Molecular dynamics simulation of alkanes and proteins methodology, prediction of properties and comparison to experimental data

Author(s):
Eichenberger, Andreas Paul

Publication Date:
2012

Permanent Link:
https://doi.org/10.3929/ethz-a-009787706

Rights / License:
In Copyright - Non-Commercial Use Permitted
Molecular dynamics simulation of alkanes and proteins: methodology, prediction of properties and comparison to experimental data

A dissertation submitted to the
ETH Zürich

for the degree of
Doctor of Sciences

presented by
ANDREAS PAUL EICHENBERGER
M. Sc. ETH
born January 8, 1982
citizen of Beinwil a. See – Switzerland

accepted on the recommendation of
Prof. Dr. Wilfred F. van Gunsteren, examiner
Dr. Lorna J. Smith, co-examiner
Prof. Dr. Hans Jürgen Herrmann, co-examiner

2012
Acknowledgement

This is it, the outcome of the last four years of my daily work. Although I am allowed to call it my thesis, there were a lot of contributions from other people, some in a direct and active way, others in a more hidden but not less important manner.

I am first addressing Wilfred van Gunsteren, my supervisor and “Doktorvater”. I am very grateful for all his support, discussions and advice concerning scientific but also non-scientific life, both of which he showed by his own example. He always managed to let me know that he believes in me and my work, even in times where I did not, and gave me a lot of freedom and trust.

In addition, I would like to thank my coexaminers Prof. Dr. Hans Jürgen Herrmann and Dr. Lorna Smith for attending my defence.

Without the IGC group, a never-ending source of new ideas and inspiration at work but also great friends and colleagues away from the ETH, this thesis would never have come off. I would like to thank Alexandra always saving a smile for me, for the uncountable text messages she wrote me while Dante was (not) falling asleep, and for showing me that haircut (and colour) may go hand in hand with clothing; Alice for her great help and patience to keep ray happy and for her never-ending optimism (at least after 10pm); Andreas for sharing his knowledge about all kind of chillies; Pitschna for supervising and introducing me to the group, while I was doing my Semesterarbeit, being a great office mate and letting me (and the rest the office) witness uncountable and amusing phone calls with her mother and other people; Bettina for teaching me maintaining Sun systems; Bruno for all the discussions at and far from ETH and for accompanying the farewell songs with his guitar tunes; Carmen for the friendly and very helpful support in administrative tasks; Chris for the many useful emails about GROMOS++ and for his helpful nature; Clara for correcting my APC++ exercises; Daan for letting me use his office place after he left IGC four years ago, I think there are still a few items from you on the shelf; Daniela for much advice concerning life in IGC and at ETH; Denise for sharing the burden of solving GROMOS++ tickets and to comfort ray in times she (ray, which is definitely a her) was missing me; Dirk for asking interesting questions I did not even know they existed; Dongqi for his extremely friendly and always helpful behaviour and open ear, even if he had no time at all; Halvor for preparing all the outings and ski weekends I enjoyed very much; Halua for the discussions about politics and other topics while waiting for our first coffee in front
of the coffee machine; Hans-Peter for mastering everything related to C4 and for the many anecdotes he told me in the floor or on the bus; Jane for the delicious diners and cakes, they were amazing, and for forgetting her group talk; Jozi for her good mood, her support at any time needed and for her email out of the middle of the night; Katharina for all the Schichtsalate and Baumkuchen, and for the statistics about how often we could see the Alps out of our office, which was always forgotten to fill out on rainy and dusty days; Lorna for all the discussions and valuable inputs during her visits; Maria for her help maintaining the systems when I had no clue what to do; Monika for keeping the printers happy and for answering my questions about lipids; Nathan for his help in all kinds of technical aspects, bringing GROMOS in shape and for accompanying me watching pointless films in cinema; Niels for being my personal post-doc, joining for jogging at reasonable daytimes and for sharing a beer on Friday evenings and days who felt like Friday evening; Noah for sharing the joy of playing ping-pong, the help in maintaining the system, and for always trying to convince me to attend the Happy Hour; Oliwia for sharing the spirit of alkanes and of maintaining ray; Pascal for mastering Exercise 2 of Informatik I while I was working on this thesis; Phil for his open ear and patience, for teaching me what “summer” really means and for respecting people (including me) how they are; Sereina for discussing and arguing about things on which we basically had the same opinion or not, for taking care of me when I got totally sea sick during whale watching in Boston, and for always being around as a confederate and friend; Stephan for his help concerning the new server and for making Ladina happy by giving her good marks at gymnastics contests; Wei for staying with me in the office (probably) till the very end, while everyone else left; Zhixiong for explaining me the difference between ALDI, Coop and Migros, and for not letting me win in ping-pong; and Zrinka for warning me of β-peptides, which I successfully avoided during my PhD study.

Many thanks go to my flatmates Thea and Tinu, two good friends from study time. Having a safe home and "family" is an invaluable treasure and a good basis for everything else. Maybe we can share another flat in about 40 years?

Further, I am very thankful to Jolande van Gunsteren-Bolt for keeping my boss happy and in a good mood, and for the numerous and really enjoyable farewells at her and Wilfred’s home as well as for the good spirit she always spread whenever we met. Further I will keep the presentation course she gave together with Wytze Visser in very good memory.

Almost at the end but not less grateful I would like to thank my parents Bernadette and Paul, and my sister Daniela. They all support me now for more than 30 years without
doubt or questioning whatever I do and did, provide a lot of freedom and confidence, the perfect environment to grow and develop. Thank you so much!

A huge "Merci" to Ladina, my secure anchor, for your love and unconditional support. You’re the one keeping me happy!
# Contents

Acknowledgement iii  
Contents vii  
Summary xiii  
Zusammenfassung xv  
Publications xix

1 Introduction 1  
1.1 Biomolecular studies ........................................ 1  
1.2 Computer simulation .......................................... 2  
  1.2.1 Atomic and molecular degrees of freedom ............... 2  
  1.2.2 Force field or inter-particle interaction function ....... 3  
  1.2.3 Sampling of the configurational space .................... 3  
  1.2.4 Boundary conditions ...................................... 5  
1.3 Free energy calculations ...................................... 6  
1.4 Analysis of simulation trajectories .......................... 7  
  1.4.1 Calculation of NMR data ................................. 7  
1.5 The protein hen egg white lysozyme .......................... 8

2 A method to apply bond-angle and dihedral-angle constraints in molecular dynamics simulation 9  
2.1 Introduction .................................................. 10  
2.2 Method ....................................................... 11  
2.3 Discussion .................................................... 16  
2.4 Appendix ...................................................... 18
3 A supra-atomic coarse-grained GROMOS force field for aliphatic hydrocarbons in the liquid phase

3.1 Introduction ................................................................. 22

3.2 Methods .......................... .................................................. 26
   3.2.1 Atomistic model ....................................................... 26
   3.2.2 Atomistic simulations ................................................ 27
   3.2.3 Analysis of FG configurations in terms of CG beads .......... 29
   3.2.4 Supra-atomic model .................................................. 30
   3.2.5 Supra-atomic simulations .......................................... 30
   3.2.6 Parametrisation strategy for the LJ interactions ............... 31
   3.2.7 Analysis ............................................................... 31

3.3 Results and discussion ................................................. 33
   3.3.1 Atomistic-level FG simulations .................................... 33
   3.3.2 Bonded interactions of the CG alkane model .................... 33
   3.3.3 Lennard-Jones interactions of the CG alkane model .......... 36
   3.3.4 Structural analysis of the CG alkanes .......................... 38
   3.3.5 Testing of the CG model ........................................... 42

3.4 Conclusion ................................................................. 43

3.5 Supplementary material .............................................. 44

4 The GROMOS++ software for the analysis of biomolecular simulation trajectories

4.1 Introduction ................................................................. 50

4.2 Trajectory quantities and basic types of analysis .................. 50
   4.2.1 Types of trajectory information handled by GROMOS++ ...... 51
   4.2.2 Types of analysis .................................................... 53
   4.2.3 Superposition of molecular structures .......................... 54
   4.2.4 Spatial boundary conditions and gathering ...................... 54
   4.2.5 Atom, vector and property specifiers ............................ 55
4.2.6 Input/Output formats .............................................. 57
4.2.7 Physical units ......................................................... 58
4.3 Observable and non-observable quantities ....................... 58
  4.3.1 Structural quantities ........................................ 59
  4.3.2 Dynamic quantities .......................................... 61
  4.3.3 Thermodynamic quantities ................................... 63
  4.3.4 Comparison to experimentally observable and derived quantities 65
4.4 Examples ....................................................................... 65
  4.4.1 Global structural properties as a function of time .......... 65
  4.4.2 Hydrogen bond analysis as function of time ................. 65
  4.4.3 Conformational cluster analysis to detect structural differences 67
  4.4.4 Configurational entropy as function of time ............... 68
4.5 Conclusion ................................................................. 68
4.6 Appendix ....................................................................... 70

5 Computation of folding free energy differences of the amide-to-ester perturbed Pin1 WW domain using enveloping distribution sampling 75
  5.1 Introduction ............................................................. 76
  5.2 Results ........................................................................ 83
    5.2.1 Structures and stability of the wild-type WW domain ...... 83
    5.2.2 Hydrogen bonding .............................................. 85
    5.2.3 Free enthalpy differences ...................................... 87
  5.3 Discussion ............................................................... 88
  5.4 Conclusion ............................................................... 98
  5.5 Computational Methods ............................................... 99
    5.5.1 Simulation program and force-field parameters .......... 99
    5.5.2 Simulation protocol ............................................ 100
    5.5.3 EDS reference-state simulations ............................ 101
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Ester-linked hen egg white lysozyme shows a compact fold in a molecular dynamics simulation: possible causes and sensitivity of experimentally observable quantities to structural changes maintaining this compact fold</td>
<td>143</td>
</tr>
<tr>
<td>8.1</td>
<td>Introduction</td>
<td>144</td>
</tr>
<tr>
<td>8.2</td>
<td>Results and discussion</td>
<td>146</td>
</tr>
<tr>
<td>8.2.1</td>
<td>Structural properties and flexibility</td>
<td>146</td>
</tr>
<tr>
<td>8.2.2</td>
<td>Comparison to experimental NMR data</td>
<td>149</td>
</tr>
<tr>
<td>8.2.3</td>
<td>Hydrogen bonding</td>
<td>157</td>
</tr>
<tr>
<td>8.3</td>
<td>Materials and methods</td>
<td>161</td>
</tr>
<tr>
<td>8.3.1</td>
<td>Simulation program and force-field parameters</td>
<td>161</td>
</tr>
<tr>
<td>8.3.2</td>
<td>Simulation set-up and protocol</td>
<td>163</td>
</tr>
<tr>
<td>8.3.3</td>
<td>Analysis</td>
<td>164</td>
</tr>
<tr>
<td>8.3.4</td>
<td>Calculation of backbone $^{15}$N-$^1$H, $^{13}$C$_{\alpha}$-$^{13}$C, and $^{13}$C-$^{15}$N residual dipolar couplings</td>
<td>166</td>
</tr>
<tr>
<td>8.4</td>
<td>Conclusion</td>
<td>167</td>
</tr>
<tr>
<td>9</td>
<td>Outlook</td>
<td>169</td>
</tr>
</tbody>
</table>

Bibliography 173

Curriculum Vitae 193
Summary

Molecular dynamics (MD) simulation is a powerful tool to investigate problems beyond the border of the area accessible to experiments and has been widely used to study the structure and dynamics of biomolecular systems. A reliable model and accurate algorithm is indispensable for a scientist to calculate and predict properties of the system under study, but treatment and post-processing of simulation trajectories is of no less importance.

Chapter 1 gives a brief introduction to MD simulations and how they may complement experimental studies, what basic decisions must be made prior to running an MD simulation, and mentions the two main limiting factors of MD simulations: insufficient configurational sampling and simulation speed.

Chapters 2 and 3 are addressing the simulation speed. In Chapter 2 a method to apply bond-angle or dihedral-angle constraints in MD simulations is presented. It constitutes an alternative to the use of distance constraints, which fail for linear molecule configurations, thereby enhancing the simulation time by allowing for a bigger integration time step.

In Chapter 3 another possibility to gain simulation speed is proposed which is based on reducing the number of degrees of freedom considered during a simulation. A supra-atomic coarse-grained (CG) model for linear alkanes in the liquid phase is presented, a possible precursor of a CG model for lipids to be used – a long-term goal – in combination with atomic-level, fine-grained (FG) protein or other biomolecules. It consists of six CG beads of sizes 2, 3 and 4 and distinguishes between beads at the chain ends (end beads) and beads between end beads (middle beads). The CG force field was derived based on experimental thermodynamic data and on mapped structural and energetic properties of corresponding FG alkane simulations.

Analysis of biomolecular simulation trajectories is considered in Chapter 4, where an overview of the GROMOS++ simulation software is given distinguishing between three types of programs: programs that calculate structural, dynamic or thermodynamic quantities from configurational simulation trajectories. GROMOS++ is a flexible and rich collection of analysis tools ready to be used for a variety of types of analysis regarding molecular simulation trajectories.

In Chapter 5, enveloping distribution sampling (EDS) is used to calculate folding free enthalpy differences between the wild-type Pin1 WW domain and 20 amide-to-ester mutants.
in analogy to an experimental study [1]. It shows that a relation between the strength of the individual hydrogen bonds and the differences in folding free enthalpy differences, as done in the experimental study [1], is not straightforward due to structural changes within the folded protein between wild-type and mutants, indicated by different hydrogen-bonding patterns in the enveloping distribution sampling (EDS) MD simulations.

In Chapter 6 the effect of 222-trifluoroethanol (TFE) upon the protein structure of hen egg white lysozyme (HEWL) is studied. Experimental NMR studies report that in 70% TFE/30% water the regions corresponding to native β-strands in HEWL retain persistent helical conformation and increase in length, while the detailed and full protein structure in TFE/water is still not known. A possible configuration largely fulfilling the available experimental NMR data was generated using MD simulations, and is the currently best possible representation of this ensemble.

Chapters 5, 7 and 8 discuss proteins with backbone amide-to-ester substitutions. Amide-to-ester mutations are widely used in experimental an theoretical studies to influence the protein structure and stability by eliminating one backbone hydrogen-bond donor (NH versus O) and weakening one backbone hydrogen-bond acceptor (amide carbonyl versus ester carbonyl). The effect of amide-to-ester replacements on the protein structure of HEWL is described in Chapters 7 and 8. Fully ester-linked HEWL, i.e. containing as many amide-to-ester substitutions as possible, shows a slight compaction while losing its native structure. However, it does not unfold completely. Partly ester-linked HEWL, in which only 34 peptide linkages that are not involved in the helical or β-strand parts of native HEWL were replaced by ester linkages showed behaviour comparable to native HEWL. The conformational changes were analysed by comparing simulation averaged values of quantities that can be measured by NMR, such as 1H–15N backbone order parameters, residual dipolar couplings, proton-proton NOE distances and 3J-couplings with the corresponding values derived from experimental NMR data for native HEWL.

Possible further development and future work is mentioned in the outlook, Chapter 9.

Eine kurze Einführung in die Methodik von Molekulardynamiksimulationen wird in Kapitel 1 gegeben. Es wird kurz aufgezeichnet, inwiefern Computersimulationen experimentelle Studien ergänzen können und welche grundsätzlichen Entscheidungen schon vor der Durchführung der Simulation getroffen werden müssen. Weiter werden die beiden Hauptgründe, welche die Aussagekraft von Computersimulationen einschränken, aufgezeichnet: mangelhafte Abtastung des Konfigurationsraums sowie die Simulationsgeschwindigkeit.


Eine andere Möglichkeit zur Beschleunigung von Computersimulationen wird in Kapitel 3 angesprochen und basiert auf der Reduktion der Freiheitsgrade, welche während der Simulation berücksichtigt werden. Es wird ein atomübergreifendes, grobkörniges Alkanmodell vorgestellt, welches als möglicher Vorläufer eines grobkörnigen Lipidmodells dienen soll, in Kombination mit Modellen für Proteine und anderen Biomolekülen auf feinkörnigem, atomarem Level. Das vorgeschlagene Modell besteht aus sechs grobkörnigen Kugeln, sogenannten “Beads” der Grösse 2, 3 und 4, und unterscheidet zwischen Beads am Ende der Alkankette (Endbeads) und solchen zwischen zwei Endbeads (Mittelbeads). Die Modellparameter des grobkörnigen Alkanmodells wurden basierend auf thermodynamischen, experimentellen Messwerten entworfen, wobei zusätzlich strukturelle und energetische Eigenschaften für das grobkörnige Modell aus feinkörnigen Alkansimulationen abgeleitet und benutzt wurden.

Das Analysieren von Simulationstrajectorien wird im Rahmen eines Überblicks über die


Der Effekt von 222-Trifluoroethanol (TFE) auf die Struktur von Hühnereiweißlysozym (HEWL) wird in Kapitel 6 studiert. Experimentelle Studien beschreiben, dass in einer 70% TFE/30% Wasser Lösung β-Faltblätter in einem bestimmten Teilbereich des Proteins zu einer beständigen α-Helixstruktur wechseln, wobei die Gesamtstruktur des Proteins in diesem Faltungszustand unbekannt ist. Mit Hilfe von Molekulardynamiksimulationen wurde ein Ensemble von HEWL Proteinkonfigurationen generiert, welches verfügbaren Daten aus der NMR Spektroskopie grösstenteils entspricht. Es ist die zur Zeit beste Darstellung dieses Ensembles.

Kapitel 5, 7 und 8 behandeln Proteine, in welchen eine oder mehrere Peptidbindungen zu Esterbindungen mutiert sind. Solche Mutationen, welche einen Protonendonator eliminieren (NH gegenüber O) und einen Protonenakzeptor schwächen (Amidkarbonyl gegenüber Esterkarbonyl), werden in Theorie und Experiment häufig benutzt, um den Einfluss von Wasserstoffbrücken auf die Stabilität und Struktur des Proteins zu untersuchen. Kapitel 7 und 8 behandeln beide den Effekt solcher Mutationen auf die Proteinstruktur von HEWL. Vollständig verestertes HEWL zeigt eine leichte Verdichtung, verliert aber die natürliche Proteinfaltung. Ein teilweise verestertes HEWL, bei dem 34 nicht in die Sekundärstruktur des Proteins involvierte Aminosäuren mutiert sind, zeigt ein Verhalten ähnlich jenem von natürlichem Lysozym. Unterschiede in der Proteinkonfiguration wurden durch den Vergleich von gemittelten, berechneten NMR Grössen sowie experimentellen Grössen zugänglich durch NMR Spektroskopie ($^{1}$H–$^{15}$N Ordnungs-
parameter, Dipol-Dipol Kopplungskonstanten, Proton-Protone NOE Distanzen und $^3$J-
Kopplungskonstanten) analysiert.

Abschließend wird in Kapitel 9 ein Ausblick auf zukünftige, weiterführende Studien,
Arbeiten und Entwicklungen gegeben.
This thesis has led to the following publications:

**Chapter 4**


**Chapter 6**

A. P. Eichenberger, W. F. van Gunsteren and L. J. Smith. „Structure of hen egg-white lysozyme solvated in TFE/water: a molecular dynamics simulation study based on NMR data”

**Chapter 7**

A. P. Eichenberger, Z. Gattin and W. F. van Gunsteren. „Molecular dynamics simulation of ester-linked hen egg white lysozyme reveals the effect of missing backbone hydrogen-bond donors on the protein structure”

**Chapter 8**

A. P. Eichenberger, L. J. Smith and W. F. van Gunsteren. „Ester-linked hen egg white lysozyme shows a compact fold in a molecular dynamics simulation: possible causes and sensitivity of experimentally observable quantities to structural changes maintaining this compact fold”
Related Publications

A. P. E. Kunz, A. P. Eichenberger and W. F. van Gunsteren.
,,A simple, efficient polarisable molecular model for liquid carbon tetrachloride”

,,Definition and Testing of the GROMOS Force Field Versions: 54A7 and 54B7”

,,Biomolecular structure refinement using the GROMOS simulation software”

,,Architecture, implementation and parallelization of the GROMOS software for biomolecular simulation”

W. Huang, A. P. Eichenberger and W. F. van Gunsteren.
,,Molecular dynamics simulation of thionated hen egg white lysozyme”

S. Riniker, A. P. Eichenberger and W. F. van Gunsteren.
,,Solvating atomic level fine-grained proteins in supra-molecular level coarse-grained water for molecular dynamics simulations”

S. Riniker, A. P. Eichenberger and W. F. van Gunsteren.
,,Structural Effects of an Atomic-Level Layer of Water Molecules around Proteins Solvated in Supra-Molecular Coarse-Grained Water”

D. Steiner, J. R. Allison, A. P. Eichenberger and W. F. van Gunsteren.
,,On the calculation of $^3J_{\alpha\beta}$-coupling constants for side chains in proteins”

,,Multi-resolution simulation in chemistry: methodological issues and exploration of the origins of differential catalysis observed between two structurally similar enzymes”
Chapter 1

Introduction

1.1 Biomolecular studies

Some people spend time on science out of pure curiosity and joy to increase their knowledge for a better understanding of the world they are living in, while others are driven by the desire to improve this world. For whatever reason, biochemistry gradually became a more popular area of research and was able to successfully explain and understand processes in living systems during the last decades. Probably one of the most popular examples is the decoding of genetic material and its function, starting with the description of the structure of DNA in 1956 [2].

Proteins are another important group of biomolecules involved in many elementary processes and, unfortunately, diseases [3, 4]. Made of only 20 L-α-amino acids this does not restrict them to be simple molecules but rather complex constructs of up to ten-thousands of amino acids as in the case of titins, which act as a molecular springs [5]. One distinguishes four different levels of protein structure: (i) the primary structure defined as the amino acid sequence, (ii) secondary structure such as helices or β-sheets, (iii) tertiary structure, defining the spatial relation of two or more secondary structure elements and therefore the overall, monomeric protein structure, and (iv) quaternary structure combining multiple protein molecules into a single protein complex. Such a construct allows for flexibility and motions at all four levels of protein structure which is not a unique, rock-solid and rigid configuration but rather a Boltzmann-weighted ensemble of protein configurations.

To understand the function of a protein it is essential to know its dominant structure and properties, usually measured by X-ray diffraction [6] or NMR spectroscopy [7]. Such methods are dependent on the corresponding resolution in space and time and yield a time- and space-averaged answer – the truth but not necessarily the full truth – to the question posed by the experimentalist. Furthermore, the quantity to be measured may not be directly observable but derived from primary, observed, experimental data based on a model. Depending on this model the quantity in question may be ambiguous or, as it is the case for hydrogen atoms in X-ray scattering experiments, may not be observed at all. Such problems can be avoided by complementing or even substituting experimental studies by experiments in silico, however, not without introducing other sources of possible errors or restrictions. Nevertheless, simulations are a powerful and useful tool to study biomolecular systems in the case where experiments cannot measure the desired quantity, provide ambiguous solutions, or are too expensive or ethically unjustifiable.
Monte Carlo (MC) and molecular dynamics (MD) simulations are two ways to generate Boltzmann-weighted ensembles of molecular configurations using computer simulation. MC is usually based on random trial moves of particles, \( m(A \rightarrow B) \), changing the state of the system from \( A \) to \( B \) with corresponding potential energies \( V(A) \) and \( V(B) \), respectively. In cases where \( V(A) \geq V(B) \) the move is accepted, while for \( V(A) < V(B) \) the move is only accepted when

\[
n_{\text{rand}} < \exp \left[ -\beta (V(B) - V(A)) \right],
\]

where \( n_{\text{rand}} \) being a random number in the interval \([0,1]\). Momentum conservation is not preserved in such an approach and the MC method is only used to model phenomena which do not depend on time. In contrast, in MD simulations Newton’s equations of motion,

\[
f_i = m_i \ddot{r}_i,
\]

are solved, where \( r_i \) is the position of and \( f_i \) the force on particle \( i \). MD simulations are more convenient to study biomolecular systems involving proteins and protein dynamics than MC simulations and have been used throughout this thesis.

### 1.2 Computer simulation

In this section we shall briefly discuss the four basic choices which have to be made to successfully target a (bio)molecular problem using computer simulations. These are: (i) the choice of atomic and molecular degrees of freedom, (ii) which model and which model parameter set (force field) has to be applied, (iii) the methods and algorithms to be used to generate a Boltzmann-weighted ensemble of configurations, i.e. how to sample the selected degrees of freedom, and (iv) the selection of boundary conditions applied during the MD simulation.

#### 1.2.1 Atomic and molecular degrees of freedom

Computer simulations are restricted by a number of limitations, one of it being the computational cost limiting the simulation time scale from ns to a few 0.1 \( \mu s \) depending on the system size. It does not help to solve a problem with maximal precision and accuracy if the calculation of a solution takes longer than the lifetime of men, or even worse, to sample a whole set of degrees of freedom which are irrelevant for the problem to be studied using simulation. In MD simulations the most fine-grained degrees of freedom are usually atoms treated as single particles. The use of coarse-grained force fields became more and more popular the past years. In such models the degrees of freedom are at the supra-atomic or supra-molecular level, the latter one being mostly used for solvents. Coarse-grained particles at the supra-atomic level have been chosen in Chapter 3 for the simulation of coarse-grained alkanes. At the other end of the resolution scale nuclei and electrons can
be separately treated which is done in quantum-chemical calculations.

1.2.2 Force field or inter-particle interaction function

In classical dynamics a system of \(N\) particles can be described by a Hamiltonian

\[
H(\mathbf{r}, \mathbf{p}) = V(\mathbf{r}) + K(\mathbf{p}) \tag{1.3}
\]

with \(\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)\) and \(\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N)\) the 3\(N\)-dimensional vectors of particle positions and particle momenta. The kinetic energy term \(K(\mathbf{p})\) can be obtained from

\[
K(\mathbf{p}) = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} = \sum_{i=1}^{N} \frac{1}{2} m_i \mathbf{v}_i^2 \tag{1.4}
\]

where \(m_i\) denotes the mass of particle \(i\) and the 3\(N\)-dimensional vector of the particle velocities, \(\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N)\) is the time derivative of the position vector, \(\mathbf{v} = \dot{\mathbf{r}}\). The potential energy term in equation Eq. (1.3) results from various bonded and non-bonded particle interactions in the system,

\[
V(\mathbf{r}) = V^b(\mathbf{r}) + V^{nb}(\mathbf{r}) \tag{1.5}
\]

The force acting on particle \(i\) is given as the negative of the spatial, partial derivative of the potential energy term \(V(\mathbf{r})\), which yields in combination with Eq. (1.2)

\[
m_i \ddot{\mathbf{r}}_i = \mathbf{f}_i = -\frac{\partial}{\partial \mathbf{r}_i} V(\mathbf{r}) = -\frac{\partial}{\partial \mathbf{r}_i} \left( V^b(\mathbf{r}) + V^{nb}(\mathbf{r}) \right) \tag{1.6}
\]

The exact nature of the bonded and non-bonded particle interactions as well as the corresponding parameters are collected and summarised in a so-called force field. These parameters for the bond- and bond-angle potential energy terms of an atomic-level fine-grained classical force field, where each particle basically represents one atom, can be obtained from X-ray diffraction or spectroscopic data. The parameters of the torsional-angle interaction term and partial charges can be obtained from quantum-chemical calculations on small molecules in the gas phase. This is different for coarse-grained models, where no experimental measurements or quantum calculations are possible and thus such models have to be parametrised based on other methods as described in Chapter 3.

1.2.3 Sampling of the configurational space

Three conditions are essential to obtain usable results or predictions from a molecular dynamics simulation: (i) the quality of the force field needs to be high enough to fully describe all important aspects of the problem, (ii) an adequate sampling of the relevant configurational space is indispensable for the calculation of some properties such as
free energy and enthalpy, and (iii) an efficient algorithm is needed to integrate Newton’s equations of motion.

The first aspect was mentioned above and will be further considered in Chapter 3. The probability of occurrence of a specific configuration of the system is given by the phase-space probability $P(r, p)$, which is defined for the canonical ensemble as

$$P_{NVT}(r, p) = \frac{e^{-\beta H(r, p)}}{\int \int e^{-\beta H(r, p)} \, dp \, dr} = \frac{e^{-\beta H(r, p)}}{\hbar^{3N} N! Z(N, V, T)}$$  \tag{1.7}

where $N$ is the number of particles within the volume $V$, $\beta$ is the inverse of the Boltzmann constant, $k_B$, multiplied by the temperature $T$, $\hbar$ denotes Planck’s constant and $Z(N, V, T)$ is the canonical partition function,

$$Z(N, V, T) = \frac{1}{\hbar^{3N} N!} \int \int e^{-\beta H(r, p)} \, dp \, dr ,$$  \tag{1.8}

and the factor $N!$ is only present for indistinguishable particles. The phase-space probability in the isothermal-isobaric ensemble, i.e. at constant temperature $T$ and pressure $p$, becomes volume dependent,

$$P(r, p, V) = \frac{e^{-\beta [H(r, p) + pV]}}{\hbar^{3N} N! Z(N, p, T)} ,$$  \tag{1.9}

with the corresponding partition function

$$Z(N, p, T) = \frac{1}{\hbar^{3N} N!} \int \int \int e^{-\beta [H(r, p) + pV]} \, dp \, dr \, dV.$$  \tag{1.10}

A quantity $Q$ can then be computed as the expectation value of the property over the whole phase space,

$$\langle Q \rangle_{NVT} = \int \int Q(r, p) P(r, p) \, dpdr$$  \tag{1.11}

and accordingly for the isothermal-isobaric ensemble. Angular brackets indicate the ensemble average. From Eq. (1.11) it becomes clear that a representative sampling of the whole phase-space is crucial for the calculation of observable quantities from simulations. This implies a criterion for a good algorithm to generate an ensemble: it should be able to sample as many configurations as possible within a reasonable amount of time.

This does not necessarily ask for a fast integration algorithm, since the time spent on integration the equations of motion is usually not the bottleneck in simulation work. Accuracy when using large time steps is much more important since these allow for visiting more configurations within the same number of integration time steps. In all simulations of this thesis the leap-frog \cite{8} algorithm was used which is second-order in the time step $\Delta t$. It computes the position $r_i$ and velocities $v_i$ of particles $i$ at different times shifted by
\[ \Delta t / 2, \]
\[ \mathbf{r}_i(t_n + \Delta t) = \mathbf{r}_i(t_n) + \mathbf{\dot{r}}_i(t_n + \frac{\Delta t}{2}) \Delta t, \quad (1.12) \]
\[ \mathbf{v}_i(t_n + \frac{\Delta t}{2}) = \mathbf{\dot{r}}_i(t_n + \frac{\Delta t}{2}) = \dot{\mathbf{r}}_i(t_n - \frac{\Delta t}{2}) + \ddot{\mathbf{r}}_i(t_n) \frac{\Delta t}{m_i}. \quad (1.13) \]

There are various methods and techniques which allow for the use of large time steps \( \Delta t \) and therefore an increase of the sampling. Supra-atomic or supra-molecular coarse-graining is one of them, where large particles are interacting via a smoother potential energy surface than in the case of atomic-level fine-grained force fields. Coarse-grained force fields are also very useful to describe parts of the simulated systems, where a high resolution is not necessarily needed, e.g. for solvent molecules far away from solute atoms, but they must not replace atomic-level ones at the core of the system to be investigated in more or full detail, i.e. the system must generally be multi-grained [9]. For this reason, and because multiple time-step integration as described in [10] is not implemented in the simulation program GROMOS [11–13], at the coarse-grained level we stick to an integration time step of 2 fs as used in the simulations of fully fine-grained systems keeping bond lengths constant by applying the SHAKE algorithm [14]. A four times smaller time step of 0.5 fs has to be chosen when SHAKE is not used in fine- or mixed-grained MD simulations. SHAKE has at least one disadvantage: it fails for linear molecules as for example acetonitrile. Other methods have to be used in such cases, e.g. constraining the bond angle as will be discussed in Chapter 2.

### 1.2.4 Boundary conditions

Most simulated systems are in reality far larger than the number of degrees of freedom that can be conveniently sampled using current computer power. Biomolecular systems consist largely of aqueous solutions of proteins and other molecules. Therefore, it is essential to apply spatial boundary conditions to mimic the presence of an infinite bulk surrounding the solute particles. This is usually done by applying periodic boundary conditions that copy the computational box to form an infinite periodic lattice. The total potential energy of such a system is then given as

\[ V_{\text{pot}}^{\text{tot}}(\mathbf{r}) = \frac{1}{2} \sum_{i,j,n} V_{i,j}^{\text{pot}}(|\mathbf{r}_{ij} + nL|), \quad (1.14) \]

where \( L \) is the box-edge length of a cubic, periodic box and \( \mathbf{n} \) is a vector of three integer numbers. However, such an approach introduces periodic contributions of long-range interactions which are not present in bulk solutions. All simulations described in this thesis used a different treatment of the non-bonded interactions, the so-called reaction-field method [15], which truncates the non-bonded interactions beyond a cut-off radius \( R_c \). The introduced error is small for Lennard-Jones interactions but non-negligible for the long-range electrostatic Coulomb interactions and thus needs a correction term assuming
Chapter 1. Introduction

a homogeneous medium of constant dielectric permittivity beyond the cut-off radius \( R_c \)

### 1.3 Free energy calculations

The free energy or Helmholtz free energy \( F \) and the free enthalpy or Gibbs free energy \( G \) can be defined as

\[
F(N,V,T) = -\frac{1}{\beta} \ln Z(N,V,T) \quad (1.15)
\]

and

\[
G(N,p,T) = F(N,V,T) + pV = -\frac{1}{\beta} \ln Z(N,p,T) \quad , \quad (1.16)
\]

respectively. An adequate computation of these properties is not feasible using MD simulations due to incomplete sampling of the phase-space. However, the computation of free enthalpy differences is possible, when the sampling is restricted to conformations that are different for the two systems or two states \( A \) and \( B \) of a system. The free enthalpy difference can be written as

\[
\Delta G(N,p,T)_{BA} = G(N,p,T)_B - G(N,p,T)_A = -\frac{1}{\beta} \ln \frac{Z_B(N,p,T)}{Z_A(N,p,T)} \quad . \quad (1.17)
\]

Different approaches have been used to computationally obtain the difference in free enthalpy between two states, e.g. thermodynamic integration (TI) [16] or one-step perturbation (OSP) [17, 18]. In Chapter 5 another, recently developed method named enveloping distribution sampling (EDS) [19, 20] has been used. The method is based on the construction of a reference state \( R \), that samples two or more states of relevance within one MD simulation. The potential energy term of the Hamiltonian of the reference state is written as

\[
V_R(r; s, \Delta E_{AB}^R) = -\frac{1}{\beta} \ln \left[ e^{-\beta(V_A(r)-E^R_A)} + e^{-\beta(V_B(r)-E^R_B)} \right] \quad , \quad (1.18)
\]

where the two parameters \( s \) and \( \Delta E_{AB}^R = E^R_A - E^R_B \) are to be adjusted to allow for an even sampling of states \( A \) and \( B \). The corresponding free enthalpy difference is then given as

\[
\Delta G_{BA} = G_{BR} - G_{AR} \\
= -\frac{1}{\beta} \ln \frac{Z_B(N,\rho,T)}{Z_A(N,\rho,T)} \frac{Z_R(N,\rho,T)}{Z_R(N,\rho,T)} = -\frac{1}{\beta} \ln \frac{\langle e^{-\beta(V_B-V_R)} \rangle_R}{\langle e^{-\beta(V_A-V_R)} \rangle_R} \quad . \quad (1.19)
\]
1.4 Analysis of simulation trajectories

The GROMOS software package for biomolecular simulations consists of two major parts: (1) MD++, containing all the tools and routines to perform energy minimisations, MD or stochastic dynamics (SD) simulations, and (2) GROMOS++, which is used for pre- and post-processing of biomolecular simulation data, i.e. preparing the input files needed for a simulation or performing analysis based on the trajectories of a simulation.

Although it is possible to calculate many properties while running the simulation, which would make the existence of GROMOS++ or similar analysis tools and packages obsolete, such a procedure is not recommended for the following two reasons:

a) Most likely one does not know all relevant properties to be calculated and needed to adequately solve the problem of interest in advance, i.e. some properties would need to be calculated by rerunning the whole simulation again or are of less importance but would slow down the simulation speed when calculated during the simulation.

b) It may not be obvious what resolution is needed to calculate averages for the properties of interest before knowing the dynamics of the system under study. A safe way would be to calculate all properties at every integration time step, but this would slow down the whole simulation and is not necessary for most properties of interest.

The part of the software performing the actual simulation and the post-processing analysis tool must use the same algorithms, for example to calculate the system energies or nearest image projections of one atom with respect to another within the periodic boundary conditions. In Chapter 4 an overview of the GROMOS++ software is given discussing the various possibilities of analysing different simulation trajectories.

1.4.1 Calculation of NMR data

GROMOS++ also provides programs to calculate data obtainable from NMR spectroscopy, such as $^1$H-$^1$H NMR NOE distances, residual dipolar couplings (RDCs), $^3$J-coupling constants and $^1$H-$^{15}$N order parameters. Such data are calculated for the protein hen egg white lysozyme (HEWL) in Chapters 6, 7 and 8. Additionally, NMR NOE proton-proton upper bound distances have been used to induce a specific fold of HEWL in TFE/water as described in Chapter 6.

Such NMR data can be used to estimate the quality of a simulation by comparing simulated averages of observable quantities to the corresponding experimentally measured ones. So-called $^3$J-coupling constants are related to structure based on the empirical Karplus relation [21, 22],

$$^3 J(\theta) = a \cos^2 \theta + b \cos \theta + c \ ,$$  

(1.20)
where \( a, b \) and \( c \) are parameters depending on the type of the four atoms involved in the coupling, and \( \theta \) is the torsional angle spanned by the four atoms. The empirical nature of the Karplus relation is a source of considerable uncertainty [23], and structural data derived using Eq. (1.20) from experimental \( ^3J \)-couplings as well as \( ^3J \)-coupling constants computed from simulations using Eq. (1.20) should not be over-interpreted.

1.5 The protein hen egg white lysozyme

A first description of HEWL was reported in 1922 [24], while the first protein structure of it was described in 1965 [25]. Its main secondary structure elements are four \( \alpha \)-helices, two \( 3_{10} \)-helices, and three \( \beta \)-strand forming one \( \beta \)-sheet (see Fig. 8.3). HEWL is well studied by NMR providing a wealth of experimental data to compare with [26–29]. Therefore it is one of the proteins of choice for testing force fields and force-field effects with respect to the secondary and tertiary structure of a protein. In Chapters 7 and 8 the influence of backbone amide-to-ester replacements on the overall protein structure is studied, while in Chapter 6 the effect of the solvent, water versus 2,2,2-trifluoroethanol/water mixture on the protein secondary structure is probed.
A method to apply bond-angle and dihedral-angle constraints in molecular dynamics simulation

A method to apply bond-angle or dihedral-angle constraints in molecular dynamics simulations of macromolecules is presented. It uses Cartesian coordinates and determines the Lagrangian multipliers necessary for maintaining the constraints iteratively. It constitutes an alternative to the use of only distance constraints between particles to maintain a particular geometry. Since the latter are not suitable to maintain particular, e.g. linear or flat, geometries of molecules, the presented method offers an alternative to the methods standardly used in such cases. It can easily handle bond-length, bond-angle and dihedral-angle constraints simultaneously, as when calculating a potential of mean force along a dihedral-angle coordinate while applying bond-length constraints throughout the macromolecule.
2.1 Introduction

In molecular dynamics (MD) simulations of large molecules, particular degrees of freedom are often kept fixed. One reason is that elimination of the highest frequency motions in a molecule, that are not essential for the properties of the molecular system to be investigated, will enhance the computational efficiency of simulating due to the larger integration time step that can be used. For macromolecules such as proteins, the application of bond-length constraints may easily reduce the computational effort by a factor of 4 [30], while the structural and dynamical properties of the protein remain unaffected by the constraints [31, 32]. The additional application of bond-angle constraints appeared to reduce the flexibility of the protein significantly [31, 32] and should be avoided in protein simulations, not only for this reason [33]. Another application of constraints in an MD simulation is in the calculation of a potential of mean force along a chosen coordinate or degree of freedom. Such a potential of mean force or free-energy profile along a degree of freedom can be computed in different ways, from standard MD simulations using a biasing umbrella potential energy term, or from MD simulations in which the coordinate of interest is kept fixed at a series of values along the degree of freedom of interest. The latter method is efficient [34] and simple to use as long as a method to impose the particular constraint of interest on to the system is available.

Different methods are available to apply distance constraints during an MD simulation [14, 35–39], of which the SHAKE method [14] is the oldest and most appropriate one for large molecules. It solves the set of quadratic equations for the Lagrange multipliers that represent the set of constraint forces by linearising these equations and solving them iteratively. If the constraint matrix containing the distance constraint lengths between atoms \( i \) and \( j \) is rather sparse, e.g. because of the few and localised bond length constraints in macromolecules such as proteins, an iterative procedure to determine the Lagrange multipliers is more efficient than the use of matrix inversion procedures [33].

Since a bond-angle constraint of an angle \( \theta = (k_1, k_2, k_3) \) can be expressed as a distance constraint \( d = (k_1, k_3) \) in case the distances \( (k_1, k_2) \) and \( (k_2, k_3) \) are also constrained, the procedure SHAKE can also be used to impose a bond-angle constraint, as for example to maintain the geometry of small solvent molecules such as water or methanol for rigid models of the compound. However, for geometrically linear molecules this way of maintaining a bond-angle constraint using SHAKE will not work properly, because the direction along which SHAKE tries to reset a constrained distance is nearly orthogonal to the direction along which it should execute the resetting in order to obtain a correctly linear configuration. This problem can be solved by formulating a bond-angle constraint in the form of an angle \( \theta_{k_1,k_2,k_3} \) constraint instead of a distance \( d_{k_1,k_2} \) constraint. Such an approach has been formulated in [40], where expressions for dihedral angles \( \varphi = (k_1, k_2, k_3, k_4) \) are given in terms of squares of \( \cos \varphi \), this in analogy of the formulation of distance constraints in terms of \( d^2 \) which avoids the appearance of square roots. Because of the use of \( \cos^2 \varphi \) in the constraint equation, the formulation of [40] cannot distinguish between positive and negative \( \cos \varphi \) values, which may lead to convergence problems if the initial \( \varphi \)-value differs...
2.2 Method

A set of \( k = 1, 2, \ldots, N_{ac} \) angle constraints between atoms with positions \( r_{k_1}, r_{k_2}, r_{k_3} \) can be written as

\[
\sigma^a_k (\theta_k(r^N); \theta^0_k) \equiv \cos (\theta_k(r^N)) - \cos (\theta^0_k) = 0 , \quad k = 1, 2, \ldots, N_{ac} \tag{2.1}
\]

where the angle \( \theta_k(r^N) \) is constrained to the value \( \theta^0_k \). The angle \( \theta_k(r^N) \) is defined in terms of the positions \( r_{k_1}, r_{k_2} \) and \( r_{k_3} \) of the three particles with indices \( k_1, k_2 \) and \( k_3 \), \( \theta_k(k_1-k_2-k_3) \), where \( k \) is a short-hand notation for \( (k_1, k_2, k_3) \), and \( r^N \equiv (r_1, r_2, \ldots, r_N) \) is a short-hand notation for the positions of the \( N \) particles in the system. The relation between \( \theta_k \) and \( r^N \) is

\[
\theta_k = \arccos \left( \frac{r_{k_1 k_2} \cdot r_{k_3 k_2}}{r_{k_1 k_2} r_{k_3 k_2}} \right) \quad 0 \leq \theta_k \leq \pi \tag{2.2}
\]

where \( r_{k_1 k_2} \equiv r_{k_1} - r_{k_2} \) and \( r_{k_1 k_2} \equiv (r_{k_1 k_2} \cdot r_{k_1 k_2})^{1/2} = ||r_{k_1 k_2}|| \) is the length of the vector \( r_{k_1 k_2} \).

Newton’s equations of motion for \( N \) particles with masses \( m_i \) including a potential energy function \( V(r^N) \) and the constraints \( \sigma^a_k \) multiplied with the Lagrange multipliers \( l^a_k(t) \) are

\[
m_i \frac{d^2 r_i(t)}{dt^2} = -\frac{\partial}{\partial r_i} \left( V(r^N) + \sum_{k=1}^{N_{ac}} l^a_k(t) \sigma^a_k (\theta_k(r^N); \theta^0_k) \right) . \tag{2.3}
\]

The Lagrange multipliers \( l^a_k(t) \) are to be determined such that condition (2.1) is satisfied. The first term on the right side in Eq. (2.3) represents the unconstrained force \( f^{uc}_i(t) \) on particle \( i \) derived from the interaction function \( V(r^N) \) and the second term represents the yet unknown constraint force

\[
f^{ac}_i(t) = -\sum_{k=1}^{N_{ac}} l^a_k(t) \frac{\partial}{\partial r_i} \sigma^a_k (\theta_k(r^N); \theta^0_k) = -\sum_{k=1}^{N_{ac}} l^a_k(t) \frac{\partial \cos (\theta_k(r^N))}{\partial r_i} \tag{2.4}
\]
Expressions for \( \partial \cos \left( \theta_k \right) / \partial r_i \) can be found in the literature [13, 41, 42],

\[
\frac{\partial \cos \left( \theta_k \left( r^N \right) \right)}{\partial r_i} = \delta_{ik_1} \frac{r_{k_1k_2}^2 r_{k_3k_2} - (r_{k_1k_2} \cdot r_{k_3k_2}) r_{k_1k_2}}{r_{k_1k_2}^3 r_{k_3k_2}} - \delta_{ik_2} \left( \frac{r_{k_1k_2}^2 r_{k_3k_2} - (r_{k_1k_2} \cdot r_{k_3k_2}) r_{k_1k_2}}{r_{k_1k_2}^3 r_{k_3k_2}} \right) + \frac{r_{k_3k_2}^2 r_{k_1k_2} - (r_{k_3k_2} \cdot r_{k_1k_2}) r_{k_1k_2}}{r_{k_1k_2}^3 r_{k_3k_2}} + \delta_{ik_3} \frac{r_{k_3k_2}^2 r_{k_1k_2} - (r_{k_3k_2} \cdot r_{k_1k_2}) r_{k_1k_2}}{r_{k_1k_2}^3 r_{k_3k_2}}
\]

(2.5)

where we have used the short-hand notation

\[
a_{k_1k_2k_3} = \frac{r_{k_1k_2}^2 r_{k_3k_2} - (r_{k_1k_2} \cdot r_{k_3k_2}) r_{k_1k_2}}{r_{k_1k_2}^3 r_{k_3k_2}} .
\]

(2.6)

The angle constraint force on particle \( i \) then becomes

\[
f_{i}^{ac}(t) = - \sum_{k=1}^{N_{ac}} t_k^a(t) \left( \delta_{ik_1} a_{k_1k_2k_3} - \delta_{ik_2} (a_{k_1k_2k_3} + a_{k_3k_2k_1}) + \delta_{ik_3} a_{k_3k_2k_1} \right) .
\]

(2.7)

The leap-frog scheme [8] to integrate Newton’s equations of motion using a time step \( \Delta t \) yields for the unconstrained positions a time \( t_n + \Delta t \),

\[
r_i^{uc}(t_n + \Delta t) = r_i(t_n) + v_i(t_n - \frac{\Delta t}{2}) \Delta t + m_i^{-1} f_i^{uc}(t_n)(\Delta t)^2 ,
\]

(2.8)

where the particle velocities are indicated by \( v_i \). The constrained positions at time \( t_n + \Delta t \) are related to the constraint forces through

\[
r_i(t_n + \Delta t) = r_i^{uc}(t_n + \Delta t) + m_i^{-1} f_i^{uc}(t_n)(\Delta t)^2 ,
\]

(2.9)

and should satisfy the constraint equations (2.1),

\[
\cos \left( \theta_k \left( r^N(t_n + \Delta t) \right) \right) - \cos \left( \theta_k^0 \right) = 0 ,
\]

(2.10)

or using Eq. (2.2)

\[
\left( \frac{r_{k_1k_2}(t_n + \Delta t) \cdot r_{k_3k_2}(t_n + \Delta t)}{r_{k_1k_2}(t_n + \Delta t) r_{k_3k_2}(t_n + \Delta t)} \right) - \cos \left( \theta_k^0 \right) = 0 .
\]

(2.11)

Since \( r_{k_1k_2}(t_n + \Delta t) \) and \( r_{k_3k_2}(t_n + \Delta t) \) are each linear in the Lagrange multipliers \( l_k^0(t_n) \),
both the numerator and the denominator of the left term in Eq. (2.11) contain powers up to two of the $l_k^n(t_n)$. Thus a set of $N_{ac}$ equations consisting of terms containing up to powers of two of the unknowns $l_k^n(t_n)$ is to be solved at every MD time step. This can be done by linearisation, i.e. neglect of terms quadratic in $l_k^n(t_n)$, followed by matrix inversion or by sequentially solving the linearised equations for each constraint, omitting the coupling between the different constraints, and iterating through all the equations until the $l_k^n(t_n)$ values converge to consistent values which satisfy Eq. (2.11). The latter method is used in the procedure SHAKE [14].

Using Eqs. (2.7) and (2.9) we find for the effect of the $k$-th constraint on the relative position of particles $i_1$ and $i_2$, and neglecting the coupling between constraints,

$$
\mathbf{r}_{i_1i_2}(t_n + \Delta t) = \mathbf{r}_{i_1i_2}^{uc}(t_n + \Delta t) + (\Delta t)^2 \left( m_{i_1}^{-1} \mathbf{f}_{i_1}^{ac}(t_n) - m_{i_2}^{-1} \mathbf{f}_{i_2}^{ac}(t_n) \right)
= \mathbf{r}_{i_1i_2}^{uc}(t_n + \Delta t) - (\Delta t)^2 \mathbf{f}_k(t_n) \cdot \left[ (\delta_{i_1k_1}m_{i_1}^{-1} + \delta_{i_2k_1}m_{i_2}^{-1}) \mathbf{a}_{k_1k_2k_3}
- (\delta_{i_1k_2}m_{i_1}^{-1} + \delta_{i_2k_2}m_{i_2}^{-1}) \mathbf{a}_{k_1k_2k_3}
+ (\delta_{i_1k_3}m_{i_1}^{-1} + \delta_{i_2k_3}m_{i_2}^{-1}) \mathbf{a}_{k_1k_2k_3} \right].
$$

(2.12)

For particles $k_1$ and $k_2$ we then get

$$
\mathbf{r}_{k_1k_2}(t_n + \Delta t) = \mathbf{r}_{k_1k_2}^{uc}(t_n + \Delta t) - (\Delta t)^2 t_k^a(t_n).
\cdot \left( m_{k_1}^{-1} \mathbf{a}_{k_1k_2k_3} + m_{k_2}^{-1} \mathbf{a}_{k_1k_2k_3} + \mathbf{a}_{k_3k_2k_1} \right)
= \mathbf{r}_{k_1k_2}^{uc}(t_n + \Delta t) - (\Delta t)^2 t_k^a(t_n) \mathbf{b}_{k_1k_2k_3},
$$

(2.13)

where we have used the short-hand notation

$$
\mathbf{b}_{k_1k_2k_3} = m_{k_1}^{-1} \mathbf{a}_{k_1k_2k_3} + m_{k_2}^{-1} \mathbf{a}_{k_1k_2k_3} + \mathbf{a}_{k_3k_2k_1}.
$$

(2.14)

For particles $k_2$ and $k_3$ we find likewise

$$
\mathbf{r}_{k_2k_3}(t_n + \Delta t) = \mathbf{r}_{k_2k_3}^{uc}(t_n + \Delta t) - (\Delta t)^2 t_k^a(t_n).
\cdot \left( m_{k_3}^{-1} \mathbf{a}_{k_3k_2k_1} + m_{k_2}^{-1} \mathbf{a}_{k_3k_2k_1} + \mathbf{a}_{k_1k_2k_3} \right)
= \mathbf{r}_{k_2k_3}^{uc}(t_n + \Delta t) - (\Delta t)^2 t_k^a(t_n) \mathbf{b}_{k_3k_2k_1},
$$

(2.15)

The scalar product in the numerator of the first term in Eq. (2.11) becomes after linearisation

$$
\mathbf{r}_{k_1k_2}(t_n + \Delta t) \cdot \mathbf{r}_{k_2k_3}(t_n + \Delta t) = \left( \mathbf{r}_{k_1k_2}^{uc}(t_n + \Delta t) \cdot \mathbf{r}_{k_2k_3}^{uc}(t_n + \Delta t) \right)
- (\Delta t)^2 t_k^a(t_n) \cdot \left( \left( \mathbf{r}_{k_1k_2}^{uc}(t_n + \Delta t) \cdot \mathbf{b}_{k_3k_2k_1} \right) + \left( \mathbf{r}_{k_3k_2}^{uc}(t_n + \Delta t) \cdot \mathbf{b}_{k_1k_2k_3} \right) \right).
$$

(2.16)

The factors in the denominator of the first term in Eq. (2.11) contain a square root which complicates solving Eq. (2.11) for $l_k^n(t_n)$. The Taylor expansion of the square root, cut-off
after the second term yields
\[
(x - y)^{1/2} = x^{1/2} \left( 1 - \frac{1}{2} \left( \frac{y}{x} \right) \right) = x^{1/2} - \frac{1}{2} y x^{-1/2} .
\]

(2.17)

Using this approximation and (2.13) and linearising with respect to \(l_k^p(t_n)\) we find
\[
r_{k_1k_2}(t_n + \Delta t) = \left( (r_{k_1k_2}^{uc}(t_n + \Delta t))^2 - 2(\Delta t)^2 l_k^p(t_n) \cdot (r_{k_1k_2}^{uc}(t_n + \Delta t) \cdot b_{k_1k_2k_3}) \right)^{1/2}
\]
\[
= r_{k_1k_2}^{uc}(t_n + \Delta t) - (\Delta t)^2 l_k^p(t_n) \frac{r_{k_1k_2}^{uc}(t_n + \Delta t) \cdot b_{k_1k_2k_3}}{r_{k_1k_2}^{uc}(t_n + \Delta t)} ,
\]

(2.18)

and likewise
\[
r_{k_3k_2}(t_n + \Delta t) = r_{k_3k_2}^{uc}(t_n + \Delta t) - (\Delta t)^2 l_k^p(t_n) \frac{r_{k_3k_2}^{uc}(t_n + \Delta t) \cdot b_{k_3k_2k_1}}{r_{k_3k_2}^{uc}(t_n + \Delta t)} .
\]

(2.19)

The linearised denominator of the first term in Eq. (2.11) is then
\[
r_{k_1k_2}(t_n + \Delta t)r_{k_3k_2}(t_n + \Delta t) = r_{k_1k_2}^{uc}(t_n + \Delta t)r_{k_3k_2}^{uc}(t_n + \Delta t) - (\Delta t)^2 l_k^p(t_n) \cdot
\]
\[
\cdot \left( r_{k_1k_2}^{uc}(t_n + \Delta t) \left( r_{k_3k_2}^{uc}(t_n + \Delta t) \cdot b_{k_3k_2k_1} \right)
\]
\[
+ r_{k_3k_2}^{uc}(t_n + \Delta t) \left( r_{k_1k_2}^{uc}(t_n + \Delta t) \cdot b_{k_1k_2k_3} \right) \right) .
\]

(2.20)

If (2.11) is written in terms of the linearised expressions of the numerator and denominator of the first term we get as equation for the \(l_k^p(t_n)\),
\[
\frac{c_1 - c_2 l_k^p(t_n)}{c_3 - c_4 l_k^p(t_n)} - \cos \left( \theta_k^0 \right) = 0
\]

(2.21)

with
\[
c_1 = \left( r_{k_1k_2}^{uc}(t_n + \Delta t) \cdot r_{k_3k_2}^{uc}(t_n + \Delta t) \right) ,
\]

(2.22)

\[
c_2 = (\Delta t)^2 \left( \left( r_{k_1k_2}^{uc}(t_n + \Delta t) \cdot b_{k_3k_2k_1} \right) + \left( r_{k_3k_2}^{uc}(t_n + \Delta t) \cdot b_{k_1k_2k_3} \right) \right) ,
\]

(2.23)

\[
c_3 = r_{k_1k_2}^{uc}(t_n + \Delta t) r_{k_3k_2}^{uc}(t_n + \Delta t) ,
\]

(2.24)

\[
c_4 = (\Delta t)^2 \left( r_{k_1k_2}^{uc}(t_n + \Delta t) \left( r_{k_3k_2}^{uc}(t_n + \Delta t) \cdot b_{k_3k_2k_1} \right)
\]
\[
+ r_{k_3k_2}^{uc}(t_n + \Delta t) \left( r_{k_1k_2}^{uc}(t_n + \Delta t) \cdot b_{k_1k_2k_3} \right) \right) ,
\]

(2.25)
which yields
\[
l_k^a(t_n) = \frac{c_1 - c_3 \cos \left( \theta_k^0 \right)}{c_2 - c_4 \cos \left( \theta_k^0 \right)} .
\] (2.26)

The precision of the values of the Lagrange multipliers \(l_k^a(t_n)\) will be affected by the linearisation of the numerator and denominator in (2.11) which made use of a Taylor expansion of the square root function. If expression (2.26) does not yield sufficiently precise values for the \(l_k^a(t_n)\), one could obtain more precise values by refining the estimate (2.26) through the following iterative procedure:

1. Calculate \(r_{k_1k_2}(t_n + \Delta t)\) and \(r_{k_3k_2}(t_n + \Delta t)\) from Eqs. (2.13) and (2.15) respectively, using the current value of \(l_k^a(t_n)\).

2. Calculate \(r_{k_1k_2}(t_n + \Delta t)\) and \(r_{k_3k_2}(t_n + \Delta t)\) as the respective sizes of these vectors.

3. Calculate \(\cos \left( \theta_k (t_n + \Delta t) \right)\) from the definition (2.2).

4. If
\[
\left( \cos \left( \theta_k(t_n + \Delta t) \right) - \cos \left( \theta_k^0 \right) \right)
\] (2.27)
differs more than a given tolerance \(\varepsilon\) from zero, use Eq. (2.26) to determine whether \(l_k^a(t_n)\) is to be increased or decreased to satisfy (2.10): if
\[
\left( \cos \left( \theta_k(t_n + \Delta t) \right) - \cos(\theta_k^0) \right) c_2 l_k^a(t_n)
\] (2.28)
is larger than zero,
\[
|l_k^a(t_n)|
\] (2.29)
is to be enlarged, while if (2.28) is smaller than zero, (2.29) is to be reduced. So in both cases only the size of \(l_k^a(t_n)\) is to be increased or decreased, while its sign is not changed, and the new value becomes the current value in the next iteration. Such a change in size of \(l_k^a(t_n)\) could e.g. start at 10% and then be halved at every iteration.

In a calculation of the potential of mean force along the angle coordinate \(\theta_k\), the value of \(\theta_k^0\) is to be changed from a value \(\theta_k^{0,A}\) to \(\theta_k^{0,B}\). This can be achieved by making \(\theta_k^0\) dependent on a so-called coupling parameter \(\lambda\) that connects the values \(\theta_k^{0,A}\) and \(\theta_k^{0,B}\) in some manner, e.g. linearly through
\[
\theta_k^0(\lambda) = (1 - \lambda) \theta_k^{0,A} + \lambda \theta_k^{0,B} .
\] (2.30)

So, \(\theta_k^{0,A}\) is the angle constraint value in state or Hamiltonian \(A\) and \(\theta_k^{0,B}\) in state \(B\). The set of angle constraints (2.1) then becomes
\[
\sigma_k^a \left( \theta_k(r^N); \theta_k^0(\lambda) \right) \equiv \cos \left( \theta_k(r^N) \right) - \cos \left( \theta_k^0(\lambda) \right) = 0 \quad k = 1, 2, \ldots, N_{ac} .
\] (2.31)
Chapter 2. Angle constraints

The derivation of the free energy $F(\lambda)$ of the system with respect to $\lambda$ for a system including angle constraints (2.30) and (2.31) becomes

$$\frac{dF}{d\lambda} = \left\langle \frac{\partial K}{\partial \lambda} \right\rangle_{\lambda} + \left\langle \frac{\partial V}{\partial \lambda} \right\rangle_{\lambda} + \left\langle \frac{\partial}{\partial \lambda} \sum_{k=1}^{N_{ac}} l_k^0 \sigma_k \left( \theta_k^0 (r^N); \theta_k^0 (\lambda) \right) \right\rangle_{\lambda}. \quad (2.32)$$

The symbol $\langle \ldots \rangle_{\lambda}$ denotes an ensemble average or trajectory average over the constrained simulation at the value $\lambda$. The first term on the right contains the contribution of the kinetic energy $K$, e.g. present when the mass of a particle depends on $\lambda$. The second term represents the contribution of the unconstrained interaction terms, and the third term the contribution of the constraint forces, which can be expressed for the $k$-th constraint as

$$\frac{dF^c_k(\lambda)}{d\lambda} = \langle l_k^0 \rangle_{\lambda} \sin \left( \theta_k^0 (\lambda) \right) \left( \theta_k^{0,B} - \theta_k^{0,A} \right). \quad (2.33)$$

The total contribution of the $N_{ac}$ constraints to the potential of mean force along the pathway defined by $\lambda$ is then

$$\frac{dF^c(\lambda)}{d\lambda} = \sum_{k=1}^{N_{ac}} \frac{dF^c_k(\lambda)}{d\lambda}, \quad (2.34)$$

and the free energy difference between states $B$ and $A$ reads

$$F^c(B) - F^c(A) = \int_0^1 \frac{dF^c(\lambda)}{d\lambda} d\lambda. \quad (2.35)$$

2.3 Discussion

For a geometrically linear molecule such as CO$_2$ or MeCN, the method presented here presents a useful alternative to the use of only distance constraints which fails to maintain particular, e.g. linear or flat, geometries [43].

Another alternative to determine the Lagrange multipliers necessary for maintaining a constrained linear geometry of particles has been proposed in [44]. It is based on a transformation of the Cartesian coordinates to a coordinate system defined by the constraints, which leads to a set of non-linear equations for the Lagrange multipliers, which is linearised by exploiting the geometric linearity of the set of particles. These equations seem more complicated than the ones presented here and less general because they only apply to geometrically linear sets of particles.

The algorithm presented here to maintain bond-angle constraints can be easily applied in conjunction with the corresponding algorithms to apply bond-length constraints [14] or dihedral-angle constraints [40], for which the formulae based on $\cos(\phi)$ instead of $\cos^2(\phi)$ constraining are given in the Appendix. A torsional, dihedral-angle constraint for particles
(k_1, k_2, k_3, k_4) is independent of the two bond-angle values for the particles (k_1, k_2, k_3) and (k_2, k_3, k_4) and also independent of the three bond-length values for the particles (k_1, k_2), (k_2, k_3), and (k_3, k_4). A bond-angle constraint for particles (k_1, k_2, k_3) is independent of the two bond-length values for the particles (k_1, k_2) and (k_2, k_3). This means that these three types of constraints may be solved separately: first the bond-length constraints, then the bond-angle constraints followed by the dihedral-angle constraints. The algorithm to maintain dihedral-angle constraints that is presented in the Appendix may thus be useful when determining a potential of mean force along a dihedral-angle degree of freedom in a macromolecule that is simulated using bond-length constraints.
2.4 Appendix

In reference [40] expressions for the constraint forces \( f^c \) and their contribution \( dF^c/d\lambda \) to the free energy \( F(\lambda) \) for dihedral-angle constraints were derived for a set of \( N_c \) dihedral-angle constraints, see Eqs. (39) and (40) of [40], that are quadratic in the cosine of the dihedral or torsional angle \( \varphi_k(r^N) \), which is defined in terms of the positions \( r_{k1}, r_{k2}, r_{k3} \) and \( r_{k4} \) of its four constituting particles \( k_1, k_2, k_3 \) and \( k_4 \), i.e. \( k_1-k_2-k_3-k_4 \). Here the corresponding expressions for a set of \( N_{dac} \) dihedral-angle constraints

\[
\sigma_k^{\text{dac}}(\varphi_k(r^N); \varphi^0_k(\lambda)) \equiv \cos(\varphi_k(r^N)) - \cos(\varphi^0_k(\lambda)) = 0 \quad k = 1, 2, \ldots, N_{dac} \tag{2.36}
\]

linear in \( \cos(\varphi_k(r^N)) \) is presented. This means that in Eqs. (39), (45) and (46) of [40] the squares of the cosines and of the first term in Eq. (46) have to be removed. Since the derivation of the expressions runs analogously to the one presented in section 5.2 of [40], we only give the modifications and some resulting expressions.

Eqn. (42) of [40] for the constraint force becomes

\[
f_i^{\text{dac}}(t) = \sum_{k=1}^{N_{dac}} l_k^{\text{da}}(t) \sin(\varphi_k(r^N)) \frac{\partial \varphi_k(r^N)}{\partial r_i} \tag{2.37}
\]

This implies that in Eqs. (44), (47–53), (55–56) and (59) the factor \( \sin(2\varphi_k(t_n)) \) is to be replaced by \( \sin(\varphi_k(t_n)) \). With this modification Eqs. (44), (47–53) can be used to derive the expression for the Lagrange multipliers \( l_k^{\text{da}}(t_n) \). We note, however, that the definitions of \( a_{ki}, a_{k2}, a_{k3}, \) and \( a_{k4} \) through Eqs. (42–44) in [40] and of \( b_{k1,k2,k3} \) through Eqs. (48–49) in [40] are different from the definitions Eq. (2.6) for \( a_{k1,k2,k3} \) and Eq. (2.14) for \( b_{k1,k2,k3} \) that are used in the present article to define bond-angle constraints. So these definitions are not to be used in the equations in this Appendix which are for dihedral-angle constraints and kept as specified in [40]. Eq. (54) of [40] is not needed for the linear constraint Eq. (2.36), which require the square root of expressions (55) and (56) of [40]. Using the approximation (2.17) we find

\[
r_{k5k2}(t_n + \Delta t) = r_{k5k2}^{\text{uc}}(t_n + \Delta t) + l_k^{\text{da}}(t_n) \sin(\varphi_k(t_n))(\Delta t)^2 \cdot \frac{(r_{k5k2}^{\text{uc}}(t_n + \Delta t) \cdot b_{k1k2k3}(t_n, t_n + \Delta t))}{r_{k5k2}^{\text{uc}}(t_n + \Delta t)} \tag{2.38}
\]

and

\[
r_{k6k3}(t_n + \Delta t) = r_{k6k3}^{\text{uc}}(t_n + \Delta t) + l_k^{\text{da}}(t_n) \sin(\varphi_k(t_n))(\Delta t)^2 \cdot \frac{(r_{k6k3}^{\text{uc}}(t_n + \Delta t) \cdot b_{k2k3k4}(t_n, t_n + \Delta t))}{r_{k6k3}^{\text{uc}}(t_n + \Delta t)} \tag{2.39}
\]
So the linearised denominator in the constraint equations

\[
\left( \frac{\mathbf{r}_{k5k2}(t_n + \Delta t) \cdot \mathbf{r}_{k6k3}(t_n + \Delta t)}{\mathbf{r}_{k5k2}(t_n + \Delta t) \mathbf{r}_{k5k3}(t_n + \Delta t)} \right) - \cos \left( \varphi_k^0(\lambda) \right) = 0
\]  

(2.40)

becomes

\[
\mathbf{r}_{k5k2}(t_n + \Delta t) \mathbf{r}_{k6k3}(t_n + \Delta t) = r_{k5k2}^{uc}(t_n + \Delta t) r_{k6k3}^{uc}(t_n + \Delta t) \\
+ t_k^{da}(t_n) \sin(\varphi_k(t_n)) (\Delta t)^2 \cdot \left( \frac{r_{k5k2}^{uc}(t_n + \Delta t)}{r_{k6k3}^{uc}(t_n + \Delta t)} \mathbf{r}_{k6k3}^{uc}(t_n + \Delta t) \cdot \mathbf{b}_{k2k3k4}(t_n, t_n + \Delta t) + r_{k6k3}^{uc}(t_n + \Delta t) \mathbf{b}_{k1k2k3}(t_n, t_n + \Delta t) \right).
\]

(2.41)

If Eq. (2.40) is written in terms of the linearised expressions of the numerator (modified Eq. (53) in [40]) and the denominator Eq. (2.41) of the first term, we get as equation for the Lagrange multipliers \( t_k^{da}(t_n) \),

\[
g_1 + g_2 t_k^{da}(t_n) - \cos \left( \varphi_k^0(\lambda) \right) = 0
\]

(2.42)

with

\[
g_1 = \left( r_{k5k2}^{uc}(t_n + \Delta t) \cdot r_{k6k3}^{uc}(t_n + \Delta t) \right),
\]

(2.43)

\[
g_2 = \sin(\varphi_k(t_n)) (\Delta t)^2 \left( r_{k5k2}^{uc}(t_n + \Delta t) \cdot \mathbf{b}_{k2k3k4}(t_n, t_n + \Delta t) \right),
\]

(2.44)

\[
g_3 = r_{k5k2}^{uc}(t + \Delta t) r_{k6k3}^{uc}(t + \Delta t),
\]

(2.45)

\[
g_4 = \sin(\varphi_k(t_n)) (\Delta t)^2 \left( r_{k5k2}^{uc}(t + \Delta t) \mathbf{r}_{k6k3}^{uc}(t + \Delta t) \cdot \mathbf{b}_{k1k2k3}(t_n, t_n + \Delta t) + r_{k6k3}^{uc}(t_n + \Delta t) \mathbf{b}_{k1k2k3}(t_n, t_n + \Delta t) \right).
\]

(2.46)

So Eq. (58) in [40] is to be replaced by

\[
t_k^{da}(t_n) = \frac{g_3 \cos(\varphi_k^0(\lambda)) - g_1}{g_2 - g_4 \cos(\varphi_k^0(\lambda))},
\]

(2.47)

If this expression for \( t_k^{da}(t_n) \) does not yield sufficiently precise values, one may apply an iterative procedure analogous to the one described for the bond-angle constraint Lagrange multipliers \( t_k^{a}(t_n) \).
Chapter 3

A supra-atomic coarse-grained GROMOS force field for aliphatic hydrocarbons in the liquid phase

A supra-atomic coarse-grained (CG) force-field for liquid \( n \)-alkanes is presented. The model was calibrated using experimental thermodynamic data and structural as well as energetic data for 18 \( n \)-alkanes from atomistic fine-grained (FG) simulations of the corresponding hydrocarbons using the GROMOS biomolecular force field. A parametrisation of the non-bonded force-field parameters to experimental values for the density and heat of vaporisation turned out to be mandatory for a correct reproduction of these data, while the bonded force-field parameters could directly be obtained from a Boltzmann-weighted fit with respect to the corresponding properties from mapped FG simulations. The model presents 6 different CG bead types for bead sizes from 2 to 4 and distinguishes between terminal and non-terminal beads within an alkane chain (end or middle). The use of bead size 3 is recommended, since larger bead sizes induce too much structure in the liquids, whereas bead size 2 delivers only moderate computational gain. The CG model was further tested by comparing predictions of the excess free energy and the self-diffusion constant for \( n \)-alkanes to experimental values. The CG model offers a thermodynamically calibrated basis for the development of CG models for lipids.
3.1 Introduction

Since the first simulation of protein motion [45] simulation of the dynamics in biomolecular systems, proteins, nucleotides, carbohydrates, lipids and the like in aqueous solution, has become a standard method of investigation over the past decades [46, 47]. Because an accurate quantum-mechanical treatment of the many nuclear and electronic degrees of freedom in such systems for a sufficiently long time period to observe functionally relevant conformational changes is as yet impossible using currently available computers, biomolecular systems are still generally simulated at the atomic level of resolution using classical-mechanical equations of motion and an empirical potential energy function that yields the potential energy of the system in terms of atomic coordinates [48]. Such a potential energy function is shortly called force field, and a number of general atomic-level interaction functions for biomolecules have been developed over the past decades, e.g. AMBER [49, 50], CHARMM [51, 52] and GROMOS [32, 53].

When simulating very large systems such as multi-protein complexes or membranes, simulation at the atomic level of resolution can only reach the nanosecond time scale which is insufficient to properly sample longer time motions in such systems. Therefore, much effort has more recently been spent on the development of more coarse-grained, supra-atomic level of resolution models for biomolecular systems, see e.g. [54–59] for reviews. Because coarse-graining involves the elimination of particular degrees of freedom from a model, some properties of the molecular system may get lost in the coarse-graining process. This implies that the choice of degrees of freedom that are to be kept in the more coarse-grained model will depend on the particular properties of the molecular system that are the focus of the investigation using computer simulations.

When can particular degrees of freedom be eliminated? They must be non-essential for the process or property of interest, and must be large in number in order to have the computational gain off-set the inevitable loss of accuracy through coarse-graining. Interactions governing eliminated degrees of freedom should be largely decoupled from those governing the degrees of freedom remaining after coarse-graining, leading to decoupled motions between these two sets of degrees of freedom, and the effective interactions governing the remaining degrees of freedom should be simply representable in the form of a fast computable potential energy function [30, 54]. For example, the motions of aliphatic hydrogen atoms in CH, CH₂ and CH₃ groups are irrelevant for most processes of biochemical interest. By treating these groups as single united atoms [60], the number of interaction sites is reduced by a factor 2 to 4. For alkanes and lipids this leads to almost a factor of 10 fewer pairwise non-bonded interactions, be it at the cost of losing the dipolar interactions of CHₙ-moieties and the van der Waals interactions of the hydrogen atoms. The intra-group motions of CHₙ moieties are largely decoupled from other motions in such molecules and the torsional interactions involving these hydrogen atoms can be effectively included in the torsional interaction function around the same bond that only involves non-hydrogen atoms. Thus, all four conditions for proper coarse-graining are basically fulfilled in this case. For this reason the GROMOS atomic-level resolution force field represents aliphatic
CH$_n$-groups as united atoms.

What molecular properties or processes are important for the process of interest and should thus be maintained while coarse-graining?

1. Molecular structure of the solute, e.g. the protein, and structure of the solvent or liquid.

2. Thermodynamic properties that carry volume and energetic information, such as density, heat of vaporisation, and excess free energy. Thermodynamic properties that characterise a response to a change in thermodynamic state point, such as compressibility and heat capacity, are of less importance.

3. Dielectric properties, such as the static dielectric permittivity $\epsilon(0)$ that governs the screening of Coulomb interactions.

4. Dynamic properties such as diffusion, viscosity and molecular relaxation times are less important, because most biomolecular processes are thermodynamically driven.

In any case, the process of coarse-graining is likely to reduce the usefulness of the model in different ways: (i) The range of thermodynamic state points at which the model may be applied is generally reduced; (ii) The transferability of model parameters between similar but not identical moieties or compounds is usually reduced; (iii) The accuracy of the reproduction of measured values for properties may be reduced; (iv) The physical basis of a particular property or process may be changed, leading to an unphysical mechanism of the process in the coarse-grained model, an example being the use of implicit solvation models [61] in which solvent degrees are omitted and their effect represented as a function of the solute coordinates only; (v) The reduction of entropy and energy in the system may lead to an unphysical balance between these two quantities in the coarse-grained model.

In short, the combined loss of usefulness on these five counts must be made up for by a much increased computational efficiency of the coarse-grained model compared to the fine-grained one. This trade-off between atomic (fine-grained) and supra-atomic (coarse-grained) levels of resolution is negative for biomolecules such as proteins, nucleotides and sugars because of their heterogeneous composition of atoms and polar versus non-polar moieties, which implies a rather large loss of specificity when coarse-graining from the atomic to the supra-atomic level of resolution in combination with a rather modest gain in computational efficiency. This trade-off is positive when coarse-graining solvent degrees of freedom from the atomic level to supra-molecular level of resolution [62, 63]. Simulating proteins in water, the use of a supra-molecular coarse-grained water model leads to an order of magnitude more efficient simulations while maintaining structural and energetic properties of the proteins [9, 64]. When coarse-graining models for lipids, the trade-off may also be positive because of their abundance in membranes.

Different supra-atomic coarse-grained models for lipids are available [65–67]. The models by Essex et al. [65] use ellipsoidal beads in conjunction with a soft sticky coarse-grained
water model, the models by Marrink et al. [66] use spherical beads in conjunction with a Lennard-Jones model for coarse-grained water, and the model by Jakobsson uses a Morse potential for the non-bonded potential energy. These models are not compatible with the atomic-level GROMOS force field which we would like to use in multi-graining simulations of atomistic-level proteins in membranes composed of supra-atomic lipids embedded in supra-molecular water. In order to develop a supra-atomic coarse-grained force field for lipids that can be made compatible with the atomic-level GROMOS force field for biomolecules, we investigated some structural and thermodynamic properties of the supra-atomic MARTINI force field for biomolecules [68] using a range of alkanes in the liquid phase [69], which served as representation of the properties of the aliphatic tails of lipids. Experimental properties such as density and heat of vaporisation were not well reproduced and bead-bead radial distribution functions appeared to be over-structured compared to the corresponding functions derived from atomic-level simulations of liquid alkanes. It was concluded that these discrepancies could be resolved by a reparametrisation of the supra-atomic force field for aliphatic chains [69]. The results of this reparametrisation are reported here.

The strategy for the development of the GROMOS force fields has been described in [55]: After having chosen a functional form for the various terms in the potential energy function, the force-field parameters have to be determined. For the bond-length and bond-angle terms this can be done using X-ray diffraction and spectroscopic data on small molecules. The parameters of the torsional-angle interaction term and partial atomic charges can be obtained from quantum-chemical calculations on small model compounds in the gas phase. However, since accurate quantum-chemical calculations for molecules in the liquid phase are not yet possible, the non-bonded interaction parameters, partial charges $q_i$ and van der Waals parameters $C_{12}(i,j)$ and $C_{6}(i,j)$ for atoms or beads $i$ and $j$, have to be further calibrated using thermodynamic data such as density, heat of vaporisation, excess free energy, and dielectric data measured for liquids of the small molecules used for parameter calibration.

This strategy that was applied to develop the atomic-level GROMOS force fields [32, 53] can also be applied to develop supra-atomic or supra-molecular force fields [62, 63]. However, in the case of supra-molecular models, some properties such as the excess free energy of the liquid, cannot be used for parametrisation and some technical issues due to the supra-molecular character of the coarse-grained beads emerge [62]. In the case of supra-atomic models, e.g. for lipids, the difficulty lies in the scarcity of structural, thermodynamic and dielectric experimental data on such molecules in the fluid phase. The surface area of bilayers of lipids in water and the C–H order parameters of CH$_2$-moieties in the aliphatic tails as derived from NMR experiments are often used to test the quality of membrane simulations and the force field used [70], but these data do not possess a high accuracy [70] and due to their low number compared to the number of degrees of freedom of a lipid molecule do not contribute data that can be used for force-field parameter calibration. For liquid alkanes there is thermodynamic experimental data which can be used in the parameter calibration. Since structural experimental data, apart from trans/gauche ratios
derived from NMR experiments [71], is lacking for these flexible molecules, one can resort to the use of structural data as generated in simulations using an atomic-level model and force field when calibrating parameters of a supra-molecular model and force field for the same compound. The caveat is here that the inaccuracies of the atomic-level force field may get propagated to the supra-atomic level one.

Based on these considerations, the strategy followed here to obtain a supra-atomic GROMOS force field for alkanes in the liquid phase is the following one.

1. Molecular dynamics (MD) simulations of 18 \( n \)-alkanes in the liquid phase at ambient temperature and pressure are carried out using the atomic-level of resolution GROMOS force field 45A3 [72] in order to obtain fine-grained configurational ensembles.

2. If these atomic-level configurational ensembles reproduce the experimental densities and heat of vaporisation of these 18 liquids, they can be used as calibration data for structural properties at the supra-atomic, coarse-grained level of modelling. To this end the atomic-level configurations are to be mapped onto supra-atomic configurations. For example, two, three, four or five atoms are mapped onto or represented by their centre of geometry or their centre of mass. Using such an atom-to-bead mapping the structural, energetic, dielectric, and dynamic properties of the atomic-level MD trajectories are represented in terms of the corresponding quantities for beads representing 2 to 5 atoms. In the present case of alkanes these quantities are e.g. distance-, angle- and torsional-angle distributions involving two to four beads, bead-bead radial distribution functions and non-bonded interaction energies.

3. Based on these atomic-level distributions of various quantities the form of the supra-atomic level model can be chosen: How many atoms are mapped onto one bead; Are beads in the middle of the aliphatic chain to be distinguished from those terminating a chain. For various reasons [54] the functional form of the supra-atomic potential energy function that depends on the coordinates of the beads is the same as that of the atomic-level GROMOS force field. A first estimate for supra-atomic force-field parameters can be obtained from this analysis of the atomic-level MD trajectories.

4. Molecular dynamics simulations of \( n \)-alkanes in the liquid phase at ambient temperature and pressure are carried out using the supra-atomic level model and estimated force-field parameters.

5. If these coarse-grained configurational ensembles do not reproduce the structural and thermodynamic target data, e.g. the density and the heat of vaporisation, a further calibration of the supra-atomic force-field parameters is required. For the \( n \)-alkanes, the \( C_{12} \) and \( C_6 \) van der Waals parameters had to be varied to reproduce the experimental densities and heats of vaporisation.

6. If the target properties are reproduced satisfactorily by one of the coarse-grained
Chapter 3. CG alkanes

models, this model and parameter set is further tested by the calculation of quantities that were not included in the set of target data.

In the current case of the development of a coarse-grained supra-atomic GROMOS force field for $n$-alkanes in the liquid phase, six bead types were defined, representing 2, 3 or 4 united atoms and distinguishing between beads within the chain having two bound neighbour beads, and beads terminating a chain having one bound neighbour bead. The use of a combination of different atom-to-bead mappings in one chain allows for a simple comparison of fine-grained and coarse-grained configurations for arbitrary chain lengths.

3.2 Methods

All fine-grained (FG) and coarse-grained (CG) molecular dynamics (MD) simulations as well as the analyses have been performed using the GROMOS software package for biomolecular simulations [11, 12, 73]. Some analysis methods concerning the calculations of CG properties calculated from mapped FG alkane configurations, such as the calculation of non-bonded bead-bead energies and the computation of radial distribution functions, needed some additional GROMOS++ code not included in the officially released GROMOS version [13].

3.2.1 Atomistic model

Interactions of the 18 FG alkanes (propane to eicosane) were modelled using the GROMOS 45A3 force-field parameter set [72].

This model consists of five atom types CH$_n$ with $n = 0, 1, \ldots, 4$, ready to be used for the simulation of aliphatic hydrocarbons in the condensed phase. All of them have zero partial charge and are non-polarisable. In the current work we only used CH$_2$ and CH$_3$ FG atoms.

A quartic bond-stretching potential energy term $V^b$ is used to calculate forces on covalently bound neighbour atoms,

$$V^b(b; K_b, b_0) = \frac{1}{4}K_b(b^2 - b_0^2)^2,$$

where $b = |\mathbf{r}_i - \mathbf{r}_j| = r_{ij}$ is the distance between atoms $i$ and $j$ and the two force-field parameters $K_b$ and $b_0$ indicate the force constant and the reference or ideal bond length. Bond lengths are generally kept rigid though. The potential energy term $V^{ba}$ for bond angles $\theta$ is defined as

$$V^{ba}(\theta; K_\theta, \theta_0) = \frac{1}{2}K_\theta(\cos(\theta) - \cos(\theta_0))^2$$
with the angle force constant $K_\theta$ and the reference bond angle $\theta_0$. GROMOS also provides the use of harmonic bond-stretching and harmonic bond-angle terms. However, we stick to the quartic and cosine-harmonic form, which is still the standard choice in GROMOS. In addition, there is a proper dihedral-angle torsional force-field term to mimic the rotational barriers along a bond,

$$V^{da}(\varphi; K_\varphi, \varphi_0, m) = K_\varphi (1 + \cos(\delta) \cos (m\varphi))$$  \hspace{1cm} (3.3)

with $\delta = 0, \pi$, $K_\varphi$ the force constant, $\varphi$ the angle of four neighbouring, covalently bound atoms and $m$ the multiplicity. The 54A3 force field [72] also offers an improper dihedral-angle potential energy term used to preserve the stereochemistry of branched alkanes.

The non-bonded van der Waals interaction between two particles $i$ and $j$ is modelled using a Lennard-Jones (12/6) functional form,

$$V^{vdW}_{ij}(r_{ij}) = \left( \frac{C_{12}(i,j)}{r_{ij}^{12}} - \frac{C_6(i,j)}{r_{ij}^6} \right) = 4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right] , \hspace{1cm} (3.4)$$

where the Lennard-Jones (LJ) parameters $C_{12}(i,j)$ and $C_6(i,j)$ for atom types $i$ and $j$ are obtained using the GROMOS combination rule

$$C_{12}(i,j) = \sqrt{C_{12}(i,i) C_{12}(j,j)} \quad \text{and} \quad C_6(i,j) = \sqrt{C_6(i,i) C_6(j,j)} . \hspace{1cm} (3.5)$$

Covalently bound first- and second-neighbour atoms are excluded from the non-bonded interaction since their behaviour is described by and would conflict with the bonded interaction terms. Third-neighbours do interact according to Eq. (3.4) but use scaled LJ interaction parameters, $CS_{12}(i,i)$ and $CS_6(i,i)$.

### 3.2.2 Atomistic simulations

If the initial configurations of the $n$-alkanes would not yield the experimentally derived averaged trans/gauche ratio, a rather long MD equilibration period would be required to obtain the proper trans/gauche ratio in the liquid phase. To reduce the equilibration time, the initial configurations of individual molecules were taken from the last $N_f$ molecule configurations 20 ps apart in a 100 ns stochastic dynamics (SD) simulation trajectory of a single alkane molecule at 298.15 K. $N_f$ corresponds to the number of molecules as indicated in Tab. 3.1 contained in the periodic computational box. They were randomly distributed in cubic boxes using the GROMOS++ program ran_box. The individual box-edge lengths were chosen between 7.9 nm (propane) and 6.7 nm (eicosane) to yield the experimental densities as indicated in Tab. 3.1. The molecular configurations were then energy minimised under periodic boundary conditions to relieve high energy contacts between individual molecules. Initial velocities were assigned from a Maxwell-Boltzmann distribution at 1 K and the temperature was continuously raised to 298.15 K during the first 500 ps of equilibration time at constant volume, followed by another 500 ps simulation
Table 3.1: Number of molecules and atoms in the FG simulations of \(n\)alkanes. Density \(\rho\) and heat of vaporisation \(\Delta H_{\text{vap}}\) for liquid alkanes at 298 K and 1 bar, unless indicated otherwise. Experimental values were taken from [74].

<table>
<thead>
<tr>
<th>alkane</th>
<th>number of molecules</th>
<th>number of atoms</th>
<th>(\rho / \text{kg m}^{-3})</th>
<th>(\Delta H_{\text{vap}} / \text{kJ mol}^{-1})</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>exp.</td>
<td>calc.</td>
<td>exp.</td>
<td>calc.</td>
</tr>
<tr>
<td>propane</td>
<td>3334</td>
<td>10002</td>
<td>493.0(^a)</td>
<td>495.1</td>
</tr>
<tr>
<td>butane</td>
<td>2500</td>
<td>10000</td>
<td>573.0(^a)</td>
<td>577.8</td>
</tr>
<tr>
<td>pentane</td>
<td>2000</td>
<td>10000</td>
<td>626.2(^b)</td>
<td>621.1</td>
</tr>
<tr>
<td>hexane</td>
<td>1667</td>
<td>10002</td>
<td>660.6</td>
<td>654.7</td>
</tr>
<tr>
<td>heptane</td>
<td>1429</td>
<td>10003</td>
<td>679.5</td>
<td>678.7</td>
</tr>
<tr>
<td>octane</td>
<td>1250</td>
<td>10000</td>
<td>698.6</td>
<td>698.0</td>
</tr>
<tr>
<td>nonane</td>
<td>1112</td>
<td>10008</td>
<td>719.2(^b)</td>
<td>712.6</td>
</tr>
<tr>
<td>decane</td>
<td>1000</td>
<td>10000</td>
<td>726.6</td>
<td>724.5</td>
</tr>
<tr>
<td>undecane</td>
<td>910</td>
<td>10010</td>
<td>740.2(^b)</td>
<td>736.4</td>
</tr>
<tr>
<td>dodecane</td>
<td>834</td>
<td>10008</td>
<td>749.5(^b)</td>
<td>743.2</td>
</tr>
<tr>
<td>tridecane</td>
<td>770</td>
<td>10010</td>
<td>756.4(^b)</td>
<td>750.4</td>
</tr>
<tr>
<td>tetradecane</td>
<td>715</td>
<td>10010</td>
<td>759.6(^b)</td>
<td>756.4</td>
</tr>
<tr>
<td>pentadecane</td>
<td>667</td>
<td>10005</td>
<td>768.5(^b)</td>
<td>762.0</td>
</tr>
<tr>
<td>hexadecane</td>
<td>626</td>
<td>10016</td>
<td>770.1</td>
<td>766.8</td>
</tr>
<tr>
<td>heptadecane</td>
<td>589</td>
<td>10013</td>
<td>778.0(^b)</td>
<td>771.4</td>
</tr>
<tr>
<td>octadecane</td>
<td>556</td>
<td>10008</td>
<td>776.8</td>
<td>774.9</td>
</tr>
<tr>
<td>nonadecane</td>
<td>527</td>
<td>10013</td>
<td>785.5(^b)</td>
<td>778.6</td>
</tr>
<tr>
<td>eicosane</td>
<td>500</td>
<td>10000</td>
<td>788.6(^b)</td>
<td>781.8</td>
</tr>
</tbody>
</table>

\(^a\) experimental density measured at a pressure > 1 bar
\(^b\) experimental density measured at a temperature of 293 K

at constant volume and constant temperature (298.15 K). Another 2 ns simulation at constant temperature and pressure (298 K, 1 atm) was appended to complete the equilibration of the different alkane boxes, before continuing the simulations under identical conditions for another 10 ns saving the atom configurations and energies every 5 ps for analysis.

The weak-coupling algorithm [75] was used for constant temperature and pressure simulations with corresponding coupling times of \(\tau_T = 0.1\) ps and \(\tau_p = 0.5\) ps, respectively, and an isothermal compressibility ranging from 0.0113 to 0.00142 (kJ mol\(^{-1}\) nm\(^{-3}\))\(^{-1}\) corresponding to experimental values. All bond lengths were kept constant using the SHAKE algorithm [14] with a relative geometric tolerance of 10\(^{-4}\), allowing for an integration time step of 2 fs when solving the equations of motion using the leap-frog algorithm [8]. Nonbonded (van der Waals) interactions were handled adopting twin-range cut-off radii: interactions within the short-range cut-off of 0.8 nm were calculated every time step from a pair list that was generated every five steps, when interactions in the range 0.8-1.4 nm
were computed. The centre of mass translation and rotation were removed every 2 ps to avoid a flying ice cube [76].

### 3.2.3 Analysis of FG configurations in terms of CG beads

Individual values for the bonded and non-bonded CG force-field parameters were obtained from an analysis of the atomic-level alkane simulation trajectories mapped onto CG beads. To this end, the atoms of the FG alkane molecules were mapped to CG bead atoms, each bead replacing 2 to 5 FG atoms. The position of the CG bead was set to the centre of mass of the replaced FG atoms,

\[
r_{CG} = \frac{1}{M} \sum_{i=1}^{N} m_i r_i,
\]

where \( N \) is the number of atoms to be replaced, \( M = \sum_{i=1}^{N} m_i \) the total mass of all replaced atoms, and \( r_i \) the position of atom \( i \). This mapping is not restricted to a fixed number of atoms per bead within one molecule chain. Bonded CG force-field parameters \( q_0 \) and \( K_q \), with \( q = b, \theta \), or \( \varphi \) in Eqs. (3.1) to (3.3) were obtained as follows. The distribution \( P_{mFG}(q_{CG}) \) of mapped \( q_{CG} \) values, i.e. bead-bead bond lengths, bond angles and torsional angles, is obtained from mapped FG simulation trajectories. The \( q_0 \) parameter for the bonded terms (Eqs. (3.1)–(3.3)) of the CG force field, \( q_{0CG} \), is taken to be the average \( \langle \ldots \rangle \) of the \( P_{mFG}(q_{CG}) \) distribution,

\[
q_{0CG} = \langle q_{CG} P_{mFG}(q_{CG}) \rangle.
\]

The \( K_q \) parameter for the bonded terms \( V_{q,CG} \) (Eqs. (3.1)–(3.3)) of the CG force field, \( K_{qCG} \), is determined from a least-squares fit of the distribution

\[
P_{CG}(q^{CG}) \propto \exp \left( -\frac{V_{q,CG}(q^{CG})}{k_B T} \right)
\]

with \( V_{q,CG}(q^{CG}) \) defined by Eqs. (3.1)–(3.3) and using \( q_{0CG} \) from Eq. (3.7) to the distribution \( P_{mFG}(q_{CG}) \) of mapped \( q_{CG} \) values as obtained from the FG simulation trajectories. The quantity \( q^{CG} \) is the bond distance, the bond angle or the torsional angle of 2 to 4 neighbouring beads, \( k_B \) the Boltzmann constant, \( T = 298.15 \) K the temperature and \( V_{q,CG} \) the corresponding potential energy function as described in Eqs. (3.1) to (3.3). Note that \( b_{0CG} \), the reference bond length at which the quartic bond stretching potential energy is zero (compare Eq. (3.1)), was chosen to correspond exactly to the average CG bond length as calculated from the mapped FG simulation trajectories.

The force-field parameters of the non-bonded LJ interaction for two beads \( I \) and \( J \), \( C_{12}(I,J) \) and \( C_6(I,J) \), were estimated from the mapped LJ interaction potential energy
according to
\[ V_{ij}^{vdW}(r_{ij}) = \left\langle \sum_{i=1}^{N} \sum_{j=1}^{N} V_{ij}^{vdW}(r_{ij}) \right\rangle, \tag{3.9} \]
where \( N \) is the number of atoms per bead, \( V_{ij}^{vdW}(r_{ij}) \) as described in Eq. (3.4), and angular brackets indicate the average over all inter-molecular bead-bead pairs.

### 3.2.4 Supra-atomic model

Since the CG model for linear alkanes is an intermediate step on the way towards a CG model for lipid bilayers to be used in biomolecular MD simulations, it is required to be compatible with existing and upcoming GROMOS force fields at the FG level for solutes as well as at the FG and CG levels for solvents. Therefore, the functional form of the bonded and non-bonded potential energy terms used in the FG model, Eqs. (3.1) to (3.5), are also chosen for the CG alkane model but applied to bonded and non-bonded interactions of beads instead of atoms. A CG bead was chosen to represent 2 to 4 FG atoms yielding 6 CG atom types: \( C_{2m}, C_{2e}, C_{3m}, C_{3e}, C_{4m} \) and \( C_{4e} \) where the indices e (end) and m (middle) indicate the position of the corresponding bead within the alkane chain.

### 3.2.5 Supra-atomic simulations

Initial configurations of the first CG alkane simulations, using the fitted bonded force-field parameters and LJ parameters directly calculated from the mapped FG alkane simulation trajectories considering inter-molecular bead interactions only, were generated by randomly distributing an alkane molecule in a cubic box. The total number of CG molecules was identical to the number of molecules in the FG simulations, compare Tab. 3.1, while the energy minimised configuration of an FG alkane molecule having the same number of atoms as the CG alkane number of beads served to obtain an initial structure, e.g. FG butane was serving as initial structure for CG dodecane with bead size 3. Box sizes were chosen to yield the experimental density of the corresponding alkane (Tab. 3.1). A steepest descent energy minimisation was performed applying periodic boundary conditions to avoid unfavourable inter-molecular bead-bead interactions.

Initial bead velocities were assigned from a Maxwell-Boltzmann distribution at 1 K and continuously raised to 298.15 K during the first 1 ns of the simulation at constant volume. Another 1 ns simulation at constant volume and temperature (298.15 K) was followed by a simulation at constant pressure and temperature (1 atm, 298.15 K) to complete the 3 ns of equilibration. The simulation was then continued for another 6 ns and the bead positions and velocities as well as the system energies saved every 5 ps for analysis.

Constant temperature and pressure were maintained using the weak-coupling algorithm \[ [75] \] with coupling times \( \tau_T = 0.1 \text{ps} \) and \( \tau_p = 0.5 \text{ps} \), respectively, and an isothermal compressibility of 4.575 (kJ mol\(^{-1}\) nm\(^{-3}\))\(^{-1}\) for all alkanes, corresponding to an averaged
value for biomolecular systems. In order to test the behaviour of the fitted, bonded force-field parameters, SHAKE was not applied in these first, initial simulations of the CG alkanes, and a time step of 0.5 ps was used when solving Newton’s equation of motion using the leap-frog algorithm [8]. Non-bonded interactions were handled using a single-range cut-off radius of 2 nm as it was applied in the GROMOS model of CG, polarisable water [62] and in mixed-grained simulations [9]. The centre of mass translation and rotation were removed every 2 ps to avoid a flying ice cube [76].

The simulations performed to parametrise the non-bonded LJ parameters were prolonging the 6 ns CG alkane simulations described above but bond lengths were held constant using the SHAKE algorithm [14] together with an integration time step of 2 fs. A much bigger time step would be possible for coarse-grained alkanes as shown in the literature [68]. However, we stuck to 2 fs since the CG model is meant for use in multi-grained simulations [9]. The simulation with the non-bonded parameters best reproducing the experimental density and heat of vaporisation was further prolonged for another 10 ns and bead configurations and system energies were saved every 5 ps for analysis.

3.2.6 Parametrisation strategy for the LJ interactions

The LJ interaction was parametrised by altering the initial $\varepsilon$ and $\sigma$ values of the LJ interaction potential energy, Eq. (3.4), which had been obtained from the mapped FG simulation trajectories, by simply scaling these values with factors $f_\varepsilon$ and $f_\sigma$, respectively, i.e. $\varepsilon_{\text{new}} = f_\varepsilon \cdot \varepsilon_{\text{init}}$.

The CX$_e$ CG particles were first parametrised using CG alkanes consisting of two of such particles only, i.e. CG butane for beads C2$_e$, hexane for beads C3$_e$, and octane for beads C4$_e$. Longer alkane chains were used to parametrise the CX$_m$ particles while keeping the $\varepsilon$ and $\sigma$ values of the end-beads constant.

Butane has a boiling temperature lower than room temperature, which made it less suitable to a proper parametrisation of C2 particles. CG alkanes consisting of beads of different bead sizes within one molecule were not used for the parametrisation of the non-bonded interactions.

3.2.7 Analysis

Radial distribution function

The radial distribution functions $g_{ee}$, $g_{em}$ and $g_{mm}$ were calculated from the mapped FG alkane simulation trajectories and from the CG simulation trajectories of the final alkane
model according to

\[ g_{\alpha\beta}(r) = \left\langle \frac{1}{4\pi \rho_{\beta} r^2 \Delta r (N_{\beta} - \delta_{\alpha\beta})} \sum_{\alpha=1}^{N_{\alpha} - \delta_{\alpha\beta}} \sum_{\beta=1}^{N_{\beta}} \delta(r_{\alpha\beta} - r) \right\rangle \]  

(3.10)

where \( N_i \) is the number of beads of type \( i = [\alpha, \beta] \), \( r_{\alpha\beta} \) is the inter-bead distance between two beads, \( \rho_{\beta} \) is the particle density of beads of type \( \beta \) and \( \Delta r \) the grid spacing. Angular brackets indicate the average over the individual configurations of the simulations trajectories.

The averages were computed for the last 2.5 ns of the FG alkane simulations and for the last 5 ns of the CG alkane simulations.

**Enthalpy of vaporisation**

The heat of vaporisation was calculated as the difference in inter-molecular potential energy between the gas phase and the liquid phase,

\[ \Delta H_{\text{vap}} = V_{\text{pot}}^{\text{intra}}(g) - (V_{\text{pot}}^{\text{intra}}(l) + V_{\text{pot}}^{\text{inter}}(l)) + RT \]

\[ = V_{\text{pot}}^{\text{intra}}(g) - V_{\text{pot}}^{\text{tot}}(l) + RT \]  

(3.11)

where \( R \) is the gas constant and \( T \) the temperature, assuming ideal conditions for the gas phase and thus no inter-molecular potential energy contribution \( V_{\text{inter}}(g) \). The intramolecular contribution in the gas phase was estimated by a stochastic dynamics (SD) simulation of one molecule.

**Self-diffusion coefficient \( D \)**

The self-diffusion coefficient \( D \) can be calculated according to the long-time limit of the mean-square displacement using the Einstein relation

\[ D = \lim_{t \to \infty} \frac{\langle (\mathbf{r}(\tau + t) - \mathbf{r}(\tau))^2 \rangle_{\tau, \text{molecules}}}{6t} \]  

(3.12)

where \( \mathbf{r} \) is the bead position at a given time \( t \). Angular brackets indicate an averaging over the time period \( \tau \) and over all molecules.
3.3 Results and discussion

3.3.1 Atomistic-level FG simulations

Tab. 3.1 lists the experimental densities and heats of vaporisation for the different alkanes as compiled in [72]. Comparison of the values obtained from the FG MD simulations to the experimental data shows a good agreement for the density as well as for the heat of vaporisation. The density did also agree for alkanes with a boiling point lower than room temperature or a melting point higher than room temperature. Since all alkane simulations were performed at 298.15 K and 1 atm, the FG model is not capable of mimicking phase transitions at the appropriate temperature. Yet the simulation data of [72] obtained for much smaller systems and shorter simulation periods could be confirmed and the FG ensembles may serve as basic data for the CG force-field calibration complementing experimental data.

3.3.2 Bonded interactions of the CG alkane model

All bonded parameters of the CG force field were estimated from the distribution of the corresponding property obtained from the analysis of the FG-to-CG mapped FG simulation trajectories. Fig. 3.1 shows the result together with the bond-length, bond-angle and torsional-angle distributions as obtained from a CG alkane simulation without bond-length constraints, i.e. not applying SHAKE [14] to investigate the distribution of the bead-bead distance. The distributions of the bond distance and the bond angles obtained from the mapped FG simulation trajectories show, especially for small bead sizes, multiple peaks. These occur due to the presence of $\text{trans}/\text{gauche}$-conformations in the underlying atomistic model simulation. For example, the distance of two mapped beads of bead size 2 is much smaller for the $\text{cis}$-configuration than for the $\text{trans}$-configuration of FG butane. For larger bead sizes this effect is less visible because the corresponding FG alkane fragment is longer and the number of combinations of $\text{trans}/\text{gauche}$-conformations with very similar bead-bead distances or angles increases, resulting in a broader and flatter distribution. It is clear that a potential energy term for a CG bond or bond-angle as described in Eqs. (3.1) and (3.2) cannot reproduce in detail the FG distribution, but the FG and CG distributions nevertheless show overlapping distributions. The FG bond-angle distributions could have been better reproduced when using an ideal bond angle $\theta_0$ of 180°, as used in the MARTINI model [68]. We decided to use a smaller ideal bond angle around the maxima of the CG distribution (140-150°) which avoids linear atom configurations within the alkane chains and allows for applying a torsional-angle potential energy term, Eq. (3.3), which has a singularity for a linear configuration of three of its constituting atoms. The maxima of the bond-angle distributions obtained from the CG simulations are therefore slightly shifted to smaller values compared to the mapped distributions. The potential energy term of the torsional angles should not be flat, which becomes clear from the third column of Fig. 3.1, but is neglected in other models [66, 69, 77]. The probabilities for small torsional angles is
Figure 3.1: Bond-length, bond-angle, and torsional-angle distributions of the coarse-grained alkanes as calculated from the mapped FG MD simulation trajectories (dashed lines) and CG MD simulation trajectories (solid lines). Colours indicate the bead size: 2 (black), 3 (red), 4 (green), and 5 (blue).
higher than for angles around 180°, which makes it impossible to mimic this effect by an adjusted, intra-molecular 1-4 bead-bead non-bonded Lennard-Jones (LJ) interaction. The computed torsional angle-distributions from the CG simulations are in good agreement with torsion-angle distributions from the mapped FG simulations except for eicosane and bead size 5.

The parameters of bonded interactions for CG alkanes are listed in Tab. 3.2. The force constants of the torsional angles in CG alkanes involving beads of size 2 and 3 are very similar to the ones involving beads of size 2 only. Therefore, the torsion-angle definition of type 2–2–2–2 was used in all mixed-bead CG simulations except those containing bead size 4.

Table 3.2: Bonded force-field parameters for coarse-grained \( n \)-alkanes. Note that there is no distinction between middle particles (\( C_{2m}/C_{3m}/C_{4m} \)) and end particles (\( C_{2e}/C_{3e}/C_{4e} \)). The parameters are defined in Eqs. (3.1) to (3.3).

<table>
<thead>
<tr>
<th>bond stretching</th>
<th>( b_0 / \text{nm} )</th>
<th>( K_b / \text{kJ mol}^{-1} \text{nm}^{-4} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2 – C2</td>
<td>0.245</td>
<td>96655</td>
</tr>
<tr>
<td>C2 – C3</td>
<td>0.299</td>
<td>35753</td>
</tr>
<tr>
<td>C3 – C3</td>
<td>0.353</td>
<td>19730</td>
</tr>
<tr>
<td>C3 – C4</td>
<td>0.405</td>
<td>11604</td>
</tr>
<tr>
<td>C4 – C4</td>
<td>0.457</td>
<td>6924</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>bond angle</th>
<th>( \theta_0 / ^\circ )</th>
<th>( K_\theta / \text{kJ mol}^{-1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2 – C2 – C2</td>
<td>150.1</td>
<td>23.3</td>
</tr>
<tr>
<td>C2 – C3 – C2</td>
<td>147.1</td>
<td>91.1</td>
</tr>
<tr>
<td>C2 – C3 – C3</td>
<td>147.9</td>
<td>58.9</td>
</tr>
<tr>
<td>C3 – C2 – C3</td>
<td>147.2</td>
<td>27.7</td>
</tr>
<tr>
<td>C3 – C3 – C3</td>
<td>146.4</td>
<td>56.6</td>
</tr>
<tr>
<td>C3 – C4 – C3</td>
<td>146.1</td>
<td>51.3</td>
</tr>
<tr>
<td>C4 – C3 – C4</td>
<td>143.7</td>
<td>47.8</td>
</tr>
<tr>
<td>C4 – C4 – C4</td>
<td>143.8</td>
<td>48.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>torsional dihedral</th>
<th>( \cos \delta )</th>
<th>( m )</th>
<th>( K_\phi / \text{kJ mol}^{-1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2 – C2 – C2 – C2</td>
<td>-1.0</td>
<td>1</td>
<td>0.68</td>
</tr>
<tr>
<td>C2 – C3 – C3 – C2</td>
<td>-1.0</td>
<td>1</td>
<td>0.67</td>
</tr>
<tr>
<td>C2 – C3 – C3 – C3</td>
<td>-1.0</td>
<td>1</td>
<td>0.69</td>
</tr>
<tr>
<td>C3 – C2 – C3 – C3</td>
<td>-1.0</td>
<td>1</td>
<td>0.68</td>
</tr>
<tr>
<td>C3 – C3 – C3 – C3</td>
<td>-1.0</td>
<td>1</td>
<td>0.74</td>
</tr>
<tr>
<td>C4 – C4 – C4 – C4</td>
<td>-1.0</td>
<td>1</td>
<td>1.20</td>
</tr>
</tbody>
</table>
3.3.3 Lennard-Jones interactions of the CG alkane model

The calculation of the $C_{12}(i, j)$ and $C_{6}(i, j)$ parameters of the LJ potential energy, as well as the third-neighbour $CS_{12}(i, j)$ and $CS_{6}(i, j)$ parameters from the mapped inter-bead LJ potential energy distribution show the need to distinguish the CG particles depending on their position in the alkane chain, especially for larger bead sizes, see Tab. 3.6 in the supplementary material. Therefore, we defined six types of CG beads, namely $C_{2e}$, $C_{2m}$, $C_{3e}$, $C_{3m}$, $C_{4e}$, and $C_{4m}$, where $CX$ indicates the mapping of $X$ FG carbon united atoms to one bead, and the indices $e$ (end) and $m$ (middle, i.e. not terminal) denote the position of the bead within an alkane chain.

A first CG alkane simulation showed bad agreement of the density and the heat of vaporisation for the different alkanes (Tab. 3.7). Thus, the LJ parameters obtained from the mapped bead-bead distance distributions needed to be refined with respect to reproduction of the experimental density and heat of vaporisation. Results of such a refinement for CG hexane are given in Tabs. 3.8 and Tab. 3.9 in the supplementary material which illustrates the behaviour of the density and the heat of vaporisation, respectively, while altering the LJ $\varepsilon$ and $\sigma$ parameters.

This refinement parametrisation yielded scaling factors $f_\varepsilon$ of sizes $1.0740$, $1.0513$ and $0.9800$ for bead sizes 2, 3, and 4 to modify the initial $\varepsilon$, and factors $f_\sigma$ of sizes $0.9964$, $1.0300$, and $1.0900$ for $\sigma$. These factors seem to be small but yet induce drastic changes in the $C_{12}$ and $C_{6}$ LJ parameters due to the powers of 12 and 6 in Eq. (3.4). The final LJ parameters are listed in Tab. 3.3 and differ in some cases by more than a factor of two from those in Tab. 3.6. The corresponding LJ potential energies as a function of $r$ are shown in Fig. 3.2, which also shows the potential energy of the mapped CG alkanes based on the FG simulation trajectories and the LJ potential energy for the FG CH$_2$ and CH$_3$ particles. The single-bead-size CG alkanes, i.e. coarse-grained by beads all of the same bead size, show a good agreement with experimental densities and $\Delta H_{\text{vap}}$ values for all chain lengths (Tab. 3.4). Note that not all possible single-bead-size CG alkanes with bead size 2 are listed in Tab. 3.4: octane, tetradecane and hexadecane are missing. This is due to the procedure SHAKE [14], which has difficulty imposing bond-length constraints in linear configurations. For these alkanes SHAKE could not set the bead-bead bond lengths to the desired values, probably due to bond angles close to 180°. The ideal bond angle for three C2 beads ($\theta_0 = 150.1^\circ$) is somewhat bigger than for the C3 and C4 beads and the bond-angle force constant smaller than for bond angles involving bigger beads, see Tab. 3.2. Therefore, a linear configuration of three beads is more likely to happen for alkanes coarse-grained with beads of size 2 than for bigger beads.

The multi-bead-size alkanes, which were not used in the non-bonded force-field parametrisation, show a good agreement for the density and heat of vaporisation compared to experiment (Tab. 3.4), indicating the standard GROMOS combination rules for the Lennard-Jones interactions (Eq. (3.5)) as used in FG force fields, can be used in the CG alkane model as well.
Table 3.3: Coarse-grained particle types and the corresponding Lennard-Jones interaction parameters for use in Eq. (3.4).

<table>
<thead>
<tr>
<th>particle name</th>
<th>bead size</th>
<th>fine-grained equivalent</th>
<th>mass [u]</th>
<th>Lennard-Jones parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(C_{12}) ([10^{-4} \text{kJ mol}^{-1} \text{nm}^{12}])</td>
</tr>
<tr>
<td>C2_m</td>
<td>2</td>
<td>2 CH₂</td>
<td>28.054</td>
<td>2.57086</td>
</tr>
<tr>
<td>C2_e</td>
<td>2</td>
<td>1 CH₃, 1 CH₂</td>
<td>29.062</td>
<td>2.95441</td>
</tr>
<tr>
<td>C3_m</td>
<td>3</td>
<td>3 CH₂</td>
<td>42.081</td>
<td>7.24940</td>
</tr>
<tr>
<td>C3_e</td>
<td>3</td>
<td>1 CH₃, 2 CH₂</td>
<td>43.089</td>
<td>10.65533</td>
</tr>
<tr>
<td>C4_m</td>
<td>4</td>
<td>4 CH₂</td>
<td>56.108</td>
<td>17.23221</td>
</tr>
<tr>
<td>C4_e</td>
<td>4</td>
<td>1 CH₃, 3 CH₂</td>
<td>57.116</td>
<td>31.82359</td>
</tr>
</tbody>
</table>
Chapter 3. CG alkanes

Figure 3.2: Lennard-Jones potential energy $V(r)$ dependent on the inter-particle distance $r$ for (i) thick lines: potential energy of the parametrised CG model for the $\text{CX}_e$ bead particles (solid line) and $\text{CX}_m$ bead particles (dashed line), as well as the CG potential energy calculated from the mapped FG simulation trajectories (thin lines). Colours indicate the bead size: 2 (black), 3 (red) and 4 (green). The Lennard-Jones potential energy of the FG $\text{CH}_2$ (dashed) and $\text{CH}_3$ (solid) particles are shown in brown.

3.3.4 Structural analysis of the CG alkanes

The radial distribution functions (RDFs) from end-bead to end-bead, $g_{ee}(r)$, end-bead to middle-bead, $g_{em}(r)$, and middle-bead to middle-bead, $g_{mm}(r)$, of CG dodecane are shown in Fig. 3.3. They are representative for all single-bead-size CG alkanes which look similar for different alkane chain lengths (see Fig. 3.5 of the supplementary material). All three different RDFs calculated from the mapped FG simulation trajectories of FG dodecane show about the same level of structure. The first peaks, as expected, are shifted to longer bead-bead distances. These mapped RDFs are to be compared to the RDFs obtained from the single-bead-size CG dodecane simulations of bead sizes 2 to 4. For bead size 2, the three RDFs are similar to the corresponding mapped RDFs, while $g_{em}(r)$ and $g_{mm}(r)$, both involving middle-beads, show a small bump around 0.7 nm. $g_{ee}(r)$ of bead size 3 has a too high and too narrow first peak but looks fine for longer distances, while $g_{em}(r)$ and $g_{mm}(r)$ have a small bump around 0.8 nm and a too high first peak, although not as bad as the one in $g_{ee}(r)$. The RDFs from bead size 4 are too structured, as already observed.
3.3. Results and discussion

Figure 3.3: End-end (ee) bead, end-middle (em) bead and middle-middle (mm) bead radial distribution function of CG dodecane. The functions are calculated from the mapped FG simulation trajectories (dashed) as well as from the CG simulation trajectories (solid). Colours indicate the bead size: 2 (black), 3 (red), and 4 (green).
### Table 3.4: Density $\rho$ and heat of vaporisation $\Delta H_{\text{vap}}$ of single-bead-size (top) and multi-bead-size (bottom) coarse-grained alkanes of bead sizes 2, 3 and 4. Experimental data are taken from [74] and measured at 298 K and 1 bar, unless specified otherwise.

<table>
<thead>
<tr>
<th>alkane</th>
<th>bead sizes</th>
<th>density / kg m$^{-3}$</th>
<th>$H_{\text{vap}}$ / kJ mol$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>exp.</td>
<td>calc.</td>
</tr>
<tr>
<td>single-bead-size CG alkanes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hexane</td>
<td>2–2–2</td>
<td>660.6</td>
<td>656.7</td>
</tr>
<tr>
<td></td>
<td>3–3</td>
<td>660.6</td>
<td>661.4</td>
</tr>
<tr>
<td>octane</td>
<td>4–4</td>
<td>698.6</td>
<td>697.9</td>
</tr>
<tr>
<td>nonane</td>
<td>3–3–3</td>
<td>719.2</td>
<td>718.7</td>
</tr>
<tr>
<td>dodecane</td>
<td>2–2–2–2–2–2–2</td>
<td>749.5</td>
<td>749.3</td>
</tr>
<tr>
<td></td>
<td>3–3–3–3</td>
<td>749.5</td>
<td>749.6</td>
</tr>
<tr>
<td></td>
<td>4–4–4</td>
<td>749.5</td>
<td>747.2</td>
</tr>
<tr>
<td>pentadecane</td>
<td>3–3–3–3–3</td>
<td>768.5</td>
<td>768.8</td>
</tr>
<tr>
<td>hexadecane</td>
<td>4–4–4–4</td>
<td>770.1</td>
<td>773.6</td>
</tr>
<tr>
<td>mixed-bead-size CG alkanes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>heptane</td>
<td>2–3–2</td>
<td>679.5</td>
<td>676.8</td>
</tr>
<tr>
<td>octane</td>
<td>3–2–3</td>
<td>698.6</td>
<td>706.1</td>
</tr>
<tr>
<td>decane</td>
<td>2–3–3–2</td>
<td>726.6</td>
<td>724.9</td>
</tr>
<tr>
<td>undecane</td>
<td>4–3–4</td>
<td>740.2</td>
<td>737.5</td>
</tr>
<tr>
<td>tridecane</td>
<td>2–3–3–3–2</td>
<td>756.4</td>
<td>752.3</td>
</tr>
<tr>
<td>tetradecane</td>
<td>3–3–2–3–3</td>
<td>759.6</td>
<td>763.0</td>
</tr>
<tr>
<td>hexadecane</td>
<td>2–3–3–3–3–2</td>
<td>770.1</td>
<td>769.7</td>
</tr>
<tr>
<td>heptadecane</td>
<td>3–3–3–3–3–3</td>
<td>778.0</td>
<td>777.7</td>
</tr>
</tbody>
</table>

* experimental density measured at a temperature of 293 K

in an earlier study of a CG alkane model with bead size 4 [69]. The narrow and high first peaks in all CG RDFs of bead sizes of 3 or larger may be a result of the deep LJ potential energy function (Fig. 3.2), which is needed to reproduce the experimental density and heat of vaporisation using the CG model, and could not be resolved as stated earlier [69]. Another possibility to correct for the structural behaviour of the CG alkanes would be to choose another interaction function than the (12/6) LJ potential energy. However, a coarse-grained model using a Morse potential energy function shows very similar structural behaviour as our model including the bump around 0.8 nm for bead size 3 and generates too much structure for bead size 4 [67].

The intra-molecular end-to-end distributions of the alkanes calculated from the CG alkane simulations of nonane, dodecane and pentane are shown in Fig. 3.4 and compared
to the corresponding mapped end-to-end distance distributions computed from the FG alkane simulation trajectories. Except for bead size 2, which we do not intend to use because of its limited computational gain, the distributions show a maximum at around

**Figure 3.4:** Intra-molecular end-to-end distance distributions for CG nonane, CG dodecane and CG pentadecane as calculated from the mapped FG alkane simulation trajectories (dashed) and from the CG alkane simulations (solid). Black: bead size 2; red: bead size 3; green: bead size 4.
Chapter 3. CG alkanes

the same positions as the mapped distributions but with a higher probability. The short end-to-end distances are better reproduced than the long-distance end-to-end distances. This is probably the price paid for using an ideal bond angle smaller than 180° in the bond-angle potential energy term as mentioned before.

3.3.5 Testing of the CG model

In order to further test the CG model excess free energies and self-diffusion coefficients were calculated. The results are shown in Tab. 3.5. The self diffusion constant is clearly dependent on the bead size, as demonstrated by the example of dodecane. The larger the bead size, the smaller the diffusion constant. The diffusion coefficient of the CG alkanes is generally smaller than the one of the FG alkanes. A comparison to available experimental data [79] for hexane to decane shows a better agreement for the CG alkanes than for the FG alkanes. The excess free energies calculated from the CG alkane simulation trajectories show good agreement to values obtained from experiment, much better than for a model described and tested earlier [69].

<table>
<thead>
<tr>
<th>alkane</th>
<th>bead sizes</th>
<th>( \Delta F_{\text{exc}} / \text{kJ mol}^{-1} )</th>
<th>( 10^5 D / \text{cm}^2 \text{s}^{-1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>pentane</td>
<td></td>
<td>14.23</td>
<td>\text{–}</td>
</tr>
<tr>
<td>hexane</td>
<td>3–3</td>
<td>16.96</td>
<td>16.59</td>
</tr>
<tr>
<td>heptane</td>
<td>2–2–2</td>
<td>19.64</td>
<td>\text{–}</td>
</tr>
<tr>
<td>octane</td>
<td>4–4</td>
<td>22.29</td>
<td>23.39</td>
</tr>
<tr>
<td>nonane</td>
<td>3–3–3</td>
<td>25.02</td>
<td>25.61</td>
</tr>
<tr>
<td>decane</td>
<td></td>
<td>27.74</td>
<td>\text{–}</td>
</tr>
<tr>
<td>undecane</td>
<td></td>
<td>30.54</td>
<td>\text{–}</td>
</tr>
<tr>
<td>dodecane</td>
<td>2–2–2–2</td>
<td>33.32</td>
<td>33.03</td>
</tr>
<tr>
<td></td>
<td>3–3–3</td>
<td>33.32</td>
<td>34.84</td>
</tr>
<tr>
<td></td>
<td>4–4–4</td>
<td>33.32</td>
<td>35.67</td>
</tr>
<tr>
<td>tridecane</td>
<td></td>
<td>36.30</td>
<td>\text{–}</td>
</tr>
<tr>
<td>tetradecane</td>
<td></td>
<td>39.23</td>
<td>\text{–}</td>
</tr>
<tr>
<td>pentadecane</td>
<td>3–3–3–3</td>
<td>42.27</td>
<td>44.14</td>
</tr>
<tr>
<td>hexadecane</td>
<td>4–4–4–4</td>
<td>45.51</td>
<td>48.16</td>
</tr>
<tr>
<td>heptadecane</td>
<td></td>
<td>\text{–}</td>
<td>\text{–}</td>
</tr>
</tbody>
</table>
3.4 Conclusion

A supra-atomic coarse-grained (CG) model for CG alkanes in the liquid phase has been parametrised based on experimental thermodynamic data and configurational ensembles obtained from atomic-level fine-grained (FG) simulations using the GROMOS biomolecular force field. It consists of 6 types of coarse-grained beads or particles representing 2 to 4 fine-grained atoms and it distinguishes between beads at the end or in the middle of an alkane chain.

Bonded interaction terms of the form of the fine-grained GROMOS 45A3 force field [72] have been used, and the force-field parameters, i.e. force constants and ideal bond lengths, bond angles, and torsional angles, were obtained from an analysis of the mapped distributions of the corresponding property as calculated from FG alkane simulation trajectories. In contrast to other CG models [66, 68] our model uses an ideal bond angle $\theta_0$ smaller than 180°, and contains a torsional dihedral-angle potential energy term.

Initial Lennard-Jones nonbonded interaction parameters were obtained from an analysis of FG simulations in terms of mapped bead-bead distances and energies. Since these parameters could not reproduce experimental densities and heats of vaporisation, they were refined with respect to reproduction of these quantities. The refined model shows good agreement for the density and heat of vaporisation compared to experimental data, independent of the chain length of the alkane.

Use of large bead sizes in a CG model would result in a larger speed-up of the simulations compared to FG models. However, the liquid becomes more structured with increasing bead size as observed from the bead-bead radial distribution functions calculated from the CG alkane simulations. On the other hand, some CG alkanes of bead size 2 could not be successfully simulated when using SHAKE to maintain fixed bead-bead bond lengths due to the occurrence of almost linear configurations of three neighbouring beads. For these reasons, beads of size 3 are recommended. Beads of size 2 and 4 can be used in combination with beads of size 3 allowing for CG alkane simulations of arbitrary chain length. Although not used in the parametrisation of the CG model, these multi-bead-size CG alkanes show good agreement between simulation and experiment with respect to the density and heat of vaporisation.

The proposed CG model was further tested by calculating the excess free energies and the self-diffusion coefficients of alkanes. The experimental values for these quantities were rather well reproduced, for the diffusion coefficient even better by the CG model than by the FG model.

The proposed CG model offers a solid basis from where a CG model for lipids can be developed.
### Supplementary material

#### Figure 3.5:
End-end (ee) bead, end-middle (em) bead and middle-middle (mm) bead radial distribution functions of the indicated CG alkanes. The functions are calculated from the mapped FG simulation trajectories (dashed lines) as well as from the CG simulation trajectories (solid lines). Colours indicate the bead size: 2 (black), 3 (red), and 4 (green).

<table>
<thead>
<tr>
<th></th>
<th>ee</th>
<th>em</th>
<th>mm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hexane</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>octane</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nonane</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>decane</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dodecane</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tetradecane</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pentadecane</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hexadecane</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**r/nm**
Table 3.6: Coarse-grained particle types and the corresponding Lennard-Jones interaction parameters as calculated from the mapping of the FG alkane simulation trajectories. The parameters are defined in Eq. (3.4).

<table>
<thead>
<tr>
<th>particle bead size fine-grained equivalent name</th>
<th>mass [u]</th>
<th>C_{12} \times 10^{-4} \text{kJ mol}^{-1} \text{nm}^{-12}</th>
<th>C_{6} \times 10^{-2} \text{kJ mol}^{-1} \text{nm}^{-6}</th>
<th>C_{12} \times 10^{-4} \text{kJ mol}^{-1} \text{nm}^{-12}</th>
<th>C_{6} \times 10^{-2} \text{kJ mol}^{-1} \text{nm}^{-6}</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_{2m}</td>
<td>2 CH_{2}</td>
<td>28.054</td>
<td>8.487855</td>
<td>2.49959</td>
<td>2.06721</td>
</tr>
<tr>
<td>C_{2e}</td>
<td>2 1 CH_{3}, 1 CH_{2}</td>
<td>29.062</td>
<td>6.08789</td>
<td>2.47332</td>
<td>2.38763</td>
</tr>
<tr>
<td>C_{3m}</td>
<td>3 CH_{2}</td>
<td>42.081</td>
<td>8.343989</td>
<td>3.890993</td>
<td>4.41650</td>
</tr>
<tr>
<td>C_{3e}</td>
<td>3 1 CH_{3}, 2 CH_{2}</td>
<td>43.089</td>
<td>18.36070</td>
<td>6.103609</td>
<td>5.61059</td>
</tr>
<tr>
<td>C_{4m}</td>
<td>4 CH_{2}</td>
<td>56.108</td>
<td>20.03889</td>
<td>8.49392</td>
<td>7.14039</td>
</tr>
<tr>
<td>C_{4e}</td>
<td>4 1 CH_{3}, 3 CH_{2}</td>
<td>57.116</td>
<td>20.30720</td>
<td>8.34992</td>
<td>7.14039</td>
</tr>
</tbody>
</table>
Table 3.7: Density $\rho$ and heat of vaporisation $\Delta H_{\text{vap}}$ obtained from experiment [74] and from the first alkane simulations using the Lennard-Jones interaction parameters directly calculated from the mapped FG alkane bead-bead distances.

<table>
<thead>
<tr>
<th>alkane</th>
<th>bead size</th>
<th>$\rho$ / kg m$^{-3}$</th>
<th>$H_{\text{vap}}$ / kJ mol$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>exp.</td>
<td>calc.</td>
</tr>
<tr>
<td>butane</td>
<td>2</td>
<td>573.00</td>
<td>587.98</td>
</tr>
<tr>
<td>hexane</td>
<td>2</td>
<td>660.60</td>
<td>678.58</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>660.60</td>
<td>819.54</td>
</tr>
<tr>
<td>octane</td>
<td>2</td>
<td>698.60</td>
<td>727.68</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>698.60</td>
<td>1135.38</td>
</tr>
<tr>
<td>nonane</td>
<td>3</td>
<td>719.20</td>
<td>862.26</td>
</tr>
<tr>
<td>decane</td>
<td>2</td>
<td>726.60</td>
<td>729.86</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>726.60</td>
<td>1640.34</td>
</tr>
<tr>
<td>dodecane</td>
<td>2</td>
<td>749.50</td>
<td>751.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>749.50</td>
<td>847.06</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>749.50</td>
<td>1129.80</td>
</tr>
<tr>
<td>pentadecane</td>
<td>3</td>
<td>768.50</td>
<td>859.03</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>768.50</td>
<td>1504.60</td>
</tr>
<tr>
<td>hexadecane</td>
<td>2</td>
<td>770.10</td>
<td>778.12</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>770.10</td>
<td>1124.32</td>
</tr>
</tbody>
</table>
Table 3.8: Density $\rho$ of CG hexane (bead size 2). $\varepsilon$ and $\sigma$ are indicated by a scaling factor $f_\sigma$ or $f_\varepsilon$, i.e. $f_\varepsilon = 1.00$ indicates an $\varepsilon$ having the same size as the one obtained from fitting to the LJ energy from the mapped FG simulation. The experimental density of hexane is $\rho = 660.6 \text{ kg m}^{-3}$. The root-mean-square deviations are listed in brackets. Different colors indicate the deviation from the experimental density: $< 0.1\%$, $< 0.5\%$, $< 1\%$, and $> 2\%$.

<table>
<thead>
<tr>
<th>prefactor $f_\varepsilon$ of $\varepsilon$</th>
<th>0.970</th>
<th>0.975</th>
<th>0.980</th>
<th>0.985</th>
<th>0.990</th>
<th>0.995</th>
<th>1.000</th>
<th>1.005</th>
<th>1.010</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.060</td>
<td>661.6</td>
<td>660.6</td>
<td>659.4</td>
<td>658.4</td>
<td>657.1</td>
<td>656.1</td>
<td>655.0</td>
<td>653.7</td>
<td>652.5</td>
</tr>
<tr>
<td>1.065</td>
<td>662.0</td>
<td>661.0</td>
<td>659.9</td>
<td>658.7</td>
<td>657.8</td>
<td>656.6</td>
<td>655.3</td>
<td>654.0</td>
<td>653.0</td>
</tr>
<tr>
<td>1.070</td>
<td>662.2</td>
<td>661.3</td>
<td>660.2</td>
<td>659.1</td>
<td>658.0</td>
<td>656.8</td>
<td>655.7</td>
<td>654.3</td>
<td>653.2</td>
</tr>
<tr>
<td>1.075</td>
<td>662.7</td>
<td>661.5</td>
<td>660.6</td>
<td>659.4</td>
<td>658.3</td>
<td>657.1</td>
<td>655.9</td>
<td>654.8</td>
<td>653.5</td>
</tr>
<tr>
<td>1.080</td>
<td>662.9</td>
<td>661.9</td>
<td>660.8</td>
<td>659.7</td>
<td>658.6</td>
<td>657.3</td>
<td>656.2</td>
<td>655.2</td>
<td>653.9</td>
</tr>
<tr>
<td>1.085</td>
<td>663.2</td>
<td>662.3</td>
<td>661.3</td>
<td>660.1</td>
<td>658.9</td>
<td>657.7</td>
<td>656.7</td>
<td>655.4</td>
<td>654.1</td>
</tr>
<tr>
<td>1.090</td>
<td>663.6</td>
<td>662.6</td>
<td>661.5</td>
<td>660.5</td>
<td>659.3</td>
<td>658.2</td>
<td>656.9</td>
<td>655.7</td>
<td>654.5</td>
</tr>
<tr>
<td>1.095</td>
<td>663.8</td>
<td>662.8</td>
<td>661.8</td>
<td>660.8</td>
<td>659.6</td>
<td>658.5</td>
<td>657.4</td>
<td>656.1</td>
<td>654.8</td>
</tr>
<tr>
<td>1.100</td>
<td>664.3</td>
<td>663.3</td>
<td>662.3</td>
<td>661.1</td>
<td>660.0</td>
<td>658.8</td>
<td>657.7</td>
<td>656.4</td>
<td>655.2</td>
</tr>
<tr>
<td>1.105</td>
<td>664.8</td>
<td>663.6</td>
<td>662.5</td>
<td>661.5</td>
<td>660.3</td>
<td>659.3</td>
<td>657.9</td>
<td>656.8</td>
<td>655.6</td>
</tr>
<tr>
<td>1.110</td>
<td>665.0</td>
<td>664.0</td>
<td>662.9</td>
<td>661.8</td>
<td>660.5</td>
<td>659.6</td>
<td>658.3</td>
<td>657.1</td>
<td>655.9</td>
</tr>
<tr>
<td>1.115</td>
<td>665.3</td>
<td>664.5</td>
<td>663.3</td>
<td>662.2</td>
<td>661.0</td>
<td>659.9</td>
<td>658.7</td>
<td>657.5</td>
<td>656.1</td>
</tr>
<tr>
<td>1.120</td>
<td>665.7</td>
<td>664.7</td>
<td>663.6</td>
<td>662.6</td>
<td>661.3</td>
<td>660.1</td>
<td>659.0</td>
<td>657.8</td>
<td>656.5</td>
</tr>
<tr>
<td>1.125</td>
<td>666.2</td>
<td>664.9</td>
<td>663.9</td>
<td>662.8</td>
<td>661.7</td>
<td>660.4</td>
<td>659.4</td>
<td>658.0</td>
<td>657.0</td>
</tr>
<tr>
<td>1.130</td>
<td>666.3</td>
<td>665.3</td>
<td>664.4</td>
<td>663.2</td>
<td>662.0</td>
<td>660.9</td>
<td>659.5</td>
<td>658.4</td>
<td>657.1</td>
</tr>
</tbody>
</table>
Table 3.9: Heat of vaporisation $\Delta H_{\text{vap}}$ of CG hexane (bead size 2).

$\varepsilon$ and $\sigma$ are indicated by a scaling factor $f_{\varepsilon}$ or $f_{\sigma}$, i.e. $f_{\varepsilon} = 1.00$ indicates an $\varepsilon$ having the same size as the one obtained from fitting to the LJ energy from the mapped FG simulation.

The experimental heat of vaporisation of hexane is $\Delta H_{\text{vap}} = 31.87 \text{ kJ mol}^{-1}$. The root-mean-square deviations are listed in brackets.

**Different colors indicate the deviation from the experimental heat of vaporisation:**

- $<0.1\%$,
- $<0.5\%$,
- $<1\%$,
- $>2\%$.

<table>
<thead>
<tr>
<th>$f_{\varepsilon}$</th>
<th>$\Delta H_{\text{vap}}$ (kJ mol$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>31.87</td>
</tr>
</tbody>
</table>

## Table 3.9: Heat of Vaporisation of CG Hexane (bead size 2)

<table>
<thead>
<tr>
<th>$f_{\sigma}$</th>
<th>$\Delta H_{\text{vap}}$ (kJ mol$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>31.87</td>
</tr>
</tbody>
</table>
Chapter 4

The GROMOS++ software for the analysis of biomolecular simulation trajectories

GROMOS++ is a set of C++ programs for pre- and post-processing of molecular dynamics simulation trajectories and as such part of the GROningen MOlecular Simulation software for (bio)molecular simulation. It contains more than 70 programs that can be used to prepare data for the production of molecular simulation trajectories and to analyse these. These programs are reviewed and the various structural, dynamic and thermodynamic quantities that can be analysed using time series, correlation functions and distributions are described together with technical aspects of their implementation in GROMOS. A few examples of the use of GROMOS++ for the analysis of MD trajectories are given. A full list of all GROMOS++ programs, together with an indication of their capabilities, is given in the appendix.
4.1 Introduction

Over the past decades, classical simulation of the dynamics of (bio)molecules in aqueous solution, in crystalline form, or embedded in a lipid membrane, has found widespread use in physicochemical, biochemical and molecular biological research [55, 80–83]. This has become possible through the availability of a number of general software packages for (bio)molecular simulation, such as AMBER [84], CHARMM [85], Desmond [86], GROMACS [87], GROMOS [88], IMPACT [89], MOLARIS [90, 91], NAMD [92], and TINKER [93], and the development of (bio)molecular force fields [94], e.g. AMBER [95], CHARMM [96–103], ECEPP [104], ENCAD [105], CFF [106], GROMOS [53, 72, 107–111], and OPLS [112, 113]. These simulation software packages generally contain a number of functions that can be used to analyse the atomic trajectories produced in a molecular dynamics (MD) simulation in terms of time series, correlation functions or trajectory averages of quantities of interest. The set of different quantities that can be analysed and the computational procedures involved differ between the various simulation packages and are generally scarcely described in the literature. This renders a great deal of research articles irreproducible because the details of the calculations cannot be recovered. Here we present the variety of functions to analyse MD (bio)molecular simulation trajectories that are part of the GROMOS software for (bio)molecular simulation [53, 88, 114–117].

The GROMOS software is written in C++ and comprises two major parts: (1) MD++, the programs that can be used to perform energy minimisation, MD or stochastic dynamics (SD) simulations, and (2) GROMOS++, the programs that can be used for pre-processing (bio)molecular data prior to a simulation and for post-processing the trajectories, coordinates, velocities, energies, etc., produced by an MD simulation. The post-processing functions, in particular, could also be used to analyse (bio)molecular trajectories generated using other (bio)molecular simulation software than GROMOS. Therefore, we only briefly review the pre-processing capabilities of GROMOS++ and describe the post-processing analysis functions in more detail.

The architecture, implementation and parallelisation of the GROMOS software is reported elsewhere [115], while the analysis functions that can be used for the analysis of NMR spectroscopic and X-ray and neutron diffraction data are described in [53] in conjunction with their use in MD simulation based on such data. Since GROMOS++ is written in an object-oriented manner, new functionality is easily implemented.

4.2 Trajectory quantities and basic types of analysis

A simulation trajectory, which allows the calculation of many different properties of the simulated system, is primary data for a computer simulation scientist. Although the simulation programs of MD++ perform the main task of generating such simulation trajectories, an appropriate set-up of a simulation and analysis of the trajectories is just as important as the simulation itself.
4.2. Trajectory quantities and basic types of analysis

In this section the pre-processing of data required by a molecular simulation is outlined, followed by a discussion of the different types of simulation trajectories that are used for analysis in combination with the GROMOS++ programs. A description of the different types of analysis is given, including technicalities such as the treatment of periodic boundary conditions or atom, vector and property specifiers implemented in GROMOS++.

There are three questions to be answered at the beginning of each molecular simulation, (1) what is the chemical structure of the system to be simulated, i.e. the number and types of elements, atoms, particles, and their connectivity, (2) which force field, containing the parameters to model the interactions in the system based on physical theories and approximations, is to be used, and (3) how is the interface with the surroundings of the molecular system to be modelled. The answers to the first two questions allow setting up a system for MD simulation using the sequence of steps specified in Tab. 4.1. A short description of the functionality of each program is given in the appendix (Tab. 4.4). A more detailed description can be found in the GROMOS manual or the electronic code documentation of GROMOS++.

### 4.2.1 Types of trajectory information handled by GROMOS++

A simulation trajectory is completely characterised by the time series of the particle coordinates and velocities, from which a variety of quantities can be calculated. However, it

<table>
<thead>
<tr>
<th>program</th>
<th>action</th>
</tr>
</thead>
<tbody>
<tr>
<td>make_top</td>
<td>build a molecular topology file holding all the parameters and specifications that characterise the molecular system</td>
</tr>
<tr>
<td>com_top</td>
<td>converts solute coordinates in Protein Data Bank (PDB) format to GROMOS format</td>
</tr>
<tr>
<td>pdb2g96</td>
<td>generates hydrogen coordinates based on geometric criteria</td>
</tr>
<tr>
<td>gch</td>
<td>is used to minimise the intermolecular solute energy to remove possible strain from the solute</td>
</tr>
<tr>
<td>MD++</td>
<td>generates the coordinates for a box with solvent molecules in GROMOS format</td>
</tr>
<tr>
<td>ran_box</td>
<td>puts the solute into a solvent box of appropriate size and removes all solvent molecules that show a given spatial overlap with solute atoms</td>
</tr>
<tr>
<td>sim_box</td>
<td>is used to minimise the system energy while keeping the solute positionally restrained to remove high-energy intermolecular contacts</td>
</tr>
<tr>
<td>ion</td>
<td>replaces solvent molecules by ions in order to neutralise the Coulomb charge of the solute or attain a given salt contents</td>
</tr>
</tbody>
</table>
is efficient to also have the possibility to save forces on the particles or energetic quantities of particular sets of atoms during an MD simulation, because the calculation of these types of quantities is computationally expensive and should not have to be repeated while post-processing MD simulation data. Depending on the input parameters, the GROMOS MD++ simulation program regularly writes specific information of the system to the corresponding trajectory file. The time interval between saved configurations should be small enough to obtain sufficient configurations for averaging, but is generally chosen much larger than the MD time step in order to avoid storage of correlated data, unless, of course, the short-time correlations happen to be of interest. Depending on the desired quantity to be calculated, a post-processing program of GROMOS++ reads one or multiple simulation trajectory files of a specific type to compute various other, more complex quantities. GROMOS++ provides more than 70 programs (see appendix) ready to be used for pre- and post-processing of molecular simulations.

The following is an overview of four different types of GROMOS simulation trajectories handled by GROMOS++. The focus is on the information content of each simulation trajectory type and not on its use for analysis. The latter should be clear after reading the sections below.

**Positions and velocities** These simulation trajectory files contain – besides the time information, i.e. current integration time step and simulated time – the three-dimensional, Cartesian atom positions and atom velocities. Additionally, the box size is indicated for every configuration of the trajectory. Information such as atom names and connectivities is generally omitted but can easily be recovered with help of the corresponding molecular topology file based on the GROMOS rule that the sequence of atoms is identical in both files.

**Forces** The force trajectory file is similar to the position and velocity trajectory files. It stores atomic Cartesian forces including those induced by application of constraints on the system, typically using SHAKE [14].

**Energies** Besides the time step information, energy trajectories consist of two main parts: (1) The total energy of the system and its components: kinetic and potential energies, bonded and non-bonded contributions, lattice-sum terms, self polarisation contributions when using a polarisable force field, and special energy terms, e.g. arising from solvent accessible surface area (SASA) implicit solvent simulations [118], from diverse restraining functions, or from enveloping distribution sampling (EDS) simulations [117]. Moreover, it contains kinetic energies for the different sets of atoms that are coupled to a thermal bath as well as the bonded, non-bonded and special energy contributions of each so-called energy group (predefined in the MD++ input file). (2) The second part of the energy trajectory contains information about the mass, the box dimensions, the temperature and the pressure of the simulated system. Some properties are stored in a $3 \times 3$ tensor.
format, namely the pressure, the virial and the molecular kinetic energy. These data can also be stored as block averages.

**Free energy data**  A free energy trajectory contains the same terms of the Hamiltonian as the energy trajectory described above, but as derivatives with respect to the coupling parameter $\lambda$. Such an approach is usually credited to Kirkwood [16] who used it to derive expressions for the chemical potential of components of mixtures. The $\lambda$-dependence is useful for the calculation of relative free energy differences $\Delta F$, e.g. via thermodynamic integration (TI),

$$\Delta F = F(\lambda_B) - F(\lambda_A) = \int_{\lambda_A}^{\lambda_B} \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_{\lambda} d\lambda , \quad (4.1)$$

where the Hamiltonian $H$ describes different systems or states of a system as a function of $\lambda$.

### 4.2.2 Types of analysis

GROMOS++ contains a number of programs that can be used in different combinations for the calculation of a variety of quantities. The majority of these programs belong to one of the following groups, depending on the type of analysis they perform: time series, distributions or time correlation functions.

**Time series** can be calculated for a variety of scalar or vector quantities $Q(t)$ or $\mathbf{Q}(t)$, respectively, which are defined in terms of Cartesian atomic coordinates or velocities. Examples of such quantities are bond angles or torsional dihedral angles or vectors defined by atoms in the molecule, inner and outer products of such vectors, etc.. GROMOS++ is able to write time series for such quantities based on position or velocity trajectories. Using energy or free energy trajectories it is possible to generate time series of temperatures, densities or energetic properties in a straightforward manner.

**Distributions** of a quantity $Q$ can be calculated using GROMOS++. Normalisation allows for the computation of probabilities $P(Q)$, while the average $\langle Q \rangle_t$ and the mean-square fluctuation $\Delta Q^2$ are accessible as the first and second moment of the distribution,

$$\Delta Q^2 = \langle (Q - \langle Q \rangle_t)^2 \rangle_t . \quad (4.2)$$

**Time correlation functions** $C_Q(t)$ or $C_{\mathbf{Q}}(t)$ are defined as

$$C_Q(t) = \langle f(Q_i(t'), Q_j(t' + t)) \rangle_{t'}
= (t_{MD} - t)^{-1} \int_0^{t_{MD} - t} f(Q_i(t'), Q_j(t' + t)) dt' . \quad (4.3)$$
Chapter 4. GROMOS++

The function \( f \) may be a simple multiplication of the two quantities \( Q_i(t') \) and \( Q_j(t' + t) \), but GROMOS++ allows a user specified definition of this function. \( C_Q(t) \) is defined in the same way, but depends on the vector quantities \( Q_i(t) \) and \( Q_j(t' + t) \) instead of \( Q_i(t') \) and \( Q_j(t' + t) \). If \( i = j \), the correlation function is an autocorrelation function, while for other cases it is a cross-correlation function. The calculation of \( Q_i \) and \( Q_j \) from a trajectory is usually done for \( N_t \) discrete, equally spaced time points \( t_n = n\Delta t \) with \( n = 0, 1, \ldots, N_t - 1 \).

The discrete equivalent of (Eq. (4.3)) is then

\[
C_Q(n\Delta t) = (N_t - n)^{-1} \sum_{k=0}^{N_t-n-1} f\left(Q_i(k\Delta t), Q_j((k + n)\Delta t)\right). \tag{4.4}
\]

If \( f(Q_i(t'), Q_j(t'+t)) = Q_i(t') \cdot Q_j(t'+t) \), GROMOS++ makes use of fast Fourier transforms instead of the more time consuming summation algorithm.

4.2.3 Superposition of molecular structures

Molecules tumble in space during a molecular simulation. This is exactly what they should do when in the liquid phase. However, if this overall motion of the solute is of no interest, GROMOS++ programs can perform an alignment of the configuration of one or multiple molecules against a reference configuration or structure, e.g. the programs \texttt{rmsd} and \texttt{rmsf}, see Section 4.6. This alignment is carried out by first superimposing the centres of mass of a defined set of atoms in each configuration followed by a rotational fit with respect to this subset of molecular atom positions. This superposition of pairs of structures is done automatically when running the programs for which it is required, while the atoms and the reference structure to be used for the superpositioning can be specified by the user.

4.2.4 Spatial boundary conditions and gathering

Cartesian position trajectories generated using (periodic) boundary conditions with the following box shapes can be handled by GROMOS++: vacuum, rectangular, truncated octahedron and triclinic. Due to the periodicity of the non-vacuum boundary conditions, the trajectory coordinates of molecules or groups of atoms may be such that covalent bonds are broken, i.e. the nearest image of an atom may not be the one saved in the trajectory file. Before calculating inter-atomic quantities such as bond lengths, bond angles, torsional dihedral angles, etc. GROMOS++ offers the possibility to select the nearest-image position with respect to a reference position, i.e. to gather a trajectory before the analysis. The different gathering methods that may be chosen by the user are listed in Tab. 4.2.
4.2. Trajectory quantities and basic types of analysis

Table 4.2: Handling of periodic boundary conditions.

<table>
<thead>
<tr>
<th>method</th>
<th>action</th>
</tr>
</thead>
<tbody>
<tr>
<td>nog</td>
<td>no gathering</td>
</tr>
<tr>
<td>glist</td>
<td>gathering with respect to a list of atoms</td>
</tr>
<tr>
<td>gtime</td>
<td>gathering with respect to the previous configuration of the trajectory</td>
</tr>
<tr>
<td>gref</td>
<td>gathering with respect to a reference structure or configuration of atoms</td>
</tr>
<tr>
<td>gltime</td>
<td>gather the first configuration of the trajectory based on a list of atoms, and the following configurations with respect to the previous configuration</td>
</tr>
<tr>
<td>grtime</td>
<td>gather the first configuration of the trajectory based on a reference configuration, and the following configurations with respect to the previous configuration</td>
</tr>
<tr>
<td>gbond</td>
<td>gathering based on the bond connectivities of the solute</td>
</tr>
</tbody>
</table>

4.2.5 Atom, vector and property specifiers

The more specific the analysis of a simulation, the higher the requirements with respect to the flexibility of the corresponding analysis tool. GROMOS++ makes use of three specifiers to define the quantity to be calculated as precisely and compactly as possible while keeping a high level of flexibility. All three specifiers are passed to the program as input parameters.

Atom specifiers offer a flexible way to access specific atoms of a system. It is even possible to specify atoms which are not present in the trajectory or molecular topology file, so-called virtual atoms, e.g. hydrogen atoms of a united CH$_x$ atom, or common properties of a group of atoms, such as the centre of geometry or the centre of mass. Atom specifiers are of the format

\[ \text{<molecule>:<atoms1>[:<atoms2>,...,<atomsN>] ,} \]

where \text{<molecule>} is the molecule number and \text{<atoms>} is one or more atoms defined in one of the following ways: the atom number within the molecule \text{<molecule>}, the atom name, or the residue and atom, both specified either by its number or name. Virtual atoms are accessed via their type and the atoms needed to construct the defined virtual atom type. For example a (virtual) aromatic hydrogen atom position is generated by the three neighbouring CH$_1$ united atom positions, based on geometric criteria [42].

A more complicated selection of atoms is accessed using the options '?', minus (<atom specifier>) and not(<atom specifier>). The wild card '?' is used to specify
groups of atoms with similar atom names, e.g. all carbon atoms (CA, CB, CG, ...) of a protein,

\[ 1:C? \]

\( \text{minus(<atom specifier>) and not(<atom specifier>) both specify atoms not to be selected. The option minus(<atom specifier>) allows atoms to be added later on while not(<atom specifier>) definitely removes the specified atoms from the selection. The following three examples all specify all CA atoms of even numbered residues of a 10 amino acid peptide:} \]

\[ 1:\text{res}(1,3,5,7,9:CA), \]

\[ 1:CA \text{ minus}(1:\text{res}(2,4,6,8,10:CA)) \],

\[ \text{not}(1:\text{res}(2,4,6,8,10:CA)) \ 1:CA \]

It is also possible to read atom specifiers from a file, e.g. written by the program atominfo, using the atom specifier file(<file name>).

**Property specifiers** Some GROMOS++ programs are able to read a property specifier, a helpful tool to describe the quantity to be calculated. The syntax for the property specifier is

\[ <\text{type}>%<\text{arg}> \]

with \(<\text{type}>\) defining the property and \(<\text{arg}>\) an atom or vector specifier providing the necessary information to calculate the desired quantity. Vector specifiers are described below. The different types of property specifiers are given in Tab. 4.3. The expression \(<\text{type}>, expr,\) is a specifier itself with the following syntax:

\[ expr%<f1>(<args1>) \ <op> \ <f2>(<args2>) \]

where \(<op>\) is an arithmetic or logical operator while \(<f1>\) and \(<f2>\) are functions with the corresponding scalar or vector arguments, \(<\text{arg}1>\) or \(<\text{arg}2>\), depending on the nature of the function. The following functions are supported: \(\sin, \cos, \tan, \text{asin}, \text{acos}, \text{atan}, \text{abs}\) (the norm of a scalar or vector), \(\text{sqrt}, \text{abs2}\) (the squared norm of a vector), dot, cross or ni (nearest image).
4.2. Trajectory quantities and basic types of analysis

Table 4.3: Property specifiers.

<table>
<thead>
<tr>
<th>specifier</th>
<th>property</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>the distance between two atoms</td>
</tr>
<tr>
<td>a</td>
<td>the angle defined by three atoms</td>
</tr>
<tr>
<td>t</td>
<td>a torsional dihedral angle</td>
</tr>
<tr>
<td>hb</td>
<td>the presence of hydrogen bonds</td>
</tr>
<tr>
<td>st</td>
<td>the stacking between two groups of atoms</td>
</tr>
<tr>
<td>o</td>
<td>the order between two vectors</td>
</tr>
<tr>
<td>op</td>
<td>the order parameter</td>
</tr>
<tr>
<td>pr</td>
<td>a pseudo rotation</td>
</tr>
<tr>
<td>pa</td>
<td>the pucker amplitude</td>
</tr>
<tr>
<td>expr</td>
<td>a quantity defined by an expression property</td>
</tr>
</tbody>
</table>

Vector specifiers Property and expression specifiers may be defined as a function of one or more vectors. In GROMOS++, a vector can be specified by its three-dimensional Cartesian coordinates,

\[ \text{cart}(<x>,<y>,<z>) \],

or in polar coordinates,

\[ \text{polar}(<r>,<\alpha>,<\beta>) \],

with \( r \) being the length of the vector and \( \alpha, \beta \) the two angles to orient the vector in space. In addition, the Cartesian coordinates of atoms may be used to specify a vector.

\[ \text{atom}(<\text{atom specifier}>) \]

refers to the coordinates of the atom defined by an atom specifier. If the atom specifier holds two atoms, the vector is specified pointing from the first to the second atom.

4.2.6 Input/Output formats

All GROMOS++ programs are called from the command line, taking the necessary additional information as command line arguments. A usual program call looks like

\[ \text{program @flag1} <\text{arg1}> [@\text{flag2} <\text{arg2}> \ldots @\text{flagN} <\text{argN}>] \]

The different arguments may also be collected in a single input file. The program call then looks like

\[ \text{program @f file} \]. \]
Since GROMOS++ is used for pre- and post-processing of molecular simulations, its input and output functionality is complex. Input flags and arguments may point to files of various formats (e.g. topology files, simulation trajectories, NOE bound specifications, etc.) which are described in more detail in the GROMOS manual. However, GROMOS++ usually writes its output to the standard output, which may be directed into a text file.

The reading and writing of coordinate files (single configuration or trajectories) is often done and is therefore optimised: compressed trajectory files can be read directly and are decompressed while reading. Additionally, the writing of coordinate files may be done in one of the following user specified formats: standard or reduced GROMOS format, PDB-format, and AMBER format.

Further, the two argument flags @inG96 and @outG96 ensure the compatibility with earlier GROMOS versions that read or write previous file formats.

4.2.7 Physical units

Physical constants define the units used in GROMOS and are therefore not hard coded in GROMOS++. If a topology file is given, the physical constants and their units are derived from the following four physical constants, defined in the molecular topology file: Boltzmann’s constant, Planck’s constant, the electric permittivity of vacuum and the speed of light. Analyses carried out without the information of a molecular topology file use physical constants which are initialised to defaults while printing a warning.

The GROMOS example files use the following basic units of the standard international system:

<table>
<thead>
<tr>
<th>Unit</th>
<th>Conversion</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>nm = 10^{-9} m</td>
</tr>
<tr>
<td>mass</td>
<td>u = atomic mass unit</td>
</tr>
<tr>
<td>time</td>
<td>ps = 10^{-12} s</td>
</tr>
<tr>
<td>temperature</td>
<td>K</td>
</tr>
<tr>
<td>charge</td>
<td>e = electronic charge = 1.6021892 \times 10^{-19} C</td>
</tr>
</tbody>
</table>

4.3 Observable and non-observable quantities

GROMOS++ consists of more than 70 individual programs for many different tasks. To describe them all in detail is beyond the scope of this work. However, we report the calculation of some structural, dynamic and thermodynamic quantities, covering the practically most relevant aspects of the post-processing functionalities of GROMOS++. The calculation of experimentally observable and derived quantities is briefly mentioned but described in more detail elsewhere [53]. The complete list of all current GROMOS++ programs with a short description is found in the appendix, tables Tab. 4.4 and Tab. 4.5.
4.3. Observable and non-observable quantities

4.3.1 Structural quantities

\textbf{rmsd, rmsf, rgyr} These programs are mainly developed for the analysis of structural properties of a peptide or protein, but may also work for other molecules. The atom \textbf{positional root-mean-square deviation} (RMSD) between two molecular structures is calculated according to

\[
RMSD(r^N_a, r^N_{a,\text{ref}}) = \sqrt{\frac{1}{N_a} \sum_{i=1}^{N_a} (r_i - r_{i,\text{ref}})^2},
\]  

(4.5)

where \(r^N_a = (r_1, r_2, \ldots, r_{N_a})\) indicates the positions of the atoms, \(N_a\) the number of atoms considered, \(r_i\) the position of atom \(i\) in the first structure and \(r_{i,\text{ref}}\) the position of atom \(i\) in the second, reference structure.

The atom \textbf{positional root-mean-square fluctuation} (RMSF) of an atom \(i\) is computed according to

\[
RMSF_i = \sqrt{\frac{1}{N_T} \sum_{t=1}^{N_T} (r_i(t) - \langle r_i \rangle)^2},
\]  

(4.6)

where \(\langle r_i \rangle\) is the time averaged position of atom \(i\), and \(N_T\) the number of configurations or time frames in the simulation trajectory.

In contrast to the RMSD and RMSF functions, which describe the positional change and mobility of atoms of a molecule, the \textbf{radius of gyration} (Rgyr) is a measure of the compactness of the structure. It can be related to light-scattering intensities and is calculated using the definition

\[
R_{gyr} = \sqrt{\frac{1}{N_a} \sum_{i=1}^{N_a} (r_i - R_{cm})^2},
\]  

(4.7)

with

\[
R_{cm} = \frac{1}{M} \sum_{i=1}^{N_a} m_i r_i
\]  

(4.8)

and

\[
M = \sum_{i=1}^{N_a} m_i ,
\]  

(4.9)

in which \(r_i\) denotes the Cartesian position of atom \(i\), \(m_i\) its mass, and \(N_a\) the number of considered atoms.

\textbf{sasa} The \textbf{solvent-accessible surface area} can be calculated as described by Lee and Richerds [119]. A spherical probe of given radius \(r\) is rolled over the surface of the solute.
The path traced out by its centre is proportional to the solvent-accessible surface area.

In GROMOS, the radii of the heavy atoms are obtained by calculating the minimum energy distance of the interaction between the solute atom and the solvent. This value is reduced by the specified probe radius $r$ to account for the radius of the solvent atom.

Program `sasa_hasel` can be used alternatively, it contains an implementation of the algorithm described by Hasel et al. [120].

**dssp** The detection of secondary structure in a protein is implemented according the rules of Kabsch and Sander [121]. An overview over the different residues with the percentage assigned to each secondary structure element over the trajectory is printed, while time series for each secondary structure element are saved in separate files. It may occur that one residue is assigned to be part of two secondary structure elements at the same time. In order to avoid ambiguous assignments, the following priority rules are applied: $\beta$-sheet/bridge > $\alpha$-helix > $\pi$-helix > $3_{10}$-helix > hydrogen bonded turn > bend.

**hbond** The occurrence of two-centre and three-centre hydrogen bonds defined using geometric criteria can be monitored over a trajectory [122]. A two-centre hydrogen bond is considered to be present if (1) the position of a hydrogen atom H connected to a donor atom D is within $d_{HA} = 0.25 \text{nm}$ from that of an acceptor atom A, and (2) the D–H–A angle is larger than $\angle_{DHA} = 135^\circ$. Occurrences of three-center hydrogen bonds are defined for a donor atom D, hydrogen atom H, and two acceptor atoms $A_1$ and $A_2$, if

1. the distances H–A$_1$ and H–A$_2$ are smaller than 0.27 nm,
2. the angles D–H–A$_1$ and D–H–A$_2$ are larger than 90$^\circ$,
3. the sum of the angles D–H–A$_1$, D–H–A$_2$ and A$_1$–H–A$_2$ is larger than 340$^\circ$ and
4. the dihedral angle defined by the planes through the atoms D–A$_1$–A$_2$ and H–A$_1$–A$_2$ is smaller than 15$^\circ$.

All distance and angle bounds used in the criteria above may be changed by the user. If a reference structure is given, only the hydrogen bonds present in the reference structure are monitored. Otherwise, all intra-molecular and/or inter-molecular hydrogen bonds between user specified subsets of atoms in the system are followed. Averages for the monitored distances, angles and occurrences are printed to the standard output while the corresponding time series are written to a file.

**tser** and **tcf** Structural quantities defined by a property specifier, e.g. distances, angles, torsional angles, and more complex, user-specified structural properties to be calculated from the atomic coordinates can be computed as a time series and/or a (normalised) distribution. The output of the program `tser` may be used for further treatment with the program `tcf`, which computes time correlation functions according to equation (Eq. (4.3)).
4.3. Observable and non-observable quantities

**dipole** The molecular electric dipole moment $\mathbf{p}$ can be calculated according to

$$
\mathbf{p}^{N_a}(\mathbf{r}^{N_a}) = \sum_{i=1}^{N_a} q_i \mathbf{r}_i,
$$

(4.10)

where $\mathbf{r}^{N_a} = (\mathbf{r}_1, \mathbf{r}_1, \ldots, \mathbf{r}_{N_a})$ indicates the atom positions, $N_a$ is the number of atoms considered, $q_i$ the charge of atom $i$ and $\mathbf{r}_i$ its Cartesian position. In the case of a total net charge $Q$ of the $N_a$ atoms, $Q = q_1 + q_2 + \cdots + q_{N_a} \neq 0$, the dipole moment is dependent on the origin of the coordinate system. Either a particular origin, i.e. the centre of geometry of the $N_a$ atoms,

$$
\mathbf{p}^{N_a}(\mathbf{r}^{N_a}) = \sum_{i=1}^{N_a} q_i (\mathbf{r}_i - \mathbf{R}_{cog}),
$$

(4.11)

with

$$
\mathbf{R}_{cog} = \frac{1}{N_a} \sum_{i=1}^{N_a} \mathbf{r}_i,
$$

(4.12)

can be chosen or a uniform background charge of $-\frac{Q}{N_a}$ is added to each atomic charge to make $\mathbf{p}^{N_a}$ origin independent.

**cluster** This program performs a conformational clustering based on a similarity matrix as calculated by the program `rmsdmat`. The clustering algorithm is described by Daura et al. [123]. Structures with RMSD values smaller than a user specified cutoff are considered to be structural neighbours. The structure with the highest number of neighbours is considered to be the central member of the cluster of similar structures forming a conformation. After removing all structures belonging to this first cluster, the procedure is repeated to find the second, third etc. most populated clusters.

A specific structure can be forced to be the central member structure of the first cluster, which can also be the reference structure. The clustering can be performed on a subset of the matrix by specifying the maximum number of structures to consider. This allows for an assessment of the development of the number of clusters over time.

4.3.2 Dynamic quantities

**diffus** This program calculates the diffusion constant $D$ of an atom or of the centre-of-geometry of a specified set of atoms. First, the mean square-displacement, $\Delta(t)$, is obtained by averaging over all considered atoms and over multiple time windows,

$$
\Delta(t) = \frac{1}{N_a} \sum_{i=1}^{N_a} \left\langle (\mathbf{r}_i(t + \tau) - \mathbf{r}_i(\tau))^2 \right\rangle_{\tau \leq t_{av} - t},
$$

(4.13)
where $r_i$ is the position of atom $i$, $N_a$ the number of atoms considered, and $t_{av}$ the averaging time. According to the Einstein relation, the diffusion constant can be estimated from the slope of $\Delta(t)$

$$D = \lim_{t \to \infty} \frac{\Delta(t)}{2N_dt} ,$$

(4.14)

where $N_d$ is the number of dimensions considered.

**rot_rel** The rotational relaxation time of a molecule can be estimated from the auto-correlation function of the Legendre polynomials of a molecular vector $v$ of unit length,

$$C_1(t) = \langle v(\tau) \cdot v(\tau + t) \rangle_\tau$$

(4.15)

$$C_2(t) = \frac{1}{2}(3 \langle v(\tau) \cdot v(\tau + t) \rangle_\tau^2 - 1) .$$

(4.16)

The program rot.rel calculates the first and second order Legendre polynomials and the time correlation functions. The output of this program can also be produced by a combination of the programs tser and tcf.

**visco** The bulk and shear viscosities can be calculated using the Einstein relation from the elements of the pressure tensor that are written to an energy trajectory. Let $P_{\alpha\beta}$ be the element of the pressure tensor, and $G_{\alpha\beta}(t)$ the time integral of $P_{\alpha\beta}$,

$$G_{\alpha\beta}(t) = \int_0^t P_{\alpha\beta}(t') dt' .$$

(4.17)

The viscosity tensor element $\eta_{\alpha\beta}$, calculated in terms of the integral (equation (Eq. (4.17))) of the pressure component $P_{\alpha\beta}$, is proportional to the mean-square change of $G_{\alpha\beta}(t)$ in the limit of infinite time,

$$\eta_{\alpha\beta} = \frac{V}{2k_BT} \lim_{t \to \infty} \frac{d}{dt} \left\langle \left( G_{\alpha\beta}(t + \tau) - G_{\alpha\beta}(\tau) \right)^2 \right\rangle_{\tau \leq t_{av} - t} ,$$

(4.18)

where $V$ denotes the volume of the (periodic) box, $k_B$ the Boltzmann constant, and $T$ is the temperature of the system. For isotropic systems, an estimate of the bulk viscosity can be obtained from the average of the three diagonal components of the pressure tensor,

$$\eta_{bulk} = \frac{1}{3}(\eta_{xx} + \eta_{yy} + \eta_{zz}) ,$$

(4.19)

while the shear viscosity is obtained by averaging the off-diagonal elements,

$$\eta_{shear} = \frac{1}{3}(\eta_{xy} + \eta_{xz} + \eta_{yz}) .$$

(4.20)
eps_field The static dielectric permittivity $\epsilon(0)$ of a liquid can be obtained by applying an external electric field during a simulation and measuring the polarisation response [124]. For an external field of strength $E_{ext}$ the permittivity is given by

$$\epsilon(0) = 1 + 4\pi \frac{\langle P_z \rangle}{E_{ext}} ,$$

where $P_z$ is the average polarisation of the system in the z-direction and $P$ is defined as

$$P(t) = \frac{M(t)}{V(t)} ,$$

where $M$ denotes the total dielectric dipole moment of the system and $V$ the volume of the simulation box.

The external electric field size should be chosen small enough to avoid saturation and large enough to induce a significant $\langle P_z \rangle_t$.

4.3.3 Thermodynamic quantities

ene_ana This program can calculate time series, time averages, root-mean-square fluctuations and statistical error estimates for properties contained in an energy or free energy trajectory file, e.g. volume, temperature, pressure, different energy terms, conformational entropy, etc.. The error estimates are calculated from block averages of different sizes. In combination with a library file, ene_ana allows the analysis of quantities that are a function of one or more properties provided by the energy and/or free energy trajectory files. We note that in principle one can teach ene_ana to read any trajectory, i.e. also force- or block-averaged energies or positions, by specifying the corresponding block names and block formats in its library file.

Free energy differences The calculation of relative free energies of ligand-protein binding, of solvation for different compounds, and of different conformational states of a (bio)molecule is of considerable interest with regard to an understanding of these processes and to the design or selection of potential inhibitors of enzymes. Since such processes in aqueous solution generally comprise energetic and entropic contributions from many molecular configurations, adequate sampling of the relevant parts of configurational space when calculating ensemble averages is required and can be reached through MD simulations. Most methods to obtain relative free energies require a particular modification of the Hamiltonian in an MD simulation, which leads to artificial forces on the atoms that enhance the sampling. The implementation in the GROMOS software of the most popular or promising of such methods, i.e. thermodynamic integration, umbrella sampling, local-elevation umbrella sampling, and enveloping distribution sampling, are described elsewhere [117]. Multiple GROMOS++ programs are available for the calculation of free
energy differences from simulations with modified Hamiltonians.

Some methods to compute relative free energies only require post-processing of trajectory data from a standard MD simulation, and are therefore mentioned here.

In the Widom particle-insertion method \cite{125}, a test or virtual atom is inserted randomly \( N_{\text{try}} \) times in each configuration of the molecular system and its free energy of ”solvation” is then calculated as

\[
\Delta F_{\text{solv}} = -k_B T \ln \left[ \frac{1}{N_{\text{try}}} \sum_{i=1}^{N_{\text{try}}} \exp \left( -\frac{V(r_N^i, r_{\text{test}})}{k_B T} \right) \right] \, ,
\]

(4.23)

where \( V(r_N^i, r_{\text{test}}) \) is the potential energy of the test particle with respect to all atoms in the system and the average is over all configurations of the \( N \) atoms of the system. The program \texttt{m.widom} calculates this free energy.

In the one-step perturbation method \cite{17, 18}, the free energy change due to a change of the reference Hamiltonian \( H_R \) into a perturbed Hamiltonian \( H_A \) is calculated as

\[
\Delta F_{AR} = F_A - F_R = -k_B T \ln \left[ \exp \left( -\frac{(H_A - H_R)}{k_B T} \right) \right]_{R} \, ,
\]

(4.24)

where the ensemble average is over the simulation based on \( H_R \). Such an ensemble average can be calculated using the program \texttt{dg.ener}.

When using a biasing potential energy term \( V_{US}(r_N) \) in the Hamiltonian \( H_{\text{bias}}(r_N) \) of the MD simulation \cite{126}, its influence has to be removed from the ensemble averages. This is done by so-called reweighting of the configurations in the averaging:

\[
\langle Q \rangle = \frac{\left\langle Q \exp \left( \frac{V_{US}}{k_B T} \right) \right\rangle_{\text{bias}}}{\left\langle \exp \left( \frac{V_{US}}{k_B T} \right) \right\rangle_{\text{bias}}} \, .
\]

(4.25)

The ensemble average \( \langle ... \rangle \) for an unbiased Hamiltonian \( H \) is expressed in terms of two ensemble averages for the biased Hamiltonian \( H_{\text{bias}} = H + V_{US} \). Reweighting can be performed using the program \texttt{reweight}.

The configurational entropy of a solute molecule can be estimated using Schlitter’s heuristic formula \cite{127}, which gives an upper bound for the true entropy,

\[
S_{\text{true}} \leq S = \frac{1}{2} k_B \ln \left[ \det \left( \mathcal{H} + k_B T \left( \frac{e}{\hbar} \right)^2 \mathbf{M} \sigma \right) \right] \, ,
\]

(4.26)

where \( \hbar \) is Planck’s constant divided by \( 2\pi \), \( e \) is Euler’s number, \( \mathbf{M} \) is the diagonal mass matrix holding on the diagonal the masses belonging to the \( 3N_a \) Cartesian degrees of freedom, and \( \sigma \) is the covariance matrix of the positional fluctuations of these degrees of
freedom. The elements of $\mathbf{\sigma}$ are

$$\sigma_{ij} = (\langle x_i - \langle x_i \rangle \rangle)(x_j - \langle x_j \rangle)$$

(4.27)

with $x_i$ being the Cartesian coordinates of the $N_a$ atoms considered for the entropy calculation after least-squares fitting of the position of a given subset of atoms of the coordinate trajectory. The program solute_entropy calculates Schlitter entropy and the quasi-harmonic entropy from molecular coordinate trajectories.

**4.3.4 Comparison to experimentally observable and derived quantities**

GROMOS++ is able to calculate a variety of experimentally observable quantities as well as quantities that are derived from experimental data, i.e. NOE intensities or atom-atom distance bounds, $^3J$-coupling constants, residual dipolar couplings (RDCs), order parameters derived from NMR experiments, small- and wide-angle X-ray scattering (SAXS/WAXS) intensities, and total neutron scattering intensities for liquids. A detailed description of the implementation of such quantities in the GROMOS software and their use in protein refinement, including examples, is described by Schmid et al. [53].

**4.4 Examples**

Here we present a few examples of the use of GROMOS++ programs to analyse MD simulation trajectories, taken from previous work.

**4.4.1 Global structural properties as a function of time**

In Fig. 4.1, the backbone atom-positional RMSD from the initial X-ray derived structure for two MD simulations of hen egg white lysozyme (HEWL) is given as function of time together with the radius of gyration, the solvent accessible surface area and the occurrence of major secondary structure elements. The trajectory of the simulation in vacuo clearly deviates more from the X-ray structure than that of the simulation in water, illustrating the effect of the solvent on protein structure. The setup of these simulations is described elsewhere [53].

**4.4.2 Hydrogen bond analysis as function of time**

In Fig. 4.2 the intra-peptide hydrogen bonding of four 7-residue peptides in methanol as observed in 100 ns MD simulations of their folding equilibria is shown. The four peptides
Figure 4.1: Atom-positional root-mean-square deviations (RMSD) from the initial X-ray structure, radius of gyration (R\text{gyr}), solvent accessible surface area (SASA), and secondary structure analysis of two HEWL simulation trajectories using the GROMOS force fields 54A7 (protein in water) and 54B7 (protein in vacuum) [53].
4.4. Examples

Figure 4.2: Occurrence of intra-solute hydrogen bonds with a population larger than 5% in the four simulations of the four hepta-peptides Val-Ala-Leu-X-Ile-Met-Phe in water: (a) X = Aib, (b) X = L-Ala, (c) X = D-Ala, and (d) X = Gly [128].

with sequence Val-Ala-Leu-X-Ile-Met-Phe differ in the central residue X, which is Aib, L-Ala, D-Ala, or Gly [128]. The presence of a central Aib residue clearly enhances hydrogen bonding, while a central Gly residue shows the least hydrogen bonding, as expected.

4.4.3 Conformational cluster analysis to detect structural differences

The four conformational ensembles for the four 7-residue peptides were compared by performing a conformational cluster analysis for three pairs of trajectories, see Fig. 4.3. It shows that the four ensembles are quite different. The Aib peptide shows a stronger propensity to bent structures, while the L-Ala peptide shows a tendency to adopt more extended conformations, and the Gly peptide shows preference for a $\beta$-turn [128].
Figure 4.3: Conformational clustering analysis over the 100 ns trajectories of the four heptapeptides (atom-positional RMSD within 0.08 nm for backbone N, C(β), C(α), and C-atoms of residues 2–6) [128]. In each panel the population of clusters observed in two joint 100 ns trajectories are shown, with the number of configurations originating from the X = Aib peptide indicated in grey.

4.4.4 Configurational entropy as function of time

In figure Fig. 4.4 the Schlitter entropy is shown as function of time for two differently protonated α- and β-peptides of similar lengths consisting purely of Ala amino acid residues and solvated in water and in methanol. On a time scale of 10 ns, the Schlitter entropy of the α-helical conformation of the 8-residue α-peptide and of the 3_14-helical conformation of the 6-residue β-peptide is well converged. Although all peptides contain the same number of 24 backbone atoms, the α-peptides show a significantly higher configurational entropy per atom than the β-peptides, irrespective of their protonation state or solvent [129].

4.5 Conclusion

An overview over the different types of analysis implemented in the GROMOS++ software has been given. Three types of analysis programs were distinguished: programs that calculate structural, dynamic or thermodynamic quantities from configurational trajectories. Additional programs are available that compute ensemble averages of experimentally
4.5. Conclusion

Figure 4.4: Build up of Schlitter entropy $S_{config}^{S}$, Eq. (4.26), as function of time. The Schlitter entropy was calculated for the $C_\alpha$ atoms during the 10 ns MD simulations of the various 24 backbone atom $\alpha$- and $\beta$-peptides in solution as labeled ($\alpha$MO- = $\alpha$-peptide, methanol, $O^-$ terminal group) [129]. In each case, the structures were first aligned according to the positions of their $C_\alpha$ atoms.

Observable quantities or quantities derived from experimental data, e.g. NMR NOE intensities or atom-atom distance bounds, $^3J$-coupling constants, residual dipolar couplings (RDCs) and order parameters, small- and wide-angle X-ray scattering (SAXS/WAXS) intensities and neutron scattering intensities for liquids [53].

Compared to self-written, individual scripts computing such quantities, the use of GROMOS++ offers the advantage that most of the building blocks required for a particular type of analysis are available and tested by a wide group of users. The composition of new analysis programs is relatively straightforward because of the implementation of atom, vector and property specifiers, which allow for a very flexible description of individual quantities to be calculated. Each GROMOS++ analysis program is described in the GROMOS manual and digital, in-code documentation. Changing the code to add new functionality does not need much effort, since GROMOS++ is written in C++ to support maximal reusability of source code [115].

In summary, GROMOS++ is a flexible and rich collection of analysis tools ready to be used for a variety of types of analysis regarding molecular simulation trajectories.
4.6 Appendix

Table 4.4: List of GROMOS++ programs for pre-processing of a molecular simulation.

<table>
<thead>
<tr>
<th>name</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bin_box</td>
<td>Creates a configuration of a condensed phase system consisting of two components.</td>
</tr>
<tr>
<td>build_box</td>
<td>Generates a configuration of a condensed phase system on a grid (only one component).</td>
</tr>
<tr>
<td>check_box</td>
<td>Checks the box dimensions of a trajectory file.</td>
</tr>
<tr>
<td>check_top</td>
<td>Checks a molecular topology for (consistency) errors.</td>
</tr>
<tr>
<td>com_top</td>
<td>Combines (multiple) molecular topology files into one.</td>
</tr>
<tr>
<td>con_top</td>
<td>Converts a molecular topology to one based on a different force-field version.</td>
</tr>
<tr>
<td>copy_box</td>
<td>Repeats/extends a simulation box along a given Cartesian axis.</td>
</tr>
<tr>
<td>cry</td>
<td>Performs (crystallographic) symmetry operations on configurations of molecules.</td>
</tr>
<tr>
<td>explode</td>
<td>Places molecules of a given box thereby expanding intermolecular distances to satisfy a specific minimum inter-molecular distance.</td>
</tr>
<tr>
<td>gca</td>
<td>Generates Cartesian coordinates for atoms from specified distances and/or (dihedral) angles for the atoms.</td>
</tr>
<tr>
<td>gch</td>
<td>Generates Cartesian coordinates for hydrogen atoms based on the coordinates of covalently bound neighbour atoms.</td>
</tr>
<tr>
<td>ion</td>
<td>Replaces solvent molecules by ions based on the local electrostatic potential or by random selection.</td>
</tr>
<tr>
<td>make_pt_top</td>
<td>Takes two or more molecular topologies and writes the differences in the perturbation topology format.</td>
</tr>
<tr>
<td>make_sasa_top</td>
<td>Adds the SASA block to a molecular topology file.</td>
</tr>
<tr>
<td>make_top</td>
<td>Creates a molecular topology file.</td>
</tr>
<tr>
<td>mk_script</td>
<td>Generates the scripts and input files to run a molecular simulation.</td>
</tr>
<tr>
<td>pdb2g96</td>
<td>Converts coordinate files from pdb to GROMOS format.</td>
</tr>
<tr>
<td>pert_top</td>
<td>Creates a perturbation topology to uniformly set interactions to given values for specified atoms.</td>
</tr>
<tr>
<td>prep_ed</td>
<td>Generates dual molecular and perturbation topologies for an EDS simulation.</td>
</tr>
<tr>
<td>pt_top</td>
<td>Combines molecular topologies and perturbation topologies to write new (perturbation) topologies.</td>
</tr>
<tr>
<td>ran_box</td>
<td>Creates a configuration for a condensed phase system of any composition with random molecule placements.</td>
</tr>
</tbody>
</table>

Table continued on the next page …
Tab. 4.4 (continued) Pre-processing programs of GROMOS++.

<table>
<thead>
<tr>
<th>Program</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ran_solvation</td>
<td>Solvates a solute by randomly placing solvent molecules around it.</td>
</tr>
<tr>
<td>red_top</td>
<td>Reduces a molecular topology to one for a subset of atoms.</td>
</tr>
<tr>
<td>sim_box</td>
<td>Solvates a solute in a solvent box removing solvent molecules that are too close to solute atoms.</td>
</tr>
</tbody>
</table>
Table 4.5: List of GROMOS++ programs for post-processing of a molecular simulation.

<table>
<thead>
<tr>
<th>name</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bilayer_dist</td>
<td>Computes the atom distribution along a bilayer normal to characterise a membrane system.</td>
</tr>
<tr>
<td>bilayer_oparam</td>
<td>Calculates order parameters for bilayer systems (membranes) with respect to a fixed orientation (usually the bilayer normal).</td>
</tr>
<tr>
<td>cluster</td>
<td>Performs a conformational clustering based on a $RMSD$ matrix, e.g. calculated by the program $rmsdmat$.</td>
</tr>
<tr>
<td>cog</td>
<td>Calculates the center of geometry or centre of mass position of all solute atoms of a simulation trajectory.</td>
</tr>
<tr>
<td>dfmult</td>
<td>Calculates free energy differences between multiple states $A$ from a simulation at a reference state $R$.</td>
</tr>
<tr>
<td>dg_ener</td>
<td>Calculates the free energy difference between two states $A$ and $B$, based on the perturbation formula. Reads the output of the program $ener$.</td>
</tr>
<tr>
<td>diffus</td>
<td>Calculates the diffusion constant for a selected set of atoms.</td>
</tr>
<tr>
<td>dipole</td>
<td>Calculates the electric dipole moment for a selected set of atoms.</td>
</tr>
<tr>
<td>dtrans</td>
<td>Monitors transitions of torsional dihedral angle rotations with respect to the potential energy.</td>
</tr>
<tr>
<td>dssp</td>
<td>Detects secondary structure elements in a protein.</td>
</tr>
<tr>
<td>eds_mult_all</td>
<td>Calculates the parameters needed for an enveloping distribution sampling (EDS) simulation from energy time series, based on an iterative scheme.</td>
</tr>
<tr>
<td>edyn</td>
<td>Performs an essential dynamics analysis over a trajectory file. The covariance matrix is calculated and diagonalised for specified atoms.</td>
</tr>
<tr>
<td>ene_ana</td>
<td>Writes a time series for specific values from a (free) energy trajectory file. Simple statistics or calculations of combined trajectory entries are possible.</td>
</tr>
<tr>
<td>ener</td>
<td>Recalculates user specified interaction energies from molecular trajectory files using the interaction parameters from the molecular topology.</td>
</tr>
<tr>
<td>eps_field</td>
<td>Calculates the relative dielectric permittivity from a trajectory of a molecular simulation in which an external electric field was applied.</td>
</tr>
<tr>
<td>epsilon</td>
<td>Calculates the relative dielectric permittivity based on a Kirkwood - Fröhlich type of equation (fluctuation formula).</td>
</tr>
<tr>
<td>filter</td>
<td>Reduces/filters a coordinate trajectory to contain only a specified set of atoms.</td>
</tr>
</tbody>
</table>

Table continued on the next page . . .
Tab. 4.5 (continued) Post-processing programs of GROMOS++.

<table>
<thead>
<tr>
<th>Program</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>follow</td>
<td>Creates a three-dimensional trace of selected atoms through time. The program takes the nearest-image with respect to the previous atom position.</td>
</tr>
<tr>
<td>gathtraj</td>
<td>Gathers a trajectory using the specified gathering method.</td>
</tr>
<tr>
<td>hbond</td>
<td>Monitors the occurrence of two- and three-centered hydrogen bonds.</td>
</tr>
<tr>
<td>int_ener</td>
<td>Recalculates the non-bonded interaction energy between two non-overlapping sets of solute atoms using the interaction parameters specified in the molecular topology.</td>
</tr>
<tr>
<td>iondens</td>
<td>Calculates the average density of ions (or other particles) in space from a molecular trajectory file.</td>
</tr>
<tr>
<td>jepot</td>
<td>Computes the $^3J$-coupling local elevation (LE) potential from a LE $^3J$-coupling restrained simulation.</td>
</tr>
<tr>
<td>jval</td>
<td>Generates time series of $^3J$-coupling constants based on a molecular trajectory.</td>
</tr>
<tr>
<td>m_widom</td>
<td>Calculates the free energy of inserting a test particle into configurations of a molecular system.</td>
</tr>
<tr>
<td>matrix_overlap</td>
<td>Calculates the overlap of two matrices (a mathematical definition of the overlap is given in the digital in-code documentation of GROMOS++).</td>
</tr>
<tr>
<td>mdf</td>
<td>For a given central set of atoms, mdf calculates the distance to the nearest atom belonging to a second set of atoms.</td>
</tr>
<tr>
<td>nhoparam</td>
<td>Calculates NH-order parameters from a simulation trajectory.</td>
</tr>
<tr>
<td>noe</td>
<td>Calculates and averages atom-atom distances. The trajectories originate either from a NOE distance restrained or free molecular simulation. The analysis may need pre-processing of data using the program prep_noe.</td>
</tr>
<tr>
<td>post_noe</td>
<td>Re-analyses of data generated by the program noe, resulting in NOE bound violations.</td>
</tr>
<tr>
<td>postcluster</td>
<td>Performs lifetime-analysis, combined clustering and writing of coordinates of (central) members of clusters, based on the output of the program cluster.</td>
</tr>
<tr>
<td>prep_noe</td>
<td>Converts X-plor NOE data formats to the GROMOS++ format (preparation for the noe program).</td>
</tr>
<tr>
<td>rdf</td>
<td>Calculates a radial distribution function for specified atoms.</td>
</tr>
<tr>
<td>rep_ana</td>
<td>Used for analysis of molecular replica exchange simulations.</td>
</tr>
<tr>
<td>rep_rewrite</td>
<td>Sorts replica exchanged trajectories according to the λ- or temperature values and writes them to different sorted files.</td>
</tr>
</tbody>
</table>

Table continued on the next page ...
### Tab. 4.5 (continued) Post-processing programs of GROMOS++.

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>reweight</td>
<td>Re-weights a time series of observed values of a quantity $X$ sampled during a simulation at state $R$, i.e. using the Hamiltonian $H_R(p, r)$, to another state $Y$ (neglecting kinetic contributions for simplicity).</td>
</tr>
<tr>
<td>rgyr</td>
<td>Calculates the radius of gyration for a specified set of atoms.</td>
</tr>
<tr>
<td>rmsd</td>
<td>Calculates the atom-positional root-mean-square deviation of a selected set of atoms.</td>
</tr>
<tr>
<td>rmsdmat</td>
<td>Calculates the positional root-mean-square deviation matrix for a set of structures. The output may be analysed by the program cluster.</td>
</tr>
<tr>
<td>rmsf</td>
<td>Computes the positional root-mean-square fluctuations for a specified set of atoms.</td>
</tr>
<tr>
<td>sasa</td>
<td>Calculates the solvent-accessible surface areas for selected atoms using the algorithm described by Lee and Richards [119].</td>
</tr>
<tr>
<td>sasa_hassel</td>
<td>Calculates the solvent-accessible surface areas using Hasel’s formula [120].</td>
</tr>
<tr>
<td>solute_entropy</td>
<td>Calculates the configurational entropy based on a coordinate trajectory.</td>
</tr>
<tr>
<td>tcf</td>
<td>Calculates distributions and time correlation functions.</td>
</tr>
<tr>
<td>tserr</td>
<td>Calculates time series of quantities.</td>
</tr>
<tr>
<td>tstrip</td>
<td>Removes solvent coordinates from a simulation trajectory.</td>
</tr>
<tr>
<td>visco</td>
<td>Calculates the bulk and shear viscosities.</td>
</tr>
</tbody>
</table>
Chapter 5

Computation of folding free energy differences of the amide-to-ester perturbed Pin1 WW domain using enveloping distribution sampling

The effect of removing a hydrogen bond donor from the backbone of the 34-residue WW domain of the protein Pin1 is investigated for 20 residues that are part of the three-stranded $\beta$-sheet fold of this protein in aqueous solution. Forty-eight molecular dynamics (MD) simulations of the wild-type protein and 20 amide-to-ester mutants started from the X-ray crystal structure and the NMR solution structure are analysed in terms of backbone-backbone hydrogen bonding and differences in free enthalpies of folding in order to provide a structural interpretation of the experimental chaotropic and thermal denaturation data available for this protein and the 20 mutants. The forty enveloping distribution sampling (EDS) simulations of the 20 mutants link the structural Boltzmann ensembles to relative free enthalpies of folding, $\Delta\Delta G_{\text{mw}}^{\text{fu}}$ between mutants and wild-type protein. The contribution of the different $\beta$-sheet hydrogen bonds to the relative stability of the mutants with respect to wild type cannot be directly inferred from thermal denaturation temperatures or free enthalpies of chaotropic denaturation for the different mutants, because some $\beta$-sheet hydrogen bonds show sizeable variation in occurrence between the different mutants.
Chapter 5. Folding free energy differences

5.1 Introduction

The prediction of the stable, dominant fold of a protein in aqueous solution from its amino-acid sequence at a particular thermodynamic state point of interest is still one of the long-standing challenges in biochemistry [130]. Although one can determine stable folded conformations by X-ray diffraction in a crystalline state or by NMR spectroscopy in solution, these stable conformations tell little about the driving forces of protein folding or about the types of interactions and motions that stabilise a particular fold under particular thermodynamic conditions of temperature, pressure, pH, ionic strength or solvent composition. Different interactions such as electrostatic and van der Waals interactions and entropy differences between folded and unfolded conformations may contribute to different extents, depending on the amino-acid composition of the protein and the composition of the solvent. For example, in aqueous solution the hydrophobic effect may drive apolar side chains out of the aqueous phase, or the excellent hydrogen-bond donor and acceptor properties of water may stabilise unfolded conformations which allow for solute-solvent hydrogen bonding relative to folded conformations in which unpaired hydrogen-bond donors and acceptors are present in the interior of a folded conformation. Although solute-solute hydrogen bonding is unlikely to drive protein folding because of the small energetic differences between solute-solute, solute-solvent and solvent-solvent hydrogen bonding for a protein in water, the existence of unpaired, i.e. not hydrogen-bonded donors and acceptors in a particular folded structure of a protein will generally destabilise such a structure compared to unfolded, solvent-exposed structures. The contribution of particular hydrogen bonds to the stability of a protein fold can be investigated by the construction of protein mutants which lack particular hydrogen bond donors or acceptors with a subsequent determination of the stability of the fold adopted by the mutant. Such an investigation can be carried out experimentally as well as computationally, and a combination of these two approaches may lead to an interpretation of the experimental data in terms of the importance of particular hydrogen bonds and local structures to the stability of the protein fold.

One way to remove a hydrogen-bond donor from the backbone of a protein is to replace an α-amino-acid residue with the corresponding α-hydroxy-acid residue, an amide-to-ester backbone substitution [131]. Such a substitution does not involve the side chain and preserves the stereochemistry of the residue. It replaces an amide NH, a hydrogen-bond donor, with an ester O, a rather weak hydrogen-bond acceptor, and an amide carbonyl, a rather strong hydrogen-bond acceptor, with an ester carbonyl, a weaker hydrogen-bond acceptor. Such amide-to-ester substitutions have been investigated experimentally [132–134] and computationally [135, 136] for different peptides and proteins.

In a systematic study of the effect of solute-solute hydrogen bonding upon the stability of the three-stranded β-sheet fold of the 34-residue WW domain of the protein Pin1 (Fig. 5.1), 20 amide-to-ester substitutions were carried out and the stability of the resulting mutants was measured in two ways, with respect to thermal denaturation and with respect to chaotrope denaturation [1]. Since the mutants need not necessarily to adopt the same
dominant folded structure as the wild-type WW domain, four experimental techniques, far-UV CD, fluorescence and 1D $^1$H NMR spectroscopy, and a ligand-binding assay, were used to establish the similarity between the stable folds of the WW domain and its 20 mutants [1]. Four mutants, R14ρ, F25ϕ, N26ν, and Q33θ, appeared to adopt structures different from the wild-type fold. They could only appear properly folded upon addition of trimethylamine N-oxide (TMAO), an osmolyte that can stabilise protein folds [137, 138]. The thermal denaturation experiments yielded midpoint temperatures $T_m$ for 16 mutants and the wild-type domain, while chaotrope denaturation experiments yielded free enthalpies of chaotrope denaturation, using guanine hydrochloride and TMAO, for 20 mutants and the wild type [1].

The processes of thermal or chaotrope denaturation of a protein are as yet inaccessible to computational methods, because the folding equilibrium of proteins cannot sufficiently be sampled in computer simulations covering 0.1 to 1 µs. Using current computer power this is only possible for short poly-peptides [141]. However, the relative stability of a particular fold for a protein and a mutant can be calculated through the use of so-called free-energy perturbation techniques in combination with a thermodynamic cycle linking the folded and unfolded conformations of the protein (Fig. 5.2). The free enthalpy differences

$$\Delta G_{fu}^x = \Delta G_x^f - \Delta G_x^u$$

(5.1)

between the folded (f) conformation and the unfolded (u) conformations for a mutant (x = m) and the wild-type (x = w) protein can be expressed in terms of free enthalpy differences

$$\Delta G_y^{mw} = \Delta G_y^m - \Delta G_y^w$$

(5.2)

between the mutant (m) and the wild-type (w) protein in the folded conformation (y = f) and the unfolded conformations (y = u),

$$\Delta \Delta G_{fu}^{mw} = \Delta G_{fu}^m - \Delta G_{fu}^w = \Delta G_{mfu}^m - \Delta G_{mu}^w ,$$

(5.3)

because the free enthalpy is a thermodynamic state function. As mentioned above, the quantities $\Delta G_{fu}^x$ are as yet computationally inaccessible for proteins, but the quantities $\Delta G_y^{mw}$ can be calculated using molecular simulation and free-energy perturbation or other techniques to compute free-energy differences between slightly different systems [142, 143].

A comparison of calculated $\Delta \Delta G_{fu}^{mw}$ values with differences in midpoint temperatures $T^m$ between mutants (m) and wild-type (w) protein,

$$\Delta T_{mw}^m = T_m^m - T_w^m ,$$

(5.4)

or with differences in chaotrope renaturation free enthalpies,

$$\Delta \Delta G_{c}^{mw} = \Delta G_c^m - \Delta G_c^w ,$$

(5.5)

is not straightforward because Eq. (5.3) contains free enthalpy differences $\Delta G_{fu}^x$ between
Chapter 5. Folding free energy differences

Figure 5.1: Position of the 11 backbone-backbone hydrogen bonds within the $\beta$-sheet region of the Pin1 WW domain (crystal structure [139]). The corresponding donor and acceptor atoms are listed in Tab. 5.1.
Table 5.1: List of the twenty mutated Pin1 WW proteins indicating the eliminated and weakened backbone hydrogen bonds (compare Fig. 5.1), separated by a “/”, the parameters $s$ and $\Delta E^R$ used in the EDS simulations based on the crystal [139] and NMR model [140] initial structures, the difference in experimentally derived midpoint of thermal denaturation curves [1] between the mutant and the native state, $\Delta T_{\text{mw}}$ Eq. (5.4), as well as experimentally ($\Delta \Delta G_{\text{mw}}$, Eq. (5.5)) [1] and computed differences of folding free energy differences between mutant and native Pin1 WW domain.

<table>
<thead>
<tr>
<th>system</th>
<th>affected H-bond</th>
<th>EDS parameters</th>
<th>calculated $\Delta \Delta G_{\text{mw}}^{\text{tri-Ala}}$ [kJ mol$^{-1}$]</th>
<th>exp. $\Delta \Delta G_{\text{mw}}^\text{c}$ $-\Delta T_{\text{mw}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$s$ $\Delta E^R$ [kJ mol$^{-1}$]</td>
<td>1PIN 2KCF</td>
<td>1PIN 2KCF</td>
</tr>
<tr>
<td>L7λ</td>
<td>-/-</td>
<td>0.134 0.130 0.2 -3.4</td>
<td>-40.2 ± 0.6 -41.7 ± 1.2</td>
<td>75.2</td>
</tr>
<tr>
<td>W11ω</td>
<td>1/-</td>
<td>0.076 0.079 -112.6 -112.5</td>
<td>73.7 ± 0.8 73.7 ± 1.3</td>
<td>38.7</td>
</tr>
<tr>
<td>E12ε</td>
<td>2/-</td>
<td>0.063 0.063 -83.2 -83.2</td>
<td>52.3 ± 1.2 49.8 ± 1.2</td>
<td>17.3</td>
</tr>
<tr>
<td>K13κ</td>
<td>-/3</td>
<td>0.060 0.046 -82.7 -259.0</td>
<td>44.6 ± 0.9 40.0 ± 2.5</td>
<td>9.5</td>
</tr>
<tr>
<td>R14ρ</td>
<td>4/-</td>
<td>0.063 0.067 -231.5 -89.7</td>
<td>34.8 ± 2.5 46.0 ± 2.2</td>
<td>11.0</td>
</tr>
<tr>
<td>M15μ</td>
<td>-/5</td>
<td>0.079 0.070 -52.3 -43.7</td>
<td>8.8 ± 0.8 10.7 ± 0.8</td>
<td>-26.3</td>
</tr>
<tr>
<td>S16σ</td>
<td>6/-</td>
<td>0.116 0.131 -143.9 -140.3</td>
<td>106.4 ± 0.5 104.2 ± 0.6</td>
<td>71.4</td>
</tr>
<tr>
<td>R17ρ</td>
<td>-/7</td>
<td>0.061 0.049 -49.4 -276.0</td>
<td>12.2 ± 1.3 -4.6 ± 2.5</td>
<td>-22.8</td>
</tr>
<tr>
<td>S19σ</td>
<td>7/-</td>
<td>0.106 0.111 -114.0 -113.4</td>
<td>83.1 ± 0.8 73.5 ± 1.4</td>
<td>38.4</td>
</tr>
<tr>
<td>V22ω</td>
<td>-/6</td>
<td>0.234 0.229 -82.0 -82.1</td>
<td>41.9 ± 0.2 42.6 ± 0.4</td>
<td>6.9</td>
</tr>
<tr>
<td>Y23ψ</td>
<td>5/-</td>
<td>0.070 0.070 -77.4 -74.4</td>
<td>70.3 ± 1.5 50.0 ± 1.1</td>
<td>15.0</td>
</tr>
<tr>
<td>Y24ψ</td>
<td>11/4</td>
<td>0.082 0.075 -92.9 -87.0</td>
<td>54.2 ± 1.8 52.3 ± 2.0</td>
<td>19.1</td>
</tr>
<tr>
<td>F25φ</td>
<td>3/10</td>
<td>0.110 0.101 -94.0 -94.1</td>
<td>50.3 ± 1.1 51.0 ± 1.0</td>
<td>15.2</td>
</tr>
<tr>
<td>N26ν</td>
<td>9/2</td>
<td>0.101 0.094 -39.7 -45.8</td>
<td>3.6 ± 0.7 4.2 ± 1.0</td>
<td>-31.5</td>
</tr>
<tr>
<td>H27η</td>
<td>-/8</td>
<td>0.078 0.083 -75.3 -79.9</td>
<td>37.8 ± 0.8 50.7 ± 1.4</td>
<td>2.7</td>
</tr>
<tr>
<td>N30o</td>
<td>8/-</td>
<td>0.088 0.100 -11.9 -14.3</td>
<td>-30.2 ± 0.6 -29.9 ± 0.9</td>
<td>-65.0</td>
</tr>
<tr>
<td>A51α</td>
<td>-/-</td>
<td>0.619 0.619 -81.5 -81.5</td>
<td>44.6 ± 0.4 44.1 ± 1.0</td>
<td>9.6</td>
</tr>
<tr>
<td>S32ρ</td>
<td>-/9</td>
<td>0.133 0.118 -145.1 -143.4</td>
<td>106.8 ± 0.3 107.0 ± 0.3</td>
<td>71.7</td>
</tr>
<tr>
<td>Q33θ</td>
<td>10/-</td>
<td>0.064 0.067 -63.2 -69.7</td>
<td>26.8 ± 1.0 29.0 ± 1.6</td>
<td>-6.0</td>
</tr>
<tr>
<td>W34ω</td>
<td>11/-</td>
<td>0.065 0.073 -115.0 -104.0</td>
<td>64.4 ± 2.0 63.3 ± 0.8</td>
<td>28.1</td>
</tr>
</tbody>
</table>
folded and unfolded conformations at one thermodynamic state point, while Eqs. (5.4) and (5.5) represent or contain free enthalpy differences of a system at different thermodynamic state points, differing in temperature in Eq. (5.4) or differing in solvent composition in Eq. (5.5). Although all three quantities $\Delta G_{fu}^{\text{mw}}$, $\Delta T_{mw}^{m}$ and $\Delta G_{c}^{\text{mw}}$ in one way or the other reflect the relative stability of the folded form of a mutant compared to wild-type, their values need not show strong correlation. This is illustrated in Fig. 5.3 where the 16 measured $\Delta T_{mw}^{m}$ values of [1] are correlated with the corresponding $\Delta G_{c}^{\text{mw}}$ values of Table 2 of [1]. The correlation coefficient is 0.87.

In the present study we report the calculation of 20 $\Delta G_{fu}^{\text{mw}}$ values for the 20 mutants of the Pin1 WW domain reported in [1]. To obtain $\Delta G_{w}^{\text{u}}$ values between two proteins, in this case an amide-to-ester mutant of the WW domain and the wild-type protein both in their folded, if stable, conformations. If the unfolded conformations of the wild-type and mutant proteins are to a large extent water exposed, the $\Delta G_{w}^{u}$ values for the 20 mutants should be very similar, because an amide-to-ester mutation does not involve an amino-acid side chain. Since it is impossible to sufficiently sample the unfolded state conformations of a 34-residue protein, we calculated as $\Delta G_{w}^{u}$ value the free enthalpy difference between an ester and an amide as the central residue of a Ala-tripeptide in water. So this approximation yields a single $\Delta G_{w}^{u}$ value for all 20 mutants. Alternatively, we treated this single $\Delta G_{w}^{u}$ value as a parameter $\Delta G_{0}^{u}$ that can be obtained from a fit of 20 calculated and experimental $\Delta G_{\text{mw}}^{\text{fu}}$ values,

$$\Delta G_{\text{fu}}^{\text{mw}} = \Delta G_{\text{mw}}^{f} - \Delta G_{0}^{u}.$$  (5.6)

Alternatively, one could assume that the free enthalpy $\Delta G_{w}^{u}$ between the mutant and wild-type proteins in the unfolded conformations is linearly related to the corresponding

**Figure 5.2:** Thermodynamic cycle for the Pin1 WW domain changing from unfolded (u) to folded (f) and wild-type (w) to mutant (m) states. The computation of the two free enthalpies changing the protein from the wild-type to the mutant state, $\Delta G_{\text{fu}}^{f}$ and $\Delta G_{\text{uw}}^{u}$, is feasible using EDS, while the two free enthalpy differences involving (un)folding of the protein cannot be estimated directly *in silico*. The difference in folding free enthalpy differences, $\Delta \Delta G_{\text{uw}}^{f}$, can be obtained using Eq. (5.3).
5.1. Introduction

Figure 5.3: Differences $\Delta T_{mw}^m$ between midpoints $T^m$ of thermal denaturation curves, versus differences $\Delta \Delta G_{mw}^c$ of folding free energy differences as obtained by chaotrope denaturation experiments [1] for the twenty mutants (m) and the wild type (w) Pin1 WW domain (Tab. 5.1).

Free enthalpy difference $\Delta G_{mw}^f$ in the folded conformations [144],

$$\Delta G_{mw}^f = \alpha \Delta G_{mw}^f + \Delta G_{1}^u,$$

(5.7)

which yields

$$\Delta \Delta G_{fw}^{fu}_{mw} = (1 - \alpha) \Delta G_{fw}^{fu}_{mw} - \Delta G_{1}^u.$$

(5.8)

In this case two parameters, $\alpha$ and $\Delta G_{1}^u$, are to be obtained from a fit of 20 calculated and experimental $\Delta \Delta G_{mw}^c$ values.

There exists not only uncertainty about the unfolded conformational ensemble, but also about the folded one. As one can see in Fig. 5.4, the crystal structure (PDB code 1PIN) of the WW domain derived from X-ray diffraction data of the whole Pin1 protein [139] and the structure of the WW domain in aqueous solution (PDB code 2KCF) derived from NMR spectroscopic data [140] are with a C$_\alpha$ backbone atom-positional root-mean-square difference (RMSD) of 0.34 nm and an all-atom RMSD of 0.42 nm not identical. Therefore, we conducted the molecular dynamics (MD) simulations of the wild type and some mutant proteins, and the EDS simulations of the 20 mutants in the folded state starting from each of these structures.
Figure 5.4: Pin1 WW configuration of a) the crystal structure (PDB entry code 1PIN [139]) and b) the final configuration of a 164.6 ns MD simulation with initial configuration based on the crystal structure as well as c) the NMR model structure (PDB entry code 2KCF [140]) and d) the final configuration after 133 ns of MD simulation starting from the NMR model structure.
The four experimental techniques used to determine whether the mutants would adopt a similar fold as the wild-type protein, CD, fluorescence and 1D $^1$H NMR spectroscopy and ligand binding, do not yield precise structural information on the mutants because these four signals are only sensitive to some average structural properties of the mutant. Thus local structural differences, e.g. in hydrogen-bonding, may not be reflected in different signals. Therefore, the structural differences between the different mutants and the wild-type protein as they occur in the simulations of these proteins were analysed. They may also contribute to differences in $\Delta \Delta G_{\text{mw}}^{\text{nu}}$ values as obtained from the EDS simulations of the 20 mutants.

5.2 Results

5.2.1 Structures and stability of the wild-type WW domain

Two simulations of the wild-type Pin1 WW domain have been performed only differing in the initial configuration used. In one case, the initial protein atom positions were derived from an X-ray structure (PDB code 1PIN [139]), while in the other simulation the initial configuration was based on a NMR model structure (PDB code 2KCF [140]). The two simulations are therefore referred to as 1PIN simulation or 2KCF simulation, respectively.

The X-ray structure and NMR model structure clearly differ from each other as shown in Fig. 5.4 and mentioned in Section 5.1 (root-mean-square deviations (RMSD) of the C$_\alpha$ atoms and all atoms of 0.34 nm and 0.42 nm, respectively). Interestingly, the two protein configurations seem to converge to similar configurations in the MD simulations (Fig. 5.4) but still show an RMSD of the C$_\alpha$ atoms of 0.29 nm between the two protein configurations after 30 ns of MD simulation.

Both, the 1PIN and 2KCF folds of the wild-type WW domain, are stable during the MD simulations and the protein does not unfold as indicated by the positional RMSD of the C$_\alpha$ or all protein atoms in Fig. 5.5. Note that the RMSD during the first ns of the equilibration period is different from zero, although the protein atoms were positionally restrained to their initial positions. This is the result of the energy minimisation of the protein in vacuo which was performed before the initialisation of the MD simulations, compare Section 5.5.2. In the first half of the 30 ns 1PIN simulation the C$_\alpha$ atoms only deviate about 0.2 nm from the X-ray structure, while the deviation from the NMR model structure is about 0.35 nm. A configurational change within the 1PIN simulation after about 16 ns leads to a similar RMSD value of the 1PIN simulation from both PDB structures, ending up with a value of about 0.35 nm. The RMSD value in the 2KCF simulation with respect to the NMR model structure is rather constant around 0.3 nm and generally smaller, about 0.07 nm, than the deviation from the X-ray structure. As expected, the RMSD values of all protein atoms are somewhat larger (about 0.1 nm) than the corresponding deviations of the C$_\alpha$ atoms only.
The evolution of secondary protein structure of the two wild-type WW domain MD simulations in time (Fig. 5.6) both show a triple-stranded $\beta$-sheet structure (residues 11-16, 21-26 and 31-33), but the 1PIN simulation maintains the structure slightly better for residues 16, 21 and 31. Residue 16 and 21 are both involved in the hydrogen bond No. 6 (Fig. 5.1).

Fig. 5.7 shows the secondary structure of six selected mutant proteins (L7$\lambda$, R14$\rho$, F25$\phi$, N26$\nu$, N30$\nu$ and Q33$\theta$) as a function of time. Only the two mutants L7$\lambda$ and N30$\nu$ fully maintain the triple-stranded $\beta$-sheet as reported in the X-ray structure [139] and seen in the end-state simulations of the wild-type WW domain (1PIN and 2KCF, see Fig. 5.6). The other four mutants either completely miss one $\beta$-strand (N26$\nu$ and Q33$\theta$) or show a reduced $\beta$-sheet structure involving much fewer residues than in the wild-type protein (R14$\rho$ and F25$\phi$). This is in agreement with spectriscopic and ligand-binding experiments which indicated that the latter four mutants were misfolded [1].
5.2. Results

Figure 5.6: Secondary structure in the two wild-type WW domain 1PIN (top) and 2KCF (bottom) MD simulations. Blue: β-strand according to [121].

5.2.2 Hydrogen bonding

Tab. 5.2 shows the occurrence of the 11 backbone hydrogen bonds of Fig. 5.1 reported in [1] during the 1PIN and 2KCF MD simulations as well as within the corresponding X-ray or NMR model structures. Hydrogen bonds No. 7 and 8 are not present in either of the experimentally derived protein structures and show a very low occurrence in both MD simulations. Hydrogen bond No. 6, which involves residues 16 and 21, and hydrogen bond No. 5, a close neighbour to hydrogen bond No. 6, show both a lower occurrence in the 2KCF simulation than in the 1PIN simulation of the wild-type, while hydrogen bond No. 1 is only moderately present in the 1PIN simulation and not present in the 2KCF simulation. Overall, only 7 out of the 11 backbone hydrogen bonds have an occurrence of at least 75% in both MD simulations, i.e. hydrogen bonds No. 2, 3, 4, 5, 9, 10, and 11. In the NMR model structure there are 6 backbone hydrogen bonds present (No. 2, 3, 6, 9, 10, and 11), while the X-ray structure shows all but hydrogen bonds No. 7 and 8.

The hydrogen-bonding behaviour calculated from the wild-type branch of the 20 mutants during the 1PIN and 2KCF EDS simulations is similar to the hydrogen-bond pattern shown in the wild-type domain end-state simulations. Hydrogen bonds No. 1, 7 and 8 are weakly or not present, while all others show generally an occurrence of 80-100%, indicating rather stable wild-type state structures of the WW domain during the EDS simulations. The occurrence of the hydrogen bonds involving the removed hydrogen-bond donors or weakened hydrogen-bond acceptors due to the reduced hydrogen-bond acceptor
### Table 5.2: Hydrogen-bond occurrence (%) of the 11 hydrogen bonds (compare Fig. 5.1) within the native branch during the EDS MD simulations starting from the two Pin1 WW domain structures (1PIN/2KCF) used in this work in the two PDB structures and in the wild-type and 20 mutant proteins. As described in Section 5.5.6 the occurrence can be >0% for hydrogen bonds which do not have a hydrogen bond donor in the indicated WW domain mutant. Bold numbers indicate hydrogen bonds not present in the mutants but still showing an occurrence >0% because calculated from the native branch of the EDS MD simulations. Weakened hydrogen-bond acceptors are in italics.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Hydrogen bond number</th>
<th>PDB structures</th>
<th>wt</th>
<th>L7i</th>
<th>W11</th>
<th>E12</th>
<th>K13</th>
<th>R14</th>
<th>M15</th>
<th>S16</th>
<th>R17</th>
<th>S19</th>
<th>V22</th>
<th>Y23</th>
<th>F25</th>
<th>N26</th>
<th>H27</th>
</tr>
</thead>
<tbody>
<tr>
<td>100/0</td>
<td>100/100</td>
<td>100/0</td>
<td>96/93</td>
<td>95/93</td>
<td>96/92</td>
<td>97/75</td>
<td>65/42</td>
<td>7/1</td>
<td>2/0</td>
<td>83/83</td>
<td>96/96</td>
<td>97/90</td>
<td>0/33</td>
<td>0/33</td>
<td>0/33</td>
<td>0/33</td>
<td>0/33</td>
</tr>
<tr>
<td>0/6</td>
<td>0/100</td>
<td>0/100</td>
<td>0/100</td>
<td>83/83</td>
<td>0/100</td>
<td>0/100</td>
<td>0/100</td>
<td>0/100</td>
<td>0/100</td>
<td>0/100</td>
<td>0/100</td>
<td>0/100</td>
<td>0/100</td>
<td>0/100</td>
<td>0/100</td>
<td>0/100</td>
<td>0/100</td>
</tr>
</tbody>
</table>
Figure 5.7: Secondary structure of the 6 indicated mutant 1PIN MD end-state MD simulations starting from the 1PIN X-ray structure. Blue: $\beta$-strand according to [121].

character of the ester carbonyl compared to the amide carbonyl indeed shows a slightly reduced occurrence compared to most other mutated WW domain proteins (bold and italic numbers in Tab. 5.2).

5.2.3 Free enthalpy differences

Tab. 5.1 shows an overview of calculated differences in folding free enthalpies, $\Delta\Delta G_{\text{hu}}^{\text{fw}}$ as well as the experimentally derived chaotropic denaturation values $\Delta\Delta G_{\text{hu}}^{\text{cmw}}$ (Eq. (5.5)) together with the differences of midpoints of thermal denaturation curves $\Delta T_{\text{hu}}^{\text{mw}}$ (Eq. (5.4)) reported in [1]. Calculated values have been obtained using Eqs. (5.3), (5.6) and (5.8).

As mentioned above, we calculated $\Delta G_{\text{hu}}^{\text{nu}} = 39.8 \pm 0.4$ kJ mol$^{-1}$ as the free enthalpy difference between an ester and an amide as the central residue of a Ala-tripeptide in water.
and used this value as an approximation of $\Delta G_{\text{mw}}^u$ in Eq. (5.3) for all unfolded mutants. The resulting $\Delta \Delta G_{\text{mw}}^{fu}$ values calculated from the EDS 1PIN and 2KCF simulation trajectories are reported in columns 7 and 8 of Tab. 5.1 and are generally much larger than the values reported in literature [1]. However, the $\Delta \Delta G_{\text{mw}}^{fu}$ values from the two individual EDS simulations (1PIN and 2KCF) are rather similar, exceptions are the mutants K13$\kappa$, R14$\rho$, R17$\rho$, S19$\sigma$, Y23$\psi$ and H27$\nu$, which show differences larger than 2.5 kJ mol$^{-1}$ ($k_B T$). These mutants can also be identified when looking at the convergence of the two $\Delta G_{\text{mw}}^f$ values during the 20 ns EDS simulations starting from the X-ray or NMR model structures as shown in Fig. 5.8. All other EDS simulations end up with generally rather converged $\Delta G_{\text{mw}}^f$ of similar size for both, the 1PIN EDS and the 2KCF EDS simulations. The convergence in the Ala-tripeptide EDS simulation is shown in Fig. 5.9.

Figs. 5.10 and 5.11 show the energy difference distributions between the amide and ester states for the reference state and the two end states for the 1PIN and 2KCF EDS simulations. Insufficient sampling of the ester or the amide state can be identified for the EDS simulations of R14$\rho$, Y23$\psi$ and W34$\omega$ (1PIN) as well as for K13$\kappa$ and R17$\rho$ (2KCF).

The reweighted end-state non-bonded solute-solute plus solute-solvent potential energy distributions are shown for all 1PIN and 2KCF EDS simulations in Figs. 5.12 and 5.13. Again, insufficient sampling of some of the states can be identified for the mutants R14$\rho$, Y23$\psi$ and W34$\omega$ (1PIN) and K13$\kappa$ and R17$\rho$ (2KCF). Additionally, the non-bonded solute-solute plus solute-solvent energy distribution as calculate from a standard (non-EDS) MD simulation of the wild-type protein (green course) is not congruent with the wild-type end-state energy distributions calculated from the EDS simulations (red curves) in a number of cases.

The positional root-mean-square deviation of the $C_\alpha$ atoms during the EDS simulation of all 20 mutants is shown in Figs. 5.14 and 5.15. All but the N30$\nu$ mutant show stable RMSDs with respect to the crystal and NMR model structures with fluctuations between 0.2 nm and 0.5 nm, which is not larger than the RMSDs of the two wild-type MD simulations shown in Fig. 5.5. The N30$\nu$ mutant ends up with a non-converged RMSD value of about 0.7 nm after 23 ns, indicating that this mutant is unfolding during the 1PIN EDS simulation. For the 2KCF EDS simulations, they are all stable and converge to an RMSD value of about 0.35 nm from the NMR model structure.

### 5.3 Discussion

The correlation between the $\Delta \Delta G_{\text{mw}}^{fu}$ values calculated using Eqs. (5.3), (5.6) and (5.8) that use three different approximations of $\Delta G_{\text{mw}}^u$, on the one hand and the $\Delta T_{\text{mw}}^m$ values inferred from chaotrope denaturation measurements or the $-\Delta T_{\text{mw}}^m$ values inferred from thermal denaturation measurements on the other hand is displayed in Figs. 5.16a–c and 5.16d–f. Comparing these correlations with the correlation between $-\Delta T_{\text{mw}}^m$ and $\Delta \Delta G_{\text{mw}}^c$ shown in Fig. 5.3, it seems that the two experimental quantities $-\Delta T_{\text{mw}}^m$ and $\Delta \Delta G_{\text{mw}}^c$ correlate slightly better with each other than with the calculated quantities $\Delta \Delta G_{\text{mw}}^{fu}$. 
Figure 5.8: $\Delta G_{\text{mol}}$ values (Eq. (5.3)) from the 1PIN (black) and 2KCF (red) EDS simulations as a function of simulation time for all 20 mutants of the Pin1 WW domain.
This is not wholly unexpected because the latter quantity is defined (Eq. (5.3)) in terms of free enthalpy differences $\Delta G_{fu}^x$ between folded and unfolded conformations of a protein or a mutant (x) at one thermodynamic state point, whereas the former two quantities are more similarly defined (Eqs. (5.4) and (5.5)) in terms of free enthalpy differences $\Delta G_c^x$ and midpoint temperatures $T_m^x$ derived for a protein or mutant (x) at different thermodynamic state points of chemical potential and temperature. Although chaotropic and thermal denaturation may affect the protein or mutant stability in different ways, weighting the number of folded versus unfolded conformations at one thermodynamic state point makes $\Delta G_{fu}^x$ a more different quantity. But, as was mentioned in the Introduction, all three quantities shed light on the differences in stability of the folded conformations of the mutants and the wild-type protein.

The quality of the calculated $\Delta \Delta G_{mw}^{fu}$ values will depend on the quality of the biomolecular force field employed and on the extent of sampling of the relevant protein conformations in the MD simulations. The standard MD simulations for the wild type and six mutant show that the triple-stranded $\beta$-sheet fold is maintained for the wild type (Fig. 5.6) and
5.3. Discussion

Figure 5.10: Energy difference distributions for the reference state, $\rho_R(\Delta V; \Delta V_{BA})$ (green), and of the two end states, $\rho_A(\Delta V; \Delta V_{BA})$ ($A =$ amide, red) and $\rho_B(\Delta V; \Delta V_{BA})$ ($B =$ ester, black), for each of the 20 amide-to-ester perturbations as obtained from the EDS simulations of the Pin1 WW domain. The initial structures of the simulations were based on the 1PIN crystal structure [139].
Figure 5.11: Energy difference distributions for the reference state $\rho_R(\Delta V; \Delta V_{BA})$, and of the two end states, $\rho_A(\Delta V; \Delta V_{BA})$ ($A =$ amide, red) and $\rho_B(\Delta V; \Delta V_{BA})$ ($B =$ ester, black), for each of the 20 amide-to-ester perturbations as obtained from the EDS simulations of the Pin1 WW domain. The initial structures of the simulations were based on the 2KCF NMR model structure [140].
Figure 5.12: Non-bonded solute-solute plus solute-solvent energy distributions of the 20 EDS mutant end states (black: ester; red: amide) obtained through reweighting from the 20 EDS reference state simulations and from an independent MD simulation of the wild-type Pin1 WW domain (green). The initial structures of the simulations were based on the 1PIN crystal structure [139].
Figure 5.13: Non-bonded solute-solute plus solute-solvent energy distributions of the 20 EDS mutant end states (black: ester; red: amide) obtained through reweighing from the 20 EDS reference state simulations and from an independent MD simulation of the wild-type Pin1 WW domain (green). The initial structures of the simulations were based on the 2KCF NMR model structure [140].
Figure 5.14: Positional root-mean-square deviation of the backbone C\textsubscript{\alpha} atoms within the 20 mutant proteins during the EDS MD simulation with respect to the crystal structure (black) [139] and the NMR model structure (red) [140]. The initial configuration of the MD simulation was derived from the crystal structure [139]. The initial 3 ns are from the equilibration using MD without EDS.
Figure 5.15: Positional root-mean-square deviation of the backbone Cα atoms within the 20 mutant proteins during the EDS MD simulation with respect to the crystal structure (black) [140] and the NMR model structure (red) [139]. The initial configuration of the MD simulation was derived from the NMR model structure [140]. The initial 3 ns are from the equilibration using MD without EDS. The initial 10 ns are from the MD simulation with EDS.
Figure 5.16: Difference of folding free enthalpy differences $\Delta\Delta G_{mw}^c$ as obtained by chaotrope denaturation experiments [1] versus differences of folding free enthalpy differences $\Delta\Delta G_{mw}^{fu}$ (top graphs) and differences $-\Delta T_{mw}^m$ between midpoints $T^m$ of thermal denaturation curves versus differences of folding free energy differences $\Delta\Delta G_{mw}^{fu}$ (bottom graphs). Calculations of the $\Delta\Delta G_{mw}^{fu}$ values have been done according to a)/d) Eq. (5.3), b)/e) Eq. (5.6), and c)/f) Eq. (5.8). Circles and crosses in graphs a) and d) distinguish between 1PIN and 2KCF simulations, respectively. Dotted lines are intended to guide the eye along $\Delta\Delta G_{mw}^c = \Delta\Delta G_{mw}^{fu}$. The correlation coefficients are (from a to e, data for 1PIN and 2KCF separated by a "/"): -0.09/-0.04, -0.06, -0.07, 0.19/0.15, 0.14, -0.15
mutants L7λ, N30ν (Fig. 5.7), whereas it is partly lost for the mutants R14ρ, F25ϕ, N26ν and Q33θ (Fig. 5.7) in agreement with the experimental observations of [1]. In all but one (N30ν starting from the X-ray structure) of the 42 EDS simulations of the wild type and mutants the atom-positional deviations from the X-ray and NMR structures (Figs. 5.14 and 5.15) is limited to 0.2-0.5 nm. This does not indicate any major force-field problems with respect to maintaining a proper fold. The analysis of the EDS simulations in terms of convergence of $\Delta G_{nw}^{u}$ values as function of time and between the pairs of simulations starting from X-ray and NMR structures (Fig. 5.8) shows converged values for 14 mutants, while for K13κ, R14ρ, R17ρ, S19σ, Y23ψ, and H27η this convergence is not satisfactory. The energy difference distributions (Figs. 5.10 and 5.11) show undersampling for R14ρ and Y23ψ in the EDS simulations starting from the X-ray structure (Fig. 5.10) and for K13κ and R17ρ for the EDS simulations starting from the NMR structure (Fig. 5.11). The comparison of non-bonded solute-solute plus solute-solvent end-state potential energy distributions for the wild type in the EDS simulations (red lines in Figs. 5.12 and 5.13) with that of the standard MD simulation of this end state (green lines in Figs. 5.12 and 5.13) show agreement except for the EDS simulations of mutants R14ρ and W34ω starting from the X-ray structure (Fig. 5.12) and of mutants K13κ and R17ρ starting from the NMR structure (Fig. 5.13). So only a few of the 40 EDS simulations of the 20 mutants show sampling and convergence problems.

The rather large differences between the $\Delta \Delta G_{nw}^{fu}$ values calculated using Eqs. (5.3), (5.6) or (5.8), i.e. using three different assumptions for the value $\Delta G_{nw}^{u}$, the free enthalpy difference between mutant and wild-type protein in the unfolded conformations, show that the assumption of an identical $\Delta G_{nw}^{u}$ value for all 20 mutants as used in Eqs. (5.3) and (5.6) is not justified. The assumption that $\Delta G_{nw}^{u}$ is proportional to $\Delta \Delta G_{nw}^{f}$ reduces the differences between the various $\Delta \Delta G_{nw}^{fu}$ values and brings them closer to the experimental $\Delta \Delta G_{nw}^{uc}$ values, see Fig. 5.16c. This figure shows that the four mutants R14ρ, F25ϕ, N26ν, and Q33θ for which experimentally no thermal denaturation midpoint temperature could be determined and which could only be kept folded by adding TMAO to the solution, display the largest discrepancy between the experimental $\Delta \Delta G_{nw}^{uc}$ values and the calculated $\Delta \Delta G_{nw}^{fu}$ values. So, the key to a precise calculation of folding free enthalpies lies in an accurate description of the unfolded state [145, 146].

5.4 Conclusion

Forty-eight standard MD and EDS simulations of the wild-type Pin1 WW domain and 20 of its amide-to ester mutants in aqueous solution have been used to shed light on the relation between protein folding free enthalpies and protein structure. Since experimental structural data were only available in the form of an X-ray crystal structure and an NMR solution structure of the wild-type protein showing roughly the same three-stranded $\beta$-sheet fold, but not for the 20 mutants, the simulated configurational ensembles could complement the experimental data by providing a structural interpretation of these.
5.5. Computational Methods

A direct comparison of experimental relative free enthalpies of folding upon chaotropic denaturation, $\Delta \Delta G^c_{mw}$, of experimental relative thermal denaturation temperatures, $-\Delta T^m_{mw}$, and of calculated relative free enthalpies of folding at a given thermodynamic state point, $\Delta \Delta G^f_{mw}$, with each other is hampered by the different definitions of the corresponding free enthalpy differences. Moreover, for a few mutants the configurational sampling within 20 ns appeared to be incomplete. Nevertheless, the simulations could identify the least stable mutants in agreement with the experimental data. The assumption of a single value for the relative free enthalpy $\Delta G^u_{mw}$ between mutant and wild type in the unfolded conformations turned out to be too crude. Assuming that this $\Delta G^u_{mw}$ is proportional to $\Delta G^f_{mw}$, the relative free enthalpy between mutant and wild type in the folded conformations, improved the agreement with the experimental data, but still seems to be a too crude assumption. Although the amide-to-ester mutation only involves the backbone, the presence of the nearby side chains may induce differences in $\Delta G^u_{mw}$ values. However, obtaining these $\Delta G^u_{mw}$ values from simulations of the mutant proteins or fragments of them in unfolded conformations lies currently beyond our computational means.

The significance of local structural variations even in the folded conformations is illustrated by the variation in backbone-backbone hydrogen bonding between X-ray and NMR structure of the wild type and between the different mutants. Hydrogen bonds 1, 5 and 6 (Fig. 5.1) involving residues 8, 11, 14, 16, 21 and 23 show sizeable variation in occurrence, and hydrogen bonds 7 and 8 involving residues 16, 19, 26 and 30 show rather low occurrence for the various mutants. This illustrates that an interpretation of the experiemtal $\Delta \Delta G^c_{mw}$ values and $-\Delta T^m_{mw}$ values in terms of the strength of particular hydrogen bonds is not straightforward. Such experimental data do not allow for a direct structural and mechanistic interpretation of the various forces stabilising the fold of a mutant or wild-type protein. Enveloping distribution sampling (EDS) simulations do provide a means to establish the relation between the Boltzmann weighted structural ensembles of the mutants and wild-type protein on the one hand and the relative free enthalpies of folding, $\Delta \Delta G^f_{mw}$, between mutants and wild type on the other hand, which makes this recently proposed method suitable for a structural interpretation of free enthalpy data on proteins and their mutants.

5.5 Computational Methods

5.5.1 Simulation program and force-field parameters

All simulations of the wild-type or mutant proteins, including the EDS simulations, have been carried out using the GROMOS simulation software [11, 12, 117] and the 53A6 force-field parameter set [110] for the wild type proteins, while for the ester linkages in the mutant proteins the parameter set 53A6_0XY [147] was used. Water was modelled using the simple point charge (SPC) model [148].

All analysis of the simulation trajectories have been done using the tools and programs
Chapter 5. Folding free energy differences

5.5.2 Simulation protocol

Sixteen standard MD simulations were performed, namely for the wild-type PIN WW domain and for the mutants L7λ, R13ρ, F25φ, N26ν, N30ν and Q33θ starting from the X-ray (PDB code 1PIN [139]) and the NMR model structure (PDB code 2KCF [140]) as well as for the Ala-tripeptide in its native and perturbed, i.e. the central residue being an α-hydroxy amino acid. In addition, 41 EDS simulations were performed to obtain the quantities \( \Delta G_{\text{nw}}^x \) for each of the 20 mutant \( x = f \) starting from the X-ray and the NMR model structure and the Ala-tripeptide \( x = u \). In this work, we name the simulations either 1PIN or 2KCF simulations, indicating which PDB structure was used to generate its initial atom positions of the solute. The initial configuration of the Ala-tripeptide had \(-133^\circ\) and \(155^\circ\) as values for the \( \phi \)- and \( \psi \)-dihedral angles.

Initial structures and MD end-state simulations

Initial coordinates of the protein atoms were derived from the PDB structures mentioned above. Missing hydrogen atom positions were generated based on geometric criteria using the GROMOS++ program gch. The protonation states of the amino acid residues were chosen corresponding to pH 7, leading to a total protein charge of +4e. Histidine 27 was protonated according its its hydrogen-bonding environment, based on the atom positions defined in the X-ray or NMR model structure: \( N_\epsilon \) for the 1PIN simulations and \( N_\delta \) for the 2KCF simulations. The different WW domain structures were energy minimised in vacuo and solvated in cubic boxes of water [148] molecules, such that the minimum distance of a non-hydrogen protein atom to the box wall was 1.4 nm. This led to box-edge lengths of 6.0/6.3 nm and 6876/7839 water molecules (1PIN/2KCF simulations). To yield an overall charge neutrality of the periodic boxes, 4 water molecules were replaced by chloride anions using the GROMOS++ program ion. The solvent was energy minimised with the positions of the heavy atoms within the protein kept fixed. Initial velocities for the simulations starting from the energy minimised PDB structures were generated based on a Maxwell-Boltzmann distribution at 1 K with position restraining of the protein atoms with an initial harmonic force constant of 25000 kJ mol\(^{-1}\). The temperature was continuously raised up to 300 K, while the force constant of the position restraining was lowered, ending in a non-restrained protein simulation at 300 K after 1 ns of simulation time. This procedure was followed by another 1 ns simulation at constant volume and 300 K and a 1 ns simulation at 300 K and at a constant pressure of 1 atm, using the weak coupling algorithm [75] with corresponding coupling times of \( \tau_T = 0.1 \) ps and \( \tau_p = 0.5 \) ps, respectively. The protein and solvent were separately coupled to the heat bath and an estimated isothermal compressibility of 4.575 \( \cdot 10^{-4} \) (kJ mol\(^{-1}\) nm\(^{-3}\))\(^{-1}\) was used. All bond lengths and the bond angles of the water molecules were kept rigid by applying the SHAKE algorithm [14] with
a relative geometric tolerance of $10^{-4}$, allowing for an integration time step of 2 fs when solving Newton’s equations of motion using the leap-frog algorithm [8]. Nonbonded interactions (van der Waals and electrostatic) were handled adopting triple-range cut-off radii: interactions within the short-range cutoff of 0.8 nm were calculated every time step from a pair list that was generated every five steps, when also interactions between 0.8 nm and 1.4 nm were computed. The long-range electrostatic forces were represented by a reaction field with a relative permittivity [149] of $\epsilon_{RF} = 61$ outside the long-range cutoff of 1.4 nm. The centre of mass translation and rotation were removed every 2 ps to avoid a flying ice cube [76].

This equilibration procedure was continued for at least 20 ns at constant temperature and pressure, using the same simulation parameters as described above. The protein atom positions, atom velocities and system energies were saved every 5 ps for analysis.

The ester and amide end-state simulations of the Ala-tripeptide were performed using the same simulation procedure and parameters as described for the Pin1 WW mutants.

### 5.5.3 EDS reference-state simulations

The single topology approach [150] was used for the EDS simulations of the mutant WW domains and the Ala-tripeptide, bearing both states, ester and amide, in the so-called reference state. The most simple approach would have been to only perturb the peptidic N atom into an ester O atom and the backbone H atom into a dummy atom, since only non-bonded solute-solute and solute-solvent energies are used to compute the free enthalpies. However, differences in the bonded interactions (force constants, bond lengths, bond angles, torsional dihedral angles and different multiplicities) lead to small changes in flexibility and to different atom configurations of the ester linkage compared to the peptide linkage and are therefore not negligible. To account for these changes the topology was branched after the C$_\alpha$ atom of the previous residue and (re)unified at the carbonyl C atom, see Fig. 5.17.

The initial coordinates for the EDS simulations were generated from the final configurations of the corresponding (1PIN or 2KCF) wild-type or Ala-tripeptide MD equilibration, see above, by copying the necessary number of atom coordinates and velocities to cover the doubly branched reference state. To find the EDS smoothness parameter $s$ and energy offset parameter $\Delta E^R$ a standard update scheme [151] was used. The resulting parameters for the different mutant EDS simulations are listed in Tab. 5.1.

After the update procedure an EDS simulation of 20 ns was performed with constant EDS parameters (Tab. 5.1). Note that these runs are not a continuation of the EDS update simulations with frozen EDS parameters, since the previously generated EDS configurations were reused. Energies were saved every 0.2 ps for further analysis and the computation of $\Delta \Delta G_{mw}$.
Chapter 5. Folding free energy differences

Figure 5.17: Single topology as used in the EDS simulations of the WW domain mutants. It governs the two states $A = \text{ester}$ and $B = \text{wild-type}$. The numbers in the upper right corner of each atom name indicate the atom charge (in e), while dashed lines separate the charge groups. Bonded and Lennard-Jones parameter types for the ester branch can be found in Tables 1 and 5 of [147]. The native branch uses the unchanged 53A6 force-field parameter set [110].

5.5.4 Computation of free-enthalpy differences

The free enthalpy of a system in the isothermal-isobaric ensemble (constant number of particles, pressure and temperature) is

$$G = -\frac{1}{\beta} \ln Z$$  \hspace{1cm} (5.9)

where $\beta$ is the inverse of the temperature $T$ multiplied by the Boltzmann constant $k_B$ and $Z$ denotes the partition function of the system. The difference in free enthalpy between two states $B$ and $A$ of a system is therefore calculated as

$$\Delta G_{BA} = G_B - G_A = -\frac{1}{\beta} \ln \frac{Z_B}{Z_A}$$  \hspace{1cm} (5.10)

Instead of perturbing the system in its state $A$ directly to its state $B$, a non-physical intermediate state $R$, a so-called reference state, can be defined, which allows to compute the free enthalpy difference

$$\Delta G_{BA} = G_{BR} - G_{AR} = -\frac{1}{\beta} \ln \frac{Z_B Z_R}{Z_A Z_R} = -\frac{1}{\beta} \ln \frac{\langle e^{-\beta(V_B-V_R)} \rangle_R}{\langle e^{-\beta(V_A-V_R)} \rangle_R}.$$  \hspace{1cm} (5.11)
from a single simulation of this reference state \( R \), where \( \langle \ldots \rangle_X \) denotes an ensemble average of state \( X \). In enveloping distribution sampling (EDS) this reference state is chosen such that its configurational ensemble envelops all important configurations of states \( A \) and \( B \). The potential energy term of the Hamiltonian of the reference state \( R \) reads

\[
V_R(r^N; s, \Delta E_{BA}^R) = -\frac{1}{\beta s} \ln \left[ e^{-\beta s (V_A(r^N)-E^R_A)} + e^{-\beta s (V_B(r^N)-E^R_B)} \right]
\]

(5.12)

where \( r^N \) denotes a configuration of the \( N \) particles in the system and \( V_X \) is the potential energy of state \( X \). Since the equations of motion depend on the energy offset difference \( \Delta E_{BA}^R = E^R_B - E^R_A \) only, and not on the absolute quantities, \( E^R_A \) is set to zero for a two-state system. Therefore, only two parameters, the smoothness parameter \( s \) and the difference in energy offset \( \Delta E_{BA}^R \), have to be determined such as to ensure sufficient sampling of both states \( A \) and \( B \) during the EDS MD simulation of the reference state \( R \) [20].

A computation of the free enthalpy difference is also possible in terms of energy difference distributions \( \rho_A(\Delta V; \Delta V_{BA}) \) and \( \rho_B(\Delta V; \Delta V_{BA}) \) sampled in the \( A \) and \( B \) ensembles, respectively [152–154],

\[
\ln \frac{\rho_A(\Delta V; \Delta V_{BA})}{\rho_B(\Delta V; \Delta V_{BA})} = +\beta \Delta V - \beta \Delta G_{BA}.
\]

(5.13)

The energy difference distributions \( \rho_A \) and \( \rho_B \) are not directly accessible from the EDS simulation of the reference state \( R \),

\[
\rho_R(\Delta V; \Delta V_{BA}) = \langle \delta[\Delta V - (V_B - V_A)] \rangle_R,
\]

(5.14)

but can be obtained by reweighing this distribution to the \( A \) and \( B \) ensembles [155, 156],

\[
\rho_X(\Delta V; \Delta V_{BA}) \propto \rho_R(\Delta V; \Delta V_{BA}) e^{-\beta (V_X - V_R)},
\]

(5.15)

where \( X \) denotes either of the two states \( A \) or \( B \).

### 5.5.5 Fitting of the calculated \( \Delta \Delta G_{mw}^{fu} \) values to experimental \( \Delta \Delta G_{mw}^{ce} \) values

The computation of the \( \Delta \Delta G_{mw}^{fu} \) values according to Eqs. (5.6) and (5.8) was done by fitting these values obtained from the Pin1 WW mutant EDS simulations to the \( \Delta \Delta G_{mw}^{ce} \) values obtained from chaotrope denaturation [1] using a least-squares fit varying one parameter \( \Delta G_0^{au} \) when using Eq. (5.6) and two parameters \( \alpha \) and \( \Delta G_1^{fu} \) when using Eq. (5.8).

The \( \Delta G_{mw}^{fu} \) values used in the fit were chosen from the two sets of EDS MD simulations on the basis of a comparison of their convergence of their \( \Delta G_{mw}^{fu} \) values (Fig. 5.8), of their energy difference distributions (Figs. 5.10 and 5.11), and of their end-state wild-type non-
bonding potential energy distributions (Figs. 5.12 and 5.13). The Pin1 simulation data were used for mutants L7, W11ω, E12ε, K13κ, M15µ, S16σ, R17ρ, V22ϖ, Y24ψ, F25φ, N26ν, H27η, A31α, S32σ, Q33θ, and W34ω, and the 2KCF simulation data were used for the mutants R14ρ, S19σ, Y23ψ, and N30ν.

The parameters (Eqs. (5.6) and (5.8)) obtained from the fits are \( \Delta G^u_0 = 74.8 \text{ kJ mol}^{-1} \), \( \Delta G^u_1 = -7.2 \text{ kJ mol}^{-1} \), and \( (1 - \alpha) = -0.01 \).

The corresponding plots \( \Delta \Delta G_{mw}^{fu} \) versus \( \Delta \Delta G_{mw}^{c} \) and \( -\Delta T_{mw}^{m} \) are shown in Fig. 5.16.

### 5.5.6 Analysis

**Atom-positional RMSD**

The atom-positional root-mean-square deviation (RMSD) between two protein structures has been evaluated based on the \( C_\alpha \) backbone atoms or all atoms according to the following formula,

\[
RMSD(r^{N_a}, r_{\text{ref}}^{N_a}) = \sqrt{\frac{1}{N_a} \sum_{i=1}^{N_a} (r_i - r_{i,\text{ref}})^2}, \tag{5.16}
\]

where \( r^{N_a} = (r_1, r_2, \ldots, r_{N_a}) \) represents the positions of the atoms. In Eq. (5.16), \( N_a \) is the number of atoms considered, \( r_i \) the position of atom \( i \) in the first structure and \( r_{i,\text{ref}} \) the position of atom \( i \) in the second, reference structure. We used the X-ray structure [139] or NMR model structure [140] as reference structure if not mentioned differently. The RMSD at the beginning of the MD simulations is not zero because the X-ray and NMR model structures were energy minimised in vacuo prior to simulation. The calculation of RMSDs was done using the GROMOS++ program \texttt{rmsd}, which also performs a superposition of the centres of mass of the two structures to be compared as well as a rotational least-squares fit based on the \( C_\alpha \) backbone atom positions.

**Hydrogen Bonds**

Hydrogen bonds were defined according to a geometric criterion: a minimum donor-hydrogen-acceptor angle of 135° and a maximum hydrogen-acceptor distance of 0.25 nm. The \texttt{hbond} program of GROMOS++ [73] has been used to detect and monitor hydrogen bonds in the Pin1 WW domains.

The EDS simulation trajectories had to be transformed to a wild-type-like trajectory before analysing hydrogen bonds, since we are not interested in hydrogen bonds from the ester branch to the wild-type branch. This was done by filtering all ester branch atoms off the EDS simulation trajectory. Therefore, in Tab. 5.2 we report an occurrence > 0% even for hydrogen bonds which are knocked out by the specific amide-to-ester mutation for some of the 20 mutants.
Detection of Secondary Structure Elements

The rules of Kabsch and Sander [121] have been applied to detect and monitor secondary structure elements in the native and ester-linked HEWL simulations. In some cases one residue may be assigned to be part of two different secondary structure elements. In order to avoid ambiguous assignments in such cases, the following priority rules were applied: β-strand/β-bridge > α-helix > π-helix > 3_{10}-helix > hydrogen bonded turn > bend.
Various experimental studies of hen egg white lysozyme (HEWL) in water and TFE/water clearly indicate structural differences between the native state and TFE state of HEWL, e.g. the helical content of the protein in the TFE state is much higher than in the native state. However, the available detailed NMR studies were not sufficient to determine fully a structure of HEWL in the TFE state. Different molecular dynamics (MD) simulations, i.e. at room temperature, at increased temperature and using proton-proton distance restraints derived from NMR NOE data, have been used to generate configurational ensembles corresponding to the TFE state of HEWL. The configurational ensemble obtained at room temperature using atom-atom distance restraints measured for HEWL in TFE/water solution satisfies the experimental data and has the lowest protein energy. In this ensemble residues 50–58, which are part of the β-sheet in native HEWL, adopt fluctuating α-helical secondary structure.
6.1 Introduction

2,2,2-Trifluoroethanol (TFE) and related co-solvents have been widely used in the study of protein structure, folding and misfolding [157]. TFE is particularly recognised for its ability to stabilise $\alpha$-helical secondary structure [158–161]. However its action depends on the peptide or protein sequence, the concentration of TFE and on the solution conditions used. Under certain conditions, TFE has also been seen to stabilise $\beta$-hairpins, turns and hydrophobic clusters, to accelerate protein folding, to facilitate the incorporation of peptides in membranes and to promote the formation of amyloid fibril aggregates [157, 162–166]. Industrially, TFE is employed as solvent in catalysis [167, 168].

TFE is thought to enhance intra-solute hydrogen bonding particularly in secondary structure regions, whilst weakening long-range tertiary structure of the protein [157, 161, 169]. Solution X-ray scattering studies suggest that an important factor is the clustering of the alcohol molecules in TFE aqueous solution [170]. NMR data, supported by MD simulations of peptides in explicit TFE/water solutions, show that this clustering results in an accumulation of TFE molecules around the peptide surface [171–174]. By coating the peptide surface they partially exclude water molecules [175]. It is suggested that this provides a local hydrophobic environment that promotes intrapeptide hydrogen bonding and maintains local hydrophobic interactions. For proteins an additional significant effect appears to be the penetration of TFE molecules into the protein interior [176].

Hen egg white lysozyme (HEWL) is one of the proteins whose conformation has been studied extensively in TFE solutions [177–182]. Low concentrations of added TFE co-solvent ($<10\%$ (v/v)) stabilise the native structure of the protein [177, 182]. However, further addition of TFE induces a cooperative transition to an intermediate state, referred to as the TFE state. The TFE state of HEWL contains substantial helical secondary in the absence of extensive persistent tertiary structure [177, 178, 182, 183]. Studies by SAXS show that the protein has an increased radius of gyration relative to the native state, although it is still relatively compact [181]. A particularly interesting feature is that the helical content of the protein in the TFE state is higher than that in the native state [177, 179]. The native state structure of HEWL consists of an $\alpha$-domain which contains four main $\alpha$-helices (A-D) and a C-terminal $3_{10}$-helix and a $\beta$-domain which contains a triple-stranded antiparallel $\beta$-sheet and a long loop. Detailed NMR studies show that in $70\%$ TFE/ $30\%$ water the regions corresponding to native helices in HEWL retain persistent helical conformation and increase in length, being extended at their termini. In addition, the C-terminal region of the $\beta$-sheet undergoes a conformational transition and adopts a more flexible helical structure for residues 50-58 [178, 179].

Despite detailed study, the NMR data available for the TFE state of HEWL are not sufficient on their own to define a structure for this state. In any case the data suggest that the TFE state will consist of an ensemble of interconverting partially folded conformers rather than a single well-defined structure [178, 180]. In this work we have therefore used an MD simulation approach to attempt to generate an ensemble of structures to describe the TFE state of HEWL. A set of MD simulations of HEWL in an explicit solvent of 70%
6.2. Materials and Methods

6.2.1 Simulation Software and Force-Field Parameters

All molecular dynamics (MD) simulations and the analysis of the simulation trajectories described in this work were performed using the GROMOS software package for (bio)molecular simulations [11, 13, 115] in combination with the GROMOS 53A6 force field [110]. Water was modelled using the simple-point-charge (SPC) model [148] and TFE using the model of Fiorini et al. [185]. Six configurational ensembles of the protein were generated by MD simulations under different conditions, see Tab. 6.1 and Fig. 6.1. The set of 50 NMR model structures that had been derived from NMR data of HEWL in aqueous solution [28] and are available in the PDB with entry code 1E8L were reformatted into GROMOS format and will be indicated as NMR\_pH7\_H2O.

6.2.2 Simulation Protocol

Initial coordinates of the the six different lysozyme simulations (see Fig. 6.1 and Tab. 6.1) were generated from the X-ray structure deposited at the Protein Data Bank (PDB) with PDB entry code 1AKI [186].

Missing hydrogen positions were generated based on geometric criteria using the GROMOS++ program gch [73]. The protonation states of the amino acid residues were chosen corresponding to pH 2 or pH 7 (Tab. 6.1 and Fig. 6.1). Histidine 15 was protonated according to its hydrogen-bonding environment and the pH value of the environment: at $N_\varepsilon$ for pH 7 and doubly protonated for pH 2 and all other HEWL structures. The four disulfide bridges were treated as covalently linked Cys residues. The different HEWL structures were energy minimised in vacuo and solvated in cubic boxes containing only water [148] molecules or a mixture of 30% water/70% TFE molecules, such that the minimum distance of a non-hydrogen protein atom to the box wall was 1.4 nm. This led to a box-edge length of 8.5 nm. At pH 7 HEWL has a net charge of +8 e, while at pH 2 this is +19 e. To yield an overall charge neutrality of the periodic boxes, 8 water molecules were replaced by chloride ions in the simulations at pH 7 and 19 water molecules at pH 2. The replacement was done using the GROMOS++ program ion [73]. The solvent was energy minimised with the positions of the heavy atoms within the protein kept fixed. Initial
### Simulation pH Solvent Initial T NOE Length

<table>
<thead>
<tr>
<th>Simulation name</th>
<th>pH</th>
<th>Solvent</th>
<th>Initial T</th>
<th>NOE</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH7 H2O</td>
<td>7</td>
<td>H2O</td>
<td>14355</td>
<td></td>
</tr>
<tr>
<td>pH2 H2O</td>
<td>2</td>
<td>H2O</td>
<td>19518</td>
<td></td>
</tr>
<tr>
<td>pH2 TFE</td>
<td>2</td>
<td>TFE</td>
<td>5411</td>
<td></td>
</tr>
<tr>
<td>pH2 HT</td>
<td>2</td>
<td>TFE</td>
<td>5411</td>
<td></td>
</tr>
<tr>
<td>pH2 HT DR</td>
<td>2</td>
<td>TFE</td>
<td>5411</td>
<td></td>
</tr>
<tr>
<td>pH2 HT DR</td>
<td>2</td>
<td>TFE</td>
<td>5411</td>
<td></td>
</tr>
</tbody>
</table>

#### Notes:
- **a** 70% TFE/30% H2O (v/v)
- **b** Energy-minimised X-ray structure 1AKJ [186]
- **c** Final configuration of simulation "MD pH2 TFE"
- **d** T is increased from 310 K to 400 K over 11.4 ns, then decreased to 310 K over the next 6.0 ns and then kept at 310 K for 2.6 ns
- **e** See Table 6.2
- **f** Configuration after 11.4 ns of simulation MD pH2 TFE HT DR
- **g** T is decreased from 400 K to 310 K over 2.0 ns and then kept at 310 K for 18.0 ns

---

**Table 6.1**: Overview the six MD simulations.
6.2. Materials and Methods

Figure 6.1: The different, sequential HEWL simulations. All initial coordinates have been generated from one of the two PDB files with PDB entry code 1AKI [186] or 1E8L [178] (black ellipses). For each simulated system (rectangles) the type of solvent and pH/protonation state/temperature/application of NOE distance restraints is indicated. Diamonds indicate the length of the simulation as well as the simulation temperature: blue: 310 K; blue→red: temperature increase from 310 K to 400 K; red→blue: cooling from 400 K to 310 K. Consecutive simulations can be identified following the flow chart. See also Tab. 6.1 for more details.

velocities for the three simulations starting from the energy-minimised X-ray structure were assigned from a Maxwell-Boltzmann distribution at 60 K with position restraining of the protein atoms with an initial harmonic force constant of 25000 kJ mol$^{-1}$ nm$^{-2}$. The temperature was stepwise raised while the force constant of the position restraining was lowered, ending in a non-restrained protein simulation at 310 K after 100 ps of simulation time. A 20 ns MD simulation followed this 100 ps thermalisation procedure. The protein configurations were saved for analysis every 0.5 ps.

All but the HT simulations were held at constant temperature and pressure (310 K, 1 atm) using the weak coupling algorithm [75] with corresponding coupling times of $\tau_T = 0.1$ ps and $\tau_p = 0.5$ ps, respectively. The protein and solvent were separately coupled to the heat bath. An estimated isothermal compressibility of $4.575 \cdot 10^{-4}$ (kJ mol$^{-1}$ nm$^{-3}$)$^{-1}$ and $2.227 \cdot 10^{-4}$ (kJ mol$^{-1}$ nm$^{-3}$)$^{-1}$ was used for systems with the pure water solvent and TFE/water mixture, respectively. All bond lengths and the bond angles of the water
molecules were kept rigid by applying constraints using the SHAKE algorithm [14] with a relative geometric tolerance of $10^{-4}$, allowing for an integration time step of 2 fs when solving the equations of motion using the leap-frog algorithm [8]. Nonbonded (van der Waals and electrostatic) interactions were handled adopting triple-range cut-off radii: interactions within the short-range cutoff of 0.8 nm were calculated every time step from a pair list that was generated every five steps, when also interactions between 0.8 nm and 1.4 nm were computed. The long-range electrostatic forces were represented by a reaction field with a relative permittivity [149] of $\varepsilon_{RF} = 61$ outside the long-range cutoff of 1.4 nm for the pure water solvent simulations. The simulations in TFE/water used a smaller relative permittivity of $\varepsilon_{RF} = 31$. The centre of mass translation and rotation were removed every 2 ps to avoid a flying ice cube [76].

The three simulations MD$_{\text{pH2}}$TFE-HT, MD$_{\text{pH2}}$TFE-DR and MD$_{\text{pH2}}$TFE-HT-DR used the same simulation parameters as described above but the initial atom positions and atom velocities were read from the final configuration of a preceding simulation. The high temperature simulations differ from the description above by a linear heating and/or cooling between 310 K and 400 K. In the DR simulations distance restraints with a harmonic force constant of 2500 kJ mol$^{-1}$ nm$^{-2}$ were applied to atom pairs from a set of NOE distance bounds derived from experiments. The six different sets of NMR NOE proton-proton distance bounds are specified in Tab. 6.2 and described below.

### 6.2.3 Sets of Derived NMR NOE Proton-Proton Distances

Two sets of NOE distance upper bounds derived from NMR experiments of HEWL in aqueous solution [28] and the other from NMR experiments on HEWL in TFE/water solution [184] were converted to GROMOS format and pseudo-atom bound corrections were added, see Tab. 6.2. A detailed description of the computational procedures with equations and distance corrections can be found in [42]. 1630 experimental NMR proton-proton upper distance bounds were taken from [28] and 1388 from [184] including pseudo-atom distance corrections as given by Wüthrich et al. [187]. Proton-proton distances were calculated using $1/r^3$ averaging, $r = (\langle r^{-3} \rangle)^{-1/3}$. The set NOE_TFE was used as distance restraints in simulations DR, Tab. 6.1. Different subsets of the two basic sets of NOE bounds were used for particular analyses.

### 6.2.4 Analysis

All analyses of the sets of protein structures have been done using the tools of GROMOS++ [73]. In particular, the following GROMOS++ programs have been used: \texttt{rmsd} (atom-positional root-mean-square deviation), \texttt{rmsf} (atom-positional root-mean-square fluctuations), \texttt{rgyr} (radius of gyration), \texttt{dssp} (detection of secondary structure in the protein), \texttt{prep_noe}, \texttt{noe} and \texttt{post_noe} (atom-atom distance bound violations).

When comparing or averaging quantities $Q$ that depend on the position of the centre
Table 6.2: Six NMR NOE proton-proton distance bound sets used as distance restraints in some simulations or for analysis of the set of NMR structures and MD generated configurational ensembles. The first three sets are specified in Table S1 and the second three in Table S2. Pairs \((i,j)\) are separated into short \((i \leq j \leq i+1)\), medium \((i+2 \leq j \leq i+4)\), and long \((j \geq i+5)\) ranged along the residue sequence indicated by \(i\) and \(j\).

<table>
<thead>
<tr>
<th>Name</th>
<th>Number of NOE distance bounds</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total</td>
<td>short</td>
</tr>
<tr>
<td>NOE$_{H_2O}$</td>
<td>1630</td>
<td>999</td>
</tr>
<tr>
<td>NOE$_{H_2O_NTFE}$</td>
<td>1209</td>
<td>622</td>
</tr>
<tr>
<td>NOE$_{H_2O_NTFE_42-60}$</td>
<td>134</td>
<td>85</td>
</tr>
<tr>
<td>NOE$_{TFE}$</td>
<td>1388</td>
<td>884</td>
</tr>
<tr>
<td>NOE$_{TFE_NH_2O}$</td>
<td>960</td>
<td>501</td>
</tr>
<tr>
<td>NOE$_{TFE_NH_2O_42-60}$</td>
<td>128</td>
<td>78</td>
</tr>
</tbody>
</table>

of mass and the spatial orientation of a particular set of atoms, the centres of mass are superimposed and a rotational least-squares fit of the positions of the set of atoms is performed before \(Q\) is calculated. The rules of Kabsch and Sander [121] have been applied to detect and monitor secondary structure elements in the HEWL simulations. In some cases one residue may be assigned to be part of two different secondary structure elements. In order to avoid ambiguous assignments in such cases, the following priority rules were applied: \(\beta\)-strand/\(\beta\)-bridge > \(\alpha\)-helix > \(\pi\)-helix > 3$_{10}$-helix > hydrogen bonded turn > bend.

6.3 Results and Discussion

Fig. 6.2 shows the radius of gyration of all seven sets of protein configurations as described in Tab. 6.1 and Fig. 6.1. The NMR model structures as well as the simulations in water and TFE/water at pH 7 or pH 2 and 310 K (NMR$_{pH7\_H_2O}$, MD$_{pH7\_H_2O}$, MD$_{pH2\_H_2O}$ and MD$_{pH2\_TFE}$) all have a radius of gyration around 1.4 nm: the NMR model structures at
Chapter 6. HEWL in TFE/water

Figure 6.2: Radius of gyration calculated for the backbone atoms of the seven sets of protein configurations. The simulation names are defined in Tab. 6.1 and Fig. 6.1. Note that for NMR\_pH7\_H2O the time scale does not have any meaning but the 50 configurational NMR model structures were evenly distributed over the 20 ns time scale length of the x-axis.

pH 7 are slightly more compact and the simulation in TFE slightly less compact than the simulation in water, while at pH 2 in water the protein starts unfolding after about 19 ns. As expected, at high temperature (HT, see Tab. 6.1) the protein becomes less compact due to partial unfolding with increased sampling of protein configurations during the heating procedure, either during the indicated simulation, MD\_pH2\_TFE\_DR, or in the previous one, as is the case for MD\_pH2\_TFE\_HT\_DR. Both sets of protein configurations reach a maximum radius of gyration of about 2 nm but do not shrink back to 1.4 nm during or after the cooling sequence. Since the configuration of HEWL in water at pH 7 is clearly different [178] from the ones in TFE/water at pH 2, it is not surprising that the simulations with applied distance restraints derived from experimental NMR NOE proton-proton distances of HEWL in TFE/water at pH 2 show a different, increased radius of gyration with proceeding simulation time too. The simulation at 310 K using distance restraints (MD\_pH2\_TFE\_DR) reaches even larger values for the radius of gyration than the free MD\_pH2\_TFE\_HT simulation. This indicates that the water state and TFE state of HEWL are structurally different or, at least, a sizeable change in the protein size is required to get from one state to the other. This increased radius of gyration in TFE/water is also observed in small-angle X-ray scattering data [181] for HEWL.

Additional information about the observed structural changes can be gained from the secondary structure analysis (see Figs. 6.3-6.5). The two simulations at pH 2 in water and TFE/water as well as the simulation in water at pH 7 (MD\_pH2\_H2O, MD\_pH2\_TFE, and MD\_pH7\_H2O, Fig. 6.4) all show the characteristic helices and $\beta$-strands in agreement with the secondary structure analysis of the NMR model structures (compare figure...
6.3. Results and Discussion

Figure 6.3: Secondary structure elements [121] calculated for the 50 configurations of the set of NMR model structures [28] derived from NMR experiments of HEWL in aqueous solution at pH 7.

Fig. 6.3). A transition from the native state of HEWL, as it is present in aqueous solution at pH 7, to the TFE state at pH 2 reported by [177, 178], especially the transformation of the \(\beta\)-sheet region (residues 42-60) to an \(\alpha\)-helix for residues 50-58 (compare figure 9 a/b of [178]), is not seen in any of the three simulations mentioned above, although in MD\(_{pH2\_H2O}\) the first \(\beta\)-strand is only intermittently present and the second one is shorter. Yet, a formation of a helix within the same residue range is not observed, even if the propensity for \(\alpha\)-helices seems to be higher in TFE/water at pH 2 than in water only, in particular at pH 7. The picture is different when enabling the expected \(\beta\)-strand-to-\(\alpha\)-helix transition using a temporally increased temperature or distance restraints based on experimentally derived NMR NOE proton-proton distance bounds of the TFE state of HEWL (MD\(_{pH2\_TFE\_HT}\) and MD\(_{pH2\_TFE\_DR}\), Fig. 6.5). The heating from 310 K to 400 K results in the complete loss of \(\beta\)-strands within residues 42–60 after about 10 ns, which happens even faster, within 0.5 ns, when applying NOE distance restraints. The heating procedure not only affects the \(\beta\)-domain of the protein but also leads to a temporary loss of the two helices within residues 110–125. However, the protein structure is not that much disrupted to inhibit a reformation of these two helices in the continuation of the simulation including the cooling procedure. Application of distance restraining dissolves the \(\alpha\)-helices around residues 30 and 110 and does not allow reformation of these helices. The loss of secondary structure is even larger when the heating/cooling procedure is followed by distance restraining (MD\(_{pH2\_TFE\_HT\_DR}\), Fig. 6.5), where all but two helices
Chapter 6. HEWL in TFE/water

Figure 6.4: Secondary structure elements [121] as a function of time calculated for (from top to bottom) the MD_pH7_H2O, MD_pH2_H2O and MD_pH2_TFE simulation trajectories. Red: α-helix; green: π-helix; black: 3_{10}-helix; blue: β-strand; yellow: β-bridge; brown: bend; grey: turn.

disappear and the transition of the β-strand region (residues 42–60) to an α-helix is not observed. This illustrates that the use of high temperature in an MD simulation strongly enhances the sampling, but it may be enhanced in the wrong or non-relevant direction in the high-dimensional conformational space, thereby leading to worse results than at low temperature.

Tab. 6.3 gives a quantitative overview of the positional root-mean-square difference between pairs out of five sets of protein configurations generated by MD simulation. The same information is qualitatively shown in Figs. 6.6–6.8. As we have seen from the analysis of the radius of gyration and the secondary structure, the two simulations at pH 2 without heating or distance restraining show the smallest deviation from each other
Table 6.3: Atom-positional root-mean-square deviation from the final configuration of one MD simulation to the final configuration of another MD simulation. Upper-right triangle: values for the C\textalpha atoms only. Lower-left triangle: values for all atoms of the protein.

<table>
<thead>
<tr>
<th></th>
<th>MD_pH2_H2O</th>
<th>MD_pH2_TFE</th>
<th>MD_pH2_TFE_HT</th>
<th>MD_pH2_TFE_DR</th>
<th>MD_pH2_TFE_HT_DR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD_pH2_H2O</td>
<td>0.00</td>
<td>0.41</td>
<td>1.43</td>
<td>1.68</td>
<td>1.40</td>
</tr>
<tr>
<td>MD_pH2_TFE</td>
<td>0.50</td>
<td>0.00</td>
<td>1.48</td>
<td>1.68</td>
<td>1.36</td>
</tr>
<tr>
<td>MD_pH2_TFE_HT</td>
<td>1.54</td>
<td>1.61</td>
<td>0.00</td>
<td>1.08</td>
<td>1.55</td>
</tr>
<tr>
<td>MD_pH2_TFE_DR</td>
<td>1.79</td>
<td>1.80</td>
<td>1.21</td>
<td>0.00</td>
<td>1.64</td>
</tr>
<tr>
<td>MD_pH2_TFE_HT_DR</td>
<td>1.50</td>
<td>1.46</td>
<td>1.67</td>
<td>1.74</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Figure 6.5: Secondary structure elements [121] as a function of time calculated for (from top to bottom) the MD_{pH2_TFE_HT}, MD_{pH2_TFE_DR} and MD_{pH2_TFE_HT_DR} simulation trajectories. Red: $\alpha$-helix; green: $\pi$-helix; black: $3_{10}$-helix; blue: $\beta$-strand; yellow: $\beta$-bridge; brown: bend; grey: turn.

(0.5 nm, see also Fig. 6.7b) and from the NMR model structure, see Figs. 6.6a and 6.6c. The difference from one of these two configurations to one of the HT or DR configurations is much larger, 1.46 nm–1.79 nm, depending on which HEWL structures are compared. Interestingly, the three HT or DR configurations have rmsd values ranging from 1.21 nm to 1.74 nm between each other, underlining the highly visible difference between these structures, compare also Fig. 6.9b. It also indicates that the different approaches in the simulations to obtain a TFE state configuration do not lead to the same result. This may reflect the wide diversity of conformations populated in the dynamic ensemble of a partially folded protein. The mentioned numbers result all from the rmsd calculations considering all protein atoms (lower-left triangle of Tab. 6.3) but the overall picture does
Figure 6.6: Final configurations of two MD simulations (in the case of the set of NMR structures the first one has been chosen) aligned and colored according to the positional root-mean-square deviation of the Cα atoms. Dark blue is good alignment, larger deviations are in red. The simulation names are defined in Tab. 6.1.

not change when basing the analysis on the Cα-atoms only in the upper-right triangle of Tab. 6.3. Therefore, the differences are not solely due to different configurations of the amino acid side chains in pure water or TFE/water, but must result from conformational changes of the protein backbone.
Figure 6.7: Final configurations of two MD simulations aligned and colored according to the positional root-mean-square deviation of the $C_\alpha$ atoms. Dark blue is good alignment, larger deviations are in red. The simulation names are defined in Tab. 6.1.

A further characterisation of the different sets of configurations of HEWL at pH 2 in TFE can be obtained using the corresponding NOE data. NOE proton-proton upper bound distances were calculated from all seven sets of HEWL protein configurations (see Tab. 6.1) and compared to six NMR NOE proton-proton distance bound sets derived by experiment, as defined in Tab. 6.2. The resulting NOE bound violations are listed in table Tab. 6.4. The protein configurations corresponding to the native state (NMR,pH7,H2O, MD,pH7,H2O) show the least violations of the H2O NOE bound set and show much more violations with
6.3. Results and Discussion

Figure 6.8: Final configurations of two MD simulations aligned and colored according to the positional root-mean-square deviation of the C$_\alpha$ atoms. Dark blue is good alignment, larger deviations are in red. The simulation names are defined in Tab. 6.1.

Figure 6.9: Final configurations of two MD simulations aligned and colored according to the positional root-mean-square deviation of the C$_\alpha$ atoms. Dark blue is good alignment, larger deviations are in red. The simulation names are defined in Tab. 6.1.
Table 6.4: Number of NOE bound violations of the seven sets of protein structures with respect to two experimental data sets of HEWL in water [28] and HEWL in a TFE/water mixture [184]. "/" separate all/short/medium/long ranged NOE bound distances as defined in Tab. 6.2. Each experimental data set is further split into “all” (containing the NOEs of the whole data set), “only” (containing the NOEs which occur in the corresponding set only) and “residues 42-60” (containing NOEs which involve atoms of residues 42-60).

<table>
<thead>
<tr>
<th>system</th>
<th>solvent</th>
<th>selection</th>
<th>number of NOE violations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>&gt; 0.1 nm</td>
</tr>
<tr>
<td>NMR, pH7, H2O</td>
<td>water</td>
<td>all</td>
<td>0/0/0/0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>only</td>
<td>0/0/0/0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>residues 42-60</td>
<td>0/0/0/0</td>
</tr>
<tr>
<td>TFE/water</td>
<td>all</td>
<td>167/50/111/6</td>
<td>70/15/49/6</td>
</tr>
<tr>
<td></td>
<td>only</td>
<td>167/50/111/6</td>
<td>70/15/49/6</td>
</tr>
<tr>
<td></td>
<td>residues 42-60</td>
<td>22/7/15/0</td>
<td>8/2/6/0</td>
</tr>
<tr>
<td>MD, pH7, H2O</td>
<td>water</td>
<td>all</td>
<td>89/12/28/49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>only</td>
<td>84/9/26/49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>residues 42-60</td>
<td>2/0/1/1</td>
</tr>
<tr>
<td>TFE/water</td>
<td>all</td>
<td>199/48/143/8</td>
<td>80/14/60/6</td>
</tr>
<tr>
<td></td>
<td>only</td>
<td>199/48/143/8</td>
<td>80/14/60/6</td>
</tr>
<tr>
<td></td>
<td>residues 42-60</td>
<td>23/6/17/0</td>
<td>7/2/5/0</td>
</tr>
<tr>
<td>MD, pH2, H2O</td>
<td>water</td>
<td>all</td>
<td>124/9/21/94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>only</td>
<td>121/7/20/94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>residues 42-60</td>
<td>10/0/4/6</td>
</tr>
<tr>
<td>TFE/water</td>
<td>all</td>
<td>179/51/121/7</td>
<td>75/14/54/7</td>
</tr>
<tr>
<td></td>
<td>only</td>
<td>179/51/121/7</td>
<td>75/14/54/7</td>
</tr>
<tr>
<td></td>
<td>residues 42-60</td>
<td>25/8/17/0</td>
<td>10/3/7/0</td>
</tr>
<tr>
<td>MD, pH2, TFE</td>
<td>water</td>
<td>all</td>
<td>101/9/17/75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>only</td>
<td>98/6/17/75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>residues 42-60</td>
<td>3/0/3/0</td>
</tr>
<tr>
<td>TFE/water</td>
<td>all</td>
<td>180/48/123/9</td>
<td>72/10/56/6</td>
</tr>
<tr>
<td></td>
<td>only</td>
<td>177/47/121/9</td>
<td>72/10/56/6</td>
</tr>
<tr>
<td></td>
<td>residues 42-60</td>
<td>27/7/20/0</td>
<td>9/3/6/0</td>
</tr>
<tr>
<td>MD, pH2, TFE, HT</td>
<td>water</td>
<td>all</td>
<td>448/32/89/327</td>
</tr>
<tr>
<td></td>
<td></td>
<td>only</td>
<td>428/23/78/327</td>
</tr>
<tr>
<td></td>
<td></td>
<td>residues 42-60</td>
<td>42/2/9/31</td>
</tr>
<tr>
<td>TFE/water</td>
<td>all</td>
<td>193/41/143/9</td>
<td>72/5/60/7</td>
</tr>
<tr>
<td></td>
<td>only</td>
<td>188/41/138/9</td>
<td>71/5/59/7</td>
</tr>
<tr>
<td></td>
<td>residues 42-60</td>
<td>13/8/5/0</td>
<td>1/0/1/0</td>
</tr>
<tr>
<td>MD, pH2, TFE, DR</td>
<td>water</td>
<td>all</td>
<td>387/20/61/306</td>
</tr>
<tr>
<td></td>
<td></td>
<td>only</td>
<td>378/14/58/306</td>
</tr>
<tr>
<td></td>
<td></td>
<td>residues 42-60</td>
<td>41/1/9/31</td>
</tr>
<tr>
<td>TFE/water</td>
<td>all</td>
<td>11/7/4/0</td>
<td>0/0/0/0</td>
</tr>
<tr>
<td></td>
<td>only</td>
<td>11/7/4/0</td>
<td>0/0/0/0</td>
</tr>
<tr>
<td></td>
<td>residues 42-60</td>
<td>0/0/0/0</td>
<td>0/0/0/0</td>
</tr>
<tr>
<td>MD, pH2, TFE, HT, DR</td>
<td>water</td>
<td>all</td>
<td>405/29/71/305</td>
</tr>
<tr>
<td></td>
<td></td>
<td>only</td>
<td>385/18/62/305</td>
</tr>
<tr>
<td></td>
<td></td>
<td>residues 42-60</td>
<td>44/4/11/29</td>
</tr>
<tr>
<td>TFE/water</td>
<td>all</td>
<td>21/10/11/0</td>
<td>1/0/1/0</td>
</tr>
<tr>
<td></td>
<td>only</td>
<td>17/9/8/0</td>
<td>0/0/0/0</td>
</tr>
<tr>
<td></td>
<td>residues 42-60</td>
<td>1/1/0/0</td>
<td>0/0/0/0</td>
</tr>
</tbody>
</table>
respect to the TFE NOE bound set. For the protein configurations corresponding to the TFE state, MD_{pH2\_TFE}, MD_{pH2\_TFE\_HT}, MD_{pH2\_TFE\_DR}, and MD_{pH2\_HT\_DR}, only for the latter three the opposite is the case. The three unrestrained simulations at 310 K (MD_{pH7\_H2O}, MD_{pH2\_H2O} and MD_{pH2\_TFE}) show that the change from pH 7 to pH 2 is raising the number of NOE bound violations with respect to the H2O NOE bounds set while only slightly reducing the number violations of the TFE NOE bound set. The secondary structure analysis showed that in MD_{pH2\_TFE\_HT} the three $\beta$-strands around residues 42-60 are dissolved and $\alpha$-helical configurations emerged. This is reflected in a significant reduction of the violations involving residues 42-60 of the TFE NOE bound set compared to the simulation MD_{pH2\_TFE}. However, the large number of NOE bound violations in regard to the NOE TFE set indicate that the configurations of MD_{pH2\_TFE\_HT} are not representative for the TFE state of the protein. The sets of configurations generated using NOE distance restraints for the TFE state do satisfy the NMR NOE TFE set of proton-proton distance bounds well, as expected, and as such both represent to some extent the configurational ensemble of the TFE state. Compared to the total number of NOE distance bounds in the two experimental NOE data sets the number of violations of the NOE TFE set for the DR configurations is even smaller than in the simulation of native HEWL in pure water at pH 7 with respect to the NOE\_H2O bound set (MD_{pH7\_H2O}, MD_{pH2\_TFE\_DR}, MD_{pH2\_TFE\_HT\_DR}: 5%, 1%, 2%). This does still not mean that we have found a unique configurational ensemble representing the TFE state of HEWL. Fig. 6.9b and Tab. 6.3 show an rmsd of 1.74 nm between the two final configurations of MD_{pH2\_TFE\_DR} and MD_{pH2\_TFE\_HT\_DR}. The number of 1388 experimental NMR NOE proton-proton upper bound distances in the NOE\_TFE set is far too low to fully determine the structure of a protein with as many degrees of freedom as HEWL.

Fig. 6.10 unifies the analysis of the NOE proton-proton distance bound violations with respect to both NOE bound sets NOE\_H2O and NOE\_TFE with the analysis of the secondary structure of the seven sets of protein configurations. The NMR model structures of HEWL at pH 7 in water do satisfy the NOE\_H2O bound set but not the NOE\_TFE bound set for the residue range 35-66 and 100-129. The simulation of HEWL at pH 7 in water shows some violations of the NOE\_H2O bound set, but reproduces the secondary structure, while showing violations of the NOE\_TFE set for about the same residue ranges as the NMR model structures. Applying NOE\_TFE distance restraints the secondary structure is partly lost, helix B is shortened or gone and the $\beta$-strands are lost, while the NOE\_TFE bounds are satisfied and the NOE\_H2O bounds are violated throughout the whole residue range.

The results from different analysis methods described above indicate that the two sets of DR configurations do satisfy the experimental data for the TFE state of HEWL. However, it is hard to decide which configuration set is the better one, since the various analysis methods do not yield a unique picture. Tab. 6.5 lists the protein-protein and protein-solvent potential energies for the seven sets of configurations.

At pH 7 MD simulation of HEWL in aqueous solution slightly lowers the solute-solute
energy compared to the set of NMR model structures. Lowering the pH from 7 to 2 results in an increase of the solute-solute energy which is due to the increase of the total charge of the solute from +8 e to +19 e. Considering the four trajectories of HEWL in TFE/water, MD.pH2.TFE has solute-solute and solute-solvent energies closest to MD.pH2.H2O. Raising the temperature leads to an increase in solute-solute energy. At higher temperature, the influence of the solvent is enhanced. Interestingly, the application of NOE distance restraints yields the lowest solute-solute energy and total energy of the protein. The restraints apparently focuses the sampling on energetically more favourable conformations of the protein. Comparing both distance restrained trajectories, the one at room temperature, MD.pH2.TFE, yields the lowest energies. So this conformational ensemble offers the best representation of the TFE state of HEWL.

Figure 6.10: NOE distance bound violations larger than 0.5 nm for atom pairs with their position within the protein residue sequence for the seven different sets of structures averaged over the final 4 ns of the simulation trajectories. Each system shows the secondary structure (*: helix, +: β-strand) of the corresponding residue. Residues with one ore more atoms involved in a violation of the experimental NOE sets are marked black. Violations of the experimental NOE bound set NOE_H2O of HEWL in water are above the secondary structure, while those of the NOE_TFE set are below the secondary structure.
6.3. Results and Discussion

Table 6.5: Protein-protein and protein-solvent potential energies for the seven sets of HEWL configurations. The energies calculated from MD simulation trajectories are averaged over the final 5 ns, while for NMR$_{\text{pH}7\text{H}_2\text{O}}$ the average is calculated over the 50 NMR model configurations.

<table>
<thead>
<tr>
<th>system</th>
<th>bonded</th>
<th>nonbonded</th>
<th>solute-solvent</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMR$_{\text{pH}7\text{H}_2\text{O}}$</td>
<td>4312</td>
<td>-11290</td>
<td>-6978</td>
<td>-</td>
</tr>
<tr>
<td>MD$_{\text{pH}7\text{H}_2\text{O}}$</td>
<td>3855</td>
<td>-11152</td>
<td>-7297</td>
<td>-12889</td>
</tr>
<tr>
<td>MD$_{\text{pH}2\text{TTE}}$</td>
<td>4454</td>
<td>-9463</td>
<td>-5009</td>
<td>-12204</td>
</tr>
<tr>
<td>MD$_{\text{pH}2\text{TTE}\text{HT}}$</td>
<td>4396</td>
<td>-9734</td>
<td>-5338</td>
<td>-11420</td>
</tr>
<tr>
<td>MD$_{\text{pH}2\text{TTE}\text{DR}}$</td>
<td>4476</td>
<td>-8405</td>
<td>-3929</td>
<td>-14064</td>
</tr>
<tr>
<td>MD$_{\text{pH}2\text{TTE}\text{HT}\text{DR}}$</td>
<td>5108</td>
<td>-11181</td>
<td>-6073</td>
<td>-13329</td>
</tr>
</tbody>
</table>

A comparison of calculated with measured [177] values of observables accessible by spectroscopic techniques such as far-UV CD or fluorescence spectroscopy for HEWL [177, 182, 183] or its constituting peptides [179, 182] would not be very conclusive for the following reasons: (i) CD spectra only indicate overall helical content, not which helices are present or not; (ii) fluorescence data only reflect modifications in the relative positions of the six Trp side chains of HEWL, but not the details with respect to the individual Trp residues; (iii) our MD simulations were done in 70% TFE in order to match the conditions of the NMR experiment [178], whereas the CD and fluorescence measurements were done at 15% [177], 50% [179], 40% [181], 0-50% [182] and 0-20% [183] TFE.

Fig. 6.11 shows the radial distribution functions from the centre of geometry of all protein atoms to the oxygen atoms of the solvent molecules for the six sets of HEWL configurations generated by MD simulations. There are almost no water molecules closer than 1 nm to the centre of the protein in pure water at pH 7. Lowering the pH to 2, the protein starts to unfold after 19 ns of simulation time (compare Fig. 6.2) and water molecules penetrate into the protein. A native-like fold in TFE/water at pH 2 (MD$_{\text{pH}2\text{TTE}}$) seems to prefer protein-TFE interaction over protein-water interaction as indicated by the peak of the first solvation layer of TFE around 0.5 nm. This is in agreement with results from previous MD simulations of peptides in explicit TFE/water solutions which showed an accumulation of TFE molecules at the peptide surface [171–174]. At high temperature or with distance restraints applied (MD$_{\text{pH}2\text{TTE}\text{HT}}$ and MD$_{\text{pH}2\text{TTE}\text{DR}}$) both solvents penetrate the protein. MD$_{\text{pH}2\text{TTE}\text{DR}\text{HT}}$ allows the highest solvent penetration, a result of the less compact and more unfolded protein configuration in this ensemble. As mentioned before, a detailed analysis of protein-TFE interactions and structural properties lies beyond the scope of the present study.
Figure 6.11: Radial distribution functions, $g(r)$, calculated from the final 5 ns of the trajectories of the six different MD simulations of HEWL. The protein-TFE (dotted lines) and protein-water (black lines) distances are the distances between the centre of geometry of all HEWL atoms and the TFE oxygen atoms or water oxygen atoms, respectively.
6.4 Conclusion

Using a set of 1388 NOE atom-atom distance bounds measured for HEWL at pH 2 in 70% TFE/30% water it was attempted to generate a configurational ensemble for this protein in its TFE state which would confirm the conversion of the $\beta$-sheet between residues 42 and 60 present in HEWL at pH 7 in water into an $\alpha$-helical segment of residues 50 to 58 in the TFE state.

Starting from a structure of HEWL at pH 7 in water and just changing the pH to 2 and the solvent to 70% TFE/30% water (called TFE) no major conformational changes were observed within 20 ns of MD simulation. The use of higher temperatures, up to 400 K, to enhance the sampling led to conformational changes and in particular cases to the emergence of $\alpha$-helical structures for residues 50-58 that had $\beta$-strand secondary structure at pH 7 in water, but the generated configuration did not satisfy the experimentally derived NOE distance upper bounds for HEWL in TFE/water. This could only be achieved by applying the set NOE_TFE of distance bounds as distance restraints in the MD simulations. The resulting sets of HEWL configurations do represent to some degree the configurational ensemble of the TFE state, but is by no means unique or complete, which is due to the limited number of NOE bounds, their distribution over the protein, to the limited sampling of configurational space within 20 ns of MD simulation and the wide conformational diversity expected in a partially folded state of this type. Yet, it represents the currently best possible representation of this ensemble.
Chapter 7

Molecular dynamics simulation of ester-linked hen egg white lysozyme reveals the effect of missing backbone hydrogen bond donors on the protein structure

The three-dimensional structure of a protein is stabilized by a number of different atomic interactions. One of these is hydrogen bonding. Its influence on the spatial structure of the hen egg white lysozyme is investigated by replacing peptide bonds (except those of the two proline residues) by ester bonds. Molecular dynamics simulations of native and ester-linked lysozyme are compared with the native crystal structure and with NOE distance bounds derived from solution NMR experiments. The ester-linked protein shows a slight compaction while losing its native structure. However, it does not unfold completely. The structure remains compact due to its hydrophobic core and a changed network of hydrogen bonds involving side chains.
7.1 Introduction

Since the first elucidation of the spatial structure of a protein, the desire to understand the particular interactions that hold a protein in its native fold has driven a great variety of investigations, both of experimental and of theoretical, and computational nature [188]. The issue is a complex one because different types of interactions play a role: hydrogen bonds between the amide (NH) and carbonyl (CO) atoms of the peptidic linkages, van der Waals interactions between tightly packed nonpolar side-chain atoms in the protein interior, hydrophobic effects due to the aqueous solvent environment, and other (de)solvation interactions [189]. Apart from these major interactions determining protein spatial structure, less important ones have been identified, such as salt bridges, hydrogen bonds involving side-chain atoms, and disulfide bridges [189]. A systematic investigation of the contribution of each of these types of interactions to the protein structure would separately eliminate each of these interactions and observe the subsequent changes in protein structure. Such a dissecting approach is easier to execute computationally than experimentally, because a computational study is less bound by chemical or physical limits. But, even computationally, it would be difficult to eliminate the hydrophobic effects without interfering with other interactions. An interaction that can be both computationally and experimentally eliminated with a minimal interference with other interactions that determine protein structure is hydrogen-bonding involving the NH and CO atoms of the peptide linkages in the polypeptide chain. Such hydrogen bonds are largely absent in \(\alpha\)-depsipeptides, naturally occurring peptides in which particular amino acid residues are replaced by hydroxy-acid residues [190], e.g., (S)-3-hydroxybutanoic acid in the case of a glycine residue, thereby replacing some of the backbone amide linkages by ester linkages. Examples of \(\alpha\)-depsipeptides are beauvericin, valinomycin, and aureobasidin A [191–193]. The amide and ester linkages both are planar due to electronic resonance and prefer the \textit{syn} orientation [194]. Their hydrogen-bonding properties are different though. While the amide (peptide) linkage may act as hydrogen bond donor (amide H-atom) as well as hydrogen bond acceptor (carbonyl O-atom), the ester linkage may exclusively function as hydrogen bond acceptor. Furthermore, the amide carbonyl O-atom is a substantially stronger hydrogen bond acceptor than the ester carbonyl O-atom [195, 196]. Thus \(\alpha\)-depsipeptides are analogues of \(\alpha\)-peptides with reduced hydrogen-bonding capabilities.

Amide-to-ester backbone substitution has been studied experimentally for the PIN WW domain, a three-stranded 34-residue protein [1, 197–199]. The folding kinetics and thermodynamics showed that the protein is most destabilized, when hydrogen bonds that are enveloped by a hydrophobic cluster are perturbed.

In the present study, we compared the structural properties of a well-known 129-residue protein, hen egg white lysozyme (HEWL), and its depsi analogue. Out of the many possible combinations of ester-with-peptide linkages, \(\sum_{k=1}^{126} \binom{126}{k}\), we chose the one with only ester linkages, in which all backbone hydrogen bond donors are eliminated. This will maximize the structural effects. When experimental data on HEWL with less than 126
ester linkages would become available, the study presented here could easily be repeated for the particular ester-linked analogue of HEWL.

A previous 3.5 ns molecular dynamics (MD) simulation of native HEWL in aqueous solution [200] is compared to that of 10 ns in which all peptide bonds, except those involving the two proline residues, have been replaced by ester bonds. Various structural properties are computed, and the sizeable effects of missing backbone-backbone hydrogen bonds clearly show up.

7.2 Computational methods

7.2.1 Molecular model

The MD simulation of ester-linked HEWL (10 ns in water) was performed using the GROMOS96 [42] simulation package and the GROMOS force-field parameter set 45A3 [72].

The initial structure was derived from a (native) crystal structure [186] by replacing the peptide bonds by ester bonds using force-field parameters as described by Gattin et al. [135]. This replacement was conducted for 126 of the 129 residues, all but 70Pro, 79Pro, and the terminal NH$_3^+$ . The system was solvated in a periodic, cubic box (box lengths: 8.54 nm) containing 19675 SPC [148] water molecules. The protonation states of protonizable amino acids corresponded to a pH of 7. This resulted in a system of in total 60228 atoms including eight Cl$^-$ ions to preserve the overall neutrality of the system.

7.2.2 Simulation protocol

At the beginning of the simulation, the velocities of the atoms were assigned from a Maxwell distribution at 300 K. After a thermalization and equilibration (100 ps), the MD simulation was performed at constant temperature (300 K) using the weak coupling algorithm [75] with corresponding relaxation time $\tau = 0.1$ ps. All bond lengths and the bond angles of the water molecules were kept rigid by applying constraints using the SHAKE algorithm [14] with a relative geometric tolerance of $10^{-4}$. This allowed an integration time step of 2 fs when solving the equations of motion of the system using the leap-frog algorithm [8]. Triple-range cutoff radii were used to treat long-range interactions (van der Waals and electrostatic): interactions within the short-range cutoff (0.8 nm) were calculated every time step from a pair list that was generated every fifth step, when also interactions between 0.8 nm and 1.4 nm were computed. The long-range electrostatic forces were represented by a reaction field [15] with a relative permittivity of $\varepsilon_{RF} = 61$ [149] outside the long-range cutoff of 1.4 nm. The centre of mass motion was removed every 2 ps. Configurations of the system were saved every 0.4 ps for analysis.
7.2.3 Analysis

All analyses were performed using the tools of the GROMOS96 simulation software [42]. They are based on two different MD simulations: the ester-linked HEWL simulation described above (10 ns) and an earlier performed (native) HEWL simulation (3.5 ns) [200] based on the same force-field parameter set (45A3) and initial structure [186].

A straightforward comparison of structures of the ester-linked protein with those of the native one is hampered by the absence of the peptide hydrogen atoms in the ester-linked protein. Therefore, we considered three slightly different procedures to generate virtual hydrogen atom positions using the positions of the ester O\(_\alpha\) atom, and its covalently bound neighbours C and C\(_\alpha\) in the ester linkage –CO–O\(_\alpha\)–C\(_\alpha\)–, where CO denotes the carbonyl group. All three procedures put the virtual hydrogen atom such that the two angles C–C\(_\alpha\)–H and H–O\(_\alpha\)–C\(_\alpha\) have the same value, and the O\(_\alpha\)–H distance is 0.1 nm. The three methods differ with respect to the plane in which the hydrogen atom will be placed. The procedure called COC places the virtual hydrogen atom in the C–O\(_\alpha\)–Ca plane, attached to the O\(_\alpha\) atom. The OCO procedure places the virtual hydrogen atom in the O–C–O\(_\alpha\) plane, and the so-called M procedure places the virtual hydrogen atom in the plane bisecting the C–O\(_\alpha\)–Ca and O–C–O\(_\alpha\) planes. If the dihedral angle \(\omega (C_\alpha – C – O_\alpha – C_\alpha)\) is 180° (or 0°), all three procedures yield the same hydrogen atom position. The dihedral angle distribution of the \(\omega (C_\alpha C Oa Ca)\) dihedral angle is shown in Fig. 7.1. It is much broader than the distribution of the \(\omega (Ca C N Ca)\) dihedral angle in native lysozyme. This means that the hydrogen atom positions generated by the three procedures may show differences for the

\[\text{Figure 7.1: The } \omega \text{ angle (native: } C_\alpha – C – N – C_\alpha, \text{ dashed line; ester-linked: } C_\alpha – C – O_\alpha – C_\alpha, \text{ solid line) is calculated from the 3.5 ns and 10 ns MD simulation of native HEWL and ester-linked HEWL, respectively. Calculations were performed for all residues except the two proline residues 70 and 79.}\]
\omega\text{-values in the tails of the distribution. However, a comparison of the three procedures with respect to structural properties obtained showed that they yield comparable results. Therefore, we only use the M procedure to place the virtual hydrogen atoms.}

The atom-positional root-mean-square differences (rmsd) between pairs of structures have been evaluated based on all backbone atoms (native: N, C, O, C\(_\alpha\); ester-linked: O\(_\alpha\), C, O, C\(_\alpha\)) of all 129 residues. The atom-positional root-mean-square fluctuations (rmsf) were also calculated for all 129 residues, but only considering the C\(_\alpha\) atoms of the backbone. In both cases (rmsd and rmsf), a translational superposition of the centers of mass of the protein and a least-squares rotational fit with respect to the crystal structure was applied. The solvent-accessible surface area (sasa) was computed using the algorithm proposed by Lee and Richards [119]. The volume of the protein was calculated in the following way: the structure of the protein (every time frame separately) was superposed on to an equilibrated box of pure SPC water. All water molecules overlapping with the atoms of the protein were removed from the box. Knowing the number of removed water molecules and the average volume of one water molecule, the protein volume can be estimated.

The radius of gyration of the protein (rgyr), a measure of the compactness of the structure that can be related to light-scattering intensity, was calculated using the definition

\[
R_{\text{gyr}} = \sqrt{\frac{1}{N_a} \sum_{i=1}^{N_a} (\mathbf{r}_i - \mathbf{R}_{\text{cm}})^2} \quad (7.1)
\]

with

\[
\mathbf{R}_{\text{cm}} = \frac{1}{M} \sum_{i=1}^{N_a} m_i \mathbf{r}_i \quad (7.2)
\]

and

\[
M = \sum_{i=1}^{N_a} m_i \quad (7.3)
\]

in which \(\mathbf{r}_i\) denotes the Cartesian position vector of atom \(i\), \(m_i\) is its mass, and \(N_a\) denotes the number of protein atoms.

The hydrogen bonds were analysed according to a geometric criterion: a minimum donor–hydrogen–acceptor angle of 135° and a maximum hydrogen–acceptor–distance of 0.25 nm [42]. Secondary structure assignments were performed using the rules defined by Kabsch and Sander [121].

Proton-proton distances were used to compare the simulations to experimental NMR data for native HEWL. These comprise upper bounds for distances between pairs of atoms that can be derived from intensities of peaks observed in NMR nuclear Overhauser effect (NOE) spectra. Out of a set of 1632 NOE upper bounds [28], 1630 were compared to the calculated \(^1\text{H}^\text{1}H\) distances averaged using \(1/r^3\) averaging, \(\langle r \rangle = (\langle r^{-3} \rangle)^{-1/3}\). Inter-proton distances involving (aliphatic) hydrogen atoms not treated explicitly in the simulations

\[
\omega = \sum_{i=1}^{N_a} m_i \mathbf{r}_i \quad (7.4)
\]
were calculated by defining virtual and pseudo-atoms [42, 187] during the analysis as described in [200, 201].

7.3 Results

7.3.1 General structural properties and flexibility

The root-mean-square deviations (rmsd) of the trajectory structures of native and ester-linked HEWL from the native X-ray structure are shown in Fig. 7.2, a. The values for the backbone and for all atoms in the native HEWL simulation slightly increase during the first 2 ns leveling off around 0.20 nm and 0.30 nm, respectively. The rmsd values of the ester-linked HEWL increase much faster over the first 5 ns of the simulation and then remain at values around 0.55 nm and 0.65 nm. This indicates that the ester-linked protein changes its spatial structure.

The convergence of the radius of gyration (rgyr), a measure of the compactness of the structure, calculated for the ester-linked HEWL simulation is also faster than that for the native HEWL simulation (Fig. 7.2, b) and seems to reach a lower value. The molecule is compacting rapidly. This is less clearly observed in the solvent-accessible surface area (sasa; Fig. 7.2, c) and protein volume (Fig. 7.2, d).

The appearance and disappearance of secondary-structure elements during the simulations of native HEWL and ester-linked HEWL is shown in Fig. 7.3. The three β-strands quickly disappear in ester-linked HEWL but show a comeback between 7.5 to 8.5 ns (residues 45, 46, and 50–53). The α-helices are preserved a bit longer but also disappear during the first 3 ns without returning later. In contrast, the π-helix (residues 111–115) survives during the whole simulation time of 10 ns, but appears to be structurally less stable than in native HEWL (occurrence in the X-ray structure/native HEWL/ester-linked HEWL: 100/71/43%). The loss of secondary structure is also illustrated in Fig. 7.4 which shows different structures along the MD trajectories.

The ϕ/ψ-angle distribution of residues 5–14, 25–36, 80–84, 89–101, 109–114, and 120–124 are shown in Fig. 7.5. These residues are the ones involved in an α-helix for at least 33% of the time during the 3.5 ns of the native HEWL simulation [201]. In the time window of 1.5–3.5 ns, 80% of all ϕ/ψ-angle combinations of the native HEWL simulation are in the region of an α-helix (Fig. 7.5, b). The ester-linked HEWL simulation shows a different ϕ/ψ-angle distribution during the same time period (Fig. 7.5, c). The distribution of the ϕ-angle is broader with its maximum value shifted from 60° to 90°. This broadening effect does not occur for the ψ-angle distribution. It becomes bimodal with a maximum at about the same position (−60°) and width as observed in the native HEWL simulation, but accompanied by a second maximum positioned at the inverse angle (60°) of the global maximum. This picture does not change during the simulation (Fig. 7.5, d).

Fig. 7.6 shows the rms fluctuations of the Cα atoms in the native and ester-linked
7.3. Results

Figure 7.2: a) Root-mean-square deviations (rmsd) of native HEWL and ester-linked HEWL from the initial X-ray structure. The values are calculated for the whole simulation time (native/ester-linked HEWL: 3.5/10 ns) considering all atoms (dotted lines) or backbone atoms only (native: N, C, O, Cα; ester-linked: Oα, C, O Cα; solid lines). b) Radius of gyration (rgyr) of backbone atoms (native HEWL: N, C, O, Cα (gray); ester-linked HEWL (black): Oα, C, O, Cα). The values are calculated through the whole simulation time of 3.5 ns (native HEWL) and 10 ns (ester-linked HEWL). c) The solvent accessible surface area (sasa), calculated for native HEWL (0 – 3.5 ns, dashed gray line) and ester-linked HEWL (0 – 10 ns, solid black line). d) Protein volume calculated for ester-linked HEWL (solid, black) and native HEWL (dashed, gray line) as well as a linear regression of the protein volume over the first 3.5 ns (native/ester-linked HEWL: dashed/solid line).

HEWL simulations. Calculations were performed over two different time windows of the same length, one from 1.5–3.5 ns (last 2 ns of the native HEWL simulation), the other from 8–10 ns (last 2 ns of the ester-linked HEWL simulation). The values are not significantly different when comparing the native HEWL to the ester-linked HEWL during the first time window. The most substantial differences appear for residues 80–100 and 116–129. The rms fluctuations calculated for the second time window (8–10 ns) show locally much
higher values than the native simulation (e.g., around residue 68), but are in general quite comparable. This confirms that the ester-linked protein is not completely unfolding and stays relatively compact.

### 7.3.2 NOE proton-proton distances

The differences in the structure between native and ester-linked HEWL can also be analyzed in terms of NOE atom-atom distance bound violations using the NOE bounds determined from NMR experiments on native HEWL [28]. The results are shown in Fig. 7.7 and compiled in Tab. 7.1.

Two sets of NOE bounds were used. Set 1 includes all 1630 NOE bounds, while set 2 contains 392 bounds between atoms in residues $i$ and $j$ with $j \geq i + 4$ only. The distance bound violations were also calculated for the crystal structure [186]. The total number of violations is much higher in the ester-linked structure than in the X-ray structure and most of the violations are observed for the long-range NOE bounds (set 2). Yet, there are only twelve violations larger than 0.5 nm. These involve only ten residues (12, 13, 17, 28, 88, 95, 98, 107, 108, and 129) and seven pairs of residues. Of these seven pairs, only two pairs (12–28 and 98–107) connect two secondary structure elements (helices A–B and C–$3_{10}$), the other five pairs connect helices (A, B, and C) with loops. Thus, less than 1%
7.3. Results

Figure 7.4: Schematic diagram showing a) the native HEWL crystal structure [186] and structures after 3.5 ns of simulations b) native HEWL, c) ester-linked HEWL and after 10 ns of simulation d) ester-linked HEWL. Colors represent: $\alpha$-helix (purple), $3_{10}$-helix (blue), $\beta$-strand (yellow), $\beta$-bridge (brown), turn (cyan) and coil (white), according to the rules implemented in VMD [202]. In the X-ray structure the four helices (A: residues 5–14; B: residues 25–36; C: residues 89–101; D: residues 109–114) are in purple, the three $3_{10}$-helices (residues 20–22, 80–84 and 120–124) in blue and the three $\beta$-sheets (residues 43–45, 51–53 and 58–59) in yellow, compare Table 1 of [201].

of the number of NOE bonds is violated by more than 0.5 nm. Or, 85% of all 1630 bounds are satisfied within 0.1 nm. Since the structure of ester-linked HEWL differs significantly from the native one (see Figs. 7.2 and 7.4), this means that a large part of the set of NOE bounds holds little spatial information regarding HEWL.

7.3.3 Hydrogen bonding

Because the ester-linked protein is not unfolding, one may ask whether intra-molecular hydrogen bonding is decreased upon replacing peptide NH groups by ester O$_\alpha$ atoms. Table S1 of the supplementary material of [203] lists the occurrence of the hydrogen bonds present in the crystal structure as well as in the native and ester-linked HEWL simulations. All calculations are based on the first 3.5 ns of the simulations. The hydrogen bonds involving the peptide NH moieties as donors and those with an occurrence of less than 10% were not listed.
Figure 7.5: Ramachandran plots based on a) the X-ray structure of native HEWL [186], b) the simulation trajectory of native HEWL from 1.5 – 3.5 ns and c) – d) the simulation trajectory of ester-linked HEWL calculated through time windows of 1.5 – 3.5 ns and 8 – 10 ns, respectively. The one-dimensional $\phi$- and $\psi$-distributions are plotted along the corresponding axes (at the outside of the graphs). Colors separate the $\phi/\psi$-combinations with respect to its populations, i.e. the grey part contains 10% of all $\phi/\psi$-combinations with the lowest populations while yellow represents the 10% with the highest population and the colors in between represent always 10% of all $\phi/\psi$-combinations with corresponding populations.

Overall, the degree of hydrogen bonding involving side chains is not changing significantly upon introduction of ester linkages (see Tab. 7.2). But, the pattern of hydrogen bonds does change. Only ca. 25% of the native hydrogen bonds is also seen in the ester-linked simulation and vice versa. Thus, one source of the loss of stability is the absence of the 126 peptide NH hydrogen bond donors in ester-linked HEWL. These are replaced by
7.4. Discussion

Figure 7.6: Root-mean-square positional fluctuations (rmsf) of $C_\alpha$ atoms through a simulation time window of 2 ns (solid line: ester-linked HEWL, 1.5 – 3.5 ns; dashed line: ester-linked HEWL, 8 – 10 ns; dotted line: native HEWL, 1.5 – 3.5 ns). The bars indicate the secondary structure of native HEWL observed in the crystal structure [186]: $\alpha$-helix (black), $\beta$-bridge (thin, black) and $\beta$-strand (grey).

126 $O_\alpha$ acceptor atoms, of which only two form a hydrogen bond with side-chain atoms (Table S1 in the supplementary material of [203]).

The result of the hydrogen bond analysis depends on the geometric criterion used. However, the use of looser distance and angle bounds turns out not to change the overall picture of the differences in hydrogen bonding between native and ester-linked HEWL.

7.4 Discussion

Tab. 7.1 and Figs. 7.2–7.7 contain information on the evolution of the native and ester-linked protein structures with time, as expressed in the form of various quantities. The picture that emerges indicates that a simulation of 10 ns is more than sufficient to equilibrate ester-linked HEWL when starting from the X-ray structure of native HEWL. An equilibration period of 1–2 ns seems sufficient. The larger changes are expected for ester-linked HEWL, but its properties during the analysis periods 1.5–3.5 and 8–10 ns appear rather similar.

The observation that ester-linked HEWL does not unfold and seems to be more compact than native HEWL is in accordance with experimental studies on the distribution of end-to-end distances and loop formation of polypeptide chains [204–206]. The slight compaction might be due to the hydrophobic effect, which may be able to reduce the hydrophobic surface area when the intra-protein hydrogen bonding is reduced due to the amide by ester-linkage replacement [1, 197], and, second, it might be due to the changed network of hydrogen bonds involving side chains. The former cause seems in contrast
Figure 7.7: Occurrence of $r^{-3}$-averaged $^1$H – $^1$H long-range NOE distances (residues $i$ to $j$ with $j \geq i + 4$) for (from top to bottom): the X-ray structure [186], the native HEWL simulation (1.5 – 3.5 ns) and the ester-linked HEWL simulation (1.5 – 3.5 ns and 8 – 10 ns). Positioning of the missing hydrogen atoms of ester-linked HEWL were done using the M procedure.

with the fact that the polypeptides studied by Kiefhaber and co-workers [187, 201, 204] are devoid of hydrophobic groups. But, compaction may have different causes for differently composed peptides. For this reason, we refrained from a detailed comparison with the results by Kelly and co-workers [1, 197–199] for the PIN WW domain. The relative stability of the $\pi$-helix may be due to its smaller surface-to-volume ratio and its smaller H-bond-to-chain length ratio compared to an $\alpha$-helix, which may make it less prone to destabilization due to reduced hydrogen bonding.

We have considered performing a conformational clustering analysis of each MD trajectory. Because the results will heavily depend on the criterion used and are thus less instructive than an analysis of hydrogen bonds, $\varphi$- and $\psi$-angles, and secondary structure, we have refrained from such an analysis.

Other perturbations than the replacement of amide linkages by ester linkages that would reduce the H-bonding capacity of the backbone have been considered. One could, for example, switch off all backbone H-to-O nonbonded interactions in the force field. However,
7.4. Discussion

Table 7.1: NOE distance bound violations in the X-ray structure of native HEWL and in the MD simulation of ester-linked HEWL with respect to the experimental NMR NOE distance bounds [28]. Two sets of NOE bounds have been used: set1 includes 1630 NOE bounds, while set2 contains only the 392 long-range (residues \( i \) and \( j \) with \( j \geq i + 4 \)) NOE bounds. In ester-linked HEWL the virtual hydrogen atom positions are generated using the M procedure. \( R_0 \) is the experimentally derived distance bound including pseudo-atom corrections [187]. \( R_E \) is the distance obtained from the MD trajectories using \( r^{-3} \) averaging.

<table>
<thead>
<tr>
<th>Averaging period (ns) or structure</th>
<th>Number of violations (set1/set2):</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( &gt; 0.1 \text{ nm} )</td>
</tr>
<tr>
<td>X-ray crystal</td>
<td>70/46</td>
</tr>
<tr>
<td>1.5 – 3.5 ns</td>
<td>241/162</td>
</tr>
<tr>
<td>8.0 – 10.0 ns</td>
<td>252/169</td>
</tr>
<tr>
<td>0.0 – 10.0 ns</td>
<td>205/150</td>
</tr>
</tbody>
</table>

List of the 12 largest NOE violations in ester-linked HEWL:

<table>
<thead>
<tr>
<th>NOE pair</th>
<th>Exp. ( R_0 ) (nm)</th>
<th>( R_E ) (nm)</th>
<th>( R_E - R_0 ) (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_N(13Lys) - H_\delta(129Leu) )</td>
<td>0.69</td>
<td>1.27</td>
<td>0.58</td>
</tr>
<tr>
<td>( H_\varepsilon(28Trp) - H_\varepsilon(88Ile) )</td>
<td>0.90</td>
<td>1.62</td>
<td>0.72</td>
</tr>
<tr>
<td>( H_\beta(95Ala) - H_\alpha(108Trp) )</td>
<td>0.40</td>
<td>1.16</td>
<td>0.76</td>
</tr>
<tr>
<td>( H_\beta(95Ala) - H_\gamma_2(108Trp) )</td>
<td>0.45</td>
<td>1.17</td>
<td>0.72</td>
</tr>
<tr>
<td>( H_\alpha(95Ala) - H_\gamma_2(108Trp) )</td>
<td>0.30</td>
<td>1.12</td>
<td>0.82</td>
</tr>
<tr>
<td>( H_\gamma_2(98Ile) - H_N(107Ala) )</td>
<td>0.70</td>
<td>1.22</td>
<td>0.52</td>
</tr>
<tr>
<td>( H_\varepsilon(12Met) - H_\varepsilon(17Leu) )</td>
<td>0.30</td>
<td>0.89</td>
<td>0.59</td>
</tr>
<tr>
<td>( H_\varepsilon(12Met) - H_\alpha(28Trp) )</td>
<td>0.45</td>
<td>1.19</td>
<td>0.74</td>
</tr>
<tr>
<td>( H_\varepsilon(12Met) - H_\varepsilon(28Trp) )</td>
<td>0.45</td>
<td>1.04</td>
<td>0.59</td>
</tr>
<tr>
<td>( H_\varepsilon(12Met) - H_\delta_1(28Trp) )</td>
<td>0.45</td>
<td>1.08</td>
<td>0.63</td>
</tr>
<tr>
<td>( H_\gamma_2(98Ile) - H_\beta(107Ala) )</td>
<td>0.75</td>
<td>1.37</td>
<td>0.62</td>
</tr>
<tr>
<td>( H_N(98Ile) - H_\gamma_2(108Trp) )</td>
<td>0.55</td>
<td>1.11</td>
<td>0.56</td>
</tr>
</tbody>
</table>

This would convert the N-H with C-O dipole-dipole nonbonded interaction into a N with C-O charge-dipole interaction, which is a larger perturbation than the amide to ester (NH to O) one. For this reason, we have refrained from purely computational perturbations and only considered those that can be physically executed. These offer the possibility of falsification of our results. To this end, experimental studies like those of [1, 197] could be performed for HEWL or observables measurable by NMR or other spectroscopic techniques could be determined for ester-linked HEWL. In particular, it would be interesting to measure proton-proton NOEs for ester-linked HEWL and compare the observed pattern of atom-atom contacts to the one reported here.
Table 7.2: Overview of the hydrogen bonds (number and occurrence during the first 3.5 ns of simulation time) not involving peptide hydrogen atoms of the backbone of HEWL. Hydrogen bonds with an occurrence lower than 10% are not considered. It is also shown how many hydrogen bonds occur in both of the two MD simulations. The occurrence is the cumulative sum over the corresponding number of hydrogen bonds, the full list is available as supplementary material of [203], Table S1.

<table>
<thead>
<tr>
<th>system</th>
<th>backbone-side chain</th>
<th>sidechains only</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>number</td>
<td>occurrence [%]</td>
</tr>
<tr>
<td>X-ray</td>
<td>18</td>
<td>1800</td>
</tr>
<tr>
<td>MD native</td>
<td>26</td>
<td>686</td>
</tr>
<tr>
<td>MD ester-linked</td>
<td>22</td>
<td>499</td>
</tr>
<tr>
<td>MD native and ester-linked</td>
<td>5</td>
<td>109/184</td>
</tr>
</tbody>
</table>

7.5 Conclusion

The role of the peptide NH groups in maintaining the three-dimensional structure of the protein HEWL has been investigated by replacing all but two (70Pro and 79Pro) peptide linkages of amino acid residues by ester linkages. The structures obtained by MD computer simulation were compared to those of a comparable, previously published MD simulation of the native protein and to the X-ray crystal structure of lysozyme.

The structure of the ester-linked protein rapidly moves away from that of the native one, but does not completely unfold. Its surface area and radius of gyration (rgyr) indicate a slight compaction compared to the native protein in solution. Due to the differences between peptide and ester linkages, the distributions of backbone \( \phi \) - and \( \psi \)-angles are different. The extent of backbone-side chain and side chain-side chain hydrogen bonding does not change upon replacement of peptide linkages by ester linkages, but the pattern of these hydrogen bonds does change: only ca. 25% of the hydrogen bonds in the native simulation is also observed in the ester-linked simulation and vice versa. The typical secondary structure of lysozyme, three \( \beta \)-strands, and four \( \alpha \)-helices and two short \( 3_{10} \)-helices, are largely lost in the ester-linked lysozyme.

The atom-positional fluctuations do not significantly increase upon esterification of the peptide bonds, but differences are observed for particular residues.

The overall picture is that ester-linked lysozyme remains folded, due to the presence of four disulphide bridges, hydrogen bonds involving side chains, and its hydrophobic core of side-chain moieties. Because it is experimentally feasible to synthesize ester-linked lysozyme, these observations from computer simulation eagerly await falsification by experiment.
Chapter 8

Ester-linked hen egg white lysozyme shows a compact fold in a molecular dynamics simulation: possible causes and sensitivity of experimentally observable quantities to structural changes maintaining this compact fold

Prediction and understanding of the folding and stability of the three-dimensional structure of proteins is still a challenge. The different atomic interactions, such as non-polar contacts and hydrogen bonding, are known but their exact relative weights and roles when contributing to protein folding and stability are not identified. Initiated by a previous molecular dynamics simulation of fully ester-linked hen egg white lysozyme (HEWL), which showed a more compact fold of the ester-linked molecule compared to the native one, three variants of this protein are analysed in this work. These are 129-residue native HEWL, partly ester-linked HEWL, in which only 34 peptide linkages that are not involved in the helical or β-strand parts of native HEWL were replaced by ester linkages, and fully (126 residues) ester-linked HEWL. Native and partly ester-linked HEWL showed comparable behaviour while fully ester-linked HEWL could not maintain the native secondary structure of HEWL in the simulation and adopted a more compact fold. The conformational changes were analysed by comparing simulation averaged values of quantities that can be measured by NMR, such as $^1$H–$^{15}$N backbone order parameters, residual dipolar couplings, proton-proton NOE distances and $^3J$-couplings with the corresponding values derived from experimental NMR data for native HEWL. The information content of the latter appeared insufficient to detect the local conformational rearrangements upon esterification of the loop regions of the protein. For fully ester-linked HEWL though a significantly reduced agreement was observed. Upon esterification the backbone-side chain and side chain-side chain hydrogen bonding pattern of HEWL changes such as to maintain its compactness and thus the structural stability of the ester-linked lysozymes.
8.1 Introduction

Proteins and the stability of their specific, active fold play an important role in molecular biology and have been extensively studied for decades. Early studies of protein denaturation reported in the 1930’s and 1940’s by Wu and Yang [207] and Anson [208] explain the stability of a native protein conformation as being due to electrostatic intramolecular interactions, e.g. hydrogen bonding. The reversibility of protein denaturation, ionic bonding and the restoration of enzymatic activity was described and discussed by Lumry and Eyring in 1954 [209]. Other causes for protein stability were considered and in 1959 Kaummann [210] proposed the “hydrophobic effect”. In 1972 the issue got a wide audience when Anfinsen received the Nobel prize “for his work on ribonuclease, especially concerning the connection between the amino acid sequence and the biologically active conformation”. He had shown that this protein could be reversibly denaturated in a test tube [130]. Anfinsen explained this observation by identifying that the protein would adapt its conformation such as to minimise its Gibbs free energy in a given environment defining the thermodynamic state point: physiological conditions favour a native fold of the protein while the unfolded conformation is preferred under different, non-physiological solvents or temperatures.

The past four decades revealed the major factors in protein folding energetics, namely the hydrophobic effect, van der Waals interactions together with peptide hydrogen bonds and solvent composition. These are accompanied by several auxiliary factors, including salt-bridges, side-chain hydrogen bonds, disulfide bridges, and propensities to form α-helices and β-structure, as reviewed by Baldwin [189]. The primary structure of a protein together with knowledge about the relative weights of these factors and the actual thermodynamic state should in theory allow a prediction of the three-dimensional (3D) structure of a protein. However, prediction of protein folding and stability remain unmet challenges since the different weights, the exact roles and the corresponding mechanisms of each factor remain largely unknown [211, 212].

A systematic investigation of the contributions of each of these types of interactions to a native protein structure would involve separately eliminating each of them and observing the subsequent changes in protein structure. Such an approach is experimentally largely impossible due to physical and chemical limitations. Even computationally it is rather challenging to eliminate some of the mentioned interactions without interfering with other interactions. However, one possibility with minimal interference is the removal of backbone hydrogen bonds, i.e. hydrogen bonds involving the NH atoms as hydrogen donor. These can be eliminated by replacing the amide bond by an ester bond, a so-called A-to-E replacement (Fig. 8.1). This eliminates one hydrogen bond donor by replacing an NH group by an O atom, and reduces the hydrogen bond acceptor ability of the carbonyl [213]. Inevitably, the backbone properties are also modified by an A-to-E replacement. The peptide Cα−C=Oα–Cα torsional angle is less restricted to planar configurations than the Cα–C−N–Cα one (see Fig. 7.1) and the dipole of the CO–Oα–Cα group is smaller and differently directed than that of a CO–NH–Cα group, which will induce long-range effects.
A-to-E replacements are experimentally feasible using a nonsense suppression technique first reported by Schultz and co-workers [131]. The chemical synthesis is straightforward and, besides the fact that a hydroxy amino acid is inserted, is very similar to the synthesis of native polypeptides. Therefore, a few studies of proteins with A-to-E replacements have been done to get further insight in protein folding and stability and/or biological activity [1, 132–134, 197–199, 214–221]. Theoretical approaches of A-to-E replaced peptides or proteins are far less abundant but not absent. A study of a coiled coil system showed that the largest structural changes observed were due to missing hydrogen bonds of the protein within the hydrophobic core [222].

Our study was suggested by a previous molecular dynamics (MD) simulation of fully ester-linked hen egg white lysozyme (HEWL), i.e. a protein built of 126 hydroxy amino acids, all but the NH$_3^+$ head group and the two prolines [203]. Here it was shown that in aqueous solution - in contrast to what one might expect - the ester analogue does not unfold but shows a different, more compact fold compared to native HEWL. The pattern of backbone-side chain and side chain-side chain hydrogen bonds did change upon replacement of the peptide linkages by ester linkages: only about 25% of the hydrogen bonds in the native simulation were also observed in the ester-linked simulation and vice versa. Despite the fact that fully ester-linked HEWL lost all its helices and $\beta$-strands, a surprising 85% of 1630 NMR NOE distance bounds derived from experiment for native HEWL were still fulfilled within 0.1 nm when calculated based on the ester-linked HEWL simulation trajectory. These two intriguing results asked for a further investigation of the effect of A-to-E replacements in HEWL on various quantities that are observable in NMR experiments.

In the present study we not only consider the two extreme cases of native HEWL (no
A-to-E replacements) and fully ester-linked HEWL (126 A-to-E replacements, as many as possible) that were considered before, but also report a simulation of partly ester-linked HEWL. Here there are 34 A-to-E replacements for amino acids not involved in helices or β-strands (Figs. 8.2 and 8.3). In addition, more quantities were analysed in more depth than in ref. [203], using longer MD trajectories. The focus of the work reported in the present paper is on two questions: which experimentally observable quantities are sensitive to the structural changes which maintain a compact fold, and why is the fold of ester-linked lysozymes still compact? To this end, different observables that can be experimentally measured by NMR techniques, i.e. NH-order parameters, backbone $^{15}$N-$^1$H, $^{13}$C$_\alpha$-$^{13}$C, and $^{13}$C-$^{15}$N residual dipolar couplings (RDCs), NOE proton-proton distances as well as $^3J_{H\alpha H\alpha}$- and $^3J_{\alpha\beta}$-coupling constants have been calculated from the MD simulation trajectories of native, partly and fully ester-linked HEWL and compared to the values for native HEWL obtained from experiment. As these observables contain only short-distance or rather local information regarding pairs of atoms, it is of interest to investigate how sensitive these measurable quantities are to configurational rearrangements that maintain a compact fold of the protein. In addition, hydrogen bond patterns have been analysed in the simulations to search for stabilisation effects in the non-native fold of ester-linked HEWL.

8.2 Results and discussion

8.2.1 Structural properties and flexibility

The positional root-mean-square deviations (RMSD) of the backbone atoms from the X-ray structure for the simulation trajectories of native, partly ester-linked and fully ester-linked HEWL are shown in Fig. 8.4, upper panel. For all three proteins the values increase during the first period of the simulation and level off around a constant value of 0.3 nm for native and partly ester-linked HEWL and about 0.5 nm for fully ester-linked HEWL after a few nanoseconds. Native and fully ester-linked HEWL reach these values after 2 and 3 ns, respectively, while the RMSD value for partly ester-linked HEWL stays constant around 0.2 nm after 1 ns but jumps up to the same value as that observed for native HEWL.

Figure 8.2: Distribution of ester bonds in partly ester-linked HEWL. White numbers on black boxes indicate a hydroxy amino acid, i.e. the connection to the previous amino acid is established by an ester bond and not by a peptide bond as in native HEWL.
8.2. Results and discussion

(a) main secondary structure elements
(b) hydrophobic box residues

Figure 8.3: (a) The main secondary structure elements of native HEWL, as they were used in the hydrogen bonding analysis, see section 8.2.3. The four $\alpha$-helices are drawn in red (A: residues 5–14; B: residues 25–34; C: residues 89–100; D: residues 109–114), the two $3_{10}$-helices in black (a: residues 80–83; b: residues 120–123) and the three $\beta$-strands in blue ($\beta1$: residues 43–45; $\beta2$: residues 51–53; $\beta3$: residues 58–59). (b) Side-chain atoms of the residues involved in the hydrophobic box (residues 17, 20, 23, 28, 98, 105, 108 and 111) are represented by green balls. Not all of the atoms in these residues are involved in the hydrophobic box of HEWL, a more precise definition can be found in literature [25].

Figure 8.4: Positional root-mean-square deviation ($RMSD$) from the initial X-ray structure [186] of the backbone atoms (N/O$_\alpha$, C$_\alpha$, C and O) and the radius of gyration (lower panel) for the three simulations of native, partly ester-linked and fully ester-linked HEWL. Black solid lines: native HEWL; grey solid lines: partly ester-linked HEWL; black dotted lines: fully ester-linked HEWL.
(0.3 nm) after 7 ns, where partly ester-linked HEWL lost its α-helix D (see below). Fully ester-linked HEWL deviates much more from the X-ray crystal structure than native and partly ester-linked HEWL. Yet, fully ester-linked HEWL becomes more compact than the other two proteins as indicated by the radius of gyration in the lower panel of Fig. 8.4. It seems that the loss of secondary structure in fully ester-linked HEWL allows it to become more spherically compact. When the secondary structure is maintained, as in partly ester-linked HEWL, the compaction is less than for native HEWL.

Fully ester-linked HEWL is more flexible than native and partly ester-linked HEWL, as seen from the root-mean-square fluctuations ($RMSF$) of the $C_\alpha$ atoms in Fig. 8.5. Partly ester-linked HEWL shows somewhat larger fluctuations than native HEWL, in particular for residues where the amide bond has been replaced by an ester bond, while the fluctuations for the other residues are more comparable. The increased flexibility is likely to be a product of a loosened hydrogen bond network throughout the molecule.

Fig. 8.6 shows the evolution of secondary structure elements of native, partly, and fully ester-linked HEWL. Partly ester-linked HEWL largely maintains the major secondary structure elements (compare Fig. 8.3) visible in native HEWL: α-helices A–C remain stable while helix D is lost after 7 ns, but it shows much less π-character instead of α-character than in native HEWL. The two $3_{10}$ helices a and b are present in both simulations but mostly in the form of an α-helix. It is known that the 45A3 force field tends to favour α over $3_{10}$-helices [201]. The three main β-strands are stable in native and partly ester-linked HEWL, but less so in the latter case. Partly ester-linked HEWL displays more structural variation than native HEWL and this higher flexibility is also reflected in the root-mean-
8.2. Results and discussion

Figure 8.6: Secondary structure elements as a function of time calculated for the three MD simulation trajectories of native, partly ester-linked and fully ester-linked HEWL (from left to right). Color code: black (3_{10}-helix), red (α-helix), green (π-helix), blue (β-strand), yellow (β-bridge), brown (bend) and grey (turn). The labelling of the secondary structure elements in native lysozyme as defined in Fig. 8.3 is indicated on the right of the figure.

square fluctuations (Fig. 8.5). Fully ester-linked HEWL shows different behaviour. All helices are lost within the first 4 ns and the β-strands are disrupted even earlier.

8.2.2 Comparison to experimental NMR data

$^1$H$-^{15}$N order parameters

A comparison of the $^1$H$-^{15}$N order parameters $S^2$ derived from experiments [26] to the ones calculated for native, partly ester-linked and fully ester-linked HEWL can be found in Fig. 8.7. The order parameters calculated from the simulations are generally smaller than those derived using the so-called model-free analysis [223] of the experimental relaxation times. The largest order parameters are found for residues in helices. The simulation of
Figure 8.7: $^1$H–$^{15}$N order parameters of HEWL derived from experiment on native HEWL (dotted line with crosses) [26] compared to calculated $^1$H–$^{15}$N order parameters for the MD simulations of native, partly ester-linked and fully ester-linked HEWL (from top to bottom). $^1$H–$^{15}$N order parameters are calculated using equation Eq. (8.6) over the whole 10 ns of simulation time but averaged over a 1 ns time window. The grey bars on top of each graph indicate the secondary structure elements of native HEWL (compare Fig. 8.3), i.e. thick bars: $\alpha$-helix, medium bars: $\beta$-strand. The state of the peptide bond in partly ester-linked HEWL, native versus ester bond, is indicated by open and filled circles, respectively, above the secondary structure bars in the plot of partly ester-linked HEWL.
native HEWL produces a pattern of \(^1\text{H}-^{15}\text{N}\) \(S^2\)-values along the backbone that is roughly similar to that of the \(S^2\)-values derived from experiment, the correlation coefficient being 0.41. Partial esterification does enlarge the deviation between calculated and experimental order parameters, illustrated by a lower correlation coefficient of 0.38, but the general pattern as a function of residue number is still similar. However, full esterification largely destroys the pattern, which is reflected by the correlation coefficient of 0.16.

**NOE proton-proton distances**

One way to measure the quality of a molecular structure from a molecular dynamics simulation is a comparison of simulated to experimentally derived NMR NOE proton-proton distances. This is generally done by evaluating whether simulated proton-proton distances are bigger than an upper limit, the NOE bound, determined based on the experimental NMR data.

A total of 1630 distance bounds, including 392 long-range NOE bounds, i.e. from residue \(i\) to residue \(j\) with \(|i-j| \geq 4\), is available from NMR experiments [28] and the corresponding \(r^{-3}\) averaged distances were calculated from each of the three simulations and compared to the NOE distance upper bounds. The number of bound violations is given in Tab. 8.1 while Fig. 8.8 shows the distribution of NOE distance bound violations for the long-range NOE atom pairs. There are less NOE bound violations in the simulation of partly ester-linked HEWL than in that of native HEWL except for the time window from 8 to 10 ns.

Interestingly, a large majority of all 1630 NOE distances calculated over the whole time frame of 10 ns is within 0.1 nm of the bounds in all three simulations: all/long-range: 96%/89%, 96%/89% and 85%/60% for native, partly ester-linked and fully ester-linked HEWL, respectively. Thus the loss of secondary structure in the fully ester-linked HEWL simulation (see Fig. 8.6) increases the number of NOE bound violations larger than 0.1 nm by only 11%/29% (all/long-range), a number which was expected to be much higher for a protein with a high proportion of its residues (54 of 129 residues, 42%) involved in the main secondary structure elements (compare Fig. 8.3). This result is particularly surprising as the atoms involved in the 1630 experimental NOE atom pairs are widely distributed over the 129 amino acid residues.

A closer look at the large (> 0.5 nm) long-range NOE distance bound violations shows that most of them involve residues of the hydrophobic box region in the core of the protein (Fig. 8.3b): Tab. 8.2 shows that 21 out of the 25 NOE bound violations larger than 0.5 nm in the 10 ns simulation of fully ester-linked HEWL involve hydrogens of the hydrophobic box residues. The hydrophobic box must be disrupted almost immediately in the fully ester-linked HEWL simulation since 19 of 20 of the long-range NOE bound violations larger than 0.5 nm already involve hydrophobic box residues during the first time window of 0–3.5 ns. This is in agreement with experimental results for A-to-E replaced peptides where the biggest structural changes due to missing hydrogen bonds were observed within
Table 8.1: NMR NOE proton-proton distance bound violations in the MD simulations of native, partly ester-linked and fully ester-linked HEWL with respect to the upper NOE distance bounds derived from experiment [28]. The calculations have been done using a $r^{-3}$-averaging for 1630 NOE bounds including 392 long-range NOE bounds, i.e. from residue $i$ to residue $j$ with $|i-j| \geq 4$.

<table>
<thead>
<tr>
<th>System</th>
<th>Number of violations (all/long-range)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt; 0.1 nm</td>
</tr>
<tr>
<td>time window: 0-3.5ns:</td>
<td></td>
</tr>
<tr>
<td>native</td>
<td>64/44</td>
</tr>
<tr>
<td>partly ester-linked</td>
<td>58/36</td>
</tr>
<tr>
<td>fully ester-linked</td>
<td>192/133</td>
</tr>
<tr>
<td>time window: 1.5-3.5ns:</td>
<td></td>
</tr>
<tr>
<td>native</td>
<td>84/52</td>
</tr>
<tr>
<td>partly ester-linked</td>
<td>70/46</td>
</tr>
<tr>
<td>fully ester-linked</td>
<td>250/159</td>
</tr>
<tr>
<td>time window: 8-10ns:</td>
<td></td>
</tr>
<tr>
<td>native</td>
<td>92/54</td>
</tr>
<tr>
<td>partly ester-linked</td>
<td>107/70</td>
</tr>
<tr>
<td>fully ester-linked</td>
<td>327/201</td>
</tr>
<tr>
<td>time window: 0-10ns:</td>
<td></td>
</tr>
<tr>
<td>native</td>
<td>69/44</td>
</tr>
<tr>
<td>partly ester-linked</td>
<td>66/42</td>
</tr>
<tr>
<td>fully ester-linked</td>
<td>237/158</td>
</tr>
</tbody>
</table>

the hydrophobic core [222]. Larger changes in other parts of the molecule start to happen later in the simulation. This is reflected in the numbers of violations during the last two ns of the simulation: the number of long-range NOE bound violations involving hydrophobic box residues increased to 27 while the total number of long-range NOE bound violations rose to 70.

Proton-proton $^3J$-coupling constants

Different sets of experimental proton-proton $^3J$-coupling constants, $^3J_{HN-H\alpha}$ and $^3J_{\alpha\beta}$, for native HEWL [29] were compared to values calculated from the simulation trajectories (see Fig. 8.9). The $^3J_{\alpha\beta}$-coupling constants were calculated only if the stereospecific H$_\beta$ assignments were available from experiment.

For native HEWL both, the $^3J_{HN-H\alpha}$- and $^3J_{\alpha\beta}$-coupling constants show a difference of $1-2$ Hz between simulated and measured values. The root-mean-square deviation ($RMSD$) of the calculated coupling constants from the experimental values for native
8.2. Results and discussion

Figure 8.8: Distribution of all 392 long range NOE bound violations (from residue $i$ to $j$ with $|i - j| \geq 4$) for native, partly ester-linked and fully ester-linked HEWL (from top to bottom) calculated from 10 ns trajectories. NOE bound violations are coloured black while the fulfilled NOE bounds appear in grey.

HEWL are 1.7, 1.4, and 2.4 Hz for native, partly and fully ester-linked HEWL. Surprisingly, partial esterification improves the agreement of the calculated $^3J_{H_NH_\alpha}$-couplings with the ones measured for native HEWL. The most apparent changes going from native or partly ester-linked HEWL to fully ester-linked HEWL are observed for residues involved in a native $\alpha$-helix (red dots in Fig. 8.9, left panels). For the $^3J_{\alpha\beta}$-coupling constants the corresponding $RMSD$ values are 2.3 Hz, 3.0 Hz, and 3.9 Hz, showing a rise with increasing esterification.

The quality of the simulated $^3J$-couplings suffers from limited sampling compared to
Figure 8.9: Comparison of simulated (0–10 ns) $^3J$-coupling constants of native (upper panels), partly ester-linked (middle panels) and fully ester-linked HEWL (lower panels) to the experimental coupling constants of native HEWL [29]. The left column shows the $^3J_{HN\alpha}$-coupling constants while the right column contains the values of the $^3J_{\alpha\beta}$-coupling constants. Colours indicate the secondary structure of the corresponding amino acid: red ($\alpha$-helix), green ($\pi$-helix), blue ($\beta$-strand), yellow ($\beta$-bridge), orange (turn), brown (bend) and grey (other).
8.2. Results and discussion

Table 8.2: Number of violations for the 124 long-range NOE distance bounds involving residues of the hydrophobic box in the MD simulations of native, partly ester-linked and fully ester-linked HEWL with respect to the experimental NMR NOE distance bounds [28]. The values are compared to the total number of the 392 long-range NOE distance bound violations in the whole protein (numbers in brackets).

<table>
<thead>
<tr>
<th>System</th>
<th>&gt; 0.1 nm</th>
<th>&gt; 0.3 nm</th>
<th>&gt; 0.5 nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>time window: 0–3.5 ns:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>native</td>
<td>19 (44)</td>
<td>1 (5)</td>
<td>0 (1)</td>
</tr>
<tr>
<td>partly ester-linked</td>
<td>17 (36)</td>
<td>5 (5)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>fully ester-linked</td>
<td>49 (133)</td>
<td>29 (53)</td>
<td>19 (20)</td>
</tr>
<tr>
<td>time window: 1.5–3.5 ns:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>native</td>
<td>22 (52)</td>
<td>9 (18)</td>
<td>0 (5)</td>
</tr>
<tr>
<td>partly ester-linked</td>
<td>22 (46)</td>
<td>5 (6)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>fully ester-linked</td>
<td>57 (159)</td>
<td>33 (82)</td>
<td>27 (37)</td>
</tr>
<tr>
<td>time window: 8–10 ns:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>native</td>
<td>28 (54)</td>
<td>7 (10)</td>
<td>0 (4)</td>
</tr>
<tr>
<td>partly ester-linked</td>
<td>25 (70)</td>
<td>7 (20)</td>
<td>2 (7)</td>
</tr>
<tr>
<td>fully ester-linked</td>
<td>77 (201)</td>
<td>51 (132)</td>
<td>27 (70)</td>
</tr>
<tr>
<td>time window: 0–10 ns:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>native</td>
<td>21 (44)</td>
<td>3 (10)</td>
<td>0 (3)</td>
</tr>
<tr>
<td>partly ester-linked</td>
<td>20 (42)</td>
<td>5 (7)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>fully ester-linked</td>
<td>60 (158)</td>
<td>35 (70)</td>
<td>21 (25)</td>
</tr>
</tbody>
</table>

the experimental time scale and from inaccuracy induced by the empirical nature of the parameters of the Karplus relation used to obtain a $^3J$-couplings from a configuration [224]. In addition, force-field induced inaccuracies may play a role. Yet, the rough correlation between simulated and measured $^3J$-couplings for native HEWL is completely lost when comparing values from the fully ester-linked HEWL simulation to the ones measured for native HEWL.

Backbone $^{15}N$-$^1H$, $^{13}C_\alpha$-$^{13}C$, and $^{13}C$-$^{15}N$ residual dipolar couplings

The quality, as represented by a so-called $Q$-value (equation (8.10)), of the backbone $^{15}N$-$^1H$, $^{13}C_\alpha$-$^{13}C$, and $^{13}C$-$^{15}N$ RDCs calculated from the simulation trajectories of native, partly ester-linked and fully ester-linked HEWL with respect to the experimental RDCs of native HEWL [27] is shown in Fig. 8.10. The calculations from the three simulations all result in a much higher $Q$-value than that one calculated from the X-ray crystal structure [186]. Interestingly, the simulation of partly ester-linked HEWL shows a better ($C_\alpha$-
Chapter 8. Native, partly ester-linked, and fully ester-linked HEWL

Figure 8.10: Q-value distribution of the calculated backbone N-H, Cα-C, and C-N RDCs of native (solid line), partly ester-linked (dashed line) and fully ester-linked (dotted line) HEWL when fitting the calculated RDCs of each configuration to the experimentally derived RDCs of native HEWL [27]. The quality of the fits from the X-ray crystal structure [186] is shown by a vertical dashed-dotted line.
C and C-N RDCs) or similar (N-H RDCs) agreement with the experimental values for native HEWL than the values obtained from the native HEWL simulation trajectory. In contrast, the $Q$-values obtained from the RDCs based on the fully ester-linked HEWL simulation trajectory are much higher.

### 8.2.3 Hydrogen bonding

The analysis of atom-positional root-mean-square differences and fluctuations, radius of gyration and NMR NOE atom-atom distances showed that even fully ester-linked HEWL did not completely unfold as one might have expected. An explanation could be the 4 disulfide bridges holding the protein together. Therefore, we repeated the simulations of the native, partly- and fully ester-linked HEWL without the 4 disulfide bridges, i.e. only the covalent interactions across the disulfide bridges were omitted [225, 226]. These trajectories did not provide a picture of the effect of esterification upon the properties analysed that is different from the one presented. Thus, the presence of disulfide bridges does not explain why 85% of all 1630 experimental NMR NOE distance bounds of native HEWL are satisfied within 0.1 nm in the simulation of fully ester-linked HEWL. What other factors or forces could keep the ester-linked protein compact?

A possible cause might be hydrogen bonding. Replacing amide by ester linkages, the number of backbone-backbone (bb-bb) and backbone-side chain (bb-sc) hydrogen bonds is massively reduced due to the missing hydrogen bond donor and reduced acceptor ability of esters compared to amides. Therefore, a lot of non-hydrogen bonded hydrogen bond acceptors were present in the beginning of the partly ester-linked and particularly in the fully ester-linked HEWL simulations, which would be able to find donor partners by local conformational rearrangement.

When analysing such a hydrogen-bond rearrangement effect, three types of hydrogen bonds were distinguished in the hydrogen bond analysis: backbone-backbone (bb-bb) hydrogen bonds to show the reduction of hydrogen bonding with an increasing number of ester linkages in the backbone, as well as backbone-side chain (bb-sc) and side chain-side chain (sc-sc) hydrogen bonds. The number of backbone-side chain hydrogen bonds is given excluding the amide as hydrogen-bond donor. This is to separate the effect due to the change in the number of donors from that due to configurational rearrangement. An increasing number of bb-sc and sc-sc hydrogen bonds upon esterification would indicate a stabilisation of the observed compact fold by them.

The upper part of Tab. 8.3 shows that the number of bb-bb and bb-sc hydrogen bonds with a total occurrence > 10% is comparable or smaller in partly ester-linked HEWL than in native HEWL and even smaller in fully ester-linked HEWL over the whole time range of the simulation (0–10 ns). The same is true for the cumulative sums of the corresponding hydrogen bond occurrences for all hydrogen bond types. However, the number of sc-sc hydrogen bonds is increasing from native over partly ester-linked to fully ester-linked HEWL. A similar behaviour is seen for the three other time windows. The picture for bb-
Table 8.3: Number of hydrogen bonds as well as the cumulative sum over the corresponding occurrence for the simulations of native, partly ester-linked, and fully ester-linked HEWL. The numbers are sorted with respect to the hydrogen bond type: backbone-backbone (bb-bb), backbone-side chain (bb-sc) and side chain-side chain (sc-sc) hydrogen bonds. Hydrogen bonds of type bb-sc involving an amide hydrogen were not considered. Upper part: only hydrogen bonds with an occurrence > 10%, during the indicated time window, have been considered while the lower part contains all hydrogen bonds with any occurrence.

<table>
<thead>
<tr>
<th>Protein Type</th>
<th>Time Window</th>
<th>Hydrogen Bonds with an Occurrence &gt; 10%</th>
<th>All Hydrogen Bonds with Any Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>bb-bb</td>
<td>occ. / %</td>
</tr>
<tr>
<td></td>
<td></td>
<td>number</td>
<td></td>
</tr>
<tr>
<td><strong>Native</strong></td>
<td>0–3.5 ns</td>
<td>90</td>
<td>4972</td>
</tr>
<tr>
<td></td>
<td>1.5–3.5 ns</td>
<td>82</td>
<td>4970</td>
</tr>
<tr>
<td></td>
<td>8–10 ns</td>
<td>85</td>
<td>5022</td>
</tr>
<tr>
<td><strong>Partly Ester-Linked</strong></td>
<td>0–3.5 ns</td>
<td>79</td>
<td>4157</td>
</tr>
<tr>
<td></td>
<td>1.5–3.5 ns</td>
<td>78</td>
<td>4154</td>
</tr>
<tr>
<td></td>
<td>8–10 ns</td>
<td>74</td>
<td>4232</td>
</tr>
<tr>
<td><strong>Fully Ester-Linked</strong></td>
<td>0–3.5 ns</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1.5–3.5 ns</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>8–10 ns</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0–10 ns</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
sc and sc-sc hydrogen bonds changes when looking at hydrogen bonds with any occurrence (Tab. 8.3, bottom part). The total occurrence of bb-sc and sc-sc hydrogen bonds is still smaller for ester-linked HEWL compared to native HEWL, except for bb-sc hydrogen bonds during the first part of the simulation where they stay rather constant. In contrast, the number of hydrogen bonds is massively growing from native to partly ester-linked to fully ester-linked HEWL, independent of the time window.

To obtain a simple picture of the changed pattern of hydrogen bonding the hydrogen bonds between secondary structure elements have been analysed. Tab. 8.4 shows the number and occurrence of hydrogen bonds connecting two secondary structure elements in native, partly ester-linked and fully ester-linked HEWL. Only hydrogen bonds linking two helices or two \( \beta \)-strands (A, B, C, D, a, b, \( \beta_1 \), \( \beta_2 \) or \( \beta_3 \), see Fig. 8.3) are considered. The connection may be intra or inter two secondary structure elements. Analysing the whole time window of 10 ns shows a clearly higher amount of bb-sc and sc-sc hydrogen bonding (number and occurrence) for partly ester-linked and fully ester-linked HEWL than for native HEWL. Therefore, longer-living bb-sc hydrogen bonds are connecting pairs of secondary structure elements, holding the protein together. This additional hydrogen bonding might not fully compensate for the loss of backbone-backbone and backbone-side chain hydrogen bonding involving the amide hydrogen atoms of native HEWL. Nevertheless, this backbone-side chain and side chain-side chain stabilisation seems to help to keep the molecule in a slightly different and more flexible compact form than displayed by native HEWL.

A similar backbone-side chain stabilisation of the hydrophobic box residues (see Fig. 8.3b) could be detected in an equivalent hydrogen bond analysis considering residues involved in the hydrophobic box (see Tab. 8.5). The changing extent of hydrogen bonding in ester-linked HEWL observed through the simulations hints at a rapid, structural change in this part of the protein.

Hydrogen bonding from the protein backbone to the solvent or vice versa is reported in Tab. 8.6 in terms of number and occurrence of amide to water NH-OW, water to amide or water to ester HW_{1/2}-O, and water to ester HW_{1/2}-O\( _\alpha \) bonds (Fig. 8.1). Although the total number of hydrogen bonds in the ester substituted proteins is somewhat bigger than in native HEWL, the corresponding occurrence decreases with increasing amount of ester-linkages within the backbone. Interestingly, the number and occurrence of protein-solvent hydrogen bonds involving the ester O\( _\alpha \) atoms of the protein backbone is larger for partly ester-linked HEWL than for fully ester-linked HEWL. Apparently, the hydrogen bond acceptors in the protein that are released from hydrogen bonding by the A-to-E replacements do prefer hydrogen bonding to other hydrogen bond donors in the protein over hydrogen bonding to water. The replacement also allows for an increase of the number of protein-protein hydrogen bonds involving backbone atoms.

The complete lists of all protein-protein hydrogen bonds during the simulation of native, partly ester-linked and fully ester-linked HEWL is given in Tabs. S4–S6 of the supplementary material in [136].
Chapter 8. Native, partly ester-linked, and fully ester-linked HEWL

**Table 8.4:** Number of hydrogen bonds involving residues in helices or β-strands as well as the cumulative sum over the corresponding occurrence for the simulations of native, partly ester-linked, and fully ester-linked HEWL. Hydrogen bonds with any occurrence have been considered but only if the hydrogen bond bridges any of the main helices or β-strands, A–D, a,b or β1–β3, compare figure Fig. 8.3, inter- or intra-structural. The numbers are sorted with respect to the hydrogen bond type: backbone-backbone (bb-bb), backbone-side chain (bb-sc) and side chain-side chain (sc-sc) hydrogen bonds. Hydrogen bonds of type bb-sc involving an amide hydrogen were not considered.

<table>
<thead>
<tr>
<th>Protein</th>
<th>bb-bb number</th>
<th>bb-bb occ. / %</th>
<th>bb-sc number</th>
<th>bb-sc occ. / %</th>
<th>sc-sc number</th>
<th>sc-sc occ. / %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Helix–helix</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Time window: 0–3.5 ns</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>55</td>
<td>1891</td>
<td>39</td>
<td>73</td>
<td>21</td>
<td>44</td>
</tr>
<tr>
<td>Partly ester-linked</td>
<td>54</td>
<td>1982</td>
<td>62</td>
<td>55</td>
<td>37</td>
<td>74</td>
</tr>
<tr>
<td>Fully ester-linked</td>
<td>0</td>
<td>0</td>
<td>136</td>
<td>118</td>
<td>52</td>
<td>66</td>
</tr>
<tr>
<td><strong>Time window: 1.5–3.5 ns</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>59</td>
<td>1942</td>
<td>50</td>
<td>97</td>
<td>27</td>
<td>40</td>
</tr>
<tr>
<td>Partly ester-linked</td>
<td>59</td>
<td>1913</td>
<td>83</td>
<td>61</td>
<td>46</td>
<td>62</td>
</tr>
<tr>
<td>Fully ester-linked</td>
<td>0</td>
<td>0</td>
<td>247</td>
<td>144</td>
<td>99</td>
<td>34</td>
</tr>
<tr>
<td><strong>Time window: 8–10 ns</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>47</td>
<td>1972</td>
<td>24</td>
<td>115</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>Partly ester-linked</td>
<td>56</td>
<td>1840</td>
<td>46</td>
<td>74</td>
<td>20</td>
<td>56</td>
</tr>
<tr>
<td>Fully ester-linked</td>
<td>0</td>
<td>0</td>
<td>97</td>
<td>114</td>
<td>31</td>
<td>7</td>
</tr>
<tr>
<td><strong>Time window: 0–10 ns</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>52</td>
<td>1930</td>
<td>26</td>
<td>72</td>
<td>18</td>
<td>25</td>
</tr>
<tr>
<td>Partly ester-linked</td>
<td>54</td>
<td>1983</td>
<td>48</td>
<td>58</td>
<td>29</td>
<td>62</td>
</tr>
<tr>
<td>Fully ester-linked</td>
<td>0</td>
<td>0</td>
<td>92</td>
<td>118</td>
<td>39</td>
<td>69</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein</th>
<th>bb-bb number</th>
<th>bb-bb occ. / %</th>
<th>bb-sc number</th>
<th>bb-sc occ. / %</th>
<th>sc-sc number</th>
<th>sc-sc occ. / %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>β-strand–β-strand</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Time window: 0–3.5 ns</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>6</td>
<td>350</td>
<td>7</td>
<td>10</td>
<td>13</td>
<td>160</td>
</tr>
<tr>
<td>Partly ester-linked</td>
<td>7</td>
<td>336</td>
<td>14</td>
<td>13</td>
<td>18</td>
<td>190</td>
</tr>
<tr>
<td>Fully ester-linked</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>47</td>
<td>18</td>
<td>151</td>
</tr>
<tr>
<td><strong>Time window: 1.5–3.5 ns</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>6</td>
<td>335</td>
<td>9</td>
<td>10</td>
<td>16</td>
<td>140</td>
</tr>
<tr>
<td>Partly ester-linked</td>
<td>7</td>
<td>316</td>
<td>24</td>
<td>15</td>
<td>27</td>
<td>177</td>
</tr>
<tr>
<td>Fully ester-linked</td>
<td>0</td>
<td>0</td>
<td>36</td>
<td>29</td>
<td>28</td>
<td>103</td>
</tr>
<tr>
<td><strong>Time window: 8–10 ns</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>5</td>
<td>287</td>
<td>6</td>
<td>11</td>
<td>8</td>
<td>136</td>
</tr>
<tr>
<td>Partly ester-linked</td>
<td>6</td>
<td>321</td>
<td>16</td>
<td>11</td>
<td>17</td>
<td>181</td>
</tr>
<tr>
<td>Fully ester-linked</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>36</td>
<td>23</td>
<td>85</td>
</tr>
<tr>
<td><strong>Time window: 0–10 ns</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>6</td>
<td>353</td>
<td>6</td>
<td>11</td>
<td>10</td>
<td>136</td>
</tr>
<tr>
<td>Partly ester-linked</td>
<td>8</td>
<td>337</td>
<td>11</td>
<td>16</td>
<td>14</td>
<td>198</td>
</tr>
<tr>
<td>Fully ester-linked</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>44</td>
<td>12</td>
<td>160</td>
</tr>
</tbody>
</table>
Table 8.5: Number of hydrogen bonds between residues of the hydrophobic box as well as the cumulative sum over the corresponding occurrence for the simulations of native, partly ester-linked, and fully ester-linked HEWL. Hydrogen bonds with any occurrence have been considered but only if the hydrogen bond bridges two residues involved in the hydrophobic box. The numbers are sorted with respect to the hydrogen bond type: backbone-backbone (bb-bb), backbone-side chain (bb-sc) and side chain-side chain (sc-sc) hydrogen bonds. Hydrogen bonds of type bb-sc involving an amide hydrogen were not considered.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Hydrophobic box–hydrophobic box</th>
<th>Hydrophobic box–hydrophobic box</th>
<th>Hydrophobic box–hydrophobic box</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bb-bb</td>
<td>bb-sc</td>
<td>sc-sc</td>
</tr>
<tr>
<td></td>
<td>number</td>
<td>occ. / %</td>
<td>number</td>
</tr>
<tr>
<td>Time window: 0–3.5 ns</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>33</td>
<td>672</td>
<td>64</td>
</tr>
<tr>
<td>Partly ester-linked</td>
<td>34</td>
<td>610</td>
<td>81</td>
</tr>
<tr>
<td>Fully ester-linked</td>
<td>0</td>
<td>0</td>
<td>182</td>
</tr>
<tr>
<td>Time window: 1.5–3.5 ns</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>34</td>
<td>610</td>
<td>84</td>
</tr>
<tr>
<td>Partly ester-linked</td>
<td>35</td>
<td>635</td>
<td>114</td>
</tr>
<tr>
<td>Fully ester-linked</td>
<td>0</td>
<td>0</td>
<td>280</td>
</tr>
<tr>
<td>Time window: 8–10 ns</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>24</td>
<td>588</td>
<td>46</td>
</tr>
<tr>
<td>Partly ester-linked</td>
<td>29</td>
<td>702</td>
<td>52</td>
</tr>
<tr>
<td>Fully ester-linked</td>
<td>0</td>
<td>0</td>
<td>114</td>
</tr>
<tr>
<td>Time window: 0–10 ns</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native</td>
<td>29</td>
<td>642</td>
<td>51</td>
</tr>
<tr>
<td>Partly ester-linked</td>
<td>28</td>
<td>580</td>
<td>69</td>
</tr>
<tr>
<td>Fully ester-linked</td>
<td>0</td>
<td>0</td>
<td>130</td>
</tr>
</tbody>
</table>

8.3 Materials and methods

8.3.1 Simulation program and force-field parameters

The molecular dynamics simulations of all three systems, native HEWL, fully ester-linked HEWL and partly ester-linked HEWL in SPC [148] water, were carried out using the GROMOS simulation software [115], and the 45A3 force-field parameter set [72, 200]. The introduction of ester bonds in the backbone of the protein required some force-field changes affecting partial charges and charge groups (Fig. 8.1) as well as bonded interactions (Tab. 8.7). No A-to-E replacements were done for the simulation of native HEWL while fully ester-linked HEWL contained 126 A-to-E replacements, all but the head amino acid and the two prolines. For the partly ester-linked HEWL simulation 34 A-to-E replacements were carried out, see Fig. 8.2. The selection of peptide bonds to be replaced by an ester bond was based on the assignment of $\alpha$-helices and $\beta$-strands as well as $\beta$-strands/bridges in the native protein as reported by Oostenbrink et al. [201] Table 1. The aim was to preserve the main secondary structure elements of native HEWL, namely the three $\beta$-strands, four...
**Table 8.6:** Number and summed-up occurrence of hydrogen bonds from a backbone atom of the protein (H, N, O, C, and α) to a water molecule (OW, HW1, HW2) from the 10 ns simulations of native, partly ester-linked and fully ester-linked HEWL.

All values are normalised, i.e. divided by the corresponding number of donor (H, O, N, α) or acceptor (O, α, C, O) atoms present in the protein backbone for the indicated protein-solvent hydrogen bond. Note that the NH₃⁺ head group has not been considered in these calculations.

<table>
<thead>
<tr>
<th>Protein</th>
<th>% / number occ.</th>
<th>% / number occ.</th>
<th>% / number occ.</th>
<th>% / number occ.</th>
<th>% / number occ.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native HEWL</td>
<td>992</td>
<td>27</td>
<td>848</td>
<td>143</td>
<td>0</td>
</tr>
<tr>
<td>Partly ester-linked HEWL</td>
<td>978</td>
<td>40</td>
<td>577</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Fully ester-linked HEWL</td>
<td>978</td>
<td>40</td>
<td>577</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein</th>
<th>% / number occ.</th>
<th>% / number occ.</th>
<th>% / number occ.</th>
<th>% / number occ.</th>
<th>% / number occ.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native HEWL</td>
<td>162</td>
<td>27</td>
<td>978</td>
<td>57</td>
<td>0</td>
</tr>
<tr>
<td>Partly ester-linked HEWL</td>
<td>152</td>
<td>27</td>
<td>947</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>Fully ester-linked HEWL</td>
<td>0</td>
<td>0</td>
<td>1460</td>
<td>50</td>
<td>0</td>
</tr>
</tbody>
</table>

*Note:* All values are normalised, i.e. divided by the corresponding number of donor (H, O, N, α) or acceptor (O, α, C, O) atoms present in the protein backbone for the indicated protein-solvent hydrogen bond. Note that the NH₃⁺ head group has not been considered in these calculations.
Table 8.7: Comparison of the force-field parameters used for the native peptidic bonds (GROMOS 45A3 [72] parameters) and for the ester bonds. The table includes all force-field changes which were made because of the ester bonds except the changes of the charges and charge groups specified in figure Fig. 8.1.

<table>
<thead>
<tr>
<th>bond</th>
<th>$b_0$ / nm</th>
<th>$K_b$ / $10^6$ kJ mol$^{-1}$ nm$^{-4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C–N</td>
<td>0.133</td>
<td>11.8</td>
</tr>
<tr>
<td>N–C$_\alpha$</td>
<td>0.147</td>
<td>8.71</td>
</tr>
<tr>
<td>C–O$_\alpha$</td>
<td>0.136</td>
<td>10.2</td>
</tr>
<tr>
<td>O$<em>\alpha$–C$</em>\alpha$</td>
<td>0.143</td>
<td>8.18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>bond angle</th>
<th>$\Theta_0$ / deg</th>
<th>$K_\Theta$ / $10^2$ kJ mol$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C–N–C$_\alpha$</td>
<td>122.0</td>
<td>700</td>
</tr>
<tr>
<td>C–O$<em>\alpha$–C$</em>\alpha$</td>
<td>109.5</td>
<td>450</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>dihedral angle</th>
<th>$\cos(\delta)$</th>
<th>$m$</th>
<th>$K_\Phi$ / kJ mol$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C$<em>\alpha$–C–N–C$</em>\alpha$</td>
<td>-1</td>
<td>2</td>
<td>33.5</td>
</tr>
<tr>
<td>C–N–C$_\alpha$–C</td>
<td>-1</td>
<td>6</td>
<td>1.00</td>
</tr>
<tr>
<td>N–C$_\alpha$–C–N</td>
<td>1</td>
<td>6</td>
<td>1.00</td>
</tr>
<tr>
<td>C$<em>\alpha$–C–O$</em>\alpha$–C$_\alpha$</td>
<td>-1</td>
<td>2</td>
<td>16.7</td>
</tr>
<tr>
<td>C–O$<em>\alpha$–C$</em>\alpha$–C</td>
<td>1</td>
<td>3</td>
<td>1.26</td>
</tr>
<tr>
<td>O$<em>\alpha$–C$</em>\alpha$–C–O$_\alpha$</td>
<td>1</td>
<td>3</td>
<td>5.29</td>
</tr>
</tbody>
</table>

$\alpha$-helices and two 3$_{10}$-helices, see Fig. 8.3.

8.3.2 Simulation set-up and protocol

The initial coordinates of the lysozyme atoms for native, fully ester-linked and partly ester-linked HEWL were derived from the native X-ray crystal structure, PDB code 1AKI [186]. The protonation states of all amino acids corresponding to a pH of 7 were used, leading to a total protein charge of 8e. Each protein was separately solvated in a cubic box of edge lengths 7.73 nm, 8.54 nm, and 8.54 nm with 14378, 19644, and 19675 SPC [148] water molecules, respectively. The overall neutrality of the box was preserved by adding 8 chloride counter ions.

Initial velocities for all atoms and simulations were assigned from a Maxwell-Boltzmann distribution at 60 K with position restraining of the protein atoms with an initial harmonic force constant of 25000 kJ mol$^{-1}$ nm$^{-2}$. The temperature was raised stepwise by 60 K and the force constant of the position restraining lowered by a factor of 10, every 20 ps, ending in a non-restrained protein simulation of 20 ps at 300 K after 80 ps of simulated time. A 10 ns simulation followed the 100 ps thermalisation procedure while saving the protein configuration every 0.5 ps for analysis.

The temperature and pressure (300 K, 1 atm) were kept constant using the weak coupling
algorithm [75] with corresponding coupling times of $\tau_T = 0.1 \text{ ps}$ and $\tau_p = 0.5 \text{ ps}$, respectively and an estimated isothermal compressibility of $4.575 \cdot 10^{-4} \text{ (kJ mol}^{-1} \text{ nm}^{-3})^{-1}$. All bond lengths and the bond angles of the water molecules were kept rigid by applying constraints using the SHAKE algorithm [14] with a relative geometric tolerance of $10^{-4}$, allowing for an integration time step of 2 fs when solving the equation of motion using the leap-frog algorithm [8]. Non-bonded (van der Waals and electrostatic) interactions were handled adopting triple-range cutoff radii: interactions within the short-range cutoff of 0.8 nm were calculated every time step from a pair list that was generated every five steps, when also interactions between 0.8 and 1.4 nm were computed. The long-range electrostatic forces were represented by a reaction field with a relative permittivity of $\varepsilon_{RF} = 61$ [149] outside the long-range cutoff of 1.4 nm. The centre of mass translation and rotation were removed every 2 ps to avoid a flying ice cube [76]. All three simulations reported here, i.e. for native, for partly ester-linked and for fully ester-linked HEWL, were carried out at constant pressure, whereas the fully ester-linked HEWL simulation reported earlier [203] has been done at constant volume. This explains the slight differences between some of the results, e.g. the RMSD values as function of time in Fig. 7.2 and 8.4.

8.3.3 Analysis

All analyses of the simulation trajectories have been done using the tools of GROMOS++ [73], which is part of the GROMOS simulation package [115]. Since the properties calculated, such as NMR NOE proton-proton distance bound violations or $^3J$-coupling constants, were compared to experimentally measured values for native HEWL, the amide hydrogen positions not present in ester-linked HEWL were generated for each protein configuration of the trajectory.

This ester-to-amide (E-to-A) transformation does nothing else than converting the O$_\alpha$ atom of the ester into an N atom and placing the missing hydrogen atom such that the N–H distance is 0.1 nm and the two angles C–N–H and H–N–C$_\alpha$ have the same value. In addition, the H atom is placed to lie on the bisecting plane of the two planes defined by the C–N–C$_\alpha$ and O–C–N atoms. An E-to-A transformed trajectory looks like a native HEWL simulation trajectory and can thus be used to detect secondary structure elements in the protein using the GROMOS++ program dssp, based on the rules of Kabsch and Sander [121] as well as for other analyses that depend on the amide hydrogen positions. E-to-A transformed trajectories were not used for the hydrogen bonding analysis of ester-linked HEWL.

When comparing or averaging quantities $Q$ that depend on the position of the centre of mass and the spatial orientation of a particular set of atoms, the centres of mass are superimposed and a rotational least-square fit of the positions of the set of atoms is performed before $Q$ is calculated.
RMSD and RMSF

The atom-positional root-mean-square deviations (RMSD) between two structures have been evaluated based on all backbone non-hydrogen atoms (N/O\textsubscript{α}, C\textsubscript{α}, C and O) according to the following formula:

\[
RMSD(r^N_a, r^N_{a,\text{ref}}) = \sqrt{\frac{1}{N_a} \sum_{i=1}^{N_a} (r_i - r_{i,\text{ref}})^2} \tag{8.1}
\]

where \(r^N_a = (r_1, r_2, \ldots, r_{N_a})\) represents the positions of the atoms. In Eq. (8.1), \(N_a\) is the number of atoms considered, \(r_i\) the position of atom \(i\) in the first structure and \(r_{i,\text{ref}}\) the position of atom \(i\) in the second, reference structure. We used the X-ray crystal structure [186] as reference structure.

The atom-positional root-mean-square fluctuations (RMSF) were calculated as

\[
RMSF_i = \sqrt{\frac{1}{N_T} \sum_{t=1}^{N_T} (r_i(t) - \langle r_i \rangle)^2} \tag{8.2}
\]

where \(i\) indicates the C\textsubscript{α}-atom of residue \(i\), \(\langle r_i \rangle\) its average position, and \(N_T\) is the number of configuration time frames in the simulation trajectory.

Radius of gyration

The radius of gyration of a protein, a measure of the compactness of the structure that can be related to light-scattering intensity, was calculated using the definition

\[
R_{gyr} = \sqrt{\frac{1}{N_a} \sum_{i=1}^{N_a} (r_i - R_{cm})^2} \tag{8.3}
\]

with

\[
R_{cm} = \frac{1}{M} \sum_{i=1}^{N_a} m_i r_i \tag{8.4}
\]

and

\[
M = \sum_{i=1}^{N_a} m_i \tag{8.5}
\]

in which \(r_i\) denotes the Cartesian position of atom \(i\), \(m_i\) its mass, and \(N_a\) denotes the number of protein atoms.
Detection of secondary structure elements

The rules of Kabsch and Sander [121] have been applied to detect and monitor secondary structure elements in the native and ester-linked HEWL simulations. In some cases one residue may be assigned to be part of two different secondary structure elements. In order to avoid ambiguous assignments in such cases, the following priority rules were applied: 

\[ \beta \text{-strand}/\beta \text{-bridge} > \alpha \text{-helix} > \pi \text{-helix} > 3_10 \text{-helix} > \text{hydrogen bonded turn} > \text{bend}. \]

Calculation of $^1\text{H}–^{15}\text{N}$ order parameters

The $^1\text{H}–^{15}\text{N}$ order parameters ($S^2$) were calculated as [227, 228]

\[
S^2 = \frac{1}{2} \left[ 3 \sum_{\alpha=1}^{3} \sum_{\beta=1}^{3} \langle \mu_\alpha \mu_\beta \rangle^2 - 1 \right]
\]

(8.6)

where $\mu_\alpha$ ($\alpha = 1, 2, 3$) are the $x$, $y$, and $z$ components of the normalized inter-atomic N–H vector, and compared to experimental data for native HEWL [26].

NMR NOE proton-proton distances and $^3J$-coupling constants

A detailed description of the calculation procedures with equations and corrections can be found in [42]. 1630 experimental NMR proton-proton upper distance bounds were taken from [28], including pseudo-atom distance corrections as given by Wüthrich et al. [187] and the $^3J$-coupling constants reported by Smith et al. [29] were used for comparison. Proton-proton distances were calculated using $1/r^3$ averaging, $\bar{r} = (\langle r^{-3} \rangle)^{-1/3}$.

8.3.4 Calculation of backbone $^{15}\text{N}$-$^1\text{H}$, $^{13}\text{C}_\alpha$-$^{13}\text{C}$, and $^{13}\text{C}$-$^{15}\text{N}$ residual dipolar couplings

The residual dipolar coupling (RDC) $D_{ij}$ between two spins $i$ and $j$ is calculated according to

\[
D_{ij} = -\frac{\gamma_i \gamma_j \mu_0 h}{8 \pi^3} \left\langle \frac{P_2(\cos(\theta_{ij}))}{r_{ij}^3} \right\rangle,
\]

(8.7)

where $\gamma_i$ and $\gamma_j$ are the gyromagnetic ratios of the two spins, $\mu_0$ is the magnetic permittivity of vacuum and $h$ is Planck’s constant. $P_2$ denotes the second-order Legendre polynomial and $\theta_{ij}$ is the angle between the inter-nuclear vector $r_{ij}$ and the static magnetic field. Since there is typically only one copy of the molecule in an MD simulation, Eq. (8.7) is reformulated so that the averaging over different orientations of a given structure of the
molecule with respect to the magnetic field is represented by an alignment tensor $A$,
\[ D_{ij} = -\frac{\gamma_i \gamma_j \mu_0 h}{8\pi^3 r_{ij}} \sum_{k \in \{x,y,z\}} \sum_{l \in \{x,y,z\}} A_{kl} \cos(\zeta_k) \cos(\zeta_l) , \]  
(8.8)
where $\zeta_x$ is the angle between the inter-nuclear vector and the $x$-axis and $\zeta_y$, $\zeta_z$ are the angles for the $y$- and $z$-axes, respectively. For every configuration in the trajectory the alignment tensor $A$ is determined by a singular-value decomposition (SVD) fit, solving the equation
\[ Ca = R , \]  
(8.9)
where the 5-dimensional vector $a = (A_{xx}, A_{yy}, A_{xy}, A_{xz}, A_{yz})$ contains the five independent elements of the $3 \times 3$ alignment tensor, and $R$ contains the $N_D$ experimental RDCs used for the fit. A more complete description of RDC calculations using GROMOS is described in reference [53].

The quality of each SVD fit is assessed by the $Q$-value,
\[ Q = \sqrt{\frac{1}{N_D} \sum_{i=1}^{N_D} (D_{i}^{exp} - D_{i}^{calc})^2} \sqrt{\frac{1}{N_D} \sum_{i=1}^{N_D} (D_{i}^{exp})^2} , \]  
(8.10)
which can be displayed as a distribution resulting from all configurations within a simulation trajectory.

**Hydrogen bonds**

Hydrogen bonds were defined according to a geometric criterion: a minimum donor-hydrogen-acceptor angle of 135° and a maximum hydrogen-acceptor distance of 0.25 nm. The hbond program of GROMOS++ [73] has been used to detect and monitor hydrogen bonds in native and ester-linked HEWL.

### 8.4 Conclusion

This study was motivated by a previous work [203] which showed a compaction of fully ester-linked HEWL, i.e. a protein bearing 126 amide-to-ester replacements, compared to native HEWL in 3.5 ns molecular dynamics simulations. In fully ester-linked HEWL 85% of all 1630 NOE proton-proton distance bounds derived from NMR experiment were still satisfied within 0.1 nm. The protein mutants showed a shifted hydrogen bonding pattern: only about 25% of the hydrogen bonds in the native simulation were also observed in the fully ester-linked HEWL simulation, and vice versa.

Here, we repeated the simulations of native and fully ester-linked HEWL, but at con-
stant pressure for 10 ns, and added a simulation of partly ester-linked HEWL, an intermediate state with 34 amide-to-ester replacements outside the helical and β-strand secondary structure elements observed in native HEWL.

It is not surprising that the simulations of native HEWL and partly ester-linked HEWL preserved the secondary structure known from native HEWL during the simulations. In contrast, fully ester-linked HEWL lost all its secondary structure during the first few nanoseconds of the simulation, doubled its structural difference from the native X-ray structure, yet showed a larger compaction during the simulation than native or partly ester-linked HEWL. Since experimental data from NMR experiments involve local information, it was of interest to investigate how sensitive the observable quantities are to local spatial rearrangements of the protein that maintain its compactness.

Overall, the simulation of partly ester-linked HEWL agreed with the experimental NMR data measured for native HEWL equally well as the simulation of native HEWL itself. $^3J_{HN-Hα}$-coupling constants, NOE proton-proton distances, and backbone residual dipolar couplings of partly ester-linked HEWL agreed better with the experimental data, while the $^1$H–$^{15}$N order parameters and the $^3J_{αβ}$-coupling constants agreed less well than in the simulation of native HEWL. This indicates that the considered observables cannot be used to distinguish local structural rearrangements. This is not too surprising considering the many approximations and model assumptions that flow into the conversion of NOE intensities or relaxation rates into distances and order parameters. For the $^3J$-couplings, the empirical character of the Karplus relation between local structure and $^3J$-coupling is a source of considerable uncertainty. Only larger structural changes, as observed here for fully ester-linked HEWL seem to be detectable by the NMR parameters considered here.

The observed compact fold of ester-linked HEWL can be explained as follows: At the beginning of the simulation, starting from the native X-ray structure, there are a lot of non-hydrogen bonded acceptors in the backbone of ester-linked HEWL. This unfavourable hydrogen-bonding configuration of the $O_α$ and $O=C$ atoms is rearranged during the simulation by forming new hydrogen bonds, preferentially within the protein and not to the solvent, which help to stabilise the protein structure. This rearrangement is reflected by the changed but not reduced hydrogen-bond pattern found previously for fully ester-linked HEWL [203] and by the increasing number and occurrence of backbone-side chain and/or side chain-side chain hydrogen bonds connecting two secondary structure elements (helix–helix or β-strand–β-strand) or two residues involved in the hydrophobic box, respectively. Thus, the compact fold of fully ester-linked HEWL is maintained by hydrogen bonds involving side chains.

It would be of interest to actually synthesise fully ester-linked HEWL to verify the conclusions reached from the simulations. Up till now, ester-linked versions of proteins are only available for shorter polypeptide chains.
Chapter 9

Outlook

Computer simulations have been proven to be a useful and important tool complementing experimental methods while investigating structural and dynamic aspects of processes in biomolecular research. They provide insight into the subject of interest at basically any level of resolution, depending on the choice of degrees of freedom considered. However, simulations are limited by the required computational effort as well as the model and methods they are based on. Computational power has continuously increased according to Moore’s law during the past decades and will further progress in the future. Therefore, simulations can become more and more complex, detailed and computationally demanding but the force fields and algorithms used, i.e. the code of the simulation software, has to be maintained and adapted by further development and testing.

In Chapter 2 a method to apply bond-angle and dihedral-angle constraints in molecular dynamics simulation in analogy to SHAKE [14] was presented. The application of SHAKE [14] to bond lengths in macromolecules or solvent molecules allows for a four times bigger integration time step when solving Newton’s equations of motion. The proposed method is not meant to replace SHAKE but offers an appropriate alternative in cases where SHAKE fails, e.g. for linear solvent molecules such as acetonitrile (MeCN). The use of such a solvent in biomolecular studies, i.e. mostly proteins in aqueous solution, is questionable but may contribute to investigation and understanding of important factors maintaining the stability and fold of proteins. Experimental studies showed that MeCN acting as a polar, aprotic cosolvent reduces the temperature of denaturation of lysozyme with increasing concentration, but a stabilisation was found for low concentration of MeCN and temperatures below room temperature, most likely due to the increase of hydrophobic interactions [229]. This calls for verification by a study using MD simulation, for which the proposed method of bond-angle constraining would be appropriate and of great use.

A similar study to the one proposed above was done on the effect of 2,2,2-triuroethanol (TFE) cosolvent upon the structure of hen egg white lysozyme (HEWL), Chapter 6. A configurational ensemble for the fold of the protein in TFE/water was generated by a sequence of MD simulations and shows a good agreement with experimental data derived by NMR spectroscopy [184]. However, the structural change of lysozyme from a native-like to a TFE fold could only be observed using enhanced sampling at higher temperatures or by applying distance restraining with respect to experimentally derived $^1$H–$^1$H NMR NOE upper bound distances [184] and not in a free simulation at 310 K under the conditions described in experimental studies [178]. Thus, the molecular model used in this study might be unable to reproduce the structural change in the protein configuration. It is well...
possible that further developed force fields for TFE and proteins, e.g. polarisable ones, will be able to reproduce such processes in the future.

The development of a supra-atomic CG GROMOS force field for aliphatic hydrocarbons in the liquid phase, calibrated on properties derived from the corresponding FG simulations and parametrised with respect to experimental data, was sketched in Chapter 3. The force field looks promising but needs to be further tested, especially its performance in mixtures with other compounds or in conjunction with atomic-level force fields, e.g. FG alkanes or CG and FG water. The proposed model uses SHAKE [14] in combination with a standard integration time step of 2fs, while it would allow for much larger time steps even without SHAKE as used in other descriptions of CG alkanes [67]. However, since the presented model is to serve as a precursor model for CG lipids to be simulated together with FG proteins or other atomic-level solute molecules, in a so-called mixed-grained simulation as recently done for FG proteins in FG and CG water [9, 64], the restriction to an integration time step of 2fs is required. The implementation of a multiple time step integration algorithm in GROMOS, e.g. as proposed in [10], would lead to an additional gain in simulation speed for mixed-grained systems. Another improvement of the CG alkane model could be established by not applying constraints and shifting the reference angle of the bond-angle potential energy term to 180°, as is done in other models [66–68]. This would correct the slightly shifted bond-angle distribution. In combination with a new description of the bond-stretching potential energy term, e.g. the sum of a harmonic potential energy overlapped with a Gaussian

\[
V_{ij}(b_0, K_0; a, b, \mu, \sigma; \mathbf{r}^N) = \frac{1}{2} K_0 (r_{ij} - b_0)^2 + a \exp \left( \frac{-(r_{ij} - \mu)^2}{2\sigma^2} \right) - d, \quad (9.1)
\]

where \(r_{ij}\) denotes the distance between two bonded particles \(i\) and \(j\), \(b_0\) and \(K_0\) are reference bond-length and force constant of the bonded interaction, and \(a\), \(\mu\) and \(\sigma\) are used to tune the height, position and width of the Gaussian function while \(d\) is used to shift the minimum potential energy, such an approach would recover the bead-bead bond length distribution as seen in the analysis of the mapped FG simulations for bead sizes up to 3 (Fig. 3.1). A similar approach is possible for the bond angles but of less relevance since the recommended bead size of 3 does not show multiple peaks in the bond-angle distribution (Fig. 3.1). However, a reference bond angle of 180° calls for trouble caused by the singularity in the definition of the torsional-angle potential energy definition. A quick solution could set the torsional-angle force to zero for linear configurations but needs to be investigated in more detail with respect to energy conservation.

In Chapter 4 an overview of the GRMOMOS++ software for the analysis of biomolecular simulation trajectories is given. It has the advantage to have been tested by the IGC group and collaborating groups all over the world and is therefore less error prone than the same routines and algorithms reprogrammed by every single user. The object-oriented design of the code avoids code duplicity and can easily be extended by new programs using the routines already present. It is and will be a major work to maintain such a code but
definitely worth the effort.

In Chapter 5, values of differences in folding free enthalpy differences, $\Delta \Delta G_{\text{fu}}^{\text{mw}}$, were computed for 20 different amide-to-ester mutants of the Pin1 WW domain using enveloping distribution sampling (EDS). Based on a thermodynamic cycle (Fig. 5.2) these values can be obtained from the difference in free enthalpy between the wild type (w) and mutant (m) of the protein in its folded (f) and unfolded (u) state, $\Delta \Delta G_{\text{fu}}^{\text{mw}} = \Delta G_{\text{uf}}^{\text{mw}} - \Delta G_{\text{fu}}^{\text{mw}}$. Accurate sampling of an unfolded protein state is beyond the means of current computer simulations. Therefore, the difference in folding free enthalpy of the unfolded state $\Delta G_{\text{uw}}^{\text{mw}}$ was estimated from a simulation of an Ala-tripeptide, thereby assuming $\Delta G_{\text{uw}}^{\text{mw}} = \Delta G_{0}^{\text{uw}}$ to be constant (Eq. (5.6)) for all mutants or linearly dependent on the difference in free enthalpy of the folded state (Eq. (5.7)). These assumptions turned out to be poor or at least incomplete, e.g. due to the neglect of possible effects due to the amino-acid side chain of the residue that is mutated or its neighbours along the residue chain. 20 simulations of short, terminally blocked peptides, i.e. 3–5 amino-acid residues, would allow for corroborating or disproving the mentioned assumptions and for testing of a more refined model for $\Delta G_{\text{uw}}^{\text{mw}}$.

In Chapters 7 and 8 the effect of backbone amide-to-ester substitutions is studied at the example of fully, i.e. performing all possible backbone amide-to-ester replacements, and partly ester-linked HEWL, i.e. amide-to-ester replacements for all residues not involved in one of the main secondary structure elements of the protein. Partly-ester linked HEWL largely maintains the secondary protein structure, while fully-ester linked HEWL does not but still keeps a compact fold. Further studies including amide-to-ester substitutions of residues involved in the secondary structure elements, e.g. every second or third residue, would be of interest to study at which level of amide-to-ester replacement the protein secondary structure breaks down.
Bibliography


177


[86] K. J. Bowers, E. Chow, Huageng Xu, R. O. Dror, M. P. Eastwood, B. A. Gregersen, J. L. Klepeis, I. Kolossvary, M. A. Moraes, F. D. Sacerdoti, J. K. Salmon, Y. Shan,

SC 2006 Proceedings Supercomputing 2006, Tampa, FL, USA.


BIBLIOGRAPHY


Curriculum Vitae

Personal Data
  Name           Andreas Paul Eichenberger
  Birth          January 8, 1982, Olten (AG)
  Nationality    Swiss

Education
  2008-2012      Doctor of Science, ETH Zürich, Switzerland
  2004-2008      Master of Science, ETH Zürich, Switzerland
  1998-2002      Matura, Kantonsschule Baden, Switzerland

Teaching
  2012           10th NCCR Practical Course and 3rd NCCR Winter Course
                  Practical course in biomolecular modelling
                  Kandersteg, Switzerland
  2010           8th NCCR Practical Course and 2nd NCCR Winter Course
                  Practical course in biomolecular modelling
                  Kandersteg, Switzerland