LUND UNIVERSITY

Fighting microbial infections with force fields: Evaluating conformational ensembles of intrinsically disordered proteins

Jephthah, Stephanie

2021

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

Link to publication

Citation for published version (APA):
Jephthah, S. (2021). *Fighting microbial infections with force fields: Evaluating conformational ensembles of intrinsically disordered proteins*. Lund University.

Total number of authors:
1

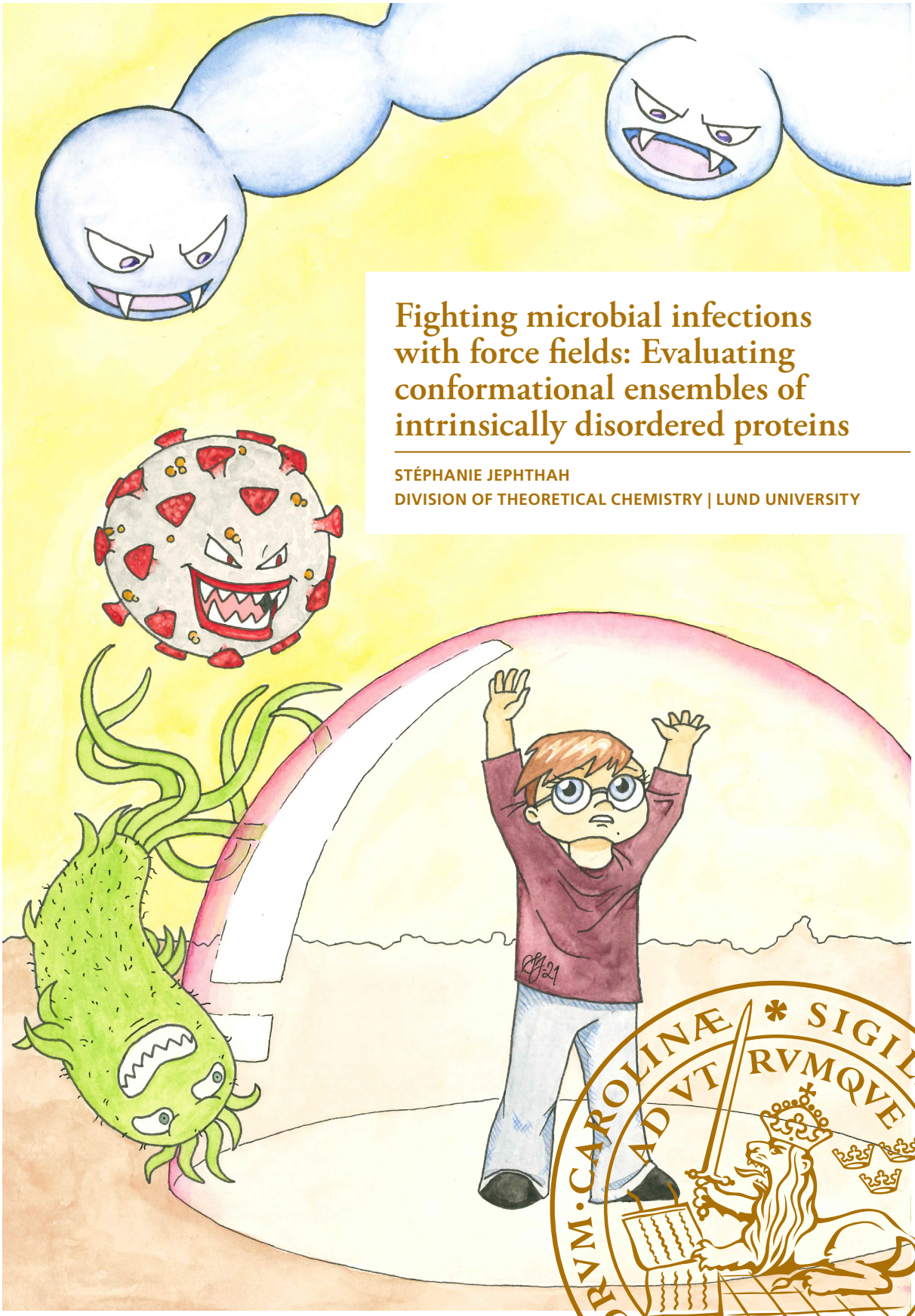
General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00



Fighting microbial infections with force fields: Evaluating conformational ensembles of intrinsically disordered proteins

STÉPHANIE JEPHTHAH
DIVISION OF THEORETICAL CHEMISTRY | LUND UNIVERSITY

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

LUND UNIVERSITY

ISBN: 978-91-7422-806-9

Theoretical Chemistry
Department of Chemistry
Faculty of Science
Lund University

9 789174 228069

Fighting microbial infections with force fields: Evaluating conformational ensembles of intrinsically disordered proteins

Fighting microbial infections with force fields: Evaluating conformational ensembles of intrinsically disordered proteins

by Stéphanie Jephthah

LUND UNIVERSITY

DOCTORAL THESIS

by due permission of the Faculty of Science, Lund University, Sweden. To be defended on Friday 28th of May 2021 at 13.00 in lecture hall A at Kemicentrum.

Faculty opponent
Dr. Robert B. Best, Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, USA.

Organization
LUND UNIVERSITY
Kemicentrum
Box 124
SE–221 00 Lund, Sweden

Document name
DOCTORAL THESIS

Date of disputation
2021-05-28

Sponsoring organization
NordForsk's Nordic Neutron Science Programme (grant number 81912)

Author(s)
Stéphanie Jephthah

Title and subtitle
Fighting microbial infections with force fields: Evaluating conformational ensembles of intrinsically disordered proteins

Abstract
The main goal of this compilation thesis has been to investigate the conformational ensemble of the intrinsically disordered protein (IDP) histatin 5 by using a mixture of experimental and computational techniques including, but not limited to, small-angle X-ray scattering (SAXS), circular dichroism (CD) spectroscopy, atomistic molecular dynamics (MD) simulations, and coarse-grained Monte Carlo (MC) simulations. Histatin 5 is a peptide of particular interest for two reasons. Firstly, histatin 5 is naturally found in the saliva, where it helps protect the teeth enamel from abrasion and degradation by being part of the dental pellicle. Additionally, it also possess candidacidal properties and thus acts as a first line of defense against fungal infections. Therefore, by studying the conformational ensemble histatin 5, we hope to aid in the development of saliva substitutes and new antimicrobial agents. Secondly, histatin 5 acts as a typical IDP that is relatively easy to work with in the lab, which makes it a perfect model IDP for evaluating different MD force fields and other simulation techniques.

In the first study of this thesis, simulations were used to investigate how the conformational ensemble of histatin 5 is affected by protonation of its histidine residues. The study also continues to investigate how this affects the interaction between histatin 5 and a negatively charged surface. This study is followed by an investigation of how histatin 5 is affected by temperature, and of how well modern simulation methods used for simulating IDPs can capture temperature-induced conformational changes. The results of this study led to another project, in which four different force fields were evaluated based on their ability to simulate histatin 5 and four other peptides that are known to possess polyproline II structure. In the final histatin 5 study, a shorter but more potent version of histatin 5 was conjugated to the polyamine spermidine, and was then investigated using a combination of experiments and simulation techniques. The aim of this study was to investigate if the increased candidacidal effect could be connected to the conformational properties of the histatin 5-spermidine conjugate. The final study of this thesis concerned a completely different peptide, namely the intrinsically disordered region of magnesium transporter A (MgtA) found in *Escherichia coli*. This study was mainly experimental and the aim was to investigate how this intrinsically disordered region contributes to the function of MgtA, and the knowledge about how bacterial proteins work might help to develop new antibiotics in the future.

Key words
molecular dynamics, force field, Monte Carlo, coarse-grained modeling, simulation, intrinsically disordered proteins, histatin 5, 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
978-91-7422-806-9 (print)
978-91-7422-807-6 (pdf)

Recipient's notes

Number of pages
214

Price

Security classification

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources the permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature _____ Date 2021-04-23

DOKUMENTDATABLAD enl SIS 61 41 21

Fighting microbial infections with force fields: Evaluating conformational ensembles of intrinsically disordered proteins

by Stéphanie Jephthah

LUND UNIVERSITY

DOCTORAL THESIS

Fighting microbial infections with force fields: Evaluating conformational ensembles of intrinsically disordered proteins

Funding information: This work is funded by NordForsk's Nordic Neutron Science Programme in the project "The structure of membrane proteins under solution conditions" (grant number 81912).

Parts of this doctoral thesis have been published before in the licentiate thesis *Two aspects of simulating disorder* by Stéphanie Jephthah (2019).

© Stéphanie Jephthah 2021

Faculty of Science, Lund University, Sweden, Kemicentrum

ISBN: 978-91-7422-806-9 (print)
ISBN: 978-91-7422-807-6 (pdf)

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

NORDIC SWAN ECOLABEL
Printed matter
3041 0903

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

"I can feel bits of my brain falling away like a wet cake"
– Bernard Black

"Hey, you know what this calls for? A party!"
– Pinkie Pie

Preface

My path though the world of chemistry might seem to have been a straight journey, but for me it feels more like I have been stumbling on a winding path. In the end I guess it has been a little bit of both. The science route started already during my lower secondary education but became more defined during the International Baccalaureate (IB) Diploma Programme, which feels like an eternity ago, although I still have nightmares about the exams. I had grand ambitions; in the beginning I thought I might become a medical doctor, but towards the end I was more interested in forensics. I decided to enroll in the bachelor's programme in chemistry and molecular biology—that way I could still qualify for studying forensics and get a broader education at the same time. However, I realized pretty soon that I was more interested in physical chemistry than in biochemistry, at least when it came to lab work and the way of how the different subjects are studied; I did not want to learn all amino acid structures by heart—I wanted to understand the concepts and principles that make chemistry work in the way it does. After a course about surface and colloid chemistry, I was informed that I could apply for working on a summer research project at the division of physical chemistry or at the division of theoretical chemistry. *Theoretical chemistry*. What does that really mean? I wanted to know more about this mysterious subject and sent an application, which was accepted. This is where I got my first impression of theoretical chemistry. Computer simulations, programming, scripting, FORTRAN, coarse-grained modeling, supercomputers—all of which were completely new concepts for me. I was lost in a new world, but also intrigued. Then it was time for the bachelor's degree project. I had yet not decided if I wanted to do it in physical or theoretical chemistry, but when I suddenly got a project offer at theoretical chemistry I was hooked. My soul now belonged to the division of theoretical chemistry and this is where I have stayed from then until now, as I am finishing my PhD project. In total, I have spent most of the last seven years at the division of theoretical chemistry in Lund (although this thesis only comprises the work of the last four and a half years) and it is finally time to let go and continue my journey elsewhere. I have learned a lot over the past years, especially during the PhD project, and I have met so many different and interesting people. Naturally, there have been hardships on the way, but I am still here and that counts for something. In the end I am happy with how far I have come during this time, and I am proud of my progress—not only within the field of chemistry, but also on a personal level. I am also grateful to everyone who has inspired me on my chemical journey.

To my *dad*, who have always supported my science interests and educational choices, and who encourages positive thinking (Foten i kläm? Jajjamensan!).

To my first chemistry teacher, *Therese Persson*, for providing an encouraging introduc-

tion to chemistry; and to my class mentor, *Anders Östling*, who taught me to take my studies more seriously.

To my IB chemistry teachers with spookily similar names, *Gunnar Nilsson*, *Gunilla Åkesson-Nilsson* and *Helene Åkesson*, for maintaining my interest of chemistry; and to my other IB teachers, *Sten Bernhardsson* and *Karin Kulka*, for inspiring me to do science.

To *João Henriques* for teaching me everything I needed to know for starting my own PhD studies; and to my supervisor, *Marie Skepö*, for everything—from the start of my bachelor's degree project to the end of this PhD project.

Stéphanie Jephthah
Lund, March–April 2021

Contents

Preface i
Populärvetenskaplig sammanfattning på svenska v
List of publications ix
Author contributions xi
Acknowledgements xiii

1 Introduction **1**

2 Background **3**
2.1 Proteins 3
2.2 Intrinsically disordered proteins 6
2.3 The protective functions of saliva 8
2.4 Xerostomia 9
2.5 Candidiasis 10
2.6 Histatin 5 and variants 11
2.7 Magnesium ion transport in bacteria 12
2.8 Simulations of IDPs 13

3 Theoretical background **15**
3.1 Statistical thermodynamics 15
3.2 The radial distribution function 16
3.3 Intermolecular interactions 16
3.3.1 Coulomb interactions 17
3.3.2 Van der Waals interactions 18
3.3.3 Hydrogen bonds 18
3.3.4 The hydrophobic effect 19

4 Simulation models **21**
4.1 The atomistic model 21
4.2 The coarse-grained model 23

5 Simulation methods **27**
5.1 The simulation box 27
5.2 Molecular dynamics simulations 28
5.2.1 Equations of motions 29

5.2.2 Constraints 30
5.2.3 Long-range interaction corrections 31
5.2.4 Temperature and pressure 31
5.3 Monte Carlo simulations 32
5.3.1 Trial moves 33

6 Simulation analyses **35**
6.1 Sampling convergence 35
6.2 Distances 38
6.3 Principal component analysis 38
6.4 DSSP analysis 39
6.5 The form factor in coarse-grained MC simulations 40

7 Experimental background **41**
7.1 Small-angle X-ray scattering 41
7.1.1 The scattering vector 41
7.1.2 The form factor 42
7.1.3 The Kratky plot 43
7.1.4 The pair distance distribution function 43
7.1.5 The Guinier approximation 44
7.2 Circular dichroism spectroscopy 44

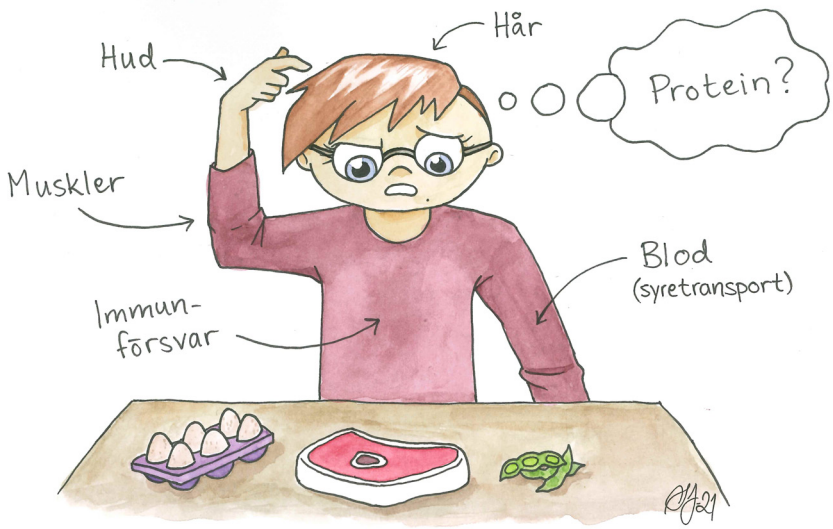
8 Research **49**
8.1 The effect of charge regulation in histatin 5 49
8.2 Temperature dependence in simulations of histatin 5 52
8.3 PPII structure in simulations of histatin 5 55
8.4 Conformational properties of Histatin 5-spermidine conjugates . . . 57
8.5 Examining KEIF 61
8.6 Outlook 63

9 References **65**

10 Scientific publications **81**
Paper I: Structural Characterization of Histatin 5–Spermidine Conjugates: A Combined Experimental and Theoretical Study 83
Paper II: Temperature dependence of IDPs in simulations: What are we missing? 85
Paper III: Physicochemical Characterisation of KEIF – The Intrinsically Disordered N-Terminal Region of Magnesium Transporter A 87
Paper IV: Force field effects in simulations of proteins with varying polyproline II propensity 89
Paper V: The effect of histidine protonation on the conformational ensemble in simulations of histatin 5 91

Populärvetenskaplig sammanfattning på svenska

Vad är ett protein? För någon som inte är särskilt insatt i kemi kanske ordet för tankarna till gymmet, där hurtbullarna häver i sig proteindrycker. Kanske går tankarna vidare till kött, ägg, soja och liknande produkter som går att hitta på tallriken. Till vardags används ordet protein främst i samband med näringsintag, och de flesta vet att det är viktigt att få i sig ordentligt med proteiner för att kunna bygga muskler. Vad folk kanske inte vet är att proteiner är inblandade i så mycket mer. Proteiner finns överallt i kroppen, och inte bara som byggstenar till våra organ. De hjälper även till med att bryta ner den mat vi äter, transportera syre i blodet, och skydda oss från sjukdomar.



Figur 1: Proteiner är inte bara näringsämnen. Proteiner är inblandade i nästan allt i kroppen: de bygger upp många olika typer av vävnader, transporterar syre och många andra viktiga atomer, hjälper till att bryta ner och omvandla molekyler, samt tar del av kroppens immunförsvar på många olika sätt.

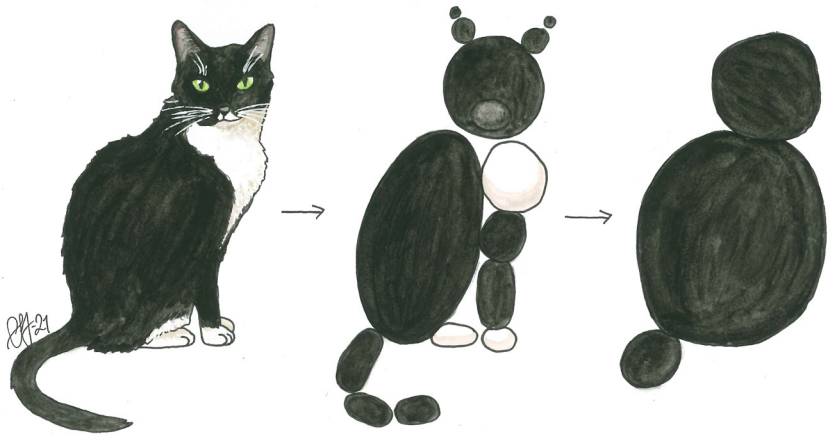
Proteiner kan liknas vid långa kedjor där varje länk är en mindre molekyl som kallas för en aminosyra. Det finns drygt tjugo olika aminosyror, och de är alla olika. Beroende på vilka aminosyror som finns i ett protein, och i vilken ordning de sitter, får proteinet specifika egenskaper. Bland annat är det vanligt att proteinet viker sig och på så sätt får en unik 3D-struktur som kan vara betydande för att proteinet ska kunna utföra sina funktioner i kroppen. Under en längre tid utgick forskare från att det var nödvändigt för proteiner att ha en bestämd 3D-struktur för att ha någon biologisk funktion. Sedan dess har vi insett att även proteiner som inte bildar sådana strukturer har viktiga funktioner i kroppen. Sådana proteiner är ofta väldigt flexibla och ändrar sin struktur konstant, och kallas därför för oordnade proteiner. Tyvärr är det också vanligt att oordnade proteiner är inblandade i otäcka sjukdomar som till exempel Parkinsons och

Alzheimers. Eftersom de oordnade proteinerna finns överallt och både kan hjälpa och stjälpa i kroppen, är det viktigt att vi lär oss hur de fungerar. Om vi kan förstå hur ett hjälpsamt protein fungerar kan vi kanske använda det i läkemedel. På likande sätt, om vi förstår hur ett skadligt protein fungerar, kan vi komma på andra läkemedel som förhindrar skadorna.

Alla experimentella metoder som vanligtvis används för att bestämma strukturen av ett protein fungerar inte för de oordnade proteinerna eftersom de byter struktur hela tiden. Därav blir man begränsad till metoder som enbart ger medelvärden av olika egenskaper som har med strukturen att göra. Till exempel går det att avgöra om de oordnade proteinerna i snitt tenderar att vara mer utsträckta eller om de hellre håller sig lite lätt vikta. Som ett komplement till att använda experimentella metoder för att studera oordnade proteiner kan man använda sig av datorsimuleringar. Med hjälp av datorsimuleringar går det att se exakt vilka strukturer proteinet växlar mellan, samtidigt som man också kan få ut samma medelvärden som de experimentella metoderna. Simuleringar kan även förutse vissa trender—till exempel hur ett protein påverkas av salthalt, pH och temperatur—vilket gör att man slipper göra vissa onödiga och dyra experiment.

Så varför gör man inte alltid simuleringar? För det första måste man vara säker på att simuleringsmetoden ger samma resultat som de experimentella metoderna. Det är relativt lätt att simulera ett vanligt protein i lösning, men ett oordnat protein fungerar annorlunda, så när de började simuleras vek dem sig alldeles för mycket och gav inte samma resultat som experimenten. Sedan dess har simuleringsmetoderna stegvis förbättrats, men ofta fungerar simuleringsmetoderna bara för ett snävt val av system. Det finns ytterligare ett problem med simuleringar: oftast går det lätt att simulera en ensam proteinkedja, men en ensam proteinkedja representerar inte alltid verkligheten. I ett naturligt system kan det finnas många andra saker i närheten av proteinet; till exempel kan det finnas fler proteiner, andra proteiner och molekyler, ytor, cellmembran, etc. Att ta hänsyn till allt detta gör simuleringarna extremt mycket mer komplicerade och tidskrävande, och därför är det i dagsläget orealistiskt att göra så komplicerade simuleringar på en större skala. Lyckligtvis finns det genvägar, bland annat genom att förenkla systemet som ska simuleras. En vanlig förenkling är att göra systemet grovkornigt. Det innebär att man antar att en grupp av atomer kan representeras av en större partikel. För grovkorniga proteiner är det vanligt att varje aminosyra representeras av en sfär, vilket gör att proteinet ser ut som ett pärlhalsband. Därav behöver man bara ta hänsyn till hur sfärerna växelverkar istället för hur alla atomer i systemet växelverkar. Att förenkla systemet på detta sätt ger såklart inte exakta strukturer men istället erhålls annan användbar information som inte går att få tag på om man enbart simulerar en ensam proteinkedja.

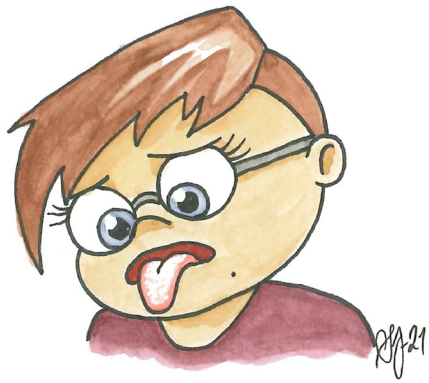
I den här avhandlingen har jag främst arbetat med att förstå hur det oordnade proteinet



Figur 2: Om jag skulle försöka göra min katt, Otto von Bismarck, grovkornig skulle det kunna gå till som på den här bilden. Det finns olika nivåer av grovkornighet; till exempel kan man göra något bara lite grovkornigt som i mittenbilden, där man fortfarande kan ana att det ska föreställa en katt. Annars kan man göra en mycket grövre approximation som i den högra bilden. Med en så pass grovkornig version är det omöjligt att avgöra om det är en katt, en hund, eller något djur över huvud taget.

histatin 5 fungerar genom att studera dess strukturella egenskaper. Histatin 5 finns vanligtvis i vår saliv, där det hjälper till att skydda tänderna genom att tillsammans med andra proteiner bilda en hinna på emaljen. Utöver det skyddar histatin 5 även mot svampinfektioner som till exempel torsk. På grund av dess intressanta biologiska egenskaper är det av intresse att förstå hur histatin 5 fungerar för att kunna optimera dess effekt och för att kunna utveckla nya och bättre läkemedel i framtiden. Histatin 5 är även intressant ur ett simuleringsperspektiv eftersom det är ett litet protein, vars strukturella egenskaper är väldigt typiska för oordnade proteiner. Därför är det fördelaktigt att använda just histatin 5 som ett modellprotein när man testar simuleringsmetoder som används specifikt för oordnade proteiner.

Till att börja med har vi studerat histatin 5 på en lite mer grundläggande nivå, genom att undersöka hur dess struktur ändras beroende på hur laddat det är. Vidare undersökte vi hur bra moderna simuleringsmetoder kan förutsäga de temperaturberoende strukturförändringar som sker för histatin 5. Resultaten av den studien gjorde oss nyfikna på att undersöka hur mycket de strukturella egenskaperna skiljs åt beroende på vilken simuleringsmetod som används. Detta gjordes inte bara för histatin 5, utan även för några andra små proteiner. Den sista studien av histatin 5 i den här avhandlingen handlar egentligen om en kortare variant av histatin 5 som är bunden till en annan molekyl för att öka den svampdödande effekten som histatin 5 har. Studien gick ut på att



Figur 3: Utan proteiner som histatin 5 i saliven får man lätt svampinfektioner. Om du drabbas av candidiasis, eller torsk som det också kallas, kan du få en vit beläggning på tungan. Hos väldigt sjuka personer, eller hos personer med nedsatt immunförsvar kan candidiasis sprida sig till blodet eller till och med benvävnaden.

försöka förstå hur strukturen av den här histatin 5-varianten bidrar till den förbättrade effekten. Avslutningsvis har vi även undersökt ett helt annat oordnat protein som egentligen är en del av ett större protein som finns i bakterier. Det är vår förhoppning att om vi kan förstå de strukturella egenskaperna hos detta protein, kan det i längden leda till nya alternativ för antibiotika.

Men vad har allt detta med kraftfält att göra? Det står "force fields" i titeln, och det betyder ju kraftfält. Dessutom är det en bild av ett kraftfält på avhandlingens omslag. Tyvärr är kraftfälten i den här avhandlingen inte lika flashiga som i science fiction. För vissa typer datorsimuleringar behövs information om alla olika parametrar och variabler som behövs för att beräkna krafterna som växelverkar mellan alla atomer i systemet—det är samlingen av all denna information som kallas för ett kraftfält. Det är på så sätt använder jag kraftfält i kampen mot infektioner.

List of publications

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

I **Structural Characterization of Histatin 5–Spermidine Conjugates: A Combined Experimental and Theoretical Study**

S. Jephthah, J. Henriques, C. Cragnell, S. Puri, M. Edgerton, M. Skepö
Journal of Chemical Information and Modeling, 2017, 57(6), pp. 1330–1341

II **Temperature dependence of IDPs in simulations: What are we missing?**

S. Jephthah, L. Staby, B. B. Kragelund, M. Skepö
Journal of Chemical Theory and Computation, 2019, 15(4), pp. 2672–2683

III **Physicochemical Characterisation of KEIF – The Intrinsically Disordered N-Terminal Region of Magnesium Transporter A**

S. Jephthah, L. K. Månsson, D. Belić, J. P. Morth, M. Skepö
Biomolecules, 2020, 10(4), pp. 623

IV **Force field effects in simulations of proteins with varying polyproline II propensity**

S. Jephthah, F. Pesce, K. Lindorff-Larsen, M. Skepö
Manuscript

V **The effect of histidine protonation on the conformational ensemble in simulations of histatin 5**

S. Jephthah, M. Polimeni, M. Lund, M. Skepö
Manuscript

All papers are reproduced with permission of their respective publishers.

Publications not included in this thesis:

Spontaneous Formation of Cushioned Model Membranes Promoted by an Intrinsically Disordered Protein

Y. Gerelli, A. E. Skog, **S. Jephthah**, R. J. L. Welbourn, A. Klechikov, M. Skepö
Langmuir, 2020, 35(15), pp. 3997–4004

Author contributions

Paper I: Structural Characterization of Histatin 5–Spermidine Conjugates: A Combined Experimental and Theoretical Study

I planned and initiated the project together with my co-authors. The simulations and the SAXS experiments, as well as the analysis of the data, were done with assistance from my co-authors. I wrote the manuscript together with my co-authors, and was involved in the submission and revision process.

Paper II: Temperature dependence of IDPs in simulations: What are we missing?

I planned the project together with my supervisor. I performed and analyzed the simulations, as well as the SAXS and the CD experiments. I wrote the manuscript with input from my co-authors, and was responsible for the submission and revision process.

Paper III: Physicochemical Characterisation of KEIF – The Intrinsically Disordered N-Terminal Region of Magnesium Transporter A

I planned and initiated the project together with my co-authors. I performed and analyzed the simulations, as well as the SAXS experiments. I wrote the manuscript together with my co-authors, and was responsible for the submission and revision process.

Paper IV: Force field effects in simulations of proteins with varying polyproline II propensity

I planned and initiated the project together with my supervisor. I performed and analyzed the simulations together with my co-authors. I wrote the first draft of the manuscript and edited the manuscript together with the co-authors.

Paper V: The effect of histidine protonation on the conformational ensemble in simulations of histatin 5

I planned and initiated the project together with my supervisor. I performed the MD simulations and I analyzed all simulations together with my co-authors. I wrote the first draft of the manuscript and edited the manuscript together with the co-authors.

Acknowledgements

To begin, I express my gratitude to my supervisor *Marie Skepö* for her support and advice, and for keeping me in her research group all the way until today. I also want to express my appreciation for all former and current colleagues in our research group and at the division. Special thanks goes to *João* for providing me with a solid foundation for performing simulations, to *Carolina* for teaching me about SAXS, and to *Sandeep* for making the experimental lab work more fun. I would also like to give extra thanks to *Maria*, *Ellen* and *Junhao*, for everything we've done together during these years—courses, conferences, experiments, discussions, etc. It has been fun. To *Erik* and *Sam*, thank you for keeping my spirits up.

Additional thanks goes to everyone in my personal life that has supported me during these years, and especially during these last few weeks. To *mamma Mia* and *pappa Marcus*; to *mormor Anette*, *farmor Lisbeth*, and *Margaretha*; to my siblings *Mathilda*, *Ottilia*, *Saga*, *Ella*, *Dante*, *Vannah*, and *Emma*; to my nerd friends, *Irene*, *David*, *Viktoria*, *Ben*, *Niklas*, *Pernilla*, *Hans*, and *Arvid*; and to my closest friends, *Jonas*, *Zoé*, *Sebastian* and *Jennifer*. I love you all.

Finally, I express my deepest gratitude to my fiancé *Martin*, for forcing me out of bed in the morning, for feeding me when I'm hangry, and for always being there for me. Du är min bästa vän och min klippa. Jag älskar dig.

1. Introduction

Summarized in this compilation thesis are four and a half years of research. The aim has been to investigate the conformational ensembles of different proteins, of which the two primary proteins are involved in microbial infections. By gaining a holistic understanding of how their conformational properties contribute to their biological functions, new alternatives for fighting microbial infections can be invented, which will advance medical research. Another aspect of the research was quickly developed by the start of this work as it was discovered that different simulation methods produced different results. For this reason, two of the studies have focused on investigation and evaluation of different simulation methods as well.

The thesis begins with a general background that introduces several basic concepts, as well as the main components that have been investigated throughout these past years. The aim of this part is to provide the reader with a basic understanding of important concepts, as well as to emphasize the importance of the study. Following the general background is a theoretical background that explains the more complicated theories that are needed to understand the underlying physics and chemistry that form the basis of the methods, and that are needed to correctly interpret the results. Next, the two different models that have been used in the simulations are presented, followed by the simulation methods. Connected to these sections are the simulation analyses that are subsequently presented. A few experimental techniques have also been utilized to complement the simulation results, and the background of these techniques, as well as different approaches for analyzing and interpreting their results, are illustrated in a section of its own. Finally, the actual research is summarized and discussed, followed by the research papers and the manuscripts.

2. Background

2.1 Proteins

Proteins are biological macromolecules that are built up by subunits called *amino acids*. There are several types of amino acids, but all amino acids have the same base structure, consisting of a central carbon known as the α-carbon, which is linked to an amino group ($-NH_2$), a carboxylic acid group ($-COOH$), a hydrogen atom (H), and a side chain (R) that is unique for each individual amino acid. At neutral pH, most amino acids exist as so-called *zwitterions*, which means that the amino group is protonated ($-NH_3^+$) and the carboxylic acid group is deprotonated ($-COO^-$). The structure of the twenty most common amino acids, as well as their respective 3- and 1-letter codes, are depicted in Figure 2.1.

Because the amino acids are designed so that there are four different groups connected to the central α-carbon, all amino acids (except glycine) are *chiral*. Chiral molecules are their own mirror images that can never be superimposed onto each other. A more relatable analogue would be our hands, which are mirroring each other but can never look exactly the same no matter how we turn them. The two chiral forms of amino acids are called the L isomer and the D isomer, but proteins are exclusively comprised of L amino acids [2]. Figure 2.2a-b depicts the amino acid structure of both chiral forms at different sides of a mirror plane.

Amino acids can bond together through *peptide bonds* to form longer chains known as *polypeptides* (also commonly referred to as peptides). If the polypeptide is long enough it is called a protein. The continuous chain structure of proteins is comprised of the α-carbons, the nitrogen atoms of the amino groups, and the carbon atoms of the carboxylic acid groups, and is commonly referred to as the protein *backbone*. Illustrations of the peptide bond and the protein backbone are depicted in Figure 2.2c. The order of the amino acids in a protein is referred to as the amino acid sequence, which constitutes the *primary structure* of the protein, see Figure 2.3a. Hydrogen bonds

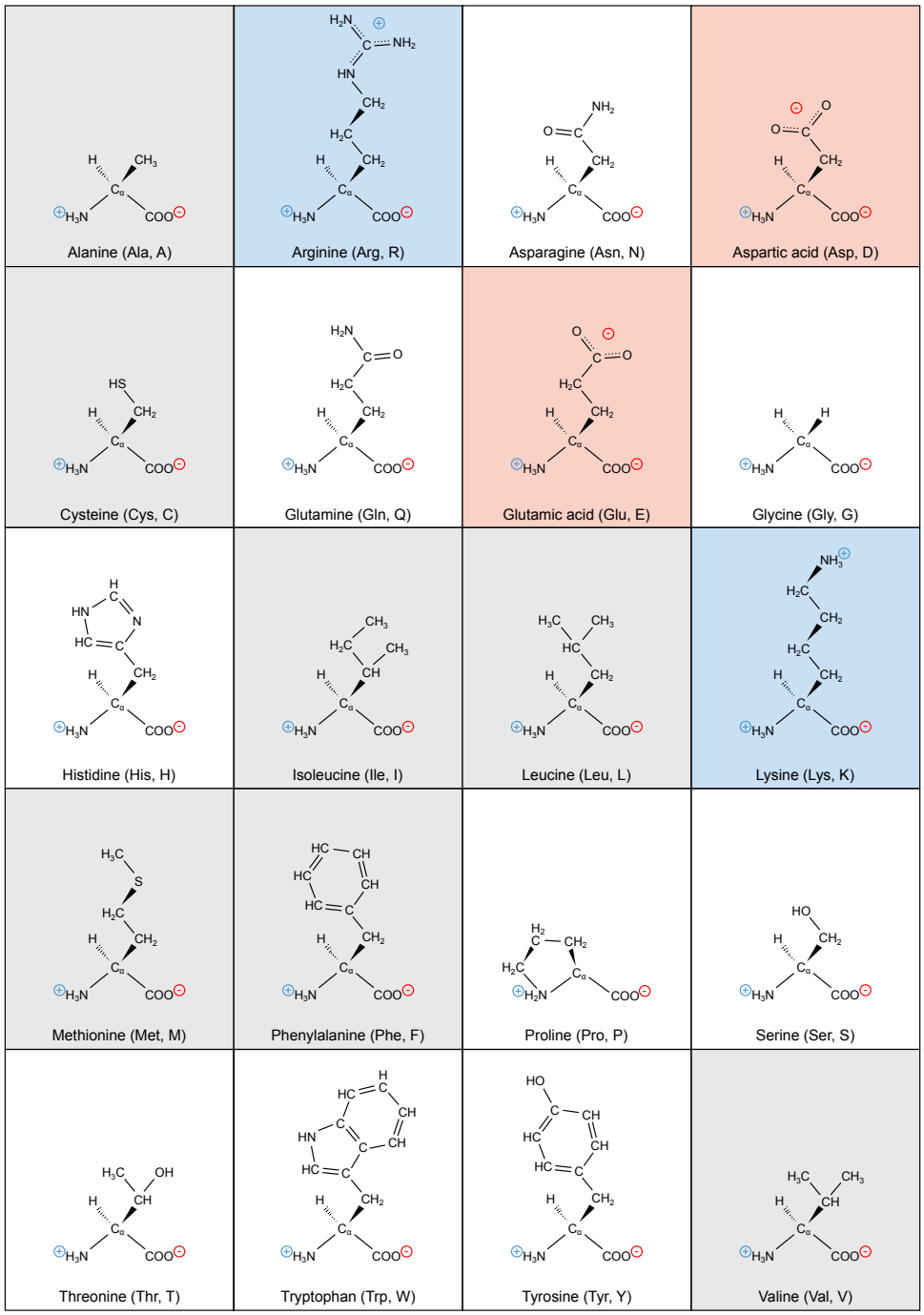


Figure 2.1: The structure of the individual amino acids at physiological pH. Their 3- and 1-letter codes are included within parenthesis after their names. Amino acids with a net positive charge ($+1e$) are shown with blue background, and amino acids with a net negative charge ($-1e$) are shown with red background. Hydrophobic amino acids (according to Kyte & Doolittle (1982) [1]) are shown with gray background.

between the amino hydrogens and the carboxylic oxygens in the protein backbone can cause local folding of the protein chain. These local rearrangements make up the *secondary structure* of the protein. There are several different types of secondary structures, depending on the relative positions of the amino acids in the sequence, and they can be defined by either (i) the hydrogen bonding pattern, or (ii) the backbone dihedral angles (ϕ, ψ). Examples of secondary structure elements are the α-helix (Figure 2.3b) and the β-sheet (Figure 2.3c). Secondary structure can be visualized in a so-called Ramachandran plot, where the dihedral angles of each amino acid residue of a protein is plotted against each other. A protein can also fold into a fixed three-dimensional structure, called the *tertiary structure*, see example in Figure 2.3d. Spatial arrangement of several protein chains into larger structures or complexes is called a *quaternary structure*. A typical example of a quaternary structure is that of hemoglobin, which consists of four protein subunits, see Figure 2.3e.

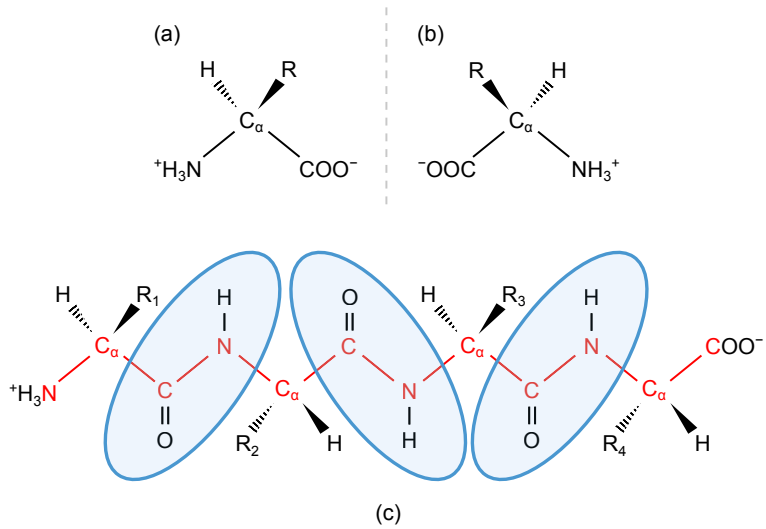


Figure 2.2: The general structure of an amino acid at neutral pH shown as (a) the L isomer and (b) the D isomer at different sides of a mirror plane. (c) A schematic polypeptide illustrating the peptide bonds (blue areas) and the protein backbone (red).

Because proteins can be created of all different sizes, and with all imaginable combinations of primary structure, they are considered to be the most versatile macromolecules. Apart from being essential nutrients and building blocks for most tissue, proteins are also responsible for a great variety of important functions in the body, such as catalysis,

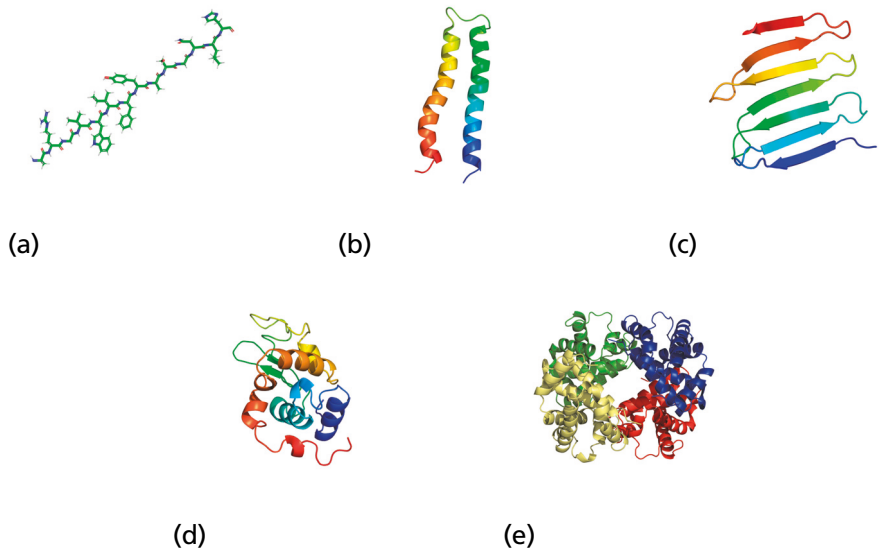


Figure 2.3: Examples of different levels of protein structure. (a) Primary structure representation, the secondary structures (b) α-helix[1], (c) β-sheet[2], (d) tertiary structure[3], and (e) quaternary structure with four different subunits[4]. All structures except for the primary structure were obtained from RCSB PDB (`https://www.rcsb.org`) [3] and are shown in cartoon representation. Some structures have been modified for visual enhancement.

molecule transport, protection, signaling, and regulation to name a few.

2.2 Intrinsically disordered proteins

For a long time, it was considered a requirement for a protein to have a well-defined folded structure in order for it to have any biological function. Nevertheless, during the last 20 years, plenty of evidence has emerged revealing proteins and protein regions that challenge this structure–function paradigm [7]. This class of proteins and protein regions are commonly referred to as *intrinsically disordered proteins* (IDPs) and *intrinsically disordered regions* (IDRs). These IPDs and IDRs are characterized by their extreme flexibility and lack of tertiary structure in aqueous solution. Additionally, they often lack secondary structure, have low hydrophobicity, and a high net charge [8, 9]. Despite being described as "unstructured", IDPs do not have to completely lack sec-

[1] 2I7U; Ma et al. (2006) Structural and dynamical analysis of a four-alpha-helix bundle with designed anesthetic binding pockets; doi: 10.2210/pdb2i7u/pdb [4].

[2] 6J48; Makabe and Hori (2019) Glycine mutation on single layer beta-sheet of OspAsm1; doi: 10.2210/pdb6j48/pdb [5].

[3] 1AKI; Carter et al. (1997) The structure of the orthorhombic form of hen egg-white lysozyme at 1.5 Angstroms resolution; doi: 10.2210/pdb1aki/pdb.

[4] 1A3N; Tame and Vallone (1998) Deoxy human hemoglobin; doi: 10.2210/pdb1a3n/pdb [6].

ondary structure. In fact, it has been discovered that several "unstructured" proteins possess a degree of local order in the form of the lesser known *left-handed poly-L-proline type II (PPII) helix* [10–13]. The PPII helix is an extended secondary structure with dihedral angles of $(\phi, \psi) = (-75°, +145°)$, a helical pitch of 9.3 Å/turn, and 3.0 residues/turn [14, 15].

The amino acid content of IDPs/IDRs has been shown to differ significantly from that of structured proteins. For example, the IDPs usually possess a low content of "order-promoting" amino acids, as for example the hydrophobic Ile, Leu, Val, as well as the aromatic Trp, Tyr, and Phe [9, 16]. Cys is also uncommon in IDPs. Instead, the content of "disorder-promoting" amino acids is much higher for IDPs. Disorder-promoting amino acids are often small or charged, with examples including Ala, Arg, Gly, Gln, Ser, Pro, Glu, and Lys [9, 16]. The structures of the order-promoting and the disorder-promoting amino acid are depicted in the previous Figure 2.1.

It is possible to predict the conformational properties of IDPs based on the number of charged residues in their amino acid sequence [17–19]. To predict the conformation, it is necessary to know the *fraction of charged residues* (FCR), and the *net charge per residue* (NCPR):

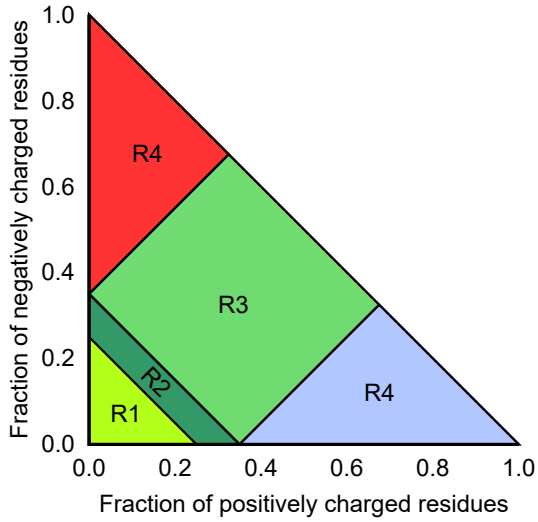
$$\mathrm{FCR} = f_+ + f_-, \tag{2.1}$$

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

where f_+ and f_- are defined as the fraction of positively and negatively charged residues, respectively. Depending on the combined values of the FCR and the NCPR, the protein is assigned to one of four predefined conformational classes, see Figure 2.4.

A common method for determining protein structure is *X-ray crystallography*. However, because of their extreme flexibility, IDPs are nearly impossible to crystallize [20]. Instead, other methods such as *circular dichroism* (CD), *Förster resonance energy transfer* (FRET), *nuclear magnetic resonance* (NMR), and *small-angle X-ray scattering* (SAXS) are more appropriate for obtaining structural information of IDPs [21–23]. Results obtained from these experiments capture averages of large ensembles of disordered conformations. Different types of radii, size distributions, general shapes, and secondary structure elements are a few examples of properties that can be obtained from experimental measurements by using these methods.

As previously stated, IDPs and IDRs are important for a variety of biological functions. A few examples are signaling, regulation, and recognition [24]. Because of their flexibility and ability to assume several different conformations, the same IDP can have several different functions in the body. While many IDPs/IDRs are essential for the body to function normally, it has been revealed that certain IDPs are involved in common diseases, such as Parkinson's disease, Alzheimer's disease, human papillomaviruses (HPVs), as well as several types of cancer [7, 24].



Class	FCR	NCPR	Expected conformation
R1	< 0.25	< 0.25	Globule
R2	$0.25-0.35$	≤ 0.35	Mix of globule and coils
R3	> 0.35	≤ 0.35	Polyampholytic coils or hairpins
R4	> 0.35	> 0.35	Polyelectrolytic semi-flexible rods or coils

Figure 2.4: Conformational classes of IDPs based on the FCR and the NCPR of the amino acid sequence according to Das et al. (2015) [18].

2.3 The protective functions of saliva

Saliva is the natural fluid that is present in the oral cavity. While the function of saliva may seem to be mainly lubricative and digestive, saliva also has several important functions involved in maintaining the oral health. Although saliva consists of more than 99 % water [25], the remaining constituents are also essential for the function of saliva. Less than 1 % of the saliva is constituted by proteins that are responsible for numerous different protective functions of saliva. For example, the lubricative function of saliva would be significantly less effective without the proteins known as *mucins* [25, 26]. Additionally, by forming a so-called dental pellicle on the enamel surface, several proteins, including mucins, *proline-rich proteins* (PRPs), *statherin*, and *histatins*, aid in protecting the teeth from abrasion and acid degradation [25, 26]. Interestingly enough, these particular saliva proteins are also classified as IDPs. Several of these IDPs are involved in the remineralization process of the teeth (PRPs and histatins), and in the inhibition of mineral precipitation and crystal formation (PRPs and statherin) [25, 26]. A less trivial function of saliva is its antimicrobial action, which protects against microbial infections. This function also arises because of different proteins that are present in the

saliva, with examples including *lysozyme*, *lactoferrin*, *peroxidases*, and histatins [25–29]. An overview of the different saliva functions and the responsible proteins is illustrated in Figure 2.5.

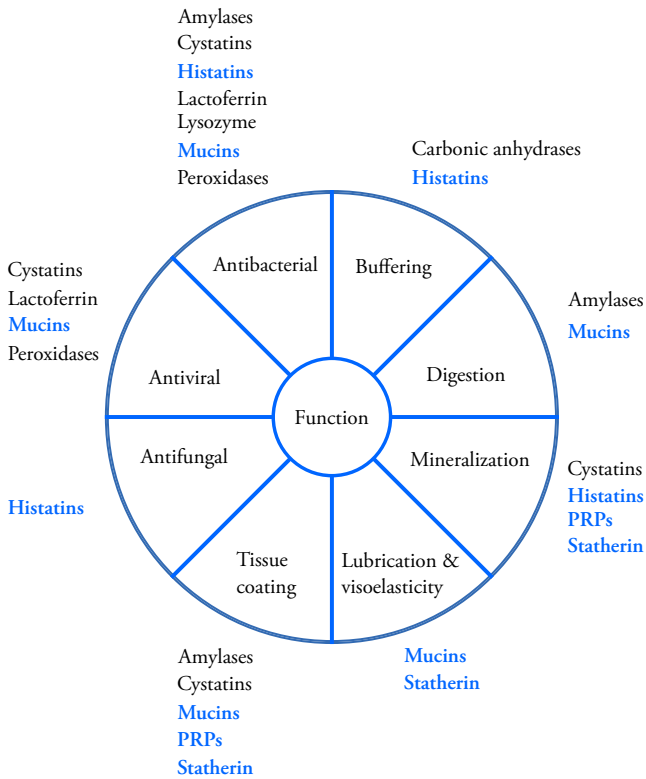


Figure 2.5: An overview of the different saliva functions and their responsible proteins. Proteins that are classified as IDPs are marked in blue. The figure is adapted from Levine (1993) [30].

2.4 Xerostomia

A change in saliva composition or a reduction of saliva flow can lead to *xerostomia*, which is also known as *dry mouth*. Xerostomia is a very common syndrome that affects up to 46 % of the human population [31]. Apart from the sensation of dryness, other symptoms may include a difficulty in swallowing, burning mouth syndrome, loss or change of taste, and increased thirst [31–35]. Additionally, a lack of saliva causes the body to become more susceptible to tooth decay (dental caries) and oral candidiasis (thrush) [32, 35]. A plentitude of different factors can cause xerostomia. For example, it can be caused by stress, anxiety, and depression [36, 37], and it is a very

common side effect of many prescription drugs, including antihistamines and antidepressants [37–39]. It can also be caused by radiation therapy in the head and neck region, which is a common treatment for some specific types of cancer [40]. Furthermore, xerostomia is a common symptom of several systematic diseases, examples including Sjögren's syndrome, human immunodeficiency virus (HIV), cystic fibrosis, and diabetes [31, 32, 34, 35, 41–43]. There is no general cure for xerostomia; the treatment is instead focused on relieving the symptoms. This can be done by avoiding drinking alcohol and smoking tobacco, by avoiding acidic and salty food and beverages, and by using salivary flow stimulants (for example sugar-free gum and hard candies) [32, 42, 44–46]. There are also saliva substitutes and oral lubricants available in the form of solutions, gels and mouth sprays [32, 45, 47]. More intensive treatment is usually long-term and demanding [42]. Further information regarding xerostomia can be found in the following references and papers therein: [31, 32, 34, 35, 42, 45].

2.5 Candidiasis

Candidiasis is a fungal infection caused by opportunistic yeast species belonging to the *Candida* genus. *Candida* spp. are part of the natural microbiota of the skin and the gastrointestinal tract [48], and are normally harmless in healthy individuals. Mucosal candidiasis can nevertheless manifest as for example thrush or vulvovaginitis [49]. These infections can normally be treated locally and with relative ease [50]. However, in immunocompromised or in critically ill individuals, *Candida* spp. can lead to life-threatening *Candida* infections known as *invasive* candidiasis. Invasive candidiasis refers to two types of infections: (i) candidemia (presence of *Candida* spp. in the bloodstream), and (ii) deep-seated tissue candidiasis such as intra-abdominal abscess (collection of infected fluid in a pocket of inflamed tissue in the abdomen), peritonitis (inflammation of the inner abdominal wall membrane), and osteomyelitis (bone infection) [51–56]. In 2015, the mortality rate of invasive candidiasis was stated to be as high as 40 % [55]. Several different *Candida* spp. can cause disease in humans, but the most prevalent species is *Candida albicans*, which by itself can account for up to 50 % of detected isolates [55, 56]. Because of the high mortality rate, there is a great interest in finding new measures for preventing and treating candidiasis. However, new antifungal resistant *Candida* spp. are constantly discovered, making the need for new antifungal alternatives even greater [56, 57].

2.6 Histatin 5 and variants

One model IDP that has been used in several studies of this work is *histatin 5*. Histatin 5 is a natural constituent of saliva, where it has several protective functions. For example, it is a part of the dental pellicle on the enamel [58] and has also shown bactericidal and fungicidal action. It is especially known for its candidacidal properties, or more specifically, its inhibition of growth and germination of *C. albicans* [59–62]. There are several types of histatins, and it has been shown that the candidacidal effect is inversely proportional to the size of the histatins [59]. Histatin 5 is the proteolytic product of histatin 3, in which it constitutes only the first 24 N-terminal amino acid residues. Thereby, histatin 5 is the most potent of the naturally occurring histatins when it comes to killing *C. albicans*.

The amino acid sequence of histatin 5 is rich in charged residues, giving the peptide a net charge of $+5e$ at physiological conditions, see Table 2.1. Additionally, $\sim$29 % of the amino acid residues are His residues, which gives histatin 5 the potential of charge regulation. It has been hypothesized that this might be a key feature in the mechanisms responsible for the antimicrobial properties of histatin 5 [63].

Because histatin 5 is a naturally occurring antimicrobial peptide, there is an interest in using it therapeutically in treatments against fungal infections [62]. However, since the saliva contains enzymes that can proteolytically degrade histatin 5, attempts have been made to find alternative histatin 5 variants that are more resistant to degradation. In these attempts, it was discovered that the candidacidal function is maintained in the shorter C-terminal fragments of histatin 5, *C14* and *C16* [60], as well as in the *P-113* fragment consisting of only 12 amino acids (residues 4-15) [64]. The amino acid sequences of all the mentioned histatin variants are presented in Table 2.1.

Table 2.1: The amino acid sequence of different histatin variants. Blue residues are positively charged, red residues are negatively charged, and histidine residues are represented in green.

Variant	Amino acid sequence
	1 5 10 15 20 25 30
Histatin 3	DSHAKRHHGYKRKFHEKHHSHRGYRSNYLYDN
Histatin 5	DSHAKRHHGYKRKFHEKHHSHRGY
C16	GYKRKFHEKHHSHRGY
C14	KRKFHEKHHSHRGY
P-113	AKRHHGYKRKFH

Although it is known *that* histatin 5 has antimicrobial properties, there is a lack of consensus regarding *how* it performs its functions. One common way for antimicrobial peptides to kill microbial cells is disrupting the cell membrane by for example forming membrane-spanning pores. However, because of its high charge and relatively low

amphipathicity, histatin 5 is unable to form such classical pores [65]. Whether or not histatin 5 can disrupt the cell membrane at all has also been debated. Although studies have shown the lack of lytic activity on membranes by histatin 5 [66, 67], more recent studies have determined that histatin 5 can cause membrane disruption at higher physiological concentrations [68], or when bonded to zinc ions [69].

It has also been established that histatin 5 targets the mitochondrion of *C. albicans* [70], but the peptide has to enter the cells to do so. Several studies have shown that histatin 5 alone has a limited ability to translocate membranes [71–73]. It has however been pointed out that this might only be the case when there is a lack of membrane potential [68], which would also explain observed lack of interactions between histatin 5 and model bilayers [72]. Kumar et al. (2011) discovered that histatin 5 can utilize the polyamine transporters Dur3 and Dur31 because of its similarity to polyamine substrates [74]. This was further proved in a study where *C. albicans* Dur3 and Dur31 was expressed in the more resistant yeast species *Candida glabrata*, which increased the histatin 5 uptake and toxicity [75]. Because of this discovery, Tati et al. (2014) conjugated the active P-113 fragment to the polyamine spermidine, which resulted an increased candidacidal effect, as well as an increased resistance against proteolytic degradation compared to the native histatin 5 [76].

2.7 Magnesium ion transport in bacteria

There are three classes of magnesium ion (Mg^{2+}) transporters that have been identified in prokaryotes. These are (i) the *CorA* magnesium transporter, (ii) *magnesium transporter E* (MgtE), and (iii) the P-type ATPase magnesium transporters, *magnesium transporter A and B* (MgtA and MgtB) [77, 78]. While the first two classes are classified as ion channels [79, 80], the P-type ATPase magnesium transporters are classified as Mg^{2+} pumps, which are close relatives of the proton pumps [81]. However, in contrast to most P-type ATPase magnesium transporters that uses ATP to mediate the *efflux* of cations against an electrochemical gradient, MgtA and MgtB instead mediate the *influx* of Mg^{2+} ions down the electrochemical gradient [77, 78].

The N-terminal regions of both MgtA and MgtB from *Salmonella typhimurium* are intrinsically disordered, and Subramani et al. (2016) predicted that this also was the case for the N-terminal region of MgtA from *Escherichia coli (E. coli)* [82]. While the study mainly highlighted that the anionic phospholipid *cardiolipin* (CL) is essential for activating MgtA *in vitro*, it also divulged the N-terminal part most likely is not required for activation by CL, nor is it involved in membrane trafficking or targeting to the CL domains. This does however give rise to the question: what is the function of the N-terminal IDR in MgtA?

The N-terminal IDR in MgtA from *E. coli* constitutes the 33 first amino acids of the MgtA sequence and is referred to as *KEIF* in this work (named after the first readable part in its amino acid sequence), see Table 2.2. At physiological pH, KEIF has a net charge of $+3e$ and the majority of the charged residues are evenly distributed in the first half of the sequence.

Table 2.2: The amino acid sequence of the N-terminal IDR of MgtA, KEIF. Blue residues are positively charged, red residues are negatively charged, and histidine residues are represented in green.

```
1   5    10   15   20   25   30
MFKEIFTRLIRHLPSRLVHRDPLPGAQQTVNTV
```

Knowing your enemy is essential in war [83], and this also applies in the battle against bacterial infections. In order to develop new antibiotics, it is necessary to not only understand the antibacterial agents, but also the bacteria it is supposed to work against. It is our hope that by gaining a holistic understanding of the biological function of KEIF in MgtA, this will aid for future discovery of new effective antibiotics.

2.8 Simulations of IDPs

Because of the relatively recent realization that IDPs have significant biological functions (see Section 2.2), interest of simulating these types of proteins has naturally followed. By performing atomistic *molecular dynamics* (MD) simulations of proteins (see Section 5.2 for more information on MD simulations), it is theoretically possible to obtain both structural and dynamical properties. Because most experimental methods used for IDP properties only capture ensemble averages, atomistic simulations become an important tool for studying proteins in more detail. For example, with the help of well-tuned simulations it is possible to obtain structural and dynamical information that is unfeasible to obtain from experiments alone. Additionally, by using simulations it is possible to predict experimental results.

Atomistic MD simulations of IDPs have historically not been straight forward. The collection of equations and potentials used for atomistic simulations are called *force fields*. There is a large diversity of force fields available, and they are usually specialized for the specific system that they are used for. Thus, initial simulations of IDPs that used the then currently available force fields for "regular" globular proteins revealed major flaws. Not only did these force fields display a strong bias for sampling α-helical or β-sheet structures in IDPs [84], but they were also predisposed to sample too compact IDP conformations [85, 86]. Consequently, several new options for atomistic IDP simulations have emerged during this past decade, mostly as completely new force fields [87–98], but an updated water model [99] has also been developed to work

for IDPs in combination with older force fields. To optimize IDP force fields, both of the previously mentioned flaws need to be dealt with. A common solution to the problem of too strong secondary structure propensities is to adjust the protein backbone dihedral parameters, by for example the addition of a two-dimensional grid-based energy correction map (CMAP) [93]. To avoid sampling too compact IDP conformations it is necessary to fine-tune the interaction between the protein and the water in the simulations [90]. Use of the Kirkwood–Buff theory of solutions has also been proposed for solving this problem [100]. More information regarding force field development for IDPs can be found in the following reviews and references within: [101–103].

Development of force fields for IDP simulations is constantly ongoing. When a new force field is tested it is unfeasible to try it for all different kinds of proteins, and often they are mainly tested for a few particular proteins of interest. Additionally, the force field may only have been validated for specific system parameters. Therefore, it is particularly important to not take any force field for granted when initiating simulations with a new IDP. To obtain accurate simulations, it is always necessary to compare the simulation results to experimental data. It is also important to investigate if the force field in use actually is able to capture the desired observables accurately.

3. Theoretical background

3.1 Statistical thermodynamics

"Som namnet föreslår, statistisk termodynamik är en blandning av stat, istik, termos och dynamik."

– Jonas Ravelid

One fundamental theory behind the methods that are used for simulating proteins is *statistical thermodynamics*, which enables calculations of macroscopic properties from interpretation of the microscopic, or molecular, properties. While non-mechanical thermodynamical properties (such as temperature, entropy, free energy, chemical potential, etc.) are treated implicitly by classical thermodynamics, the mechanical properties (pressure, energy, volume, etc.) need to be calculated as *ensemble averages*. An ensemble is a collection of systems that are identical on the macroscopic level but can be very different on the molecular level. The systems in the ensemble are representations of all the possible states available for the thermodynamic system. Thus, properties can be calculated for each individual system, but these may vary significantly. Instead, it is more useful to calculate an ensemble average, which is the mean value calculated from each system at any instant of time.

The *first postulate of statistical mechanics* states that an ensemble average can be used in place of the time average of a real system, provided that the ensemble contains a large number of systems [104]. In the *second postulate of statistical mechanics*, which is also known as the principle of *a priori* probabilities, it is stated that the probability of finding a system in a specific state is the same for all possible states in the ensemble. The combination of the two postulates is called the *ergodic hypothesis*, which implies that an isolated system in an ensemble spends equal amounts of time in each available state [104, 105].

Thermodynamic systems with macroscopic constraints define the statistical ensembles. For example, an isolated system of fixed volume, V, containing a constant number of particles, N, while having a constant total energy, E, is called a *micro-canonical ensemble*, or an *NVE* ensemble. Other constraints give different ensembles such as the *grand canonical ensemble (μVT)*, where the constant variables instead are the chemical potential, μ, and the temperature, T, in addition to the fixed volume. In the work presented in this thesis, the *canonical ensemble (NVT)* and the *isobaric-isothermal ensemble (NpT)* were used. The number of particles, the volume, and the temperature is kept constant in the canonical ensemble, while the pressure, p, is constant instead of the volume in the isobaric-isothermal ensemble.

3.2 The radial distribution function

To evaluate the internal structure of particle systems, it is common to use the *radial distribution function*, $g(r)$. The radial distribution function is a measure of how the density of particles, ρ, in the system deviates from the bulk density as a function of the distance, r, from a reference particle [104]. Thus, $g(r) \to 0$ for $r \to 0$, and $g(r) \to 1$ for $r \to \infty$. Schematic illustrations of the radial distribution function are depicted in Figure 3.1. Thermodynamic properties such as energy, pressure, and chemical potential can also be obtained from the radial distribution function [105], although the function itself cannot be determined from thermodynamic measurements [104]. It is however possible to calculate the radial distribution function from scattering experiments, where the relation between the distance distribution function and the structure factor, $S(q)$, is given by [105–107]:

$$S(q) = 1 + \int_0^\infty 4\pi r^2 \rho (g(r) - 1) \frac{\sin(qr)}{qr} dr, \tag{3.1}$$

where q is the magnitude of the scattering vector. More theory about scattering and the structure factor is presented in section 7.1.

3.3 Intermolecular interactions

There are two categories of forces responsible for interactions on the molecular level: (i) the *intra*molecular forces that acts *within* molecules and that can connect atoms to form molecules, and (ii) the *inter*molecular forces, which are responsible for interactions *between* atomic and molecular species without creating chemical bonds between them. The intermolecular forces are central when trying to explain the structural and conformational properties of proteins. There are several types of intermolecular forces,

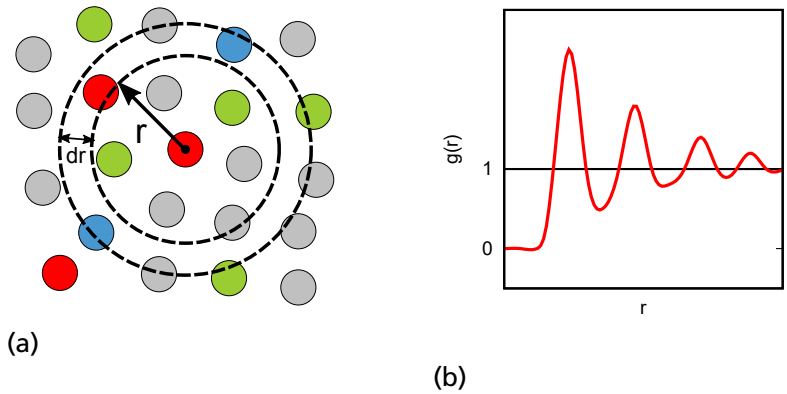


Figure 3.1: (a) A 2D representation of how the radial distribution function is obtained, and (b) a schematic plot of the radial distribution function of a liquid.

and they can be classified and grouped in numerous ways. The interactions that are relevant for this work are summarized in the following subsections, and all equations in these subsections are based on those given by Israelachvili (2011) [108].

3.3.1 Coulomb interactions

Charged particles in any medium give rise to electric fields, that act on the other charged particles in the system by the *Coulomb force*. For two charges, q_1 and q_2, at distance r, the magnitude of the force, $F(r)$, is given by:

$$F(r) = \frac{q_1 q_2}{4\pi\varepsilon_0\varepsilon_r r^2}. \tag{3.2}$$

The vacuum permittivity is given by ε_0, and ε_r is the dielectric constant. From this equation, an expression for the free energy of the *Coulomb interaction* between the two charges can be obtained:

$$U(r) = \int_0^\infty -F(r)dr = \frac{q_1 q_2}{4\pi\varepsilon_0\varepsilon_r r} = \frac{z_1 z_2 e^2}{4\pi\varepsilon_0\varepsilon_r r}. \tag{3.3}$$

In solution, q_i is exchanged for the ionic valency, z_i, multiplied by the elementary charge, e. Charges can also interact with *electric dipoles*. Electric dipole moments are present in many molecules, implying that they have an unequal distribution of electrons, that is, a separation of charges. The energy of a charge–dipole interaction is given by:

$$U(r) = -\frac{zeu\cos\theta}{4\pi\varepsilon_0\varepsilon_r r^2}, \tag{3.4}$$

where $u = ql$ is the dipole moment of a polar molecule with distance l between its charges. The angle, θ, is the angle of the dipole moment. Molecules that are neither charged nor dipoles, can instead be polarized by the electric field from other molecules or particles, that is, they become *induced dipoles*. The energy of the charge–induced dipole interaction also has a separate equation associated to it:

$$U(r) = -\frac{(ze)^2}{2(4\pi\varepsilon_0\varepsilon_r)^2 r^4}\left(\alpha_0 + \frac{u^2}{3kT}\right), \tag{3.5}$$

where α_0 is the polarizability of a nonpolar molecule, k is the Boltzmann constant and T is the temperature. All atoms and molecules are polarizable to different degrees, which is important for some of the interactions presented in the following subsection.

3.3.2 Van der Waals interactions

The *van der Waals interactions* are a collection of three different types of interaction that all possess a $1/r^6$ distance dependence. These three interactions arise from (i) the dipole–induced dipole interaction known as the induction force or the *Debye interaction*, (ii) the orientation force, which is a dipole–dipole interaction known as the *Keesom interaction*, and finally (iii) the *London forces*, also known as *dispersion forces*. The total van der Waals interaction energy between two molecules is given by:

$$\begin{aligned} U(r) &= -[U_{Debye} + U_{Keesom} + U_{London}] \\ &= -\left[(u_1^2\alpha_{02} + u_2^2\alpha_{01}) + \frac{u_1^2 u_2^2}{3kT} + \frac{3\alpha_{01}\alpha_{02}h\nu_1\nu_2}{2(\nu_1+\nu_2)}\right] \Big/ (4\pi\varepsilon_0)^2 r^6, \end{aligned} \tag{3.6}$$

where h is the Planck constant, and ν_i is the orbiting frequency of each electron. As mentioned in section 3.3.1, all atoms and molecules are polarizable. Although the time average of the dipole moment of a neutral molecule is zero, it is possible for it to possess an instantaneous dipole moment due to the movement of the electrons. This instantaneous dipole moment can give rise to induced dipoles in surrounding molecules, which causes interactions in the system. This is the basis of the dispersion forces, and although these forces usually are attractive, they can give rise to repulsion in many-body systems. The dispersion interactions are present between all atoms and molecules.

3.3.3 Hydrogen bonds

When a hydrogen atom is covalently bonded to an electronegative atom, it becomes depleted of electrons. Because the electron-depleted hydrogen is so small, it allows for other electronegative atoms, such as oxygen, nitrogen, or fluoride, to get close enough

to form a relatively strong bond referred to as a *hydrogen bond*. The hydrogen bond is primarily of electrostatic character and can be compared to a strong dipole–dipole interaction. Additionally, the bond is directional and can orient surrounding molecules. Although listed here under "intermolecular interactions", the hydrogen bond can also act as a non-bonding intramolecular interaction. Examples of hydrogen bonds are depicted in Figure 3.2.

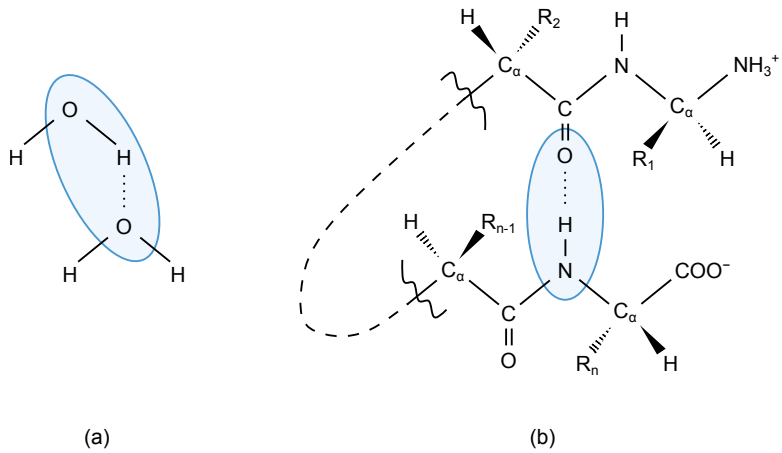


Figure 3.2: Illustrations of hydrogen bonds (dotted lines between hydrogen and oxygen atoms) as (a) an intermolecular interaction between two water molecules, and (b) as a non-bonding intramolecular interaction within a polypeptide.

3.3.4 The hydrophobic effect

The structure of water can be described as a network of hydrogen bonds. When introducing a hydrophobic molecule to an aqueous solution, it will disrupt the hydrogen bond network. To minimize this disruption, the water molecules have to rearrange themselves around the hydrophobic molecule. For small molecules, this can be done without breaking any of the hydrogen bonds. Nevertheless, the hydrogen bonds may need to deviate from their favored alignment, which is associated with a small energy increase. A rearrangement of the water molecules is entropically unfavorable, but it is generally more favorable than isolating water molecules completely from the hydrogen bond network. This tendency of water to maintain as much of the hydrogen bond network as possible is called the *hydrophobic effect*.

4. Simulation models

4.1 The atomistic model

As the name suggests, the atomistic model includes *all* atoms of the modeled molecules, including the solvent. An atomistic representation of histatin 5 is depicted in Figure 4.1. While the results from atomistic simulations can be of high detail, the computational cost of the simulations increases quickly as the size of the simulated systems increases. The model provided by the GROMACS package (version 4.6.7) [109–112] has been used as the atomistic simulation model for this work.

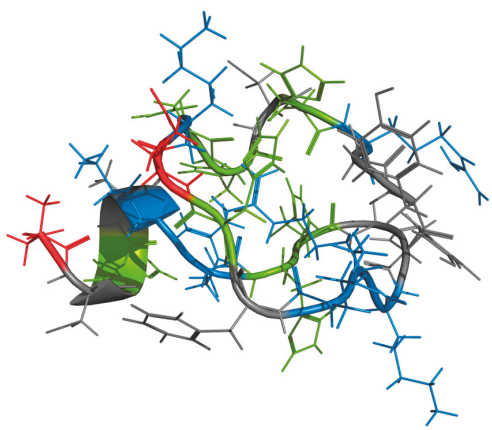


Figure 4.1: An atomistic representation of histatin 5. The backbone is also shown as a cartoon representation. Red color represents negative residues, blue represents positive residues, and green represents His residues. Surrounding atomistic water molecules are not included for visual simplicity.

The total potential energy in the atomistic model is constituted by intermolecular (non-

bonded) and intramolecular (bonded) interaction potentials. The bonded and the non-bonded potentials include four and two types of interactions respectively:

$$\begin{aligned} U_{tot} &= U_{bonded} + U_{non\text{-}bonded} \\ &= U_{bond} + U_{angle} + U_d + U_{id} + U_{LJ} + U_{el}. \end{aligned} \tag{4.1}$$

All of the interaction potentials are summed over all groups of atoms involved in the respective interaction. The first bonded potential is the bond stretching, that is, a two-body harmonic stretching potential between covalently bonded atoms:

$$U_{bond} = \sum \frac{1}{2} k_b (r_{ij} - r_0)^2, \tag{4.2}$$

where r_{ij} is the distance between two bonded atoms with an equilibrium bond length r_0, and k_b is the force constant. A bond-angle vibration is included as a three-body harmonic potential between the atoms:

$$U_{angle} = \sum \frac{1}{2} k_\theta (\theta_{ijk} - \theta_0)^2, \tag{4.3}$$

where k_θ is a force constant, θ_{ijk} is the bond angle, and θ_0 is the equilibrium bond angle. The last two bonded potentials are torsion potentials with dihedral angles. A dihedral angle is an angle between two intersecting planes. In this model, the *proper dihedral angle* is defined according to the IUPAC/IUB convention [113]. For a proper dihedral, the angle, ϕ_{ijkl}, spans between the *ijk* and *jkl* planes formed by four linearly bonded atoms. The proper dihedral potential displays periodic behavior with the periodicity n and phase ϕ_s:

$$U_d = \sum k_\phi [1 + \cos(n\phi_{ijkl} - \phi_s)], \tag{4.4}$$

where k_ϕ is a force constant. To preserve the planar structure of planar groups, and to maintain chirality, the *improper dihedral angle* is introduced. Similar to the proper dihedral, the improper dihedral interaction is based on an angle, ξ_{ijkl}, between two planes. The difference is that the four interacting atoms do not need to be bonded linearly, and that this potential is harmonic in nature according to:

$$U_{id} = \sum \frac{1}{2} k_\xi (\xi_{ijkl} - \xi_0)^2, \tag{4.5}$$

with k_ξ as the force constant. All bonded interactions are illustrated in Figure 4.2.

The non-bonded interaction potentials are described by (i) a Lennard-Jones interaction, and (ii) a Coulomb interaction. Both potentials are pairwise additive and centro-symmetric. The Lennard-Jones interaction contains both a repulsion and a dispersion

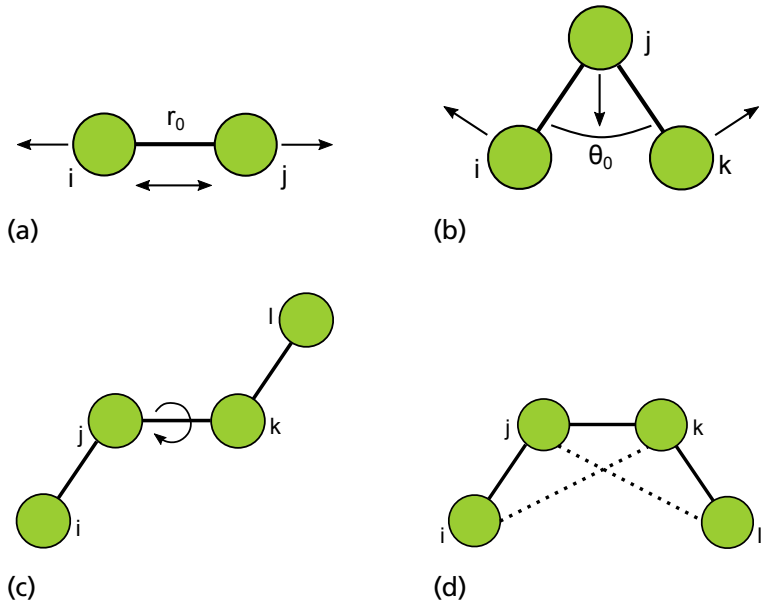


Figure 4.2: Examples of the bonded interactions in the MD simulations: (a) harmonic bond stretching, (b) bond-angle vibration, (c) proper dihedral, and (d) improper dihedral.

term:

$$U_{LJ} = \sum_{i<j} 4\epsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right), \tag{4.6}$$

where ϵ_{ij} is the potential well depth, σ_{ij} is the distance at which the potential becomes zero. The Coulomb potential represents the electrostatic interactions that act between two particles with the charges q_i and q_j:

$$U_{el} = \sum_{i<j} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}}. \tag{4.7}$$

4.2 The coarse-grained model

The computational cost can easily become high when performing atomistic simulations of proteins, mainly because of the large amount of solvent that is needed. Two ways of reducing this cost are (i) to *coarse-grain* the protein and (ii) to treat the solvent implicitly through the dielectric constant. Coarse-graining is a way to simplify the simulation model by representing a collection of atoms by larger structures. For the model used in this work [114], each amino acid is represented by a single hard sphere (bead), see Figure 4.3. The N- and C-termini are also explicitly represented as a single sphere each. Each

sphere can be neutral or assigned a point charge, which is set based on the amino acid sequence at physiological pH. The spheres are connected to each other, according to the amino acid sequence, by harmonic bonds in a so-called bead-necklace configuration. Any counter-ions in the system are treated explicitly as charged spheres, whereas the salt concentration is treated implicitly by adjusting the Debye screening length.

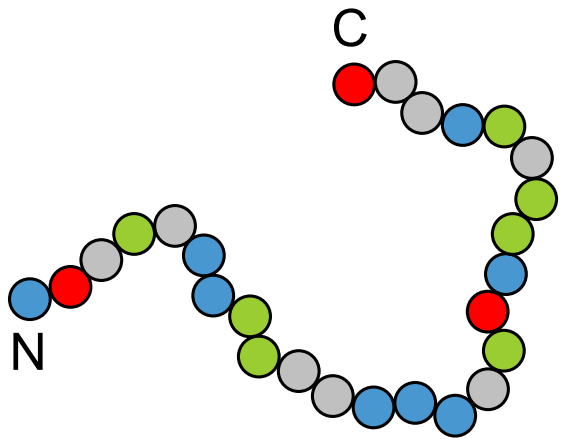


Figure 4.3: A coarse-grained representation of histatin 5—compare to the atomistic representation in Figure 4.1. Red beads have a charge of $-1e$, blue beads have a charge of $+1e$, and green beads represent His residues.

The total potential energy of the coarse-grained model is also separated into non-bonded and bonded interaction potentials [114, 115]. These potentials can be described as a sum of several different contributions:

$$\begin{aligned} U_{tot} &= U_{bonded} + U_{non\text{-}bonded} \\ &= U_{bond} + U_{hs} + U_{el} + U_{short}. \end{aligned} \tag{4.8}$$

For this model, the bonded potential was also comprised by a harmonic bond potential similar to that of the atomistic model:

$$U_{bond} = \sum_{i=1}^{N-1} \frac{1}{2} k_{bond} (r_{i,i+1} - r_0)^2, \tag{4.9}$$

where N is the total number of beads, $r_{i,i+1}$ is the distance between two bonded beads with an equilibrium bond length r_0, and k_{bond} is the force constant. An angular component can also be included in the bonded potential to impose restrictions on the chain flexibility but was not applied for this work.

The non-bonded potential comprises the three latter terms of the expression in Equation 4.8. A hard sphere potential, U_{hs}, is used to include the excluded volume of the beads:

$$U_{hs} = \sum_{i<j} u_{ij}^{hs}(r_{ij}), \tag{4.10}$$

where the hard sphere potential between the beads, $u_{ij}^{hs}(r_{ij})$ is given by:

$$u_{ij}^{hs}(r_{ij}) = \begin{cases} 0, & r_{ij} \geq R_i + R_j, \\ \infty, & r_{ij} < R_i + R_j, \end{cases} \tag{4.11}$$

where R is the radius of a bead. An extended Debye-Hückel potential is used to model the electrostatic potential, U_{el}:

$$U_{el} = \sum_{i<j} u_{ij}^{el}(r_{ij}) = \sum_{i<j} \frac{Z_i Z_j e^2}{4\varepsilon_0 \varepsilon_r} \frac{\exp[-\kappa(r_{ij} - (R_i + R_j))]}{(1 + \kappa R_i)(1 + \kappa R_j)} \frac{1}{r_{ij}}, \tag{4.12}$$

where Z is the bead charge, and κ is the inverse Debye screening length. The third and final nonbonded potential is a short-ranged attraction that acts between all beads through the following sum:

$$U_{short} = -\sum_{i<j} \frac{\varepsilon}{r_{ij}^6}, \tag{4.13}$$

where ε is a force constant.

5. Simulation methods

5.1 The simulation box

When performing computer simulations of solutions in a box, there will undoubtedly be particles interacting with the surface of the box. However, since most simulations of solutions are trying to capture realistic bulk behavior, such surface effects are sub-optimal and need to be minimized. One way to solve this problem could be by simply simulating an extremely large box with a huge number of particles in it. Unfortunately, this is not feasible since such a simulation would use up an unreasonable amount of time and computer resources. Thus, another approach is needed to deal with this problem. The most common alternative is to introduce *periodic boundary conditions* [105]. In this method, the simulation box is replicated in a surrounding lattice in such a way so that when the particles in the original box move, the replicated particles will move in exactly the same directions. Thus, if a particle were to move out from the original box on one side, one of its replicated images will enter on the opposite side. By using periodic boundary conditions, the system has been converted from a simulation box of finite size, to a system of infinite size. An illustrative example of periodic boundary conditions is presented in Figure 5.1. Although using periodic boundary conditions might seem great for simulating solutions in bulk conditions, it is not entirely problem free. A bulk solution is non-periodic, and by using the periodic boundary conditions, a periodicity is introduced to the system, which can give rise to erroneous anisotropic correlations. The smaller the system is, the risk of this error becomes larger. This error is however expected to be smaller than the error that would arise from not using any periodic boundary conditions. To minimize the anisotropies in a periodic system, it is important to test and compare different box sizes.

The simplest simulation box has a cubic shape, but just because it is simple does not mean that it necessarily is the best alternative. One inconvenience with the cubic shape is that more solvent molecules are needed to fill box because of the corners of the cube. To minimize the solvent volume, it would be ideal to have a sphere. Unfortunately,

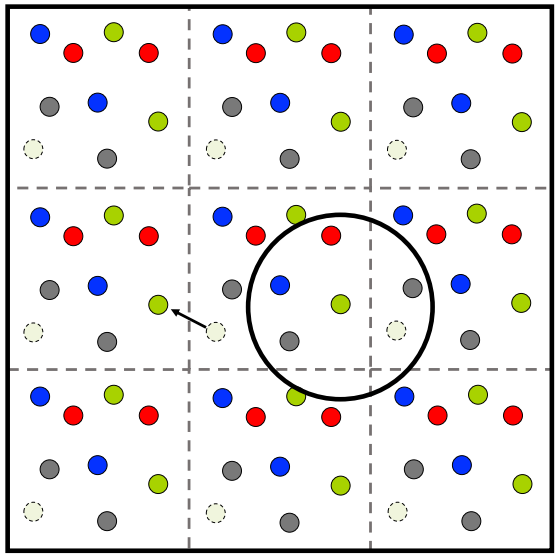


Figure 5.1: An illustration of periodic boundary conditions in two dimensions. The large circle represents a possible cutoff around the green particle that conforms to the minimum convention.

spheres cannot be organized in the lattice structure used for periodic boundary conditions. It is however possible to use other structures that are closer to a sphere in shape, but still available for lattice formation. Examples of such shapes are the truncated octahedron and the rhombic dodecahedron.

Periodic boundary conditions are normally used in combination with the *minimum image convention*, which states that each particle may only interact with the closest image of the remaining particles in the periodic lattice [105,116]. This is usually applied to the short-ranged nonbonded interactions by the use of a cubic or a spherical cutoff. The cutoff distance should be large enough to include all the important interactions but must not exceed half of the box length for a cubic box (this value depends on the box shape) to conform to the minimum image convention. This method is not always applicable for long-ranged interactions, which instead requires other corrections to the interaction potentials.

5.2 Molecular dynamics simulations

Molecular dynamics (MD) is a method that utilizes classical mechanics to describe particle displacement in computer simulations of molecular ensembles. All particles in the system are given initial coordinates and velocities. The initial conditions, to-

gether with the interaction potentials, enable calculations of forces in the system, and the particles will move accordingly. A basic MD algorithm [117] is shown in Example 5.1 below.

Example 5.1: A basic molecular dynamics algorithm

1. Generate the starting configuration with positions, velocities, and potential interactions of all atoms in the system.
2. Compute forces.
3. Update configuration.
4. Write output.
5. Repeat from step 2.

According to Allen and Tildesley (1989) [105] there are a few requirements that need to be fulfilled for an optimal simulation algorithm. These include that the algorithm should (i) be fast, (ii) allow for the use of a long time step, (iii) accurately reproduce the classical trajectory, (iv) follow the laws of conservation of energy and momentum, (v) be time-reversible, and finally, (vi) be simple. Some of these requirements are more important than others, depending on what type of simulation method that is used. For the MD method, it is more important that the time step is long than the algorithm being quick. The importance of the time step and how to increase it will be discussed in section 5.2.2. Since the method samples ensemble averages, exact classical trajectories are not necessary, and it is of greater importance to ensure the conservation of energy.

5.2.1 Equations of motions

The movement of the particles in an MD simulation is derived from the forces acting on each particle. The forces, $\mathbf{F}_i$, are given by the gradient of the potential energy at each particle position, $\mathbf{r}_i$:

$$\mathbf{F}_i = -\frac{\partial U}{\partial \mathbf{r}_i}. \tag{5.1}$$

According to *Newton's equations of motion* in classical mechanics, the force is given by:

$$\mathbf{F}_i = m_i \cdot \mathbf{a} = m_i \cdot \frac{\partial^2 \mathbf{r}_i}{\partial t^2}, \tag{5.2}$$

where m_i is the particle mass, and t is the time. The new positions are obtained by integrating Newton's equations of motion. The integration methods are usually based

on finite differences and solved at each simulation time step [118]. There are several integration methods available for MD simulations, but the integrator used in this work is called the *leap-frog algorithm* [118]. The name originates from that the positions and the velocities, $\mathbf{v}_i$, of the system are calculated alternately in time according to the following equations:

$$\mathbf{v}_i\left(t+\frac{1}{2}\Delta t\right) = \mathbf{v}_i\left(t-\frac{1}{2}\Delta t\right) + \frac{\Delta t}{m_i}\mathbf{F}_i(t), \tag{5.3}$$

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}_i(t) + \Delta t \cdot \mathbf{v}_i\left(t+\frac{1}{2}\Delta t\right). \tag{5.4}$$

An illustration of the leap-frog algorithm is depicted in Figure 5.2. If the equations of motions are solved correctly in the simulations, the energy conservation law will apply. In addition, these equations are reversible in time.

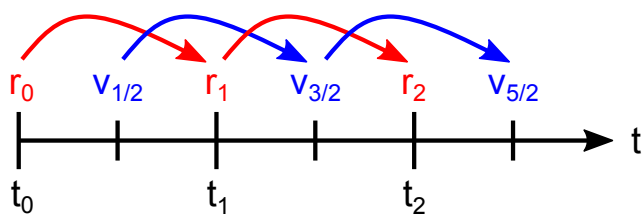


Figure 5.2: A schematic illustration of the leap-frog algorithm. The calculations of the coordinates (red) and the velocities (blue) are leaping past each other in time.

5.2.2 Constraints

How far the particles in a simulation move depends on the time step used in the simulations. The time step is usually of the same magnitude as molecular vibrations, that is, on the femtosecond scale. When studying proteins, some processes are comparatively slow, and it is therefore necessary to perform simulations on the microsecond scale. Performing long simulations with short time steps result in a large amount of simulation steps, which can become problematic considering available computer resources. Thus, it may be desirable to increase the time step. One way of achieving this is by imposing constraints on the relative coordinates in the simulated system. This is usually done by first solving the equations of motion without any constraints, and then performing a correction to the particle coordinates. One of the most common methods for applying bond constraints is the *LINCS algorithm* [119], which projects the new bond onto the old bond, followed by a correction for the change in bond length that arises due to rotation of the bond, see illustration in Figure 5.3.

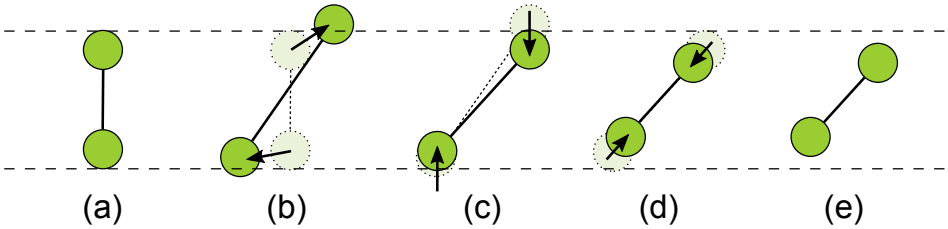


Figure 5.3: A schematic example of how bond constraints are applied by using the LINCS algorithm. State (a) shows the initial bond configuration. In (b) the particles are moved, followed by a correction by projection onto the original bond (c). The bond length is then adjusted accordingly (d) before reaching the final bond configuration (e).

5.2.3 Long-range interaction corrections

It is possible for the cutoff distance used for the non-bonded interactions in the MD simulations to interfere with the long-ranged interactions of the system. For those particular cases, it is necessary to apply long-range corrections. The correction of the Lennard-Jones potential is done assuming that the cutoff is long enough for ignoring the repulsion term, and that the system is homogeneous beyond the cutoff distance. The long-ranged electrostatic interactions are corrected by using the *particle-mesh Ewald* (PME) method [120], which is an improved version of the original Ewald summation method. In the Ewald summation, the potential energy is converted into two quickly converging terms: (i) a direct space sum, and (ii) a reciprocal space sum. Originally, the reciprocal sum was associated by a high computational cost, making the Ewald summation unfeasible for large systems. Instead, the PME method utilizes interpolation and fast Fourier transform to speed up the calculation of the reciprocal sum significantly for larger systems.

5.2.4 Temperature and pressure

The MD simulations of this work have been performed in an isobaric-isothermal ensemble, which requires both the temperature and the pressure to be constant throughout the simulations. This is done by applying temperature and pressure couplings. There are several different options for couplings available for MD simulations, but for this work, the *velocity-rescaling temperature coupling* [121] and the *Parinello-Rahman pressure coupling* [122] have been used. The velocity-rescaling coupling is based on another coupling method called the Berendsen temperature coupling [123]. When using the Berendsen algorithm, any deviations from the reference temperature, T_0, is corrected by an exponential decay,

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau}, \tag{5.5}$$

where τ is a time constant. The problem with this algorithm is that it suppresses kinetic energy fluctuations, which can lead to incorrect ensemble averages. By instead using the velocity-rescaling temperature coupling, the problem is solved by the addition of a kinetic energy correction term. When using the Parinello-Rahman barostat, an additional term containing the simulation box vectors is included in the equations of motion to represent changes in the pressure.

5.3 Monte Carlo simulations

Contrary to the MD method is the *Metropolis Monte Carlo* (MC) method [105], which is based on random displacements instead of physical moves. Due to the unphysical nature of the particle displacement in the MC algorithm, it is necessary to first perform a trial move to check if the move is viable. This is done by calculating the energy of the new configuration and compare it to that of the old configuration. The trial move is always accepted if the new energy is lower than the old energy. However, to ensure that the MC simulation satisfies the ergodic hypothesis, it is also necessary to accept some moves that will lead to a configuration of higher energy. The probability of such a move is determined by a Boltzmann distribution. A basic MC algorithm is shown in Example 5.2 below.

Example 5.2: A basic Monte Carlo algorithm

1. Generate a random starting configuration.
2. Select a particle at random and calculate its energy, $U(o)$.
3. Displace the particle by performing a trial move.
4. Calculate the new energy, $U(n)$.
   (a) If $U(n) \leq U(o)$, accept the move.
   (b) If $U(n) > U(o)$, generate a random number, $X \in [0, 1]$.
   (c) If $X < e^{-[U(n)-U(o)]/kT}$, accept the move.
   (d) Otherwise reject the move and restore the previous configuration.
5. Repeat from step 2.

5.3.1 Trial moves

A large variety of moves can be applied in the MC algorithm, and they can be either single-particle moves, or multi-particle moves. All moves have a probability associated with them. In this section, only the moves used in this work will be summarized.

Single particle translation is the only single-particle move that has been used in the MC simulations of this work. For this move, a single particle is displaced a short distance, see Figure 5.4a. Usually, this distance is random in both magnitude and direction, although the magnitude cannot exceed a set maximum value.

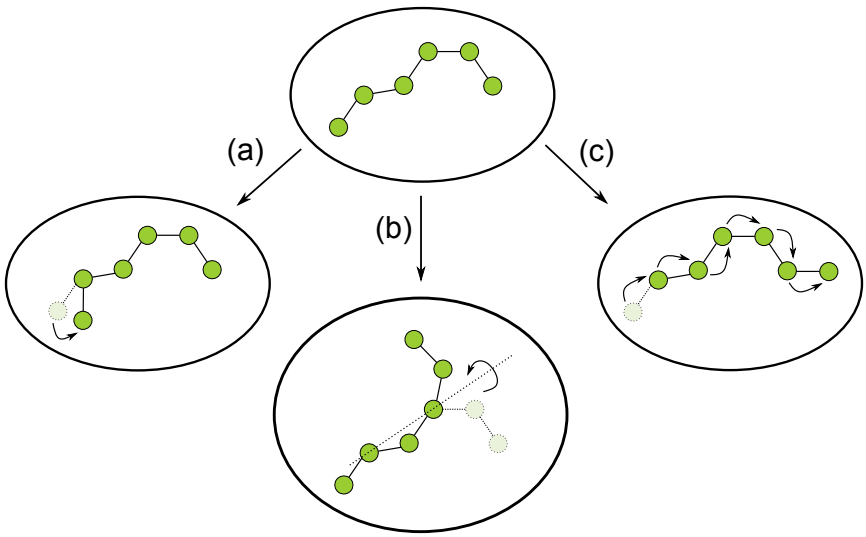


Figure 5.4: A schematic illustration of a few trial moves used in the MC simulations: (a) single particle translation, (b) Pivot rotation, and (c) slithering chain move.

The most basic group move used in the simulations is the *chain translation*. When using this move, the entire chain is displaced in the simulation box. Although this move does not alter the chain conformation and is thus not particularly important in simulations of single chains with implicit salt and solvent, it has significant impact on simulations with several proteins or other particles, and while using explicit salt and solvent.

Another group move that has been used is the *Pivot rotation*. When performing a Pivot rotation, one end of the simulated chain is rotated around an axis created by two randomly selected particles in the chain, see Figure 5.4b.

The fourth and last trial move is the *slithering chain move*, also known as a reptation move. Both names are related to reptiles, as the motion of a moving snake has been used as inspiration for the move. The move is initiated by the displacement of one of the particles at the end of the chain. The remaining particles in the chain will follow

the movement in the trail of the previous configuration, analogous to the slithering move of a snake, see Figure 5.4c.

6. Simulation analyses

6.1 Sampling convergence

When performing simulations, it is of uttermost importance to make sure that the results are reliable for making quantitative predictions. Grossfield and Zuckerman (2009) [124] claimed that there are two main reasons for erroneous molecular simulations: (i) model inaccuracy, and (ii) insufficient sampling. While the accuracy of the model should be relatively easy to evaluate by comparing with experimental references, evaluating the sampling of a simulation is not as straightforward. A very common method of assessing the sampling of a simulation is to investigate the convergence of a single observable, such as the radius of gyration. Using this method alone may however be unreliable, especially for dynamical systems that possess one or more smaller states [124]. An illustration of single observable convergence of a globular protein compared to a flexible protein is given in Example 6.1.

Example 6.1: Convergence of globular vs. flexible proteins

A simulation of a protein is initiated in the conformation of an extended straight chain. The natural state of the protein is a globular conformation and thus, as the simulation progresses, the radius of gyration decreases until it converges to a constant value, see example in Figure 6.1a. For a system like this, the initial non-converged part is removed before performing further analyses of the system. This does however not apply for flexible proteins because being fully extended is one of the natural conformations, and the radius of gyration will vary a lot throughout the entire simulation, see Figure 6.1b. Therefore, it is impossible to determine convergence of flexible proteins, such as IDPs, by using this method only.

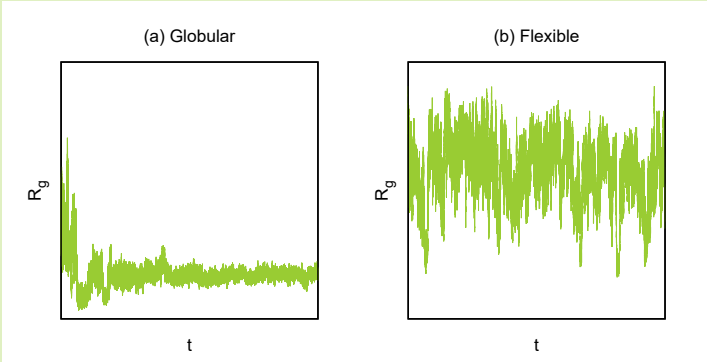


Figure 6.1: The radius of gyration as a function of simulation time for (a) a globular protein, and (b) a flexible protein. Both plots are shown on the same time scale.

Although the convergence of a single observable might be unreliable in dynamic simulations, it is possible to estimate the correlation and uncertainty of the observable. This can be done in two ways: (i) time correlation analysis, and (ii) block averaging. Time correlation is a statistical measure of correlation between different observables, A and B, at different time steps. Since the correlation is observed as a function of time, t, it is also known as a time correlation function, $C_{AB}(t)$:

$$C_{AB}(t) = \langle A(t' + t)B(t)\rangle. \tag{6.1}$$

If the correlated observables are identical ($A = B$), this measure is instead referred to as an *autocorrelation function* (ACF), $C(t)$. The integral of the ACF is called the correlation time, τ:

$$\tau = \int_0^\infty C(t)dt, \tag{6.2}$$

which describes the length of the simulation time required for an observable to no longer have any correlation with its previous values. The correlation time may be used to estimate the quality of the sampling of a single observable by $N = t_{tot}/\tau$, where t_{tot} is the total simulation time, and $N \gg 1$ indicates good sampling [124].

By using the block averaging method, it is possible to obtain error estimates of a single observable. This method is especially useful for estimating single observable convergence since it is based on information from the entire simulated trajectory [124]. To obtain the error estimate, the simulated trajectory is divided into n segments referred to as "blocks", and an average value of the observable, B_i, is calculated for each block. The *block-averaged standard error* (BSE), $\varepsilon(n)$, is then calculated for each block according

to the following equation [117, 124]:

$$\varepsilon(n) = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (B_i - \langle B \rangle)^2}, \tag{6.3}$$

where $\langle B \rangle$ is the total average. These errors are then plotted as a function of the block size, and the resulting curve increases with the block size until it converges to a final BSE. A converged BSE is usually a good indicator of a converged observable. More information regarding this type of error estimation can be found in the following references and references within: [124–126]. Example 6.2 below includes a discussion and illustrations of how ACF and BSE can be applied when analyzing convergence of simulations of flexible proteins.

Example 6.2: ACF and BSE

Although the radius of gyration may vary a lot for flexible proteins (see Example 6.1), it can still be more or less converged. Figure 6.2 shows examples where the radius of gyration of three different simulations of flexible proteins have been analyzed through their respective ACF and BSE. Already by visual inspection of both plots it can be observed that the radius of gyration of Sim 1 (dark green) seems to be considerably less converged compared to Sim 2 (light green) and Sim 3 (blue). This is also evident from the correlation time, which is only 9 ps for Sim 3 and 1607 ps for Sim 2, but as long as 23824 ps for Sim 1. For the latter case, this corresponds to slightly more than 2 % of the total simulation time. The BSE of Sim 1 converges at 0.058 nm, Sim 2 at 0.013 nm, and Sim 3 at 0.009 nm.

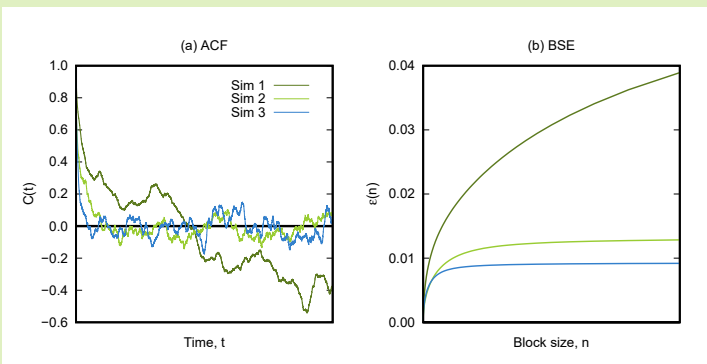


Figure 6.2: (a) Examples of ACFs with different correlation times, and (b) examples of BSEs of different levels of convergence. All curves are obtained from simulations with a total simulation time of 1 μs each.

6.2 Distances

Because of the flexible nature of IDPs, it is only reasonable to characterize them by their average properties. The *radius of gyration*, R_g, is an average measure of the compactness of a structure. In a macromolecule, the radius of gyration is defined as the root mean square distance between the mass elements and the center of mass [127]. The radius of gyration is calculated using the following equation:

$$R_g = \sqrt{\frac{\sum_{i=1}^{n} m_i \|\mathbf{r}_i\|^2}{\sum_{i=1}^{n} m_i}}, \tag{6.4}$$

where n is the total number of mass elements, m_i is the mass of each mass element, and $\mathbf{r}_i$ is the position of the mass elements relative to the center of mass of the macromolecule. For polymers it is common to treat each monomer as individual mass elements. This also extends to proteins, for which the amino acids are considered to be the individual mass elements.

Another property that is useful to study is the *end-to-end distance*, R_{ee}. This distance is calculated as the root mean square of the distance between the two ends of a linear chain:

$$R_{ee} = \sqrt{\|\mathbf{r}_1 - \mathbf{r}_n\|^2}. \tag{6.5}$$

The ends of a chain are defined differently depending on the model used. In the atomistic model, the distance is calculated between the α-carbons in the terminal amino acid residues of the protein. In the coarse-grained model, the distance is instead calculated between the center of mass of the terminal beads.

To ensure that the protein is not interacting with itself during a simulation, it is also important to keep track of the minimum distance between the protein and its periodic images. In the MD simulations, the minimum distance should never be larger than the cutoff distance of the nonbonded interactions. For the MC simulations, the box size is always at least double the theoretical maximum extension of the modeled chain.

6.3 Principal component analysis

A more extensive study of conformational variations can be done using *principal component analysis* (PCA) based on the method proposed by Campos and Baptista (2009) [128]. PCA is a mathematical method used to reduce the dimensionality of a data set without losing information about variations [129]. Applied to our protein system, the aim is to map the conformational space onto a low-dimensional representation, and to

study the conformations in a multidimensional free energy landscape. All of the conformational information could technically be found in a complete energy landscape, but this is generally undesired due to two reasons. Firstly, a complete energy landscape provides much more information than is actually needed. Secondly, the use of a complete energy landscape includes a vast number of dimensions, scaling as $(3N-6)$ with the number of atoms, N, in the system, which is unfeasible for most systems. The PCA analysis transforms a set of correlated variables into a set of uncorrelated variables, or coordinates, called principal components (PCs). Most of the variation is captured by the first PC, and most of the remaining variation is accounted for in the second PC, and so on. To obtain an analysis of the conformational space of high accuracy, it is necessary to use several PCs. However, since the strongest correlations should be included in the first PCs, it should be sufficient to only use the first two to get an adequate estimation of the most significant conformational information of the system. Furthermore, when performing a PCA on a protein system, a few dimensions can be excluded by ignoring all atoms that are not part of the peptide backbone. When performing this type of PCA on proteins, structures from each time step in a simulation are aligned and compared to a reference structure. Deciding the reference structure is difficult for IDPs because of their intrinsic flexibility. Thus, it should be noted that PCA might not be the most optimal analysis for IDPs. In our case, the reference structure has been chosen from a *root mean square deviation* (RMSD) analysis that selects the most "average" structure in the simulated trajectory, that is, the structure that differs the least from all other structures in the simulation. Because of this, PCA of IDPs is still useful for comparing replicates of the same protein simulations (and can thus help in assessing sampling convergence) and can be used to identify different conformational classes.

6.4 DSSP analysis

In 1983 W. Kabsch and C. Sander presented the *Define Secondary Structure of Proteins* (DSSP) program for objective determination of the secondary structure in proteins [130]. For the program to work properly it needs a 3D structure input, from which the most likely secondary structure is computed. Instead of using backbone dihedrals, the algorithm utilizes structure recognition of hydrogen bonding patterns. The program defines the hydrogen bond by a generous energy cutoff that is estimated by assuming the hydrogen bond to be a purely electrostatic interaction. Although the definition of the hydrogen bond can be disputed, Kabsch and Sander claimed that their approximation is a suitable compromise for the intended purpose of the program. The secondary structures that can be identified by the algorithm are the 3_{10}-helix, the α-helix, the π-helix, the isolated β-bridge, the extended strand (β-sheet), as well as hydrogen bonded turns and bends. Unfortunately, the classic DSSP program is unable

to identify the PPII helix. Thus, Mansiaux et al. (2011) developed a new method called *DSSP-PPII* [131], which is able to perform PPII assignment by utilizing the DSSP program.

6.5 The form factor in coarse-grained MC simulations

To obtain a scattering profile from the coarse-grained MC simulations used in this work, the individual particles (beads) are considered to be *point scatterers*. A point scatterer scatter equally at all angles and thus has a constant form factor. With a constant form factor, the total scattering intensity is directly proportional to the structure factor. In a system containing N identical beads, the structure factor is given by:

$$S(q) = \left\langle \frac{1}{N} \left| \sum_{j=1}^{N} \exp(i\mathbf{q} \cdot \mathbf{r}_j) \right|^2 \right\rangle. \tag{6.6}$$

However, in a system using different bead types, i and j, partial structure factors can be obtained by using the following equation:

$$S(q)_{ij} = \left\langle \frac{1}{(N_i N_j)^{1/2}} \left[\sum_{i=1}^{N_i} \exp(i\mathbf{q} \cdot \mathbf{r}_i) \right] \left[\sum_{j=1}^{N_j} \exp(-i\mathbf{q} \cdot \mathbf{r}_j) \right] \right\rangle, \tag{6.7}$$

where N_i and N_j are the total number of beads i and j, respectively. The total structure factor can then be calculated from the partial structure factors by using the relation given by:

$$S(q) = \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \frac{(N_i N_j)^{1/2}}{N} S(q)_{ij}. \tag{6.8}$$

In a system with a single coarse-grained peptide chain, all scattering arises from intra-particle interference, which means that the calculated scattering profile is, in fact, the form factor. Because the beads are considered to be point scatterers, system differences are only observed at short distances, that is, at high q-values.

The method presented in this section has previously been described in a paper by Cragnell et al. (2018) [114]. Because the method does not differentiate between beads that are within a chain, and beads that belong to different chains, it is impossible to obtain the peptide form factor in a system containing more than one peptide chain. Further scattering theory is presented in Section 7.1.

7. Experimental background

Since the intrinsic flexibility of IDPs causes them to assume a wide range of conformations, their structure cannot be captured experimentally by, for example, crystallographic techniques. Fortunately, there are other options available to experimentally study the structure of IDPs. A few of the most common techniques are SAXS, NMR, and single molecule FRET, which all provide ensemble averages and/or distance distributions of IDP structures in solution. In the following subsections, some of the most important experimental techniques used for this thesis work will be presented.

7.1 Small-angle X-ray scattering

Small-angle X-ray scattering is an experimental technique used to obtain the average structure, shape, and flexibility of particles in the size range of 1-100 nm. It can be applied to samples in both solid and liquid states and is thus also suitable for IDPs in solution. The results obtained from SAXS are representative, but only give the ensemble average of the sample. For this reason, it is usually advantageous to combine SAXS with other complementary techniques.

7.1.1 The scattering vector

The fundamental principle of SAXS is, as the name suggests, the scattering of X-rays. When a sample is irradiated by X-rays, some of the radiation will be scattered by all the atoms in the sample. If the scattering is elastic, that is, it has the same energy as the incident radiation, the scattered waves from the atoms will interfere with each other. When the interference reaches a detector, it is interpreted as a 2D interference pattern, from which the internal structure of the sample can be deduced.

The momentum transfer, also known as the *scattering vector*, $\mathbf{q}$, is defined as the difference between the incident wave vector, $\mathbf{k}_0$, and the scattered wave vector, $\mathbf{k}_s$, see Figure

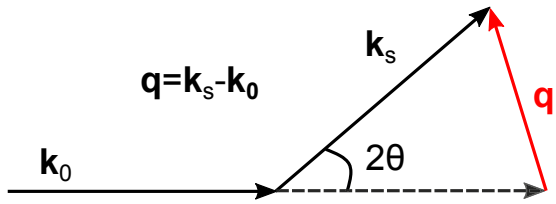


Figure 7.1: A schematic representation of the wave vectors and the resulting scattering vector (red).

7.1. The angle between the two wave vectors is defined as 2θ. Since the scattering is elastic, the magnitude of the two wave vectors is the same, that is, $|\mathbf{k}_0| = |\mathbf{k}_s| = 2\pi/\lambda$, where λ is the wavelength of the incident wave. Thus, the magnitude of the scattering vector, q, can be obtained accordingly:

$$q = |\mathbf{q}| = \frac{4\pi \sin(\theta)}{\lambda}. \tag{7.1}$$

7.1.2 The form factor

The *scattering intensity* is normally expressed as a function of the magnitude of the scattering vector: $I(q)$. The scattering intensity is dependent on the particle volume and the electron density, which means that larger particles give rise to larger scattering intensities. When considering scattering from particles, the scattering intensity can also be expressed by the following relation:

$$I(q) = P(q)S(q), \tag{7.2}$$

where $P(q)$ is known as the *form factor*, and $S(q)$ is the so-called *structure factor*. The form factor represents the *intra*-particle distances, which are derived from the interference pattern that arises due to scattering from individual particles. Similarly, the structure factor represents the *inter*-particle distances. Properties such as average shape and size distributions can be obtained from the form factor. To obtain the pure form factor, it is necessary to have very dilute samples. However, dilute samples are extremely sensitive to contaminations, since the scattering from small particles can be completely dominated by the scattering from larger particles. Thus, it is important to keep the samples pure, and to perform several measurements for averaging when trying to obtain the form factor. Examples of different IDP form factors are shown in Figure 7.2a.

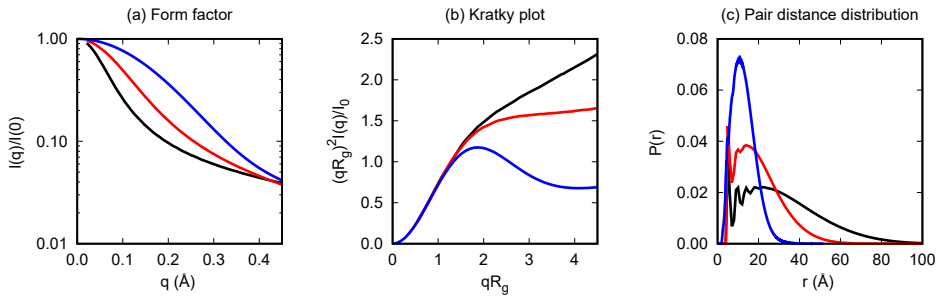


Figure 7.2: Examples of different plots obtainable from SAXS. (a) The form factor, (b) the Kratky plot, and (c) the pair distance distribution function. The blue curve represents a more globular shape, the red curve represents a flexible random chain, and the black curve represents a more rigid and extended random chain.

7.1.3 The Kratky plot

The shape and flexibility of a protein in solution can be determined by transforming the SAXS spectra into a *Kratky plot*. A Kratky plot is done by plotting $q^2 I(q)$ vs. q [22]. To make the plot dimensionless, $I(q)$ is normalized by the *forward scattering*, $I(0)$, and q is normalized by the radius of gyration. The shape of the resulting curve gives a good indication of the shape and flexibility of the protein. A well-defined globular protein will show as a bell-like shape in the Kratky plot, while a flexible, random chain will increase and reach a plateau. Further increase of the plateau at high q-values indicates a more rigid and extended structure. Examples of these different shapes are illustrated in the Kratky plot in Figure 7.2b.

7.1.4 The pair distance distribution function

By studying the form factor, it is possible to obtain the shape and the intra-particle distances expressed as a *pair distance distribution function*, $P(r)$. This is done by applying a Fourier transformation on the entire scattering profile [132]:

$$P(q) = 4\pi \int_0^\infty P(r) \frac{\sin(qr)}{qr} dr. \tag{7.3}$$

The shape of the pair distance distribution function also gives the shape of the particles in the studied solution. It is possible to obtain the radius of gyration and the forward scattering from the pair distance distribution function as well [22, 133]:

$$R_g^2 = \frac{\int P(r) r^2 dr}{2 \int P(r) dr}, \tag{7.4}$$

$$I(0) = 4\pi \int_0^{D_{max}} P(r) dr. \tag{7.5}$$

The pair distance distribution function is assumed to be zero at $r = 0$ and at D_{max}, which is the maximum extension of the protein chain. Examples of the pair distance distribution for different types of peptides are depicted in Figure 7.2c. Other shapes, as for example disks, hollow spheres, and dumbbells, would result in a very different appearance of the pair distance distribution function, which is nicely illustrated in Figure 5 in the paper by Svergun and Koch (2003) [106].

7.1.5 The Guinier approximation

Another way of obtaining the radius of gyration from the form factor is by using the *Guinier approximation* [22,106,133], which states that at sufficiently low values of q, the scattering profile can be approximated to:

$$I(q) = I(0) \exp\left(-\frac{q^2 R_g^2}{3} \right), \tag{7.6}$$

or rather:

$$\ln\big(I(q)\big) = \ln\big(I(0)\big) - \frac{R_g^2}{3} q^2. \tag{7.7}$$

By making a straight-line fit in the so-called *Guinier region* the forward scattering and the radius of gyration can be obtained as the y-axis intercept and the slope, respectively, see example in Figure 7.3. The approximation is restricted to $qR_g < 1.3$ for well-folded, globular proteins, but the Guinier region is even smaller for extended molecules, including many IDPs. For the analysis of IDPs, a value of $qR_g < 0.8$ has been recommended [22], although this value may vary depending on the nature of the IDPs.

In contrast to the pair distance distribution method, which utilizes the entire scattering profile, using such a restricted part of the curve may result in only a few data points, which reduces the reliability of the resulting radius of gyration. For this reason, utilizing the pair distance distribution function is considered to be more precise for determining the radius of gyration [133]. The pair distance distribution method is, however, very sensitive and consequently more difficult to reproduce. Therefore, it is advantageous to compare and report the radius of gyration from both methods.

7.2 Circular dichroism spectroscopy

By utilizing *circular dichroism* (CD) spectroscopy, it is possible to detect the secondary structure of proteins in solution. Although IDPs often lack secondary structure, it is still an important tool for detecting changes in the conformational ensemble as an

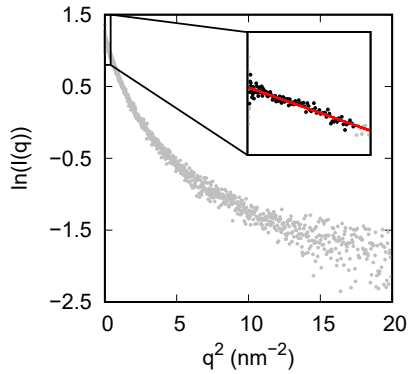


Figure 7.3: An example of a Guinier plot of an IDP. To obtain the forward scattering and the radius of gyration, linear regression (red line) is done on the black data points, which belongs to the Guinier region ($qR_g < 0.8$). The remaining scattering profile is shown in gray.

effect of changes in the solution environment. The basic principle behind the CD technique is the difference in absorption of left- and right-handed circularly polarized light. Differently polarized light is absorbed to different extents by optically active molecules in solution. The *difference in absorption*, ΔA, is often expressed in terms of *ellipticity*, θ, in degrees (deg) instead, and the two quantities are related through the following equation: $\theta = 32.98 \cdot \Delta A$ [134].

Studies of the secondary structure of proteins on a conventional CD instrument are done in a wavelength range of 180-260 nm. To obtain more reliable results it is necessary to measure in the far-UV region (170 nm and below), but this requires more elaborate techniques and synchrotron radiation. It is the absorbance of the peptide bond that makes it possible to determine the secondary structure of proteins. The characteristic spectra arise due to a strong $\pi \rightarrow \pi*$ transition around 190 nm, and a weaker $n \rightarrow \pi*$ transition around 220 nm [134]. Depending on the secondary structure of the studied protein, a characteristic CD spectrum will be observed. Different secondary structures give rise to different shapes in the spectra, see Figure 7.4. For example, an α-helix is characterized by negative bands around 210 and 220 nm, as well as a positive band between 190-195 nm. A negative band close to 220 nm and a positive band at 195 nm are characteristic for antiparallel β-sheets. IDPs that lack secondary structure are attributed a random coil structure that is recognized by a single negative band at low wavelengths in the CD spectrum. Less known is the extended PPII helix [13] that can be identified for some IDPs. This structure is defined by a weak positive band close to 220 nm and, a strong negative band around 190 nm [135].

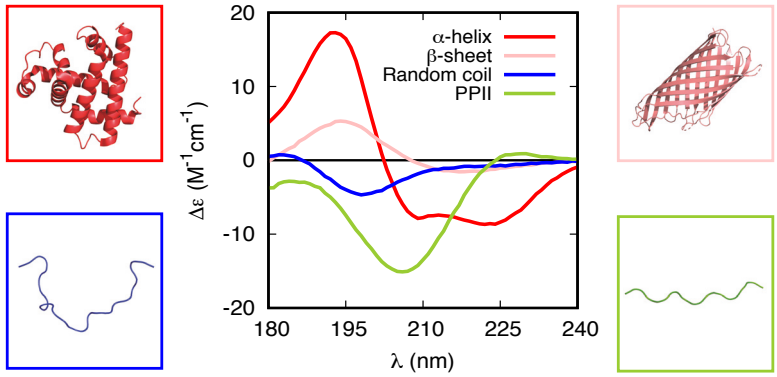


Figure 7.4: Examples of selected secondary structures that can be identified in a CD spectrum. All the plotted data was obtained from the Protein Circular Dichroism Data Bank (PCDDB) [136]. The protein structures in the corresponding boxes are examples; they are not necessarily the same proteins that are shown in the graph, but their secondary structure content is of the same type. The typical spectrum of the α-helix and its corresponding example are shown in red[1]. Similarly, the β-sheet examples are shown in pink[2], the random coil examples in blue[3], and the PPII helix examples in green[4].

To compare CD spectra, it is advantageous to normalize the spectra with respect to molar concentrations. This is done by converting the ellipticity to *molar ellipticity* or *mean residue ellipticity*. For proteins, the mean residue ellipticity is calculated from the mean residue weight, *MRW*, of the peptide bonds:

$$MRW = \frac{M}{N-1}, \tag{7.8}$$

where M is the molecular weight of the protein, and N is the number of amino acid residues in the protein chain. The mean residue ellipticity, $[\theta]$, at wavelength λ is then given by:

$$[\theta] = \frac{MRW \cdot \theta_\lambda}{10 \cdot d \cdot c}, \tag{7.9}$$

where θ_λ is the observed ellipticity at wavelength λ, d is the pathlength in cm, and c is the concentration in g/mL. The mean residue ellipticity is usually given in deg cm^2 dmol^{-1}. If the CD data is expressed as the difference in absorption, it is instead normalized by converting ΔA to the *molar differential extinction coefficient*, $\Delta\varepsilon$:

$$\Delta\varepsilon = \frac{\Delta A}{C \cdot d}. \tag{7.10}$$

[1]Spectrum CD0000048000 Lees et al. (2006) [137], and structure 1MNK; Krzywda and Wilkinson (1995) Interactions among residues CD3, E7, E10 and E11 in myoglobins: attempts to simulate the O2 and CO binding properties of *Aplysia* myoglobin; doi: 10.2210/pdb1mnk/pdb [138] from RCSB PDB [3].

[2]Spectrum CD0000116000, Abdul-Gader et al. (2011) [139], and structure 1UYO; Oomen et al. (2004) Translocator domain of autotransporter NalP from *Neisseria meningitidis*; doi: 10.2210/pdb1UYO/pdb [140] from RCSB PDB [3].

[3]Spectrum CD0006121000, Tolchard et al. (2018) [141] and in-house generated structure example.

[4]Spectrum CD0004553000, Lopes et al. (2014) [142] and in-house generated structure example.

The variable C is the molar concentration, which gives $\Delta\varepsilon$ the unit of $M^{-1}\,cm^{-1}$. Because of the previously mentioned relation between the ellipticity and the difference in absorption, conversion between the mean residue ellipticity and the molar differential extinction coefficient is done according to the following equation: $[\theta] = 3298 \cdot \Delta\varepsilon$.

The resolution obtained from CD measurements is quite low as it only provides general information about structural averages. The technique is also sensitive to the choice of solution components and contaminants. Several buffers and ions have shown to possess high absorption in the far-UV range. To minimize the absorbance from other species than the protein of interest, it is recommended to for example use low buffer concentrations, and to exchange chloride ions with sulphate or fluoride ions [134]. It is also possible to use a shorter cell pathlength to reduce the effect of disturbing species.

8. Research

This chapter summarizes the research and results that have been procured during my PhD studies. Most of the studies in this thesis concern histatin 5, which has been thoroughly studied by using several different theoretical and experimental techniques. Here, I first present histatin 5 while discussing its conformational properties as a single chain in bulk conditions. The first section deals with the effect of charge regulation in histatin 5, followed by a section investigating its temperature dependence. This section also concerns investigation and evaluation of different simulation methods. Further research concerning different MD force fields is presented in the subsequent section, where the simulated secondary structure of histatin 5 (and a few other peptides) is investigated. Toward the end of this chapter, the focus is shifted back to the conformational ensemble, but for a more potent variant of histatin 5, that is, the shorter P-113 peptide conjugated to the polyamine spermidine. Histatin 5 and its variants are of interest because of their antimicrobial properties, but we have also studied the bacterial KEIF peptide, which is summarized in a section of its own. Apart from characterizing the conformational ensemble of KEIF, the study also covered investigation of the interactions between the peptide and model membranes. The chapter is concluded with research implications and a discussion about future research connected to this work.

8.1 The effect of charge regulation in histatin 5

Because histatin 5 is comprised by such a large portion of His residues, it is important to consider charge regulation. It is known that histatin 5 interacts with negatively charged surfaces [143–145] and membranes (see Section 2.6), and it would be naive to assume that the His residues of histatin 5 are completely neutral in the vicinity of such structures under physiological conditions. Thus, histatin 5 has been investigated by simulations of its two extreme cases: (i) with all His residues neutral, and (ii) with all His residues protonated, that is, with a charge of $+1e$. In the latter case, histatin 5 has a net charge of $+12e$ instead of $+5e$. By considering the charge distribution of histatin 5, the con-

formational properties can be predicted. With FCR = 0.38 and NCPR = 0.21, histatin 5 with neutral His residues has been found to belong to the conformational class R3 and is thus predicted to be a polyampholytic IDP (see Figure 2.4, Section 2.2 for reference). When all the His residues are protonated, the peptide instead belongs to the R4 class (FCR = 0.67 and NCPR = 0.50) and thus behaves as a polyelectrolyte.

By performing MD simulations of histatin 5 with all His residues neutral (Hst5_HIE) and with all His residues protonated (Hst5_HIP), the conformational properties were investigated and compared. The radius of gyration was found to be larger for Hst5_HIP, and it was observed to sample more extended conformations than Hst5_HIE, see Figure 8.1a. Both distance distributions had similar shapes, with a FWHM of 0.54 nm and 0.51 nm for Hst5_HIE and Hst5_HIP respectively. Representative structures of the two histatin 5 variants are displayed in Figure 8.1b.

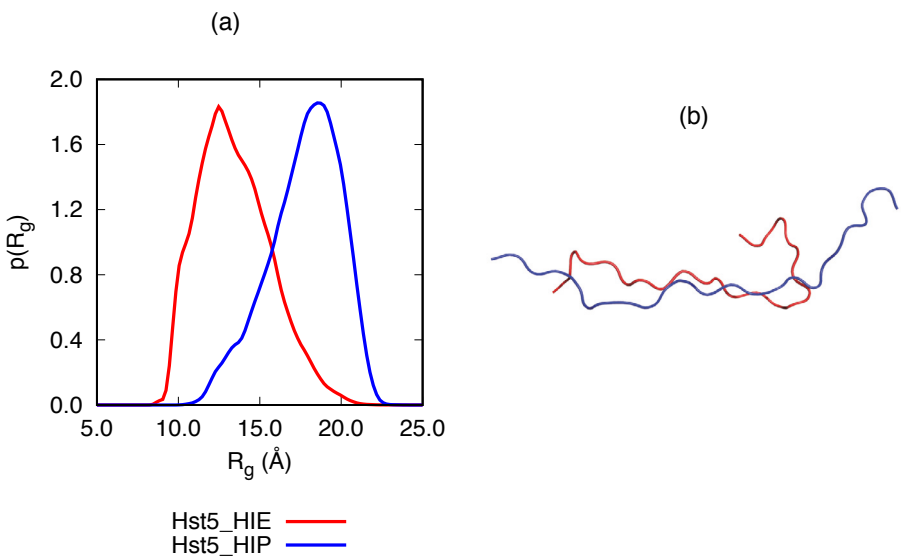


Figure 8.1: (a) Distance distribution of the radius of gyration, and (b) cartoon representations of the largest cluster conformer of the two Hst5 variants. Hst5_HIE is shown in red and Hst5_HIP in blue.

The secondary structure was estimated based on the backbone dihedral angles of each amino acid residue, and their relative populations in different regions of a Ramachandran map. Both variants of the peptide had large population of the PPII-helical and β-sheet regions of the Ramachandran map, and both of these regions were more populated for Hst5_HIP, see Figure 8.2. While Hst5_HIE displayed some α-helical structure elements, this was almost absent in Hst5_HIP. Thus, it seems like protonation of the HIS residues caused the PPII and β-sheet content of histatin 5 to increase at the expense of α-helical and random coil structures.

These results are extremely important for other studies of histatin 5 at for example lower

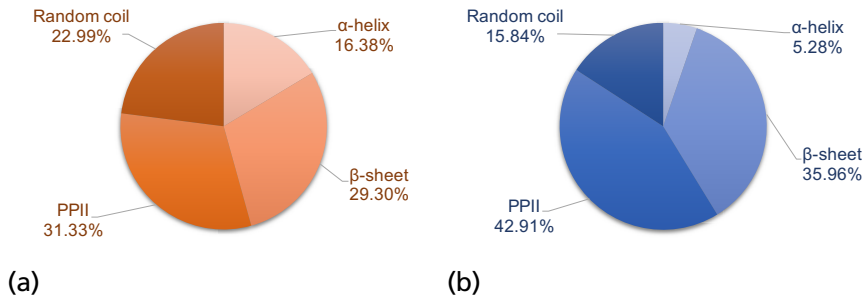


Figure 8.2: The secondary structure content of (a) Hst5_HIE and (b) Hst5_HIP from MD simulations.

pH values or in systems containing other charged chemical species, such as surfaces or membranes. Thus, further investigation was performed, based on the paper by Hyltegren et al. (2020) [146], in which surface adsorption of histatin 5 was studied using a combination of MC simulations and a library of structures obtained from MD simulations. In their paper, only histatin 5 with neutral HIS residues were investigated. Therefore, we decided to perform similar simulations but with both Hst5_HIE and Hst5_HIP, where the previously presented MD simulations were used to build the libraries of MD structures.

The MC simulations of Hst5_HIE and Hst5_HIP in the vicinity of negatively charged surfaces yielded similar results to what was obtained in the MD simulations, both for the radius of gyration and its distance distribution, as well as for the average secondary structure content per amino acid. It was also discovered that these properties were nearly unaffected by changes in the salt concentration. When it came to surface adsorption, clear differences were found between the two histatin 5 variants, and these results were affected by the salt concentration. Three regions were defined for determining adsorption: (A) bulk region (A $>$ 30 Å from the surface), (B) intermediate region (6 Å $<$ B $\leq$ 30 Å from the surface), and (C) adsorbed region (C $\leq$ 6 Å from the surface). A peptide was considered to belong to a region if its center of mass was found within the limits of that region. At low salt concentration, the majority of the simulated frames were located in region B for both Hst5_HIE and Hst5_HIP, whereas at high salt concentration there were most often located in the bulk (region A). Although higher salt concentrations seemed to screen the electrostatic attraction between the peptides and the surface, a larger fraction of Hst5_HIP was always found close to the surface compared to Hst5_HIE. There also seemed to be a small increase in the average radius of gyration of Hst5_HIP when going across the three regions from A to C. However, further investigation is needed to determine if this trend is of statistical significance.

8.2 Temperature dependence in simulations of histatin 5

During 2016 trial measurements and short simulations of histatin 5 at different temperatures had been performed and were indicating contradictory results for different methods. Thus, longer simulations using different techniques, as well as several different experimental techniques were performed to further investigate these discrepancies. SAXS, as well as atomistic MD simulations and coarse-grained MC simulations were used to study the change in the radius of gyration with temperature. However, because atomistic MD simulations are computationally expensive and SAXS beam time is precious, only a few carefully chosen temperatures were selected for these two techniques: 10 °C, 20 °C, 37 °C, and 50 °C. The selection was limited by the temperature range of the SAXS equipment (4-60 °C), and the same temperatures were selected for both techniques for easier comparison.

Atomistic MD simulations of histatin 5 were performed with three different combinations of force fields and water models: (i) Amber ff99SB-ILDN [89] and TIP3P [147] (MD(A3)), (ii) Amber ff99SB-ILDN and TIP4P-D [99] (MD(A4)), and (iii) CHARMM36m [93] with TIP3P (MD(C3)). Although the first combination, MD(A3), was known to give too compact structures for IPDs [86, 99], it was unknown if the temperature-induced conformational trends would be the same as for MD(A4). Thus, MD(A3) was still considered to be interesting for the study and was included, nevertheless. The combination of MD(A4) was previously known to work well for histatin 5 at room temperature and was therefore an obvious choice to use in this study. Simulations with the MD(C3) setup have previously been proven to work well for IDPs, which made it a good alternative for expanding the study to include a "new", different force field. The coarse-grained MC simulations of histatin 5 (MC(IW)) were performed using an implicit water model and were thus computationally cheap to perform compared to the MD simulations. This made it possible to perform the MC(IW) simulations at a wider selection of temperatures, although the selected temperatures were still chosen within the temperature range of the previous techniques (10-50 °C).

Two additional experimental techniques were also included in the study: CD spectroscopy and NMR spectroscopy. The CD spectroscopy was performed at the same four selected temperature that were used in the SAXS measurements and the MD simulations, and the aim of the measurements was to detect any temperature-induced changes to the secondary structure of histatin 5. The NMR spectroscopy of histatin 5 was performed in a temperature range of 5-50 °C. The aim of the NMR measurements was to obtain hydrodynamic radii of histatin 5 at the different temperatures, which were then converted to radii of gyration by a compaction factor of 1.1 that was obtained using the method presented by Marsh and Forman-Kay (2010) [148]. Re-evaluation of the paper after publication, and by using the method by Nygaard et al. (2017) [149]

as intended, revealed that a conversion factor of $\sim 0.91 \pm 0.05$ would have been a more reasonable choice. Although this would change the obtained values, it would not significantly affect the trend that is discussed in the paper, and thus not affect the conclusions either.

The resulting average radii of gyration, $\langle R_g \rangle$, at different temperatures are depicted in Figure 8.3. Interestingly enough, all the different methods used gave different temperature trends. No trend could be found from the SAXS measurements, and although some correlations were found from the MD simulations, they were not statistically significant enough to draw any conclusion. Only the NMR and the MC(IW) simulations displayed strong and significant temperature correlations, although in the opposite direction of each other. Further investigation of the SAXS data revealed that no obvious differences between the results at different temperatures could be discerned for any of the systems.

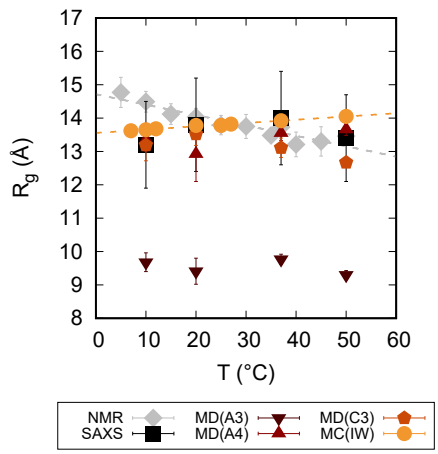


Figure 8.3: The radius of gyration, $\langle R_g \rangle$, of histatin 5 as a function of temperature of the six different methods. The dashed lines are linear regression trend lines. All data points include error bars, although the errors are too small to be visible for some data points.

A more obvious temperature-induced conformational change was observed in the CD spectra, which are depicted in Figure 8.4. Negative bands around 191 and 240 nm, as well as a positive band around 222 nm, were observed for all temperatures, although the absolute intensites of the bands were found to decrease with increased temperature. The shape of the observed spectra was a clear indication of PPII structure. By studying the change in ellipticity at 191 and 222 nm, a significant linear correlation with temperature was found, which indicated a destabilization of the PPII structure.

Secondary structure analysis was performed on representative structures obtained from conformation clustering of each simulated MD trajectory by utilizing a DSSP-PPII

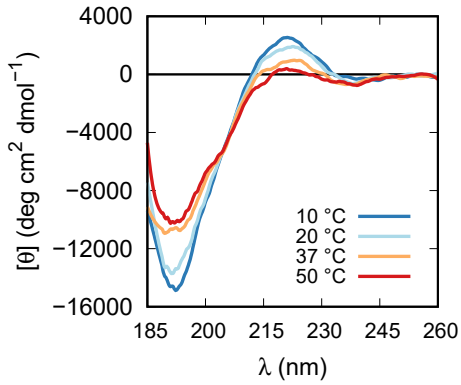


Figure 8.4: Normalized CD spectra of histatin 5 at the four different temperatures.

program similar to the program developed by Mansiaux et al. (2011) [131] and to the software used in the PolyPrOnline database [150]. To our surprise, the analysis revealed no PPII structure for any of the simulated structures. With the exception of the MD(A3) simulation, which possessed some degree of helical structures, the MD simulations were dominated by bends, turns and random coils instead. Initial speculation of why no PPII structure was identified mainly revolved around the problem of sampling; maybe enhanced sampling or longer simulations would be needed to capture the correct secondary structure of IDPs in MD simulations. However, by comparing MD(A4) after 5 µs to after 7 µs provided identical representative structures from the conformation cluster analysis, which strongly indicated that the sampling was sufficient for this simulation.

Because the temperature trends so far had been all over the place in the MD simulations, the temperature dependence of the two MD water models were investigated. MD simulations of pure water (TIP3P and TIP4P-D), were performed at different temperatures using the Amber ff99SB-ILDN force field. The dielectric constant was then computed compared to reference values [151–154]. The dielectric constant of both water models was found to decrease linearly with increasing temperature, although the actual magnitudes of the values did not match the reference values; TIP3P generated values larger than the reference, and TIP4P-D generated values smaller than the reference. However, since both water models displayed a correct temperature dependence in the temperature range of interest, it was concluded that the deviating behavior of histatin 5 at different temperatures in the MD simulations is highly unlikely to be caused by the water models.

The dielectric constant was also investigated in the coarse-grained MC model by performing additional simulations of histatin 5 but by using the dielectric constants obtained from the MD simulations in place of the standard reference values. Interestingly enough, this did barely affect the resulting radius of gyration at all. The results were initially surprising because the largest difference between the simulated dielectric constants at a single temperature can be compared to the difference between the dielectric constant of water and the dielectric constant of for example ethanol. Nevertheless, the results indicated that the temperature-induced conformational changes of histatin 5 in the MC simulations are completely independent of electrostatic interactions in the system. To further investigate this result, coarse-grained MC simulations were also performed of two other chains with equal length as histatin 5 but consisting completely of (i) neutral beads and (ii) beads with a charge of $+0.2e$ each (net change $+5e$). Although this led to different values of the radius of gyration, the temperature trend was revealed to be more or less the same for all three chains. From these results it was clear that the temperature dependence was not electrostatically driven in the coarse-grained MC model.

8.3 PPII structure in simulations of histatin 5

It was unexpected that the previous analysis of the simulated histatin 5 trajectories did not display any PPII structure. Thus, further investigation was performed using a different approach. By simply analyzing the backbone dihedral angles of the trajectories, Ramachandran plots were made, which clearly showed that the simulated histatin 5 in fact possessed substantial PPII content. Subsequently, this also revealed problems with the DSSP-PPII program, which was immediately updated to properly assign PPII structure. It was also discovered that the different force fields that had been used (Amber ff99SB-ILDN and CHARMM36m) had propensities for different secondary structure elements, although both provided PPII structure. This sparked an interest for further comparing different force fields and their ability to sample PPII structure in histatin 5 and in other small peptides that are known to possess PPII structure.

Four different force field that are known work for simulating IDPs were chosen: (A) Amber ff99SB-disp [96], (B) Amber ff99SB-ILDN [89,99], (C) CHARMM36IDPSFF [97, 98], and (D) CHARMM36m [93]. The force fields were used in combination with their recommended water models. Histatin 5 and four other peptides, as well as four variants of one of the peptides, were simulated for at least a total of 5 µs each. All the chosen force fields provided similar average values of the radius of gyration, although the values of CHARMM36IDPSFF were slightly smaller compared to the rest. By using PCA, force fields were shown to sample the same conformational spaces for each individual peptide. This was however done with different probabilities, and

CHARMM36IDPSFF seemed to be most adept at exploring all of the conformational space at the chosen temperature. The secondary structure estimation of this study was performed in the same way as it was in the study of Section 8.1. All force field sampled PPII structure but to different extents. Amber ff99SB-disp had the highest population in the PPII region for most of the peptides, whereas Amber ff99SB-ILDN and CHARMM36m were more prone to sample β-sheet structure, and CHARMM36IDP often sampled more "random coils". The force fields were more directly compared by investigating the conformation clustering in the simulations of histatin 5. For this analysis, all simulations with the different force fields were concatenated to a single trajectory on which the conformation clustering was performed. The clusters were then analyzed to obtain the relative population of the individual force fields, see Figure 8.5. These results revealed that the force fields populated the different clusters unevenly, which further indicates that there are slight biases toward different conformations for the different force fields.

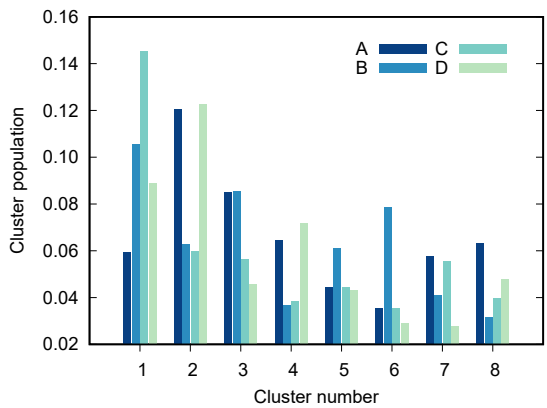


Figure 8.5: Weighted cluster population from the individual force fields in the top eight clusters of histatin 5. A = Amber ff99SB-disp, B = Amber ff99SB-ILDN, C = CHARMM36IDP, and D = CHARMM36m.

An effort was made to determine which of the force fields were more accurate. This was done by making CD predictions using SESCA [155]. Unfortunately, the algorithm was unable to provide any simulated spectrum that could be matched with any of the experimental spectra.

As a final part of this study, the correlation between the number of Pro residues and the PPII content was investigated for P_{13} and four variants containing only Ala and Pro, see Figure 8.6. A linear regression was performed, which revealed that all force fields provided linear correlations of statistical significance ($p < 0.05$). This correlation does however need to be investigated further to see how it is affected by for example chain length and the relative position of the Pro residues within the sequences.

(a)

Name	Sequence	#P
P_{13}	PPPPPPPPPPPPP	13
V1: P_6AP_6	PPPPPPAPPPPPP	12
V2: $(PA)_6P$	PAPAPAPAPAPAP	7
V3: $(P_3A_2)_2P_3$	PPPAAPPPAAPPP	9
V4: $(A_3P_2)_2A_3$	AAAPPAAAPPAAA	4

(b)

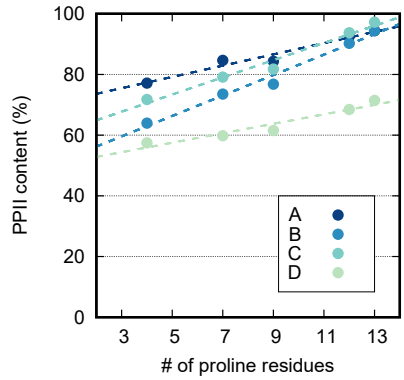


Figure 8.6: (a) The amino acid sequences and the total number of Pro residues (#P) in each sequence of P_{13} and its four variants. (b) PPII content as a function of the total number of proline residues for the four different force fields. A = Amber ff99SB-disp, B = Amber ff99SB-ILDN, C = CHARMM36IDP, and D = CHARMM36m.

8.4 Conformational properties of Histatin 5-spermidine conjugates

Because the candidacidal effect was shown to be increased when the shorter histatin 5 fragment, P-113, was conjugated to the polyamine spermidine, we decided to investigate its conformational ensemble to see how it contributes to this increased candidacidal effect. By considering the charge distribution of histatin 5 and its variants, the conformational properties were predicted, see Table 8.1. The predictions are based on the amino acid charges at physiological pH (pH 7.4), assuming positively charged N-terminal residues and negatively charged C-terminal residues, as well as completely neutral histidine residues.

Table 8.1: Net charge, FCR, NCPR, and the corresponding conformational class for the protein sequences in aqueous solution at physiological pH.

Name	q/e	FCR	NCPR	Class
Histatin 5	+5	0.38	0.21	R3
P-113	+5	0.58	0.42	R4
CSpd	+6	0.40	0.40	R4
NSpd	+4	0.40	0.27	R3

The shorter histatin 5 variant, P-113 (also known as $Hst5_{4-15}$ in Paper 1), was pre-

dicted to belong to the R4 class, which is reasonable since it has the same net charge as histatin 5 but distributed over half the number of amino acid residues. When conjugated to spermidine, a glycine linker consisting of three glycine residues in a row has been added to the P-113 chain, which makes the peptide part of the conjugates increase from twelve to fifteen amino acid residues. Since one terminus each of NSpd and CSpd is connected to spermidine, they only contain one charged terminal residue each, see Figure 8.7, which was accounted for in the predictions. Conjugation of spermidine at the C-terminus of P-113 did not seem to impose a large difference for the prediction, as the protein part of CSpd also was predicted to belong to the R4 class. Conjugation at the N-terminus, however, seems to predict slightly more flexibility as NSpd instead was predicted to belong to the R3 class. In summary, based on the charge distribution only, histatin 5 and the peptide part of NSpd are expected to form polyampholytic coils or hairpins, whereas P-113 and the peptide part of CSpd are expected to form polyelectrolytic rods/coils (see section 2.2 for reference) in aqueous solution at physiological pH.

Figure 8.7: A schematic illustration of the charge distribution of (from top to bottom) histatin 5, P-113, CSpd, and NSpd. The N-termini and the C-termini are explicitly included as NT and CT respectively. Succinic acid and spermidine are denoted by SAC and SPD respectively. Gray residues are neutral, blue residues are positive ($+1e$), red residues are negative ($-1e$), and the green spermidine has a charge of $+2e$.

SAXS curves representing the histatin 5-spermidine conjugates were obtained from both experimental measurements and from MD simulations. Initial comparisons of the experimental and the computational data revealed a large discrepancy between the experimental and computational data, as the SAXS curves did not overlap properly. Further investigation proceeded, and after reevaluating the system and the methods it was hypothesized that the spermidine could not be detected properly in the SAXS measurements. Calculation of the scattering length density of spermidine was performed and revealed to be close to that of the bulk used in the experiments (20 mM Tris and 140 mM NaCl in aqueous solution), a case which is known for giving a very low contrast in the measurements. Hence, the experimental SAXS curve of the histatin 5-spermidine conjugate was compared to that of the P-113 fragment only. The comparison revealed nearly identical SAXS curves, thus confirming the hypothesis. To further support the hypothesis, SAXS curves representing only the peptide part of the simulated conjugates were also obtained and compared to the experimental SAXS curve of the histatin 5-spermidine conjugate (see Figure 8.8). Finally, the experimental and the simulated results were in reasonable agreement and the previously stated hypothesis was again confirmed.

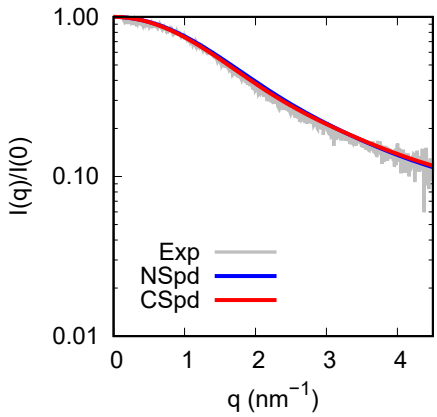


Figure 8.8: SAXS spectrum of the histatin 5-spermidine conjugate from experiment (gray) compared to the SAXS spectra representing only the peptide part of the simulated conjugates (red and blue).

The conclusion that the spermidine is basically "invisible" in the SAXS measurements provided an additional important piece of information: the conformational properties of P-113 are unaffected by conjugation to spermidine. This was also observed by investigating the distance distributions of the radius of gyration (and the end-to-end distance) of the simulated P-113 and the histatin 5-spermidine conjugates, where the peptide part of the histatin 5-spermidine conjugates were seen to sample the same distributions as P-113, see Figure 8.9. Similarly, PCA of the protein backbone of both P-113 and the histatin 5-spermidine conjugates proved that the same conformational space was sampled in all different systems.

Evaluation of the SAXS curves representing the histatin 5-spermidine conjugates showed that the protein part of the conjugates possessed extended flexible conformations, identified by the characteristic shape of the Kratky plots. For the conjugates in their entirety, a slightly more extended and rigid average conformation was indicated by the shape of the Kratky plot and the $P(r)$ of the simulated data, see Figure 8.10.

When comparing the results of the two types of conjugates from the simulations, CSpd was found to be more rigid and extended relative to NSpd, despite the latter having a longer contour length because of its succinic acid linker. In fact, the presence of the succinic linker could be an important factor causing the observed differences. Conjugation at the N-terminus decreases the charge density of the entire NSpd conjugate, both by increasing the contour length of the molecule, and also by effectively removing one positive charge by not having any N-terminus. In contrast, CSpd has a shorter contour length as well as no negative C-terminal charge, which leads to a higher charge density and increased intramolecular electrostatic repulsion. This could also be related

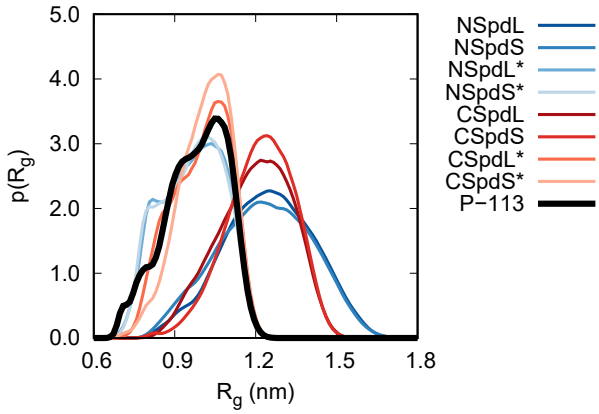


Figure 8.9: Distance distribution functions of the radius of gyration of the simulated conjugates and the P-113 fragment. The conjugates have been analyzed in their entirety (NSpdL, NSpdS, CSpdL and CSpdS) and by considering their peptide part only (NSpdL*, NSpdS*, CSpdL* and CSpdS*).

to the fact that CSpd has been shown to have slightly superior fungicidal activity compared to NSpd [76]. It was hypothesized that the more rigid and extended structure of CSpd exposes the spermidine further to the solvent, which makes it more accessible for recognition by the polyamine transporters.

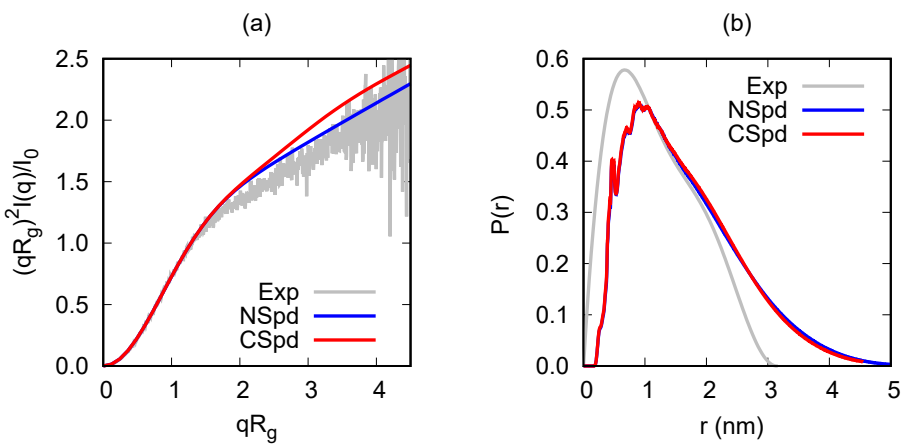


Figure 8.10: (a) Kratky plot and (b) P(r) of the histatin 5-spermidine conjugate from experiments (gray), as well as from the simulated conjugates (blue and red).

The secondary structure of histatin 5 and the histatin 5-spermidine conjugates were investigated by performing CD measurements. The resulting spectra (see Figure 8.11) showed expected α-helical structure for histatin 5, but no characteristic features associated with established secondary structure could be seen for the conjugates, which

were thus considered to be completely disordered in this medium. The experiments were performed in trifluoroethanol (TFE) because, at the time of the study in Paper 1, histatin 5 was considered to possess no secondary structure in aqueous solution. Later, we learned that histatin 5 actually possess some degree PPII structure in aqueous solution (see Section 8.2). Therefore, it does not seem unreasonable for the histatin 5-spermidine conjugates to also have the ability to assume PPII structure. In fact, since PPII structure is quite extended, the histatin 5-spermidine conjugates might even have a higher tendency to form PPII structure compared to histatin 5.

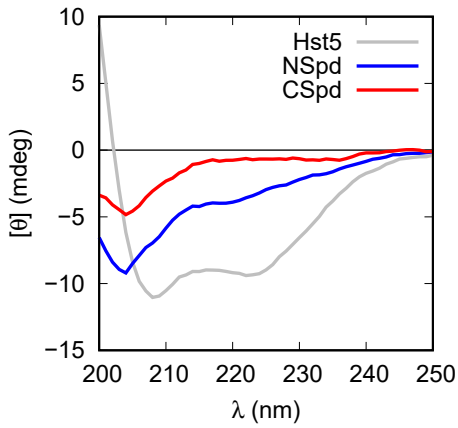


Figure 8.11: CD spectra of histatin 5 (gray) and the two conjugates (blue and red) in TFE.

8.5 Examining KEIF

The interest of KEIF sparked from the suggestion that it might be intrinsically disordered, and from the fact that its function in MgtA was unknown. Because of our previous experience with small IDPs, and because of the limited computer resources, it was decided that we would study KEIF only, that is, without the rest of the MgtA protein. Initial investigation was done based on the primary structure of the peptide, which predicted that KEIF would assume globular conformations in aqueous solution and that it was unlikely to be able to traverse a cell membrane on its own because of its low hydrophobicity.

Bulk properties of KEIF were investigated using SAXS, CD spectroscopy, and atomistic MD simulations. Both the SAXS results and the simulations agreed on the conformational properties of KEIF, which were found to be a fully flexible and slightly extended peptide, despite being predicted to form globules in aqueous solution. Additionally,

ensemble optimization method (EOM) [156, 157] was used to fit theoretical scattering intensities to the experimental SAXS data. There was good correspondence between the scattering curves from the MD simulations and the EOM. The SAXS curves are depicted in Figure 8.12, where both the from factor (a) and the Kratky plot (b) is shown.

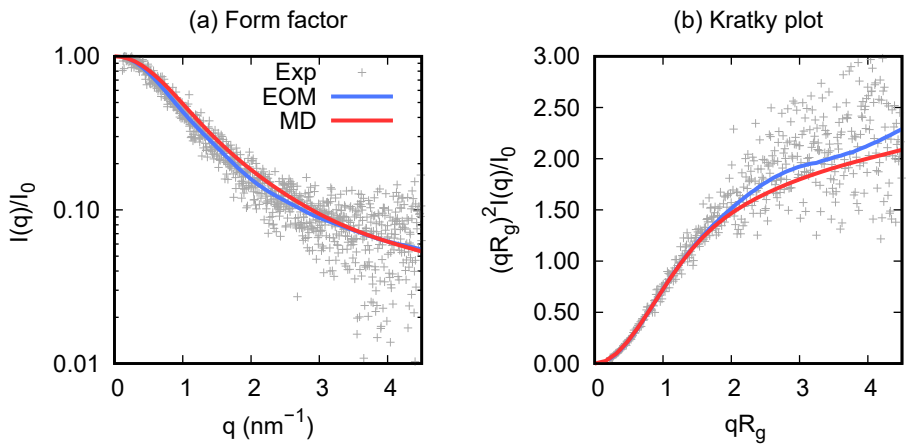


Figure 8.12: (a) KEIF form factor, and (b) Kratky plot of the experimental SAXS results (grey) compared to EOM (blue) and MD simulations (red).

The CD spectroscopy showed that KEIF was dominated by β-sheet structure and "random coils" in aqueous solution. Analysis of the spectra was done using BeSt-Sel [158, 159], which does not include the PPII helix, although the shape of the spectra indicates the presence of PPII structure. Similarly to histatin 5, KEIF also adopt more α-helical conformations in TFE. By analyzing the backbone dihedral angles from the simulated trajectory, the simulations confirmed that KEIF possess a significant amount of PPII helical structure under bulk conditions, see Figure 8.13. The simulations also revealed some β-structure but only little α-helical structure.

Because KEIF is part of a membrane protein, it was also of interest to study its interactions with model membranes. This was investigated exclusively by experimental methods where interactions between KEIF and neutral and anionic vesicles were studied using *dynamic light scattering* (DLS), *laser Doppler velocimetry* (LDV), *cryogenic transmission electron microscopy* (cryo-TEM), and CD spectroscopy. The results suggested that KEIF does not possess any means for significant interactions with neutral model membranes. Interaction with anionic vesicles were however observed by all methods, which together suggested that KEIF adsorbs to the surface of anionic membranes. Additionally, the CD spectroscopy revealed a conformational change in KEIF in the presence of the anionic vesicles; the α-helical content increased to 10 % from previously being completely absent. Thus, there is a possibility that KEIF actually is significant

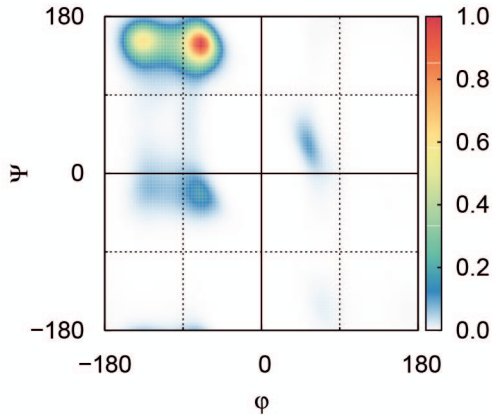


Figure 8.13: A Ramachandran plot obtained from analysis of the backbone dihedral angles in the trajectory from MD simulations of KEIF. The plot is normalized for a maximum intensity of 1.

for the function of MgtA in the bacterial cells. For example, it was suggested that KEIF could be part of the mechanism that anchors the MgtA protein in the bilayer as illustrated in Figure 8.14, although further research is needed to confirm this hypothesis.

8.6 Outlook

The research of this work has mainly focused on studying the conformational properties of single chain IDPs, in particularly histatin 5 and KEIF. Because of its role as a saliva protein, and because of its antimicrobial properties, this research concerning histatin 5 might, in the long run, aid the development of new saliva substitutes and new antimicrobial drugs. On a similar theme, but in a different way, an understanding of how bacterial proteins like MgtA work will also aid the development of new antibiotics.

Although the single chain properties are important, it is also important to know how the conformational properties are affected by higher concentrations and how they might change in the presence of surfaces and membranes. The effect of changing the conditions accordingly may be observed by applying different experimental techniques. However, because of the flexible nature of the IDPs, computational studies will still be vital to elucidate the conformational details of the IDPs, which in turn are needed to gain a holistic understanding of how the IDPs perform their biological functions. New force fields for MD simulations of IDPs are continuously being developed and it is important to remain updated on this field to ensure the chosen methods work well for the specific systems.

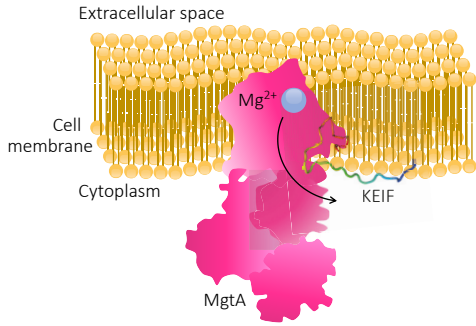


Figure 8.14: A schematic cross-section of MgtA (pink) in a cell membrane (yellow), illustrating how KEIF (rainbow cartoon) could aid in anchoring MgtA to the cell membrane.

Histatin 5 is a well-studied peptide that is frequently involved in new research studies. Nevertheless, its complete mechanism still needs to be defined for its full potential to be unlocked. Apart from being a useful antimicrobial agent, it has also been shown to possess other interesting properties, such as being able to spontaneously forming cushioned model membranes [160]. Therefore, further studies should be performed investigating surface and membrane interactions of histatin 5 and its derivatives. One presently ongoing computational project in our group concerns the interaction of the P-113 peptide with model membranes.

In our study of KEIF, we hypothesized that this IDR might be part of a mechanism that anchors the MgtA protein to the bacterial cell membrane. This hypothesis should be further explored by for example performing experiments of MgtA with and without the KEIF region, in the vicinity of model membranes.

9. References

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

[2] J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*. W. H. Freeman and Company, 7 ed., 2012.

[3] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, 2000.

[4] D. Ma, N. R. Brandon, T. Cui, V. Bondarenko, C. Canlas, J. S. Johansson, P. Tang, and Y. Xu, "Four-α-helix bundle with designed anesthetic binding pockets. part I: structural and dynamical analyses," *Biophysical journal*, vol. 94, no. 11, pp. 4454–4463, 2008.

[5] Y. Hori, H. Fujiwara, W. Fujiwara, and K. Makabe, "Grafting a short chameleon sequence from αb crystallin into a β-sheet scaffold protein," *Proteins: Structure, Function, and Bioinformatics*, vol. 87, no. 5, pp. 416–424, 2019.

[6] J. R. Tame and B. Vallone, "The structures of deoxy human haemoglobin and the mutant Hb Tyrα42His at 120 K," *Acta Crystallographica Section D: Biological Crystallography*, vol. 56, no. 7, pp. 805–811, 2000.

[7] V. N. Uversky, C. J. Oldfield, U. Midic, H. Xie, B. Xue, S. Vucetic, L. M. Iakoucheva, Z. Obradovic, and A. K. Dunker, "Unfoldomics of human diseases: linking protein intrinsic disorder with diseases," *BMC genomics*, vol. 10, no. 1, pp. 1–17, 2009.

[8] V. N. Uversky, J. R. Gillespie, and A. L. Fink, "Why are "natively unfolded" proteins unstructured under physiologic conditions?," *Proteins: structure, function, and bioinformatics*, vol. 41, no. 3, pp. 415–427, 2000.

[9] P. Tompa, "Intrinsically unstructured proteins," *Trends in biochemical sciences*, vol. 27, no. 10, pp. 527–533, 2002.

[10] S. Krimm and M. L. Tiffany, "The circular dichroism spectrum and structure of unordered polypeptides and proteins," *Israel Journal of Chemistry*, vol. 12, no. 1-2, pp. 189–200, 1974.

[11] H. J. Dyson and P. E. Wright, "Intrinsically unstructured proteins and their functions," *Nature reviews Molecular cell biology*, vol. 6, no. 3, pp. 197–208, 2005.

[12] F. Zhu, J. Kapitan, G. E. Tranter, P. D. Pudney, N. W. Isaacs, L. Hecht, and L. D. Barron, "Residual structure in disordered peptides and unfolded proteins from multivariate analysis and ab initio simulation of Raman optical activity data," *Proteins: Structure, Function, and Bioinformatics*, vol. 70, no. 3, pp. 823–833, 2008.

[13] A. A. Adzhubei, M. J. Sternberg, and A. A. Makarov, "Polyproline-II helix in proteins: structure and function," *Journal of molecular biology*, vol. 425, no. 12, pp. 2100–2132, 2013.

[14] S. Kakinoki, Y. Hirano, and M. Oka, "On the stability of polyproline-I and II structures of proline oligopeptides," *Polymer Bulletin*, vol. 53, no. 2, pp. 109–115, 2005.

[15] Z. Shi, K. Chen, Z. Liu, and N. R. Kallenbach, "Conformation of the backbone in unfolded proteins," *Chemical reviews*, vol. 106, no. 5, pp. 1877–1897, 2006.

[16] V. N. Uversky, "Intrinsically disordered proteins and their environment: effects of strong denaturants, temperature, pH, counter ions, membranes, binding partners, osmolytes, and macromolecular crowding," *The protein journal*, vol. 28, no. 7-8, pp. 305–325, 2009.

[17] 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.

[18] 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.

[19] K. M. Ruff, "Predicting conformational properties of intrinsically disordered proteins from sequence," in *Intrinsically Disordered Proteins*, pp. 347–389, Springer, 2020.

[20] P. Bernadó and D. I. Svergun, "Structural insights into intrinsically disordered proteins by small-angle X-ray scattering," in *Instrumental analysis of intrinsically disordered proteins: Assessing structure and conformation*, pp. 451–476, John Wiley and Sons Hoboken, New Jersey, 2010.

[21] E. V. Kuzmenkina, C. D. Heyes, and G. U. Nienhaus, "Single-molecule FRET study of denaturant induced unfolding of RNase H," *Journal of molecular biology*, vol. 357, no. 1, pp. 313–324, 2006.

[22] V. Receveur-Bréchot and D. Durand, "How random are intrinsically disordered proteins? A small angle scattering perspective," *Current Protein and Peptide Science*, vol. 13, no. 1, pp. 55–75, 2012.

[23] M. R. Jensen, R. W. Ruigrok, and M. Blackledge, "Describing intrinsically disordered proteins at atomic resolution by NMR," *Current opinion in structural biology*, vol. 23, no. 3, pp. 426–435, 2013.

[24] V. N. Uversky, V. Dave, 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," *Chemical reviews*, vol. 114, no. 13, pp. 6844–6879, 2014.

[25] S. P. Humphrey and R. T. Williamson, "A review of saliva: normal composition, flow, and function," *The Journal of prosthetic dentistry*, vol. 85, no. 2, pp. 162–169, 2001.

[26] L. C. Schenkels, E. C. Veerman, and A. V. Nieuw Amerongen, "Biochemical composition of human saliva in relation to other mucosal fluids," *Critical reviews in oral biology & medicine*, vol. 6, no. 2, pp. 161–175, 1995.

[27] J. Tenovuo, "Clinical applications of antimicrobial host proteins lactoperoxidase, lysozyme and lactoferrin in xerostomia: efficacy and safety," *Oral diseases*, vol. 8, no. 1, pp. 23–29, 2002.

[28] H. Välimaa, M. Waris, V. Hukkanen, M. Blankenvoorde, A. Nieuw Amerongen, and J. Tenovuo, "Salivary defense factors in herpes simplex virus infection," *Journal of dental research*, vol. 81, no. 6, pp. 416–421, 2002.

[29] Y. Yamaguchi, M. Semmel, L. Stanislawski, A. Strosberg, and M. Stanislawski, "Virucidal effects of glucose oxidase and peroxidase or their protein conjugates on human immunodeficiency virus type 1.," *Antimicrobial agents and chemotherapy*, vol. 37, no. 1, pp. 26–31, 1993.

[30] M. J. Levine, "Development of artificial salivas," *Critical Reviews in Oral Biology & Medicine*, vol. 4, no. 3, pp. 279–286, 1993.

[31] J. W. Millsop, E. A. Wang, and N. Fazel, "Etiology, evaluation, and management of xerostomia," *Clinics in Dermatology*, vol. 35, no. 5, pp. 468–476, 2017.

[32] J. Guggenheimer and P. A. Moore, "Xerostomia: etiology, recognition and treatment," *The Journal of the American Dental Association*, vol. 134, no. 1, pp. 61–69, 2003.

[33] H. Mese and R. Matsuo, "Salivary secretion, taste and hyposalivation," *Journal of oral rehabilitation*, vol. 34, no. 10, pp. 711–723, 2007.

[34] P. C. Fox, "Xerostomia: recognition and management," *Dental Assistant*, vol. 77, no. 5, p. 18, 2008.

[35] J. M. Plemons, I. Al-Hashimi, and C. L. Marek, "Managing xerostomia and salivary gland hypofunction: executive summary of a report from the American Dental Association Council on Scientific Affairs," *The Journal of the American Dental Association*, vol. 145, no. 8, pp. 867–873, 2014.

[36] S. S. Anttila, M. L. Knuuttila, and T. K. Sakki, "Depressive symptoms as an underlying factor of the sensation of dry mouth," *Psychosomatic medicine*, vol. 60, no. 2, pp. 215–218, 1998.

[37] M. Bergdahl and J. Bergdahl, "Low unstimulated salivary flow and subjective oral dryness: association with medication, anxiety, depression, and stress," *Journal of dental research*, vol. 79, no. 9, pp. 1652–1658, 2000.

[38] L. M. Sreebny and S. S. Schwartz, "A reference guide to drugs and dry mouth–2nd edition," *Gerodontology*, vol. 14, no. 1, pp. 33–47, 1997.

[39] C. Scully Cbe, "Drug effects on salivary glands: dry mouth," *Oral diseases*, vol. 9, no. 4, pp. 165–176, 2003.

[40] J. Karbach, C. Walter, and B. Al-Nawas, "Evaluation of saliva flow rates, Candida colonization and susceptibility of Candida strains after head and neck radiation," *Clinical oral investigations*, vol. 16, no. 4, pp. 1305–1312, 2012.

[41] P. A. Moore, J. Guggenheimer, K. R. Etzel, R. J. Weyant, and T. Orchard, "Type 1 diabetes mellitus, xerostomia, and salivary flow rates," *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, vol. 92, no. 3, pp. 281–291, 2001.

[42] M. Tanasiewicz, T. Hildebrandt, and I. Obersztyn, "Xerostomia of various etiologies: A review of the literature.," *Advances in clinical and experimental medicine: official organ Wroclaw Medical University*, vol. 25, no. 1, p. 199, 2016.

[43] R. M. López-Pintor, E. Casañas, J. González-Serrano, J. Serrano, L. Ramírez, L. de Arriba, and G. Hernández, "Xerostomia, hyposalivation, and salivary flow in diabetes patients," *Journal of diabetes research*, vol. 2016, 2016.

[44] R. Billings, H. Proskin, and M. Moss, "Xerostomia and associated factors in a community-dwelling adult population," *Community dentistry and oral epidemiology*, vol. 24, no. 5, pp. 312–316, 1996.

[45] D. Greenspan, "Xerostomia: diagnosis and management.," *Oncology (Williston Park, NY)*, vol. 10, no. 3 Suppl, p. 7, 1996.

[46] M. Rad, S. Kakoie, F. N. Brojeni, and N. Pourdamghan, "Effect of long-term smoking on whole-mouth salivary flow rate and oral health," *Journal of dental research, dental clinics, dental prospects*, vol. 4, no. 4, p. 110, 2010.

[47] S. F. Cassolato and R. S. Turnbull, "Xerostomia: clinical aspects and treatment," *Gerodontology*, vol. 20, no. 2, pp. 64–77, 2003.

[48] R. Cohen, F. Roth, E. Delgado, D. Ahearn, and M. Kalser, "Fungal flora of the normal human small and large intestine," *New England Journal of Medicine*, vol. 280, no. 12, pp. 638–641, 1969.

[49] J. A. Vazquez and J. D. Sobel, "Mucosal candidiasis.," *Infectious disease clinics of North America*, vol. 16, no. 4, p. 793, 2002.

[50] P. Eggimann, J. Garbino, and D. Pittet, "Epidemiology of Candida species infections in critically ill non-immunosuppressed patients," *The Lancet infectious diseases*, vol. 3, no. 11, pp. 685–702, 2003.

[51] V. J. Fraser, M. Jones, J. Dunkel, S. Storfer, G. Medoff, and W. C. Dunagan, "Candidemia in a tertiary care hospital: epidemiology, risk factors, and predictors of mortality," *Clinical Infectious Diseases*, vol. 15, no. 3, pp. 414–421, 1992.

[52] D. P. Lew and F. A. Waldvogel, "Osteomyelitis," *The Lancet*, vol. 364, no. 9431, pp. 369–379, 2004.

[53] P. Vergidis, C. J. Clancy, R. K. Shields, S. Y. Park, B. N. Wildfeuer, R. L. Simmons, and M. H. Nguyen, "Intra-abdominal candidiasis: the importance of early source control and antifungal treatment," *PLoS One*, vol. 11, no. 4, p. e0153247, 2016.

[54] S. I. Blot, K. H. Vandewoude, and J. J. De Waele, "Candida peritonitis," *Current opinion in critical care*, vol. 13, no. 2, pp. 195–199, 2007.

[55] B. J. Kullberg and M. C. Arendrup, "Invasive candidiasis," *New England Journal of Medicine*, vol. 373, no. 15, pp. 1445–1456, 2015.

[56] P. G. Pappas, M. S. Lionakis, M. C. Arendrup, L. Ostrosky-Zeichner, and B. J. Kullberg, "Invasive candidiasis," *Nature Reviews Disease Primers*, vol. 4, no. 1, pp. 1–20, 2018.

[57] A. A. Cleveland, L. H. Harrison, M. M. Farley, R. Hollick, B. Stein, T. M. Chiller, S. R. Lockhart, and B. J. Park, "Declining incidence of candidemia and the shifting epidemiology of candida resistance in two US metropolitan areas, 2008–2013: results from population-based surveillance," *PLoS one*, vol. 10, no. 3, p. e0120452, 2015.

[58] W. Siqueira, H. Margolis, E. Helmerhorst, F. Mendes, and F. Oppenheim, "Evidence of intact histatins in the in vivo acquired enamel pellicle," *Journal of dental research*, vol. 89, no. 6, pp. 626–630, 2010.

[59] F. Oppenheim, T. Xu, F. McMillian, S. Levitz, R. Diamond, G. Offner, and R. Troxler, "Histatins, a novel family of histidine-rich proteins in human parotid secretion. Isolation, characterization, primary structure, and fungistatic effects on Candida albicans.," *J. Biol. Chem.*, vol. 263, no. 16, pp. 7472–7477, 1988.

[60] P. A. Raj, M. Edgerton, and M. Levine, "Salivary histatin 5: dependence of sequence, chain length, and helical conformation for candidacidal activity," *J. Biol. Chem.*, vol. 265, no. 7, pp. 3898–3905, 1990.

[61] H. Tsai, P. A. Raj, and L. A. Bobek, "Candidacidal activity of recombinant human salivary histatin-5 and variants.," *Infect. Immun.*, vol. 64, no. 12, pp. 5000–5007, 1996.

[62] H. Tsai and L. Bobek, "Human salivary histatins: promising anti-fungal therapeutic agents," *Crit. Rev. Oral Biol. Med.*, vol. 9, no. 4, pp. 480–497, 1998.

[63] A. Kurut, J. Henriques, J. Forsman, M. Skepö, and M. Lund, "Role of histidine for charge regulation of unstructured peptides at interfaces and in bulk," *Proteins: Struct., Funct., Bioinf.*, vol. 82, no. 4, pp. 657–667, 2014.

[64] D. M. Rothstein, P. Spacciapoli, L. T. Tran, T. Xu, F. D. Roberts, M. Dalla Serra, D. K. Buxton, F. G. Oppenheim, and P. Friden, "Anticandida activity is retained in P-113, a 12-amino-acid fragment of histatin 5," *Antimicrob. Agents Chemother.*, vol. 45, no. 5, pp. 1367–1373, 2001.

[65] E. J. Helmerhorst, W. Van't Hof, P. Breeuwer, E. C. Veerman, T. Abee, R. F. Troxler, A. V. Nieuw Amerongen, and F. G. Oppenheim, "Characterization of histatin 5 with respect to amphipathicity, hydrophobicity, and effects on cell and mitochondrial membrane integrity excludes a candidacidal mechanism of pore formation," *Journal of Biological Chemistry*, vol. 276, no. 8, pp. 5643–5649, 2001.

[66] M. Edgerton, S. E. Koshlukova, T. E. Lo, B. G. Chrzan, R. M. Straubinger, and P. A. Raj, "Candidacidal activity of salivary histatins identification of a histatin 5-binding protein on candida albicans," *Journal of Biological Chemistry*, vol. 273, no. 32, pp. 20438–20447, 1998.

[67] S. E. Koshlukova, T. L. Lloyd, M. W. Araujo, and M. Edgerton, "Salivary histatin 5 induces non-lytic release of ATP from Candida albicans leading to cell death," *Journal of Biological Chemistry*, vol. 274, no. 27, pp. 18872–18879, 1999.

[68] A. B. Mochon and H. Liu, "The antimicrobial peptide histatin-5 causes a spatially restricted disruption on the Candida albicans surface, allowing rapid entry of the peptide into the cytoplasm," *PLoS Pathog*, vol. 4, no. 10, p. e1000190, 2008.

[69] H. L. Norris, R. Kumar, C. Y. Ong, D. Xu, and M. Edgerton, "Zinc binding by histatin 5 promotes fungicidal membrane disruption in C. albicans and C. glabrata," *Journal of Fungi*, vol. 6, no. 3, p. 124, 2020.

[70] E. J. Helmerhorst, P. Breeuwer, W. Van't Hof, E. Walgreen-Weterings, L. C. Oomen, E. C. Veerman, A. V. Nieuw Amerongen, and T. Abee, "The cellular target of histatin 5 on Candida albicans is the energized mitochondrion," *Journal of Biological Chemistry*, vol. 274, no. 11, pp. 7286–7291, 1999.

[71] A. L. den Hertog, H. W. W. F. Sang, R. Kraayenhof, J. G. Bolscher, W. Van't Hof, E. C. Veerman, and A. V. Nieuw Amerongen, "Interactions of histatin 5 and histatin 5-derived peptides with liposome membranes: surface effects, translocation and permeabilization," *Biochemical Journal*, vol. 379, no. 3, pp. 665–672, 2004.

[72] A. L. den Hertog, J. van Marle, H. A. van Veen, W. Van't Hof, J. G. Bolscher, E. C. Veerman, and A. V. Niew Amerongen, "Candidacidal effects of two antimicrobial peptides: histatin 5 causes small membrane defects, but LL-37 causes massive disruption of the cell membrane," *Biochemical Journal*, vol. 388, no. 2, pp. 689–695, 2005.

[73] A. Ruissen, J. Groenink, W. Van't Hof, E. Walgreen-Weterings, J. Van Marle, H. Van Veen, W. Voorhout, E. Veerman, and A. Nieuw Amerongen, "Histatin 5 and derivatives: Their localization and effects on the ultra-structural level," *Peptides*, vol. 23, no. 8, pp. 1391–1399, 2002.

[74] R. Kumar, S. Chadha, D. Saraswat, J. S. Bajwa, R. A. Li, H. R. Conti, and M. Edgerton, "Histatin 5 uptake by Candida albicans utilizes polyamine transporters Dur3 and Dur31 proteins," *J. Biol. Chem.*, vol. 286, no. 51, pp. 43748–43758, 2011.

[75] S. Tati, W. S. Jang, R. Li, R. Kumar, S. Puri, and M. Edgerton, "Histatin 5 resistance of candida glabrata can be reversed by insertion of candida albicans polyamine transporter-encoding genes DUR3 and DUR31," *PloS one*, vol. 8, no. 4, p. e61480, 2013.

[76] S. Tati, R. Li, S. Puri, R. Kumar, P. Davidow, and M. Edgerton, "Histatin 5-spermidine conjugates have enhanced fungicidal activity and efficacy as a topical therapeutic for oral candidiasis," *Antimicrobial agents and chemotherapy*, vol. 58, no. 2, pp. 756–766, 2014.

[77] D. G. Kehres and M. E. Maguire, "Structure, properties and regulation of magnesium transport proteins," *Biometals*, vol. 15, no. 3, pp. 261–270, 2002.

[78] M. E. Maguire, "Magnesium transporters: properties, regulation and structure," *Front Biosci*, vol. 11, pp. 3149–3163, 2006.

[79] V. V. Lunin, E. Dobrovetsky, G. Khutoreskaya, R. Zhang, A. Joachimiak, D. A. Doyle, A. Bochkarev, M. E. Maguire, A. M. Edwards, and C. M. Koth, "Crystal structure of the corA Mg 2+ transporter," *Nature*, vol. 440, no. 7085, pp. 833–837, 2006.

[80] M. Hattori, Y. Tanaka, S. Fukai, R. Ishitani, and O. Nureki, "Crystal structure of the MgtE Mg 2+ transporter," *Nature*, vol. 448, no. 7157, pp. 1072–1075, 2007.

[81] W. Kühlbrandt, "Biology, structure and mechanism of P-type ATPases," *Nature reviews Molecular cell biology*, vol. 5, no. 4, pp. 282–295, 2004.

[82] S. Subramani, H. Perdreau-Dahl, and J. P. Morth, "The magnesium transporter A is activated by cardiolipin and is highly sensitive to free magnesium in vitro," *Elife*, vol. 5, p. e11407, 2016.

[83] S. Tzu and M. Mushashi, *The Art of War & The Book of Five Rings*. BN Publishing, 2007. Translated from Chinese by Lionel Giles and D.W.

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

[85] 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," *Current opinion in structural biology*, vol. 24, pp. 98–105, 2014.

[86] 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.

[87] R. B. Best and G. Hummer, "Optimized molecular dynamics force fields applied to the helix- coil transition of polypeptides," *The journal of physical chemistry B*, vol. 113, no. 26, pp. 9004–9015, 2009.

[88] R. B. Best and J. Mittal, "Protein simulations with an optimized water model: cooperative helix formation and temperature-induced unfolded state collapse," *The Journal of Physical Chemistry B*, vol. 114, no. 46, pp. 14916–14923, 2010.

[89] 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: Struct., Funct., Bioinf.*, vol. 78, no. 8, pp. 1950–1958, 2010.

[90] 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.

[91] 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," *The journal of physical chemistry B*, vol. 118, no. 25, pp. 6983–6998, 2014.

[92] 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," *Journal of chemical theory and computation*, vol. 11, no. 8, pp. 3696–3713, 2015.

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

[94] E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J. Y. Xiang, L. Wang, D. Lupyan, M. K. Dahlgren, J. L. Knight, *et al.*, "OPLS3: a force field providing broad coverage of drug-like small molecules and proteins," *Journal of chemical theory and computation*, vol. 12, no. 1, pp. 281–296, 2016.

[95] D. Song, W. Wang, W. Ye, D. Ji, R. Luo, and H.-F. Chen, "ff14IDPs force field improving the conformation sampling of intrinsically disordered proteins," *Chemical biology & drug design*, vol. 89, no. 1, pp. 5–15, 2017.

[96] P. Robustelli, S. Piana, and D. E. Shaw, "Developing a molecular dynamics force field for both folded and disordered protein states," *Proceedings of the National Academy of Sciences*, vol. 115, no. 21, pp. E4758–E4766, 2018.

[97] H. Liu, D. Song, H. Lu, R. Luo, and H.-F. Chen, "Extensive tests and evaluation of the CHARMM36IDPSFF force field for intrinsically disordered proteins and folded proteins," *Chemical biology & drug design*, vol. 92, no. 4, pp. 1722–1735, 2018.

[98] 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," *Physical Chemistry Chemical Physics*, vol. 21, no. 39, pp. 21918–21931, 2019.

[99] 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.

[100] D. Mercadante, S. Milles, G. Fuertes, D. I. Svergun, E. A. Lemke, and F. Gräter, "Kirkwood–Buff approach rescues overcollapse of a disordered protein in canonical protein force fields," *J. Phys. Chem. B*, vol. 119, no. 25, pp. 7975–7984, 2015.

[101] S.-H. Chong, P. Chatterjee, and S. Ham, "Computer simulations of intrinsically disordered proteins," *Annual Review of Physical Chemistry*, vol. 68, pp. 117–134, 2017.

[102] R. B. Best, "Computational and theoretical advances in studies of intrinsically disordered proteins," *Current Opinion in Structural Biology*, vol. 42, pp. 147–154, 2017.

[103] P. S. Nerenberg and T. Head-Gordon, "New developments in force fields for biomolecular simulations," *Current opinion in structural biology*, vol. 49, pp. 129–138, 2018.

[104] T. L. Hill, *An introduction to statistical thermodynamics*. Courier Corporation, 1986.

[105] M. Allen and D. Tildesley, *Computer simulation of liquids*, vol. 18. Oxford university press, 1989.

[106] D. I. Svergun and M. H. Koch, "Small-angle scattering studies of biological macromolecules in solution," *Reports on Progress in Physics*, vol. 66, no. 10, p. 1735, 2003.

[107] D. F. Evans and H. Wennerström, *The colloidal domain: where physics, chemistry, biology, and technology meet*. Wiley-Vch New York, 2 ed., 1999.

[108] J. N. Israelachvili, *Intermolecular and surface forces*. Academic press, 2011.

[109] 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.

[110] E. Lindahl, B. Hess, and D. van der Spoel, "GROMACS 3.0: a package for molecular simulation and trajectory analysis," *J. Mol. Model.*, vol. 7, no. 8, pp. 306–317, 2001.

[111] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. Mark, and H. Berendsen, "GROMACS: fast, flexible, and free," *J. Comput. Chem.*, vol. 26, no. 16, pp. 1701–1718, 2005.

[112] 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.

[113] G. Moss, "Basic terminology of stereochemistry (IUPAC Recommendations 1996)," *Pure and applied chemistry*, vol. 68, no. 12, pp. 2193–2222, 1996.

[114] C. Cragnell, E. Rieloff, and M. Skepö, "Utilizing coarse-grained modeling and monte carlo simulations to evaluate the conformational ensemble of intrinsically disordered proteins and regions," *Journal of molecular biology*, 2018.

[115] J. Reščič and P. Linse, "MOLSIM: A modular molecular simulation software," *Journal of computational chemistry*, vol. 36, no. 16, pp. 1259–1274, 2015.

[116] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The journal of chemical physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[117] D. van Der Spoel, E. Lindahl, B. Hess, and the GROMACS development team, *GROMACS User Manual version 4.6.7*. www.gromacs.org, 2014.

[118] H. Berendsen and W. Van Gunsteren, "Practical algorithms for dynamic simulations," *Molecular-dynamics simulation of statistical-mechanical systems*, pp. 43–65, 1986.

[119] B. Hess, H. Bekker, H. Berendsen, and J. Fraaije, "LINCS: a linear constraint solver for molecular simulations," *J. Comput. Chem.*, vol. 18, no. 12, pp. 1463–1472, 1997.

[120] 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, p. 10089, 1993.

[121] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," *J. Chem. Phys.*, vol. 126, p. 014101, 2007.

[122] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *J. Appl. Phys. (Melville, NY, U. S.)*, vol. 52, p. 7182, 1981.

[123] H. J. Berendsen, J. v. Postma, W. F. van Gunsteren, A. DiNola, and J. Haak, "Molecular dynamics with coupling to an external bath," *The Journal of chemical physics*, vol. 81, no. 8, pp. 3684–3690, 1984.

[124] A. Grossfield and D. M. Zuckerman, "Chapter 2 Quantifying uncertainty and sampling quality in biomolecular simulations," in *Annual reports in computational chemistry*, vol. 5, pp. 23–48, Elsevier, 2009.

[125] H. Flyvbjerg and H. G. Petersen, "Error estimates on averages of correlated data," *The Journal of Chemical Physics*, vol. 91, no. 1, pp. 461–466, 1989.

[126] B. Hess, "Determining the shear viscosity of model liquids from molecular dynamics simulations," *The Journal of chemical physics*, vol. 116, no. 1, pp. 209–217, 2002.

[127] R. Stepto, T. Chang, P. Kratochvíl, M. Hess, K. Horie, T. Sato, and J. Vohlídal, "Definitions of terms relating to individual macromolecules, macromolecular assemblies, polymer solutions, and amorphous bulk polymers (IUPAC Recommendations 2014)," *Pure and Applied Chemistry*, vol. 87, no. 1, pp. 71–120, 2015.

[128] S. 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.

[129] I. Jolliffe, "Principal component analysis," *Technometrics*, vol. 45, no. 3, p. 276, 2003.

[130] 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.

[131] 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, no. 3, p. e18401, 2011.

[132] H. Schnablegger and Y. Singh, "The SAXS guide: getting acquainted with the principles," *Austria: Anton Paar GmbH*, 2011.

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

[134] S. M. Kelly, T. J. Jess, and N. C. Price, "How to study proteins by circular dichroism," *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, vol. 1751, no. 2, pp. 119–139, 2005.

[135] V. N. Uversky and A. K. Dunker, *Intrinsically Disordered Protein Analysis: Volume 1, Methods and Experimental Tools*. Springer, 2012.

[136] L. Whitmore, A. J. Miles, L. Mavridis, R. W. Janes, and B. A. Wallace, "PCDDB: new developments at the protein circular dichroism data bank," *Nucleic acids research*, vol. 45, no. D1, pp. D303–D307, 2017.

[137] J. G. Lees, A. J. Miles, F. Wien, and B. Wallace, "A reference database for circular dichroism spectroscopy covering fold and secondary structure space," *Bioinformatics*, vol. 22, no. 16, pp. 1955–1962, 2006.

[138] S. J. Smerdon, S. Krzywda, A. M. Brzozowski, G. J. Davies, A. J. Wilkinson, A. Brancaccio, F. Cutruzzola, C. T. Allocatelli, M. Brunori, R. E. J. Brantley, *et al.*, "Interactions among residues CD3, E7, E10, and E11 in myoglobins: attempts to simulate the ligand-binding properties of Aplysia myoglobin," *Biochemistry*, vol. 34, no. 27, pp. 8715–8725, 1995.

[139] 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, no. 12, pp. 1630–1636, 2011.

[140] C. J. Oomen, P. van Ulsen, P. Van Gelder, M. Feijen, J. Tommassen, and P. Gros, "Structure of the translocator domain of a bacterial autotransporter," *The EMBO journal*, vol. 23, no. 6, pp. 1257–1266, 2004.

[141] J. Tolchard, S. J. Walpole, A. J. Miles, R. Maytum, L. A. Eaglen, T. Hackstadt, B. A. Wallace, and T. M. Blumenschein, "The intrinsically disordered Tarp protein from chlamydia binds actin with a partially preformed helix," *Scientific reports*, vol. 8, no. 1, pp. 1–11, 2018.

[142] J. L. 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 Science*, vol. 23, no. 12, pp. 1765–1772, 2014.

[143] C. Richardson, M. Johnsson, P. Raj, M. Levine, and G. Nancollas, "The influence of histatin-5 fragments on the mineralization of hydroxyapatite," *Archives of oral biology*, vol. 38, no. 11, pp. 997–1002, 1993.

[144] A. Yin, H. Margolis, J. Grogan, Y. Yao, R. Troxler, and F. Oppenheim, "Physical parameters of hydroxyapatite adsorption and effect on candidacidal activity of histatins," *Archives of oral biology*, vol. 48, no. 5, pp. 361–368, 2003.

[145] M. Yoshinari, T. Kato, K. Matsuzaka, T. Hayakawa, T. Inoue, Y. Oda, K. Okuda, and M. Shimono, "Adsorption behavior of antimicrobial peptide histatin 5 on PMMA," *Journal of Biomedical Materials Research Part B: Applied Biomaterials: An Official Journal of The Society for Biomaterials, The Japanese Society for Biomaterials, and The Australian Society for Biomaterials and the Korean Society for Biomaterials*, vol. 77, no. 1, pp. 47–54, 2006.

[146] K. Hyltegren, M. Polimeni, M. Skep⊠, and M. Lund, "Integrating all-atom and coarse-grained simulations–toward understanding of idps at surfaces," *Journal of chemical theory and computation*, vol. 16, no. 3, pp. 1843–1853, 2020.

[147] 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, p. 926, 1983.

[148] J. A. Marsh and J. D. Forman-Kay, "Sequence determinants of compaction in intrinsically disordered proteins," *Biophysical journal*, vol. 98, no. 10, pp. 2383–2390, 2010.

[149] M. Nygaard, B. B. Kragelund, E. Papaleo, and K. Lindorff-Larsen, "An efficient method for estimating the hydrodynamic radius of disordered protein conformations," *Biophysical journal*, vol. 113, no. 3, pp. 550–557, 2017.

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

[151] C. Malmberg and A. Maryott, "Dielectric constant of water from 00 to 1000 C," *J. Res. Nat. Bur. Stand*, vol. 56, pp. 369131–8, 1956.

[152] B. B. Owen, R. C. Miller, C. E. Milner, and H. L. Cogan, "The dielectric constant of water as a function of temperature and pressure," *The Journal of Physical Chemistry*, vol. 65, no. 11, pp. 2065–2070, 1961.

[153] M. Uematsu and E. Frank, "Static dielectric constant of water and steam," *Journal of Physical and Chemical Reference Data*, vol. 9, no. 4, pp. 1291–1306, 1980.

[154] D. P. Fernández, A. Goodwin, E. W. Lemmon, J. Levelt Sengers, and R. Williams, "A formulation for the static permittivity of water and steam at temperatures from 238 K to 873 K at pressures up to 1200 MPa, including derivatives and

Debye–Hückel coefficients," *Journal of Physical and Chemical Reference Data*, vol. 26, no. 4, pp. 1125–1166, 1997.

[155] G. Nagy, M. Igaev, N. C. Jones, S. V. Hoffmann, and H. Grubmüller, "SESCA: predicting circular dichroism spectra from protein molecular structures," *Journal of chemical theory and computation*, vol. 15, no. 9, pp. 5087–5102, 2019.

[156] P. Bernadó, E. Mylonas, M. V. Petoukhov, M. Blackledge, and D. I. Svergun, "Structural characterization of flexible proteins using small-angle X-ray scattering," *J. Am. Chem. Soc.*, vol. 129, no. 17, pp. 5656–5664, 2007.

[157] G. Tria, H. D. Mertens, M. Kachala, and D. I. Svergun, "Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering," *IUCrJ*, vol. 2, no. 2, pp. 207–217, 2015.

[158] A. Micsonai, F. Wien, L. Kernya, Y.-H. Lee, Y. Goto, M. Réfrégiers, and J. Kardos, "Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 24, pp. E3095–E3103, 2015.

[159] A. Micsonai, F. Wien, É. Bulyáki, J. Kun, É. Moussong, Y.-H. Lee, Y. Goto, M. Réfrégiers, and J. Kardos, "BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W315–W322, 2018.

[160] Y. Gerelli, A. Eriksson Skog, S. Jephthah, R. J. Welbourn, A. Klechikov, and M. Skepö, "Spontaneous formation of cushioned model membranes promoted by an intrinsically disordered protein," *Langmuir*, vol. 36, no. 15, pp. 3997–4004, 2020.

10. Scientific publications