S. Ham, "Computer simulations of intrinsically disordered proteins," *Annu. Rev. Phys. Chem.*, vol. 68, no. 1, pp. 117–134, 2017.

[99] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, 1953.

[100] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*. San Diego, CA, USA: Academic Press, 2nd ed., 2002.

[101] J. Reščič and P. Linse, "MOLSIM: A modular molecular simulation software," *J. Comput. Chem.*, vol. 36, no. 16, pp. 1259–1274, 2015.

[102] M. Allen and D. Tildesley, *Computer Simulation of Liquids*. Oxford University Press, 1989.

[103] M. Abraham, B. Hess, D. van der Spoel, and E. Lindahl, *GROMACS Reference Manual version 2018.4*. The GROMACS development teams, www.gromacs.org.

[104] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N·log(N) method for ewald sums in large systems," *J. Chem. Phys.*, vol. 98, no. 12, pp. 10089–10092, 1993.

[105] C. W. Hopkins, S. Le Grand, R. C. Walker, and A. E. Roitberg, "Long-time-step molecular dynamics through hydrogen mass repartitioning," *J. Chem. Theory Comput.*, vol. 11, no. 4, pp. 1864–1874, 2015.

[106] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "Lincs: A linear constraint solver for molecular simulations," *J. Comput. Chem.*, vol. 18, no. 12, pp. 1463–1472, 1997.

[107] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," *J. Chem. Phys.*, vol. 126, no. 1, p. 014101, 2007.

[108] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *J. Appl. Phys.*, vol. 52, no. 12, pp. 7182–7190, 1981.

[109] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *J. Chem. Phys.*, vol. 81, no. 8, pp. 3684–3690, 1984.

[110] D. Svergun, C. Barberato, and M. H. J. Koch, "Crysol– a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates," *J. Appl. Crystallogr.*, vol. 28, no. 6, pp. 768–773, 1995.

[111] J. Henriques, L. Arleth, K. Lindorff-Larsen, and M. Skepö, "On the calculation of SAXS profiles of folded and intrinsically disordered proteins from computer simulations," *J. Mol. Biol.*, vol. 430, no. 16, pp. 2521–2539, 2018. Intrinsically Disordered Proteins.

[112] P. Chen and J. Hub, "Validating solution ensembles from molecular dynamics simulation by wide-angle X-ray scattering data," *Biophys. J.*, vol. 107, no. 2, pp. 435–447, 2014.

[113] Y. Hayashi, M. Ullner, and P. Linse, "Complex formation in solutions of oppositely charged polyelectrolytes at different polyion compositions and salt content," *J. Phys. Chem. B*, vol. 107, no. 32, pp. 8198–8207, 2003.

[114] H. Arkın and W. Janke, "Gyration tensor based analysis of the shapes of polymer chains in an attractive spherical cage," *J. Chem. Phys.*, vol. 138, no. 5, p. 054904, 2013.

[115] M. Kenward and M. D. Whitmore, "A systematic monte carlo study of self-assembling amphiphiles in solution," *J. Chem. Phys.*, vol. 116, no. 8, pp. 3455–3470, 2002.

[116] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[117] R. Chebrek, S. Leonard, A. G. de Brevern, and J.-C. Gelly, "PolyprOnline: polyproline helix II and secondary structure assignment database," *Database*, vol. 2014, 11 2014. bau102.

[118] D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment," *Proteins*, vol. 23, no. 4, pp. 566–579, 1995.

[119] W. Humphrey, A. Dalke, and K. Schulten, "VMD – Visual Molecular Dynamics," *J. Mol. Graph.*, vol. 14, pp. 33–38, 1996.

[120] Y. Zhang and C. Sagui, "Secondary structure assignment for conformationally irregular peptides: Comparison between DSSP, STRIDE and KAKSI," *J. Mol. Graph. Model.*, vol. 55, pp. 72–84, 2015.

[121] L. Mavridis and R. W. Janes, "PDB2CD: a web-based application for the generation of circular dichroism spectra from protein atomic coordinates," *Bioinformatics*, vol. 33, pp. 56–63, 09 2016.

[122] G. Nagy, M. Igaev, N. C. Jones, S. V. Hoffmann, and H. Grubmüller, "Sesca: Predicting circular dichroism spectra from protein molecular structures," *J. Chem. Theory Comput.*, vol. 15, no. 9, pp. 5087–5102, 2019.

[123] E. Fagerberg, L. K. Månsson, S. Lenton, and M. Skepö, "The effects of chain length on the structural properties of intrinsically disordered proteins in concentrated solutions," *J. Phys. Chem. B*, vol. 124, no. 52, pp. 11843–11853, 2020.

[124] S. Jephthah, F. Pesce, K. Lindorff-Larsen, and M. Skepö, "Force field effects in simulations of flexible peptides with varying polyproline II propensity," *J. Chem. Theory Comput.*, 2021.

[125] P. Wernet, D. Nordlund, U. Bergmann, M. Cavalleri, M. Odelius, H. Ogasawara, L. Å. Näslund, T. K. Hirsch, L. Ojamäe, P. Glatzel, L. G. M. Pettersson, and A. Nilsson, "The structure of the first coordination shell in liquid water," *Science*, vol. 304, no. 5673, pp. 995–999, 2004.

[126] S. R. R. Campos and A. M. Baptista, "Conformational analysis in a multidimensional energy landscape: Study of an arginylglutamate repeat," *J. Phys. Chem. B*, vol. 113, no. 49, pp. 15989–16001, 2009.

[127] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philos. Trans. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, 2016.

[128] A. Grossfield and D. M. Zuckerman, "Chapter 2 quantifying uncertainty and sampling quality in biomolecular simulations," vol. 5 of *Annual Reports in Computational Chemistry*, pp. 23–48, Elsevier, 2009.

[129] A. Grossfield, P. N. Patrone, D. R. Roe, A. J. Schultz, D. Siderius, and D. M. Zuckerman, "Best practices for quantification of uncertainty and sampling quality in molecular simulations [article v1.0]," *Living Journal of Computational Molecular Science*, vol. 1, p. 5067, Oct. 2018.

[130] B. J. H. Kuipers and H. Gruppen, "Prediction of molar extinction coefficients of proteins and peptides using uv absorption of the constituent amino acids at 214 nm to enable quantitative reverse phase high-performance liquid chromatography–mass spectrometry analysis," *J. Agric. Food Chem.*, vol. 55, no. 14, pp. 5445–5451, 2007.

[131] E. Mihalyi, "Numerical values of the absorbances of the aromatic amino acids in acid, neutral, and alkaline solutions," *J. Chem. Eng. Data*, vol. 13, no. 2, pp. 179–182, 1968.

[132] D. I. Svergun, M. H. J. Koch, P. A. Timmins, and R. P. May, *Small Angle X-ray and Neutron Scattering from Solutions of Biological Macromolecules*. Oxford, UK: Oxford University Press, 1st ed., 2013.

[133] J. Pérez and P. Vachette, *A Successful Combination: Coupling SE-HPLC with SAXS*, pp. 183–199. Singapore: Springer Singapore, 2017.

[134] Guinier, André, "La diffraction des rayons x aux très petits angles : application à l'étude de phénomènes ultramicroscopiques," *Ann. Phys.*, vol. 11, no. 12, pp. 161–237, 1939.

[135] V. Receveur-Bréchot and D. Durand, "How random are intrinsically disordered proteins? a small angle scattering perspective," *Curr. Protein Pept. Sci.*, vol. 13, pp. 55–75, 2012.

[136] D. Orthaber, A. Bergmann, and O. Glatter, "SAXS experiments on absolute scale with Kratky systems using water as a secondary standard," *J. Appl. Crystallogr.*, vol. 33, pp. 218–225, Apr 2000.

[137] D. Durand, C. Vivès, D. Cannella, J. Pérez, E. Pebay-Peyroula, P. Vachette, and F. Fieschi, "NADPH oxidase activator p67^{phox} behaves in solution as a multidomain protein with semi-flexible linkers," *J. Struct. Biol.*, vol. 169, no. 1, pp. 45 – 53, 2010.

[138] O. Glatter, "Data evaluation in small angle scattering: calculation of the radial electron density distribution by means of indirect fourier transformation," *Acta Phys. Austriaca*, vol. 47, no. 1-2, pp. 83–102, 1977.

[139] D. I. Svergun, "Determination of the regularization parameter in indirect-transform methods using perceptual criteria," *J. Appl. Crystallogr.*, vol. 25, no. 4, pp. 495–503, 1992.

[140] D. A. Jacques and J. Trewhella, "Small-angle scattering for structural biology—expanding the frontier while avoiding the pitfalls," *Protein Sci.*, vol. 19, no. 4, pp. 642–657, 2010.

[141] F. Jin and F. Gräter, "How multisite phosphorylation impacts the conformations of intrinsically disordered proteins," *PLoS Comput. Biol.*, vol. 17, no. 5, p. e1008939, 2021.

[142] A. Miles and B. Wallace, "Chapter 6 - circular dichroism spectroscopy for protein characterization: Biopharmaceutical applications," in *Biophysical Characterization of Proteins in Developing Biopharmaceuticals* (D. J. Houde and S. A. Berkowitz, eds.), pp. 109 – 137, Amsterdam: Elsevier, 2015.

[143] S. M. Kelly, T. J. Jess, and N. C. Price, "How to study proteins by circular dichroism," *Biochim. Biophys. Acta, Proteins Proteomics*, vol. 1751, no. 2, pp. 119 – 139, 2005.

[144] L. Whitmore, A. J. Miles, L. Mavridis, R. W. Janes, and B. Wallace, "PCDDB: new developments at the Protein Circular Dichroism Data Bank," *Nucleic Acids Res.*, vol. 45, pp. D303–D307, 09 2016.

[145] A. Abdul-Gader, A. J. Miles, and B. A. Wallace, "A reference dataset for the analyses of membrane protein secondary structures and transmembrane residues using circular dichroism spectroscopy," *Bioinformatics*, vol. 27, pp. 1630–1636, 04 2011.

[146] J. L. S. Lopes, A. J. Miles, L. Whitmore, and B. A. Wallace, "Distinct circular dichroism spectroscopic signatures of polyproline II and unordered secondary structures: Applications in secondary structure analyses," *Protein Sci.*, vol. 23, no. 12, pp. 1765–1772, 2014.

[147] J. Tolchard, S. J. Walpole, A. J. Miles, R. Maytum, L. A. Eaglen, T. Hackstadt, B. A. Wallace, and T. M. A. Blumenschein, "The intrinsically disordered tarp protein from chlamydia binds actin with a partially preformed helix," *Sci. Rep.*, vol. 8, no. 1, p. 1960, 2018.

[148] N. Sreerama and R. W. Woody, "Computation and analysis of protein circular dichroism spectra," in *Numerical Computer Methods, Part D*, vol. 383 of *Methods in Enzymology*, pp. 318 – 351, Academic Press, 2004.

[149] B. Schuler, A. Soranno, H. Hofmann, and D. Nettels, "Single-molecule fret spectroscopy and the polymer physics of unfolded and intrinsically disordered proteins," *Annu. Rev. Biophys.*, vol. 45, no. 1, pp. 207–231, 2016.

[150] J. A. Riback, M. A. Bowman, A. M. Zmyslowski, K. W. Plaxco, P. L. Clark, and T. R. Sosnick, "Commonly used fret fluorophores promote collapse of an otherwise disordered protein," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 116, no. 18, pp. 8889–8894, 2019.

[151] M. Carballo-Pacheco and B. Strodel, "Comparison of force fields for alzheimer's a : A case study for intrinsically disordered proteins," *Protein Sci.*, vol. 26, no. 2, pp. 174–185, 2017.

[152] G. H. Zerze, W. Zheng, R. B. Best, and J. Mittal, "Evolution of all-atom protein force fields to improve local and global properties," *J. Phys. Chem. Lett.*, vol. 10, no. 9, pp. 2227–2234, 2019.

[153] E. W. Martin, A. S. Holehouse, C. R. Grace, A. Hughes, R. V. Pappu, and T. Mittag, "Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation," *J. Am. Chem. Soc.*, vol. 138, no. 47, pp. 15323–15335, 2016.

[154] E. Bienkiewicz and K. Lumb, "Random-coil chemical shifts of phosphorylated amino acids," *J. Biomol. NMR*, vol. 15, no. 3, pp. 203–206, 1999.

[155] R. Zangi, R. Zhou, and B. J. Berne, "Urea's action on hydrophobic interactions," *J. Am. Chem. Soc.*, vol. 131, no. 4, pp. 1535–1541, 2009.

[156] L. Costantino, G. D'Errico, P. Roscigno, and V. Vitagliano, "Effect of urea and alkylureas on micelle formation by a nonionic surfactant with short hydrophobic tail at 25 °c," *J. Phys. Chem. B*, vol. 104, no. 31, pp. 7326–7333, 2000.

[157] N. Homeyer, A. H. C. Horn, H. Lanig, and H. Sticht, "Amber force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine," *J. Mol. Model.*, vol. 12, pp. 281–289, Feb 2006.

[158] T. Steinbrecher, J. Latzer, and D. A. Case, "Revised amber parameters for bioorganic phosphates," *J. Chem. Theory Comput.*, vol. 8, no. 11, pp. 4405–4412, 2012.

[159] G. A. Naganagowda, T. L. Gururaja, and M. J. Levine, "Delineation of conformational preferences in human salivary statherin by 1h, 31p nmr and cd studies: Sequential assignment and structure-function correlations," *J. Biomol. Struct. Dyn.*, vol. 16, no. 1, pp. 91–107, 1998.

[160] K. T. Debiec, A. M. Gronenborn, and L. T. Chong, "Evaluating the strength of salt bridges: A comparison of current biomolecular force fields," *J. Phys. Chem. B*, vol. 118, no. 24, pp. 6561–6569, 2014.

[161] M. C. Ahmed, E. Papaleo, and K. Lindorff-Larsen, "How well do force fields capture the strength of salt bridges in proteins?," *PeerJ*, vol. 6, p. e4967, June 2018.

[162] N. Errington and A. J. Doig, "A phosphoserine–lysine salt bridge within an α-helical peptide, the strongest α-helix side-chain interaction measured to date," *Biochemistry*, vol. 44, no. 20, pp. 7553–7558, 2005.

[163] E. Fagerberg, S. Lenton, and M. Skepö, "Evaluating models of varying complexity of crowded intrinsically disordered protein solutions against SAXS," *J. Chem. Theory Comput.*, vol. 15, no. 12, pp. 6968–6983, 2019.

[164] C. Cragnell, L. Staby, S. Lenton, B. B. Kragelund, and M. Skepö, "Dynamical oligomerisation of histidine rich intrinsically disordered proteins is regulated through zinc-histidine interactions," *Biomolecules*, vol. 9, no. 5, 2019.

[165] J. R. Long, W. J. Shaw, P. S. Stayton, and G. P. Drobny, "Structure and dynamics of hydrated statherin on hydroxyapatite as determined by solid-state NMR," *Biochemistry*, vol. 40, no. 51, pp. 15451–15455, 2001.

[166] G. Goobes, R. Goobes, O. Schueler-Furman, D. Baker, P. S. Stayton, and G. P. Drobny, "Folding of the c-terminal bacterial binding domain in statherin upon adsorption onto hydroxyapatite crystals," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, no. 44, pp. 16083–16088, 2006.

[167] A. Amano, H. T. Sojar, J. Y. Lee, A. Sharma, M. J. Levine, and R. J. Genco, "Salivary receptors for recombinant fimbrillin of porphyromonas gingivalis," *Infect. Immun.*, vol. 62, no. 8, pp. 3372–3380, 1994.

Printed by Media-Tryck, Lund 2021 NORDIC SWAN ECOLABEL 3041 0903

Coarse-grained and atomistic modelling of phosphorylated intrinsically disordered proteins

In this thesis, computational and experimental methods are applied to study the conformational ensembles of intrinsically disordered proteins. The main goals have been to investigate the relation between sequence and structure, focusing on the impact of phosphorylation, and to investigate different models applicable for studying intrinsically disordered proteins.

ISBN: 978-91-7422-828-1

Department of Chemistry
Faculty of Science
Lund University