A Molecular View on the Organizational Complexity of Proteins in Membranes

Djurre H. de Jong

Foar Emma

Cover: Artist impression of cluster of proteins (red) being shuttled into the liquid ordered domain of a lipid bilayer by GM1 ganglioside lipids (blue). The bilayer consists of unsaturated lipids (white) forming the liquid disordered domain and saturated lipids (grey) and cholesterol (green) forming the liquid ordered domain.

Typesetting by LaTeX
Software used for cover: VMD version 1.9, GIMP version 2.6

Printed by:

ISBN (printed version): 978-90-367-6075-1
ISBN (digital version): 978-90-367-6122-2

rijksuniversiteit
groningen

A Molecular View on the Organizational Complexity of Proteins in Membranes

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschapen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
24 mei 2013
om 14:30 uur

door

Djurre Hendrik de Jong

geboren op 8 januari 1985
te Joure

Promotor:	Prof. dr. S.J Marrink
Copromotor:	Dr. X Periole
Beoordelingscommissie:	Prof. dr. J.A Killian Prof. dr. P.R Onck Prof. dr. D.P Tieleman

ISBN (printed version): 978-90-367-6075-1
ISBN (digital version): 978-90-367-6122-2

Contents

Preface

In the four years since I joined the MD-group[1] it has always been a very "dynamic" bunch. Dynamic in the sense of the many events that were always organized (including but not limited to: groupoutings, parties, spontaneous dinners, the many conferences, WABSes, VrijMiBo's and, very important, lunchtime running.), but also dynamic in the almost infinite number of people that have joined the group for shorter or longer periods. It was very much fun to work with all of them and many have contributed in some way to the realization of this thesis. It will be virtually impossible to thank everybody and not forget a single person, but I'll try. In case you find your name missing, I've left over some space for this to be corrected. Please help yourself out and don't feel obliged to be honest.

Dear ....., you are one of the most ..... people I have ever known. Without your contributions this thesis would have been nonexistent/marginalized/much better*. Via this way would like to express my gratitude and I hope we will have many inspiring discussions/sportive moments/wild nights* to come.

*Cross what's not applicable.

Allereerst Siewert Jan, als mijn promotor niet geheel onbelangrijk voor dit proefschrift. Siewert, ontzettend bedankt voor de vrijheid en het vertrouwen dat je mij altijd hebt gegeven. Daardoor wordt werken in de MD-group een van de leukste dingen die er zijn. Bedankt!

Xavier, thank you for everything you have learned me as my daily supervisor and copromotor. Thank you for learning me how to be a critical and thorough scientist, how to properly drink coffee, the importance of recycling coffee cups, how to be grumpy without being serious, how to be serious without being grumpy and most of all, thank you for bringing all the fun into the office.

Lars, you really kick-started my PhD and were the perfect source to learn about MD. Without you there would definitely be a few chapters less in this thesis. Ik hoop dat het oefenen van het Nederlands voor jou net zo nuttig was.

Alex, als (bijna?) oneindige bron van kennis heb jij meestal een antwoord op al mijn vragen over MD, en anders wel een boek om het in op te zoeken.

[1] It has actually been 7 years since I first joined to group, counting my bachelor and master projects in the NMR-group.

Helgi, if I ever get fat and obese it will be because of all the chocolate cookies you brought to the lab. If you have someone in your office with a strong appetite for good food, who also happens to have a well founded opinion on virtually anything, life at the office becomes a real treat.

Clement, from listening to your stories I understand all the French are bad-mooded, rebelling criminals. From knowing you I discovered the opposite is true. With you, everything one does seems to be fun and relaxed. Special thanks from Sir Ritchfield!

Jaakko, even though you only joined the group a year ago, we have been running and discussing together quite a bit. I really enjoyed it and I hope we ever get the opportunity train for a marathon that we both finish below 3:30h. Next time Ill beat you!

Manel, when thinking of you two things come to mind: stokslobberjen and food. I'm glad I have contributed a little bit to your knowledge of the Frisian language and you've made me try some excellent (Portugese) dishes.

Cesar, since you were the PdD-student that started in the group before me, you were a great example in many ways. I could use your work in many places (hence our faraday paper), I could ask you advice when writing this thesis and, very important, I could get a feel for how it is to be a young father.

De meest ingewikkelde problemen die een PdH-student moet oplossen zijn de in te vullen reisdeclaraties en andere bureaucratische rompslomp. Hilda, jij was onmisbaar in het oplossen van die puzzels, bedankt!

Floris, als "Photosystem"-expert doe jij veel van het werk dat ik misschien eigenlijk had moeten doen. Ik hoop dat we daardoor nog veel kunnen samenwerken in de komende maanden. Bedankt dat je de organisatie van de Dutch MD Day van me hebt overgenomen, als het zo blijft groeien als nu is dat in no time een begrip!

Juan Juan, you have barely joined our group but are already involved in so many things, amongst which the continuation of the work on the softer potentials in this thesis. You spotted some serious error in that chapter just hours before I was going to send it to the printer. Thanks for that! I hope you haven't proven it completely useless by now...

Ruud en Pieter, ik vind het altijd erg gezellig als jullie langs komen, wat helaas niet erg vaak gebeurt (maar toch altijd nog vaker dan dat ik bij jullie langs kom). Misschien kun je meer koffie gaan drinken (Ruud) of vaker de spectrometer bij ons in het gebouw komen bijvullen (Pieter)?

Alia, we seem to be doing things in parallel all the time. We did our master project in the MD-group at the same time, we started our PhD at the same time and now you are expecting ("gave birth" by the time you read this) at the same time as Jorieke. I find it truely heroic how you manage to finish your PhD and prepare to become a mom at the same time!

Tsjerk, dat wittenskip net oan in taal bûn is bewize wy eltse dei: Frysk, Hollânsk of Ingelsk it makket foar us diskusjes neat út. Ik hoopje dast snel in mear "stabyl" karrière paad fynst en o ja, you're a SIMPLETON!! (Smart and Infintely More Python-oriented Explorer of Tons of Original New scripts),

Marcelo, you've been cast in mysteries. The first few years we never met, although we were formally part of the same group (right?), later it was always a question at which floor you would be. Besides that it never became clear to me whether or not you speak Dutch.

Andrzej, I really enjoyed the runs with you, which were good because of the running but most of all because of the company. I specifically remember two runs: one in a blazing snowstorm in Groningen and two weeks later in San Fransisco to celebrate Chinese new year!

Tom, als mijn student heb jij me letterlijk werk uit handen genomen. Jouw master project in de MD-group vormt de basis van een van de hoofdstukken in deze thesis. Bedankt en veel succes met je eigen promotie in Amsterdam!

Many more people have been part of group while I was there but left longer ago. I would dearly like to thank: Daniela Mueller, Durba Sengupta, Egon Rypkema (ik hoop dat we nog veel elfstedentochten kunnen fietsen), Frans Mulder, Hein Wijma, Herman Berendsen, Jelger Risselada, Kamil Tamiola, Klaas Dijkstra, Marc Fuhrmans, Martti Louhivuiri, Monica Bulacu, Nicu Goga, Renee Otten and Srinivasa Gopal. From the people that visited Groningen for a shorter period, I would like to mention Felipe Franco (who is from Colombia and not Bolivia), Iwona Siuda, Jonathan Barnoud, Lucca Monticelli, Martina Pannuzo and Zofie Sovova. Thank you all very much!

I would also like to thank the people asking stupid questions at the Martini forum. People often say there is no such thing as a stupid question, but believe me... Luckily there were also a lot of interesting, useful and fun-to-answer questions. And if you are reading this: your question was ofcourse one of the latter.

Besides the people in the MD-group in Groningen, there are also a number of people from all over the world I worked with on the various papers (Some of which I have never met and who will probably never read this). There are Norbert Kucerka, Drew Marquardt, Thad Harroun, Mu-Ping Nieh, Stephen Wassall and John Katsaras for the 2010 *Biochemistry* paper. Andrea Holt, Bert Poolman and Antoinette Killian for our PNAS paper. Next there is Helmut Grubmüller for the JCC paper on equilibrium constants: I really enjoyed working with you and my visit to you lab, even though it was only one day. With Mingyang Hu and Markus Deserno I had a lot of fun working on the Kappa-Bar paper. The time difference between Pittsburg and Groningen is perfect for efficient e-mail collaborations! And most recently Gurpreet Singh, Drew Bennett and Peter Tieleman for working together on the medley of improvements on Martini, called Martini 2.2.

De leescommissie, bestaande uit Antoinette Killian, Peter Tieleman en Patrick Onck, wil ik graag bedanken voor het kritisch doorlezen van dit proefschrift.

Gelukkig was er naast het werk ook nog tijd voor ontspanning en ook dat heb ik altijd in roemrucht gezelschap kunnen doen. Allereerst wil ik iedereen in ons "elftal" (Anne, Astrid, Els, Hilde, Jorieke, Jurrit Jan, Klaas, Marjo, Robert en Wolter) bedanken dat zich jaar in, jaar uit een weekend lang laat op sluiten in een hutje op de hei. Onze weekendjes weg en alle andere momenten (O&N, Wampex, etc) zijn fantastisch en ik hoop dat we nog lang door kunnen gaan. Jurrit en Robert wol ik tiiche dank sizze foar us avontoeren op de fyts. Langs de Alve stden, yn Switserlan of der tusken yn. Henk en Alma, bedankt voor de gezellige diner avondjes. Vervolgens wil ik de leden van de XL-lustrum commissie bedanken voor hun masochistische voorkeur voor slechte kroegen en dito drank. Ik hoop dat we nog lang door gaan met het steunen van de firma Baccardi.

Myn family: myn heit en mem, Jelte en Elsje, Berber en Ruurd en eltsenyn oars. It is tiige fyn te witten dat d'r altyd minsken binnen dyt d'r foar je binne. Jelte en Berber, ik bin ek tiige fereare dat jim tydens de ferdigening myn paranimfen wêze wolle.

De laatste woorden zijn natuurlijk voor Jorieke. Jij was er altijd de afgelopen jaren: als het goed ging en leuk was, maar vooral ook als het tegenzat. Zonder jou had ik waarschijnlijk dubbel zo veel tijd aan mijn proefschrift besteed, maar was het niet half zo goed geworden. Ik hoop dat we nog heel lang van de wereld kunnen genieten, niet meer met z'n tweeën maar met z'n drieën.

1

Introduction to Molecular Dynamics: A brief history and overview

The field of *Molecular Dynamics* (MD) can be defined as the study of the interactions between and the movements of molecules. However, in recent decades, and certainly in this thesis, the term molecular dynamics has become a synonym for *molecular dynamics computer simulations*. The interactions between and movements of molecules have been studied extensively since the 19^{th} century, for example in the fields of thermodynamics and statistical mechanics by people like Clausius, Gibbs, Boltzmann and Van der Waals. The field of computational molecular dynamics, due to its dependence on extensive, iterated computation, had to wait for the invention of the electronic computer halfway of the 20^{th} century to be feasible. Despite this relative short history, the field has expanded at an amazing pace and found applications in a wide range of research areas. Several papers giving a short historical overview of the field have been written. Both Levitt[1] and Karplus[2], have written down the developments in light of their personal careers. Furthermore, some papers provide a combination of a short historical overview and an outlook to the future[3, 4]. However, no comprehensive work on the history of molecular dynamics has been written, but that is also not what this chapter aspires to. It pinpoints those studies that played a pioneering role in the field treated in the rest of this thesis: molecular dynamics simulations of bio-molecules.

1.1 First molecular dynamics computer simulations

In 1955 Fermi, Pasta and Ulam published the first results of what one could call a Molecular Dynamics simulation:

> *"A one-dimensional dynamical system of 64 particles with forces between neighbors containing nonlinear terms has been studied on the Los Alamos computer MANIAC I."[5]*

A one-dimensional string of beads can hardly be considered a set of molecules and thus the step to higher dimensional systems was quickly made. In 1957 Alder and Wainwright published the results of a two-dimensional system of 'spheres'[6], which is often considered to be the first MD-paper. The first extensive description of the MD-method and the algorithms used, were published two years later[7] by the same authors. The method of (computational) molecular dynamics was preceded by

a few years by the very related method of Monte Carlo simulations[8, 9]. In those early studies, the method was mostly applied at studying physical phenomena such as the phase transitions and equations of state of simple spheres. In those early papers, many of the present concepts (Periodic Boundary Conditions, Lennard Jones potentials) were already present, as well as the present concerns:

> *"It is sometimes a worry whether a system is in equilibrium or whether it is in a meta stable state[6]."*

1.2 Towards realistic molecules

The first simulations of actual molecules were published in 1964 by Rahman, who simulated a system of 864 liquid argon atoms[10]. Seven years later it was also Rahman who reported the first molecular dynamics simulation of water[11]. Again, MD was preceded by two years by Monte Carlo simulations by Barker and Watts[12]. The interactions in their water model were described as:

> *"a) the electrostatic interaction of four charges on one molecule with four on another, b) a Lennard-Jones 6:12 function of the distance between the charge centers of the molecules, with parameters* $\epsilon = 0.707$ *kcal/mole, sigma* $= 2.725$ *Å. The charges are* $+0.3278e$ *at* $0.5944a \pm 0.76166b$ *and* $-0.3278e$ *at* $\pm 3.2539c$*; e is the electronic charge and a, b and c are orthogonal unit (Å) vectors fixed in the molecule. The origin is the oxygen nucleus and the charge center is* $0.2922a$*. We imposed a hard-sphere cut-off: at* $2Å$ *and neglected interactions beyond* 6.2 *Å."*

A significantly improved model was introduced in 1974[13] dubbed the ST2[1] model, which was widely used until the introduction in 1981 of the two most used water models of today: the *Simple Point Charge* (SPC) model[15] and the *Transferable Inter-molecular Potential functionS* (TIPS)[16] by the groups of Berendsen and Jorgensen, respectively. Both apply a simple, 3-site topology and fairly well reproduce properties of water such as density and oxygen radial distribution function. In later years different refinements were published for both the SPC and TIP series[17, 18, 19], further tweaking the properties of the water model.

One of the motivations behind developing potentials for water models, was the desire to simulate bio-molecules, like proteins[20]. The first simulations of proteins were lacking solvent water, mostly due to large computational cost related to the still very limited computational power. The first atomistic simulation of a protein in 1977 contained four water molecules, strongly bound inside the structure[21]. In this study McCammon *et al.* used an atomistic interaction potential, derived in earlier work to minimize the energy of X-ray structures, to simulate the 58 amino acid Bovine Pancreatic Trypsin Inhibitor (BPTI) for 8.8 ps. In the potential no explicit hydrogen atoms were present (a so called 'united atom' model) and bonds, bond angles, dihedral angles, hydrogen bonds, van der Waals, and electrostatic interactions were all explicitly described.

[1] No explanation for the acronym is given in the original article. However, the ST2 model is a successor of the BNS model [14]. These letters are the first letters of the name of the authors of that paper (Ben-Naim and Stillinger). Apparently the first letters of the authors of the second paper (SR, Stillinger and Rahman) were deemed less suitable then the initials of the first author.

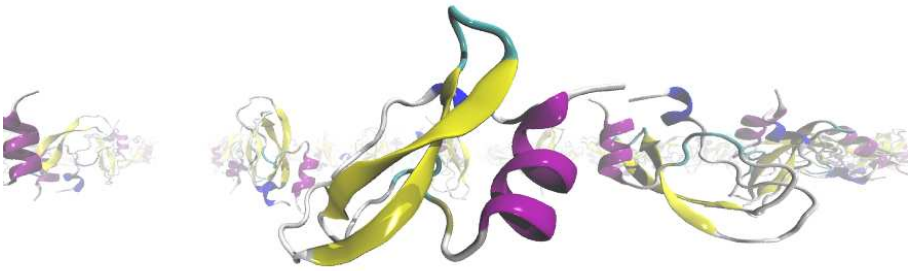

Figure 1.1: Randomly oriented copies of Bovine Pancreatic Trypsin Inhibitor (BPTI). Different secondary structure elements have been colored differently.

The simulated protein, BPTI (Figure 1.1), was very often the subject of choice in the early days of MD computer simulations. First because of its small size and second because it was one of the first proteins for which a high resolution crystal structure was available. In another study using BPTI, two years before McCammon published his atomistic simulations, Levitt *et al.* published[22] the results of what is arguably an even earlier MD simulation of a protein. However, in this study every residue was represented by only one or two interaction sites, instead of one interactions site for every atom. This does make it the first 'coarse grain' simulation of a protein, though.

Different papers might be honored as the first simulation of a water solvated protein, all of them using BPTI. The 1988 paper by Levitt [23] is often quoted to be the first, simulating BPTI solvated in 2607 water molecules for 210 ps. However, Van Gunsteren *et al.* simulated BPTI in a truncated octahedron surrounded by 1467 water molecules for 20 ps already in 1984[24] and Wong *et al.* simulated BPTI in 4785 water molecules for 42 ps in 1987[25]. In his paper Levitt dismisses the other two papers by arguing that the simulations were either to short and showed a too high root mean square deviation of the atoms positions (2.7 Å compared to 1.1 Å in his simulations) in the case of the work by Van Gunsteren, or they where not properly analyzed and gave a too high diffusion constant for bulk water in the case of the work by Wong.

Around the same time as these simulations of proteins, the first simulations of another type of bio-molecule were published: lipids and the complex structures they form. In 1976 a first onset to the simulations of lipid bilayers was given by Cotterill[26] using a highly abstract model of a monolayer: a two dimensional system of 240 dumbbell shaped 'molecules'. These abstract simulation were only a crude exploration and were followed in 1980 by the first atomistic simulation of a monolayer[27] and of the first atomistic simulation of a bilayer in 1982[28], both omitting explicit water. In 1986 the first study of a micelle in explicit water was published by Jönsson *et al.*[29] They simulated 15 sodium octanate molecules surrounded by 1094 water molecules for 260 ps (taking 19 minutes of CPU time per picosecond simulation time on a Cray-1 computer). Two years later Egberts *et al.* reported the first bilayer with explicit water: 52 decanoate, 76 decanol, and 526 water molecules, and 52 sodium ions were simulated for 180 ps[30]. All these systems were rather simple model systems

for lipid membranes. The first study of an atomistic phospholipid (DPPC) bilayer was published in 1994[31][2]. These bilayers were pre-assembled at the start of the simulation. Only seven years later, in 2001, the first simulation of the self-assembly of a lipid bilayer was published. Marrink *et al.* [33] described how in a simulation of 64 randomly oriented DPPC-molecules a bilayer spontaneously formed. Additionally, they studied other lipids (POPC, DOPC and DOPE) and larger systems, all forming stable bilayers.

A logical next step, combining lipids and proteins, was made by Edholm *et al.* in 1995 by simulating bacteriorhodopsin in a lipid bilayer[34]. Although this was the first simulation of an integral membrane protein, several examples of (integral) membrane peptides and soluble proteins interacting with a membrane were published in the preceding years. Good reviews on membrane (protein) simulations are given in [35, 36].

One of the main focus areas of molecular dynamics simulations has been the (ab-initio) folding of proteins. The large interest in this area makes it hard to define the first successful atomistic simulation of a protein in explicit solvent at physiological temperature. There are many examples of protein unfolding, partial folding or folding of small peptides, however the first simulation successfully folding a natural and complete protein is hard to pinpoint. The earlier mentioned work by Levitt [22] used a coarse grain model and was only marginally successful in folding, but is the first *attempt* to fold a protein using molecular dynamics. In 1998, Daura *et al.* simulated the reversible folding a β-heptapeptide in methanol at atomistic resolution, and the obtained results were in good agreement with experiment[37, 38]. However it concerned a designed, non-natural peptide folded into only a very short left-handed 3_1-helix. Also in 1998 Duan *et al.* published the results of the first 1 μs simulation, showing the folding of the Villin Headpiece[39]. The peptide studied is considerably larger then the heptapeptide, it concerns a domain of a natural protein and was simulated in water. However, Villin Headpiece is not a complete protein and, in the words of the authors, only *"A marginally stable state, which has a lifetime of about 150 nanoseconds, a favorable solvation free energy, and shows significant resemblance to the native structure"* is reached.

In 2002, Snow *et al.* reported on studies of the folding of a 23 residue (designed) protein[40], consisting of a β-hairpin, a turn and an α-helix (using implicit solvation). In this study the concept of *distributed computing* was used, where not one single long simulation was done to see multiple folding and unfolding events, but tens of thousands of simulations of 5-20 ns that yielded several folding events. The simulations were done at the 30,000 PC's of volunteers all over the world participating in the folding@home project[41]. The protein simulated was "complete" in that it contained all the major secondary structure elements, however it still concerned a designed protein and was simulated in implicit water.

In 2011, in the first molecular dynamics study reaching the millisecond for a single trajectory[42], the reversible folding was reported, using a 23 residue WW-domain designed to have fast folding and unfolding times. In the same study BPTI could only be studied in its folded state, since no complete unfolding events took place

[2] In 1993 Marrink *et al.* published a paper[32] describing full atomistic bilayer simulations. However this paper refers 'back' to the 1994 paper by Egberts *et al.*[31] (as "in press"). The work in this second paper was already part of the thesis of Egberts published in 1988[32]. This makes the work described in the 1994 paper the first simulations of an atomistic bilayer

within one millisecond (in agreement with experiment). The long time scales assessed in this study could be reached by the use of a computer specifically designed for MD simulations[43][3], affectionately called *Anton* (named after the inventor of the microscope, Antonie van Leeuwenhoek).

In later papers[46, 47] from the two previously mentioned groups, they utilize their vast computer power (either distributed in *folding@home* or specifically built as *Anton*) to fold several larger, fast folding proteins with varying success. If anything, this shows that although the first atomistic simulations of the folding of small peptides were published over 14 years ago, the folding of real proteins by MD-simulations is not an every day activity yet and we have yet to fully understand protein folding.

Besides water, proteins and lipids, MD computer simulations have been extensively used to simulate many other (biological) molecules, such as DNA[48], carbohydrates[49] and small (organic) molecules. Since these molecules play only a minor role in this thesis, the reader is referred to the given references for reviews on the work in these fields.

1.3 Ingredients of a MD computational model.

In the 70s and 80s more and more effort started going into the derivation of the interaction potentials of different molecules and the development of new algorithms. Because of this the force fields and software grew more sophisticated and complex, however the main components never changed. Figure 1.2 schematically shows the ingredients brought together by combining software and forcefields: the molecular coordinates (top left) are inputted by the user (this might be a protein, bilayer or any other molecule). The force field (bottom left) describes the interactions between all atoms in the molecules. The software has implemented algorithms to integrate the laws of motion and evolve the system time (bottom right) while maintaining system properties such as temperature and pressure (top right).

The force field (Figure 1.2, bottom left) consists of a set of parameters for interaction potentials of a specific form. Some of the most notable and most used atomistic force fields are AMBER (Assisted Model Building with Energy Refinement), CHARMM (Chemistry at HARvard Macromolecular Mechanics), GROMOS (GROningen MOlecular Simulation) and OPLS (Optimized Potentials for Liquid Simulations). These force fields all date back to the 1980s and different versions have been published since then. Despite the many differences, they also share some basic characteristics: interaction sites are positioned at atoms only, interacting with other, non-bonded atoms via a Lennard-Jones potential (Van der Waals interactions) and a $1/r$ electrostatic term (charge-charge interactions). Interactions with atoms in the same molecule are propagated via harmonic springs (bonds, for first neighbor interactions and angles, for second neighbors), or cosine based functions (dihedral angles, for third neighbors). It is beyond the scope of this chapter to give an extensive review of the inner workings of, and differences between those force fields. A comparison is given by Guvench *et al.*[50].

[3] Although most MD simulations are these days ran on general purpose (super-)computers, Anton is not the first machine specifically designed for MD simulations. Notable examples are the MDGRAPE[44] developed by the RIKEN institute in Japan and the GROMACS[45] (Groningen Machine for Chemical Simulations) built at the university of Groningen.

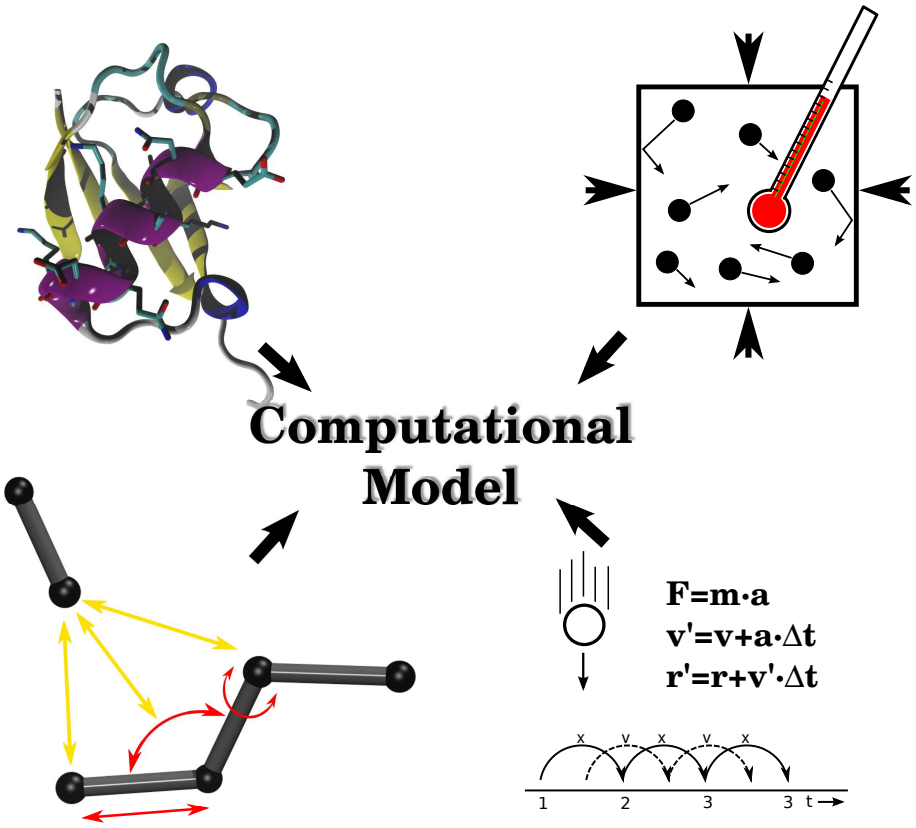

Figure 1.2: Main ingredients of molecular dynamics computation model. Starting from the upper left in counter-clockwise direction: **Atomistic positions.** The positions of all atoms of all molecules, including solvent. Protein positions are often taken from the PDB-database. **Forcefield.** Definition of the interactions between atoms. Red and yellow arrows indicate bonded (bond-stretching, angle-bending and dihedral rotations) and non-bonded interactions. **Physics algorithms.** The laws of physics and the algorithms handling them, ranging from Newtons laws of motion and the leap frog integrator to schemes to handle electrostatics and alchemical calculations. **Input parameters.** Simulations conditions, such as temperature, pressure and boundary conditions.

The laws of physics and the algorithms handling them are combined in simulation packages. These packages mostly appeared around the same time as the force fields and are often co-developed. The software contains implementations of the most used algorithms to run MD simulations together with programs to setup and analyze simulations. Well known examples are GROMACS, GROMOS, NAMD (Not (just) Another Molecular Dynamics program), AMBER and CHARMM. An overview of these packages is given in the same review[50]. These software packages, together with the force fields make it possible to perform molecular dynamics studies without any knowledge of programming algorithms and parametrization protocols.

The sampling during simulations is subject to the ergodicity principle: the properties of a system are the same when averaged over time as averaged over many systems. The set of all states of the system is called an "ensemble", the mathematical space in which it is described is called "phase-space". Depending on which constraints are applied to the system different ensembles can be obtained. Some examples are the microcanonical ensemble (with the particle number (N), system volume (V) and total energy (E) imposed), the canonical ensemble (with N, V and system temperature (T) imposed) and the isobaric-isothermal ensemble (with N, system pressure (p) and T imposed)[51].

The ergodicity principle also implies that, for systems in equilibrium, running one long simulation, will give the same result as running multiple short simulations. This way one can do multiple simulations at the same time at different computers. However, this assumes that the separate systems are in equilibrium, which for some processes might not be feasible in short simulations. For these cases most software packages mentioned above have implemented parrallelization schemes (e.g see [52]) that allow a single simulation to run on multiple processors, by distributing the calculations. In this way longer simulations of larger systems can be simulated in the same time by using very large computers.

Another possibility to improve sampling is to influence which parts of phase-space are sampled, by adding terms to the potential. Thus one can limit the time a system spends on sampling certain areas of phase-space. One of the simplest examples is a so called "pulling experiment": in order to determine the interaction strength between two molecules, the molecules are constrained (or restrained) at a defined distance. This is much more effective than letting the molecules sample space freely and only use the frames where they are at the distance of interest. (see chapter 2 and 3 for extensive description and comparison of both methods).

Another example to circumvent slow sampling is an alchemical simulation, where the potential is altered in a way that would be physically impossible, for example by turning off the interactions between two specific molecules. By doing this stepwise following an interaction parameter λ and integrating over the derivative of the energy to this interaction parameter, the energy cost of going from the starting state ($\lambda = 0$) to the final state ($\lambda = 1$) is obtained. The method, called Thermodynamic Integration (TI)[53], can be used to calculate energies differences between two states of a system, even if the difference is very large (which would require long sampling in a "free" simulation.) An excellent formal description of the method is given in [51]. Even more efficient alternatives to these simple examples of sampling schemes, pulling and thermodynamic integration, are constantly being developed. Examples are algorithms like reaction pathway sampling[54] or local elevation[55]. An extensive overview and categorization of many of these methods is given in [56].

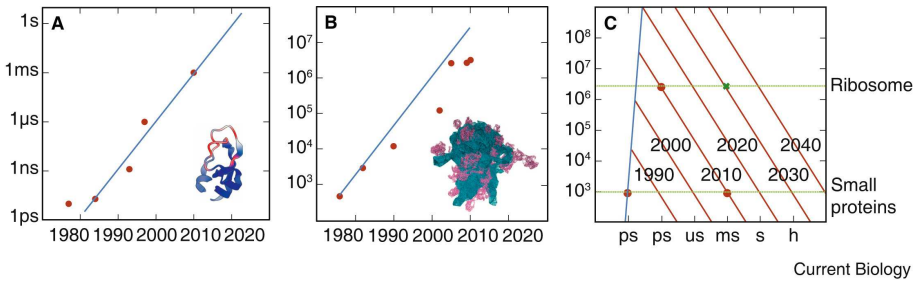

Figure 1.3: (A) Growth in the timescale accessible to molecular dynamics simulations of proteins. (B) Growth in the size (in number of atoms) of the protein systems studied by all-atom molecular dynamics simulations. The blue line is a theoretical limit derived by assuming that the growth in system size goes as the square root of the increase in timescale, as the number of interactions scales as the square of the number of atoms. (C) Change with time in the limits of the timescales and system sizes accessible by molecular dynamics simulations. The blue and red diagonal lines define the boundaries at given times. Figure from [4], reproduced with permission.

1.4 Moore's law.

As described at the beginning of this chapter MD computer simulations are strongly dependent on the availability of computational power. It is in this light interesting to compare the progress of the field to "Moore's law"[57]. It was postulated by Gordon E. Moore in 1965 and is often referred to as:

> *"The performance of computer chips at an affordable price will double every 24 months."*

If the length of molecular dynamics simulations doubles faster than every 24 months, this means new developments in the field add extra value on top of the ever increasing speed of computers. Vendruscolo and Dobson looked at exactly that in 2011[4], after the publication of the first millisecond all-atom simulations, by analyzing the length and system size of simulations published over the last three decades. Figure 1.3 shows the trends they found. The first panel shows the increase in length of simulations against time, whereas the second panel looks at the increase of system size. It can be seen that the length of computer simulations doubles approximately every year, while the system size (number of particles) that can be simulated doubles every two years. Combining the growth rate of simulation size and length (third panel), the authors predict that the first bacterial cell can be simulated on relevant timescales around 2050. This is significantly later than predicted by Van Gunsteren *et al.*, who in a similar manner predicted that an (*E. coli*) bacteria can be simulated in 2034[58]. In both cases the increase of simulation length has been faster than that predicted by Moore's law.

It is however important to note that with the growing of system size, longer time scales are necessary to observe phenomena of interest. Deserno[59] has elegantly quantified this point for a bilayer patch: if one has studied a lipid bilayer of size

20×20 nm and now wants to study a bilayer with ten times longer sides, it takes 100 times more computer power to obtain the same simulation time. However, undulation modes present in a membrane patch, decay with λ^4, where λ is the wave length. In a box of 10 times the size, ten times longer wave lengths are possible and thus one has to simulate 10,000 times longer to properly sample all the modes. The total computational time needed to obtain interesting results for a system of ten times the width and length is thus 1,000,000 times more. Similar effects will occur when simulated proteins grow in size, or multiple different components are combined in one system, allowing for more different mutual interactions.

The above comparison to Moore's law is a simple way to monitor the progress in MD simulation length. However, improvements have not only been made in the length and size of simulations. First, the quality[4] of the simulations has improved, partially due to the developments in water models and force fields described earlier and partially due to the development of new, more accurate algorithms, such as integration schemes, thermo- and barostats and new models for non-bonded interactions[5]. Second, large improvements have been made in the form of methods that allow for better sampling without increasing the required simulations time and thus computer power. In this category fall the sampling methods discussed earlier and a different kind of force fields, known as coarse grain (CG) force fields. Since the development of the different algorithms plays only a minor role in the rest of this thesis, the reader is referred to the references mentioned above for an overview. Coarse graining will be treated in the next section.

1.5 Coarse Graining.

Coarse graining concerns omitting irrelevant degrees of freedom of a system, in order to simplify the model. Every potential used for MD simulations is in some sense a coarse grain potential. For example, in all-atom force fields the dynamics of electrons and nuclei are not explicitly considered and united-atom force fields do not explicitly consider carbon-bound hydrogens. Here however, we will refer to CG force fields as those that do not explicitly consider every heavy atom as a separate interactions site. This is also reflected by the term *coarse grain*, as it implies the existence of a *fine grain* resolution. In other words, the term *coarse grain* can only be used for these force fields in comparison to atomistic models, which are still the gold standard in the field of molecular simulations. This does not mean that CG force fields are always derived from an atomistic force field.

As discussed earlier, CG force fields have been around since the early days of MD simulations: Levitt used a protein model with one or two interaction sites per amino acid residue[22] and Cotterill represented lipids by simple dumbbell shaped objects[26]. These examples show that CG force fields can have a wide range of different resolutions (the number of heavy atoms mapped to one CG site) and very different methods can be used to construct the potential energy function. An excellent compendium of many different methods, force fields and applications is given in [60].

Roughly three categories of methods to obtain a CG potential can be distinguished: those that are entirely based upon an atomistic force field, those that are entirely inde-

[4] Quality here relates to how faithfully a model reproduces nature.

[5] New algorithms can also greatly influence the speed of simulation, both in a positive and a negative manner.

pendent of any atomistic force field and combinations of these two. The first category includes methods like *inverse Boltzmann*[61] and *force matching*[62, 63], where both bonded and non-bonded terms are directly derived from atomistic simulations. The second category includes force fields that have been purely parametrized to match macroscopic properties of the system to experimental quantities, like bilayer properties or protein packing data derived from the Protein Data Bank. Examples of this type of force fields are described in [64, 65]. Methods in the third category often derive non-bonded interaction from experimental quantities, like partitioning free energies or liquid properties, while they (partially) rely on atomistic simulations for bonded interactions. Examples of this last category are the force field derived by DeVane and coworkers[66] and the Martini force field, which is more extensively described in the next section.

1.6 Martini force field

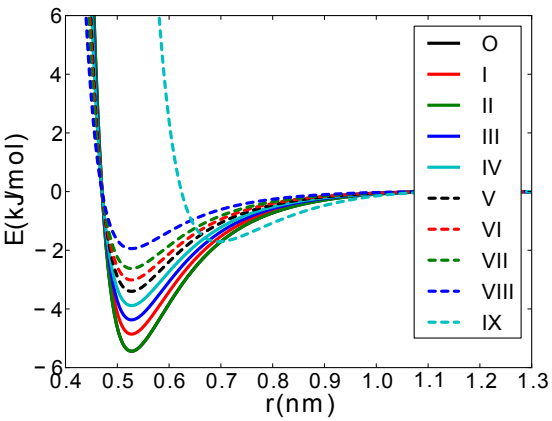

Figure 1.4: Plots of all 10 Martini interactions between. The different interaction levels are: O, ϵ=5.6 kJ mol^{-}1; I, ϵ=5.0 kJ mol^{-}1; II, ϵ=4.5 kJ mol^{-}1; III, ϵ=4.0 kJ mol^{-}1; IV, ϵ=3.5 kJ mol^{-}1; V, ϵ=3.1 kJ mol^{-}1; VI, ϵ=2.7 kJ mol^{-}1; VII, ϵ=2.3 kJ mol^{-}1; VIII, ϵ=2.0 kJ mol^{-}1; IX, ϵ=2.0 kJ mol^{-}1. σ = 0.47 nm for all levels except IV for which σ = 0.62nm.

The Martini CG force field applies a mapping of (on average) four heavy atoms to one CG interaction site. Four CG bead types have been defined: charged (Q), polar (P), non-polar (N) and apolar (C), which in turn are subdivided in four or five levels, giving a total of 18 bead types. For the interactions between these bead types, 10 different interactions levels are defined (O-IX)[67]. The interactions between these beads are Lennard-Jones[68] potentials at ten different levels, corresponding to different *'well depths'* as defined by the interaction parameter ϵ. Figure 1.4 shows the shifted LJ potentials for the 10 different levels. Which interaction is used between two beads has been parametrized by matching the free energies of vaporization, hydration and partitioning between water and apolar solvents for model compounds. These model compounds are chemical entities covering a wide range of molecules. They are each represented by one bead type and serve as building blocks for larger molecules.

This modular setup and relatively small number of pre-parametrized bead types make it easy to build new molecules compatible with the Martini force field. The initial publication contained parameters for water, organic solvents, surfactants, lipids

and cholesterol[69, 67]. In later publications parameters were added for more types of molecules, including proteins[70], carbohydrates[71], dendrimers[72], and fullerenes[73].

The loss of resolution in the Martini model compared to atomistic force fields brings along different limitations and challenges while parameterizing. Except for the obvious loss in structural detail, there are a few problems especially worth mentioning.

First, the grouping together of 4 heavy atoms reduces the entropy in a molecule. In order to obtain correct free energies, this is corrected for by adapting the enthalpic interactions[74]. As a result, the balance between entropy and enthalpy will be disturbed and separating these two contributions has to be done with the greatest care.

Second, a mapping of four heavy atoms to one CG interaction site, means four water molecules are grouped together in one bead. This works fine, unless one wants to study the behavior of single water molecules, for example water inside water channels in proteins or interfacial water at protein interfaces. Note that grouping four water molecules to one bead leads to a decrease in the number of interaction sites by a factor of 12 and is also responsible for the mass used for beads in Martini: 72 amu.

Third, coarse grain beads do not explicitly represent a possible polar nature of the underlying molecular building block. The polar nature is mimicked by a slightly stronger interaction between polar beads and polar and charged beads. In the case of a four water cluster also the polarizability is lost. For water this has been solved by introducing polarizable Martini water[75] and for polar amino acids a solution is introduced in chapter 3.

Fourth, for proteins the secondary and ternary structure are not stable due to the lack of detail in the (backbone) interactions and the resulting lack of polarity mentioned above. For the secondary structure this has been solved by constraining α-helices and β-strands by angles, dihedrals and local elastic networks[70]. The ternary structure can be kept stable by using an global elastic network, the so called Elnedyn approach, which connects backbone beads in different parts of the protein by long, elastic bonds[76].

Compared to united atom models, the coarse grain mapping decreases the number of interaction sites by four, and for water even by 12. The resulting, total speed-up of a simulation, as compared to a united-atom force field is between a factor 100 to 1000. There are three contributions to the speed-up. First, a reduced number of interactions sites means less calculations to be performed per time step. Notice that this goes roughly with the square of the reduction in the number of particles, since not only the number of particles for which the interactions have to be calculated is reduced, but also the number of interactions to other particles. For Martini, this would be on average $4^2 = 16$, although for water it is a lot more ($12^2 = 144$). Second, coarse graining leads to a smoother energy landscape due to the omission of fast modes in the molecules. The smoother energy landscape allows for larger time steps without making large integration errors. For Martini, typically a time step between 20 and 40 fs can be applied[77], 4 to 40 times longer than those used in atomistic simulations. Last, the smoother energy surface is also responsible for a faster progress of dynamic processes, for example diffusion. In others words, in the same simulation time a system will sample a larger part of phase space.

To test the actual speed-up a comparison between two 10 ns simulations of a 128 lipid DOPC bilayer with 37.75 waters per lipid at 303 K was made. One simulation was run using the CG Martini force field, the other using (united-atom) GROMOS parameters. The number of particles in both systems was 3000 and 21387, a factor

of 7 difference. For the CG system a time step of 40 fs was used, with a neighbor update every 10 steps and a neighbor list cutoff of 1.2 nm. Atomistically, a 2 fs time step was used, with a twin-range cutoff: the neighbor list extended up to 1.4 nm and was updated every 5 steps, forces between 0.9 and 1.4 nm were only recalculated every 5 steps. The simulations were run sequentially on the same machine[6] using Gromacs (version 4.5.5) with 8 threads for each simulation. The runs produced 4935 and 17.2 ns per day for the CG and atomistic simulations, respectively, equal to a speed-up of 287 times. The diffusion speeds (calculated over longer trajectories) were 0.053 cm^2s^{-1} and 0.005 cm^2s^{-1} for CG and atomistic, respectively, a speed-up of approximately 10 times. Thus the total speed-up is 2870 times.

Table 1.1: Diffusion rate of different molecules obtained using Martini (CG) and atomistic level (AL) simulations and experiments (Exp). All values are given in cm^2s^{-1}. The values for DOPC:DOPE mixtures and the transmembrane peptides concern lateral diffusion. Simulation values have been obtained using the Martini or Gromos 45A3 force fields at 300 K and parameters similar to those described before, unless otherwise indicated. **a.** One coarse grain bead accounts for 4 real water molecules, therefore the obtained value has been multiplied by 4 to obtain the water diffusion[78]. **b.** Value from [15]. **c.** Values from [67]. **d.** Simulations were ran in NVT ensemble to ensure the correct liquid density. **e.** Value extrapolated from temperature dependent data in reference [79]. **f.** Values from references [80, 81, 82]. **g.** Bilayer thickness of 2.38 and 2.2 nm for experimental and simulation, respectively. **h.** Bilayer thickness of 3.31 and 3.4 nm for experimental and simulation, respectively. **i.** Values from [83]. In the paper simulations values have been multiplied by the often applied factor 4, to correct for the speed-up of Martini simulations. Here the factor 4 is removed.

Molecule	CG	AL	Exp
Water	5.6[a]	3.6[b]	2.3[c]
Butane	6.1	6.3	>5[c]
Octane	2.3	2.9	2[c]
Dodecane	1.2	1.3	-
Hexadecane	0.7	0.7	-
Propanol	2.4[d]	1.7[d]	0.24[e]
Octanol	2.2	0.2	0.1[c]
1:0 DOPC:DOPE	0.053	0.005	0.006-0.017[f]
3:1 DOPC:DOPE	0.048	0.005	-
1:1 DOPC:DOPE	0.054	0.004	-
1:3 DOPC:DOPE	0.051	0.003	-
WALP23[g]	0.048[i]	-	0.0078[i]
WALP27[g]	0.035[i]	-	0.0065[i]
WALP23[h]	0.016[i]	-	0.0042[i]
WALP27[h]	0.013[i]	-	0.0036[i]

The computational speed-up will depend on the composition of the system, since the ratio of atoms per bead differs slightly per molecule. Also the speed-up of the diffusion depends on the type of molecule. Table 1.1 gives an overview of the diffusion speeds calculated for different molecules. The speed-up will not only depend on the force field, also the choice of integrator and thermostat will be important. The values reported here are all simulated with a leap frog integrator and a Berendsen[84] or 'Velocity Rescale'[85] thermostat. The factor four often used to correct diffusion rates obtained with the Martini model, has not been used here: the actually measured diffusion rates are reported (values taken from literature that did use the conversion factor, were converted back to the measured diffusion rates).

The speed-up ranges between 1.2 and 22 comparing Martini to experiments and 0.79 and 17 comparing Martini to atomistic simulations. The speed-up is dependent on the type of molecule, e.g.: Martini alkanes have a small speed-up compared to experiments or even a slow-down compared to atomistic simulations, whereas alcohols show large speed-up factors). However, what defines the speed-up of a CG model is not understood quantitatively, despite recent studies[86]. The values in ta-

[6] Intel i7 920 with four multithreaded CPU cores running at 2.67 GHz. The machine was not otherwise used during the runs.

ble 1.1 may be used to derive some qualitative rules of thumb. CG polar liquids, like propanol and octanol, show a large speed-up factor (≈10). This might be caused by the lack of explicit hydrogen bonds in the CG model. Water, although highly polar, only shows a relatively lower speed-up (≈2.5), possibly due to the bundling together of four molecules. CG apolar liquids show no speed-up. Lipids show a large speed-up, suggesting that their diffusion is mainly controlled by the polar head groups. For small transmembrane peptides the speed-up is essentially that of the (lipid) solvent. If water soluble proteins also follow the speed-up of water remains to be seen.

In summary, if one wants to study diffusion related processes in the Martini model, this can only be done in a qualitative way or quantitatively by comparing to a reference (e.g. the diffusion of a solute compared to the solvent) or by calibrating the timescales by comparison to experiment or atomistic simulations.

1.7 Thesis Overview

The rest of this thesis is organized in three main topics. In the first two chapters we test the performance of the Martini forcefield by investigating the dimerization free energies of amino acid side chains. Chapter 2 describes a method to calculate binding free energies of molecules in molecular dynamics systems, which is subsequently applied in chapter 3 to compare these binding free energies of amino acid sidechains in the Gromos, OPLS and Martini force fields. The next two chapters are applications of the Martini force field. Chapter 4 studies the segregation of simple peptides between different coexisting membrane domains. In chapter 5 this segregation is further explored by adding different membrane constituents, such as anchored peptides, larger proteins and GM1 lipids, to the system. The final three chapters concern new force field developments. The chapter 6 describes the parameterization of thylakoid cofactors in the Gromos and Martini 2.1 force fields. Chapter 7 describes a new set of parameters for Martini amino acids, dubbed Martini 2.2. The final chapter (chapter 8) explores different potential shapes that might be used for future releases of the Martini force field (Martini 3.x).

1.8 Bibliography

[1] M Levitt. The birth of computational structural biology. *Nat. Struc. Mol. Biol.*, 8:392–393, 2001.

[2] M Karplus. Molecular dynamics of biological macromolecules: A brief history and perspective. *Biopolymers*, 68:350–358, 2003.

[3] T Schlick, R Collepardo-Guevara, L.A Halvorsen, S Jung, and X Xiao. Biomolecular modeling and simulation: a field coming of age. *Quart. Rev. Biophys.*, 44:191–228, 2011.

[4] M Vendruscolo and C.M Dobson. Protein dynamics: Moore's law in molecular biology. *Curr. Biol.*, 21:R68–R70, 2011.

[5] E Fermi, J Pasta, and S Ulam. *Studies of nonlinear problems.* Los Alamos Report LA-1940, 1955.

[6] B.J Alder and T.E WainWright. Phase transition for a hard sphere system. *J. Chem. Phys.*, 27:1208–1211, 1957.

[7] B.J Alder and T.E WainWright. Studies in molecular dynamics. i. general method. *J. Chem. Phys.*, 31:1–9, 1959.

[8] N Metroplis, A.W Rosenbluth, M.N Rosenbluth, and A.H Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1986–1092, 1953.

[9] B.J Alder, S.P Frankel, and V.A Lewinson. Radial distribution function calculated by the montecarlo method for a hard sphere fluid. *J. Chem. Phys.*, 23:417, 1955.

[10] A Rahman. Correlations in the motion of atoms in liquid argon. *Phys. Rev.*, 136:405–411, 1964.

[11] A Rahman and F.H Stillinger. Molecular dynamics study of liquid water. *J. Chem. Phys.*, 55:3336, 1971.

[12] J.A Barker and R.O Watts. Structure of water; a monte carlo calculation. *Chem. Phys. Lett.*, 3:144–145, 1969.

[13] F.H Stillinger and A Rahman. Improved simulation of liquid water by molecular dynamics. *J. Chem. Phys.*, 60:1545, 1974.

[14] A Ben-Naim and F.H Stillinger. *Structure and Transport of Processes in Water and Aqueous Solutions*, chapter Aspects of the Statistical-Mechanical Theory of Water, pages 295–330. Wiley, 1972.

[15] H.J.C Berendsen, J.P.M Postma, W.F Van Gunsteren, and J Hermans. Interaction models for water in relation to protein hydration. *Intermolecular forces*, 331, 1981.

[16] W.L Jorgensen. Quantum and statistical mechanical studies of liquids. 10. transferable intermolecular potential functions for water, alcohols, and ethers. application to liquid water. *J. Am. Chem. Soc.*, 103:335–340, 1981.

[17] W.L Jorgensen, J Chandrasekhar, J.D Madura, R.W Impey, and M.L Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926, 1983.

[18] H.J.C Berendsen, J.R Grigera, and T.P Straatsma. The missing term in effective pair potentials. *J. Phys. Chem.*, 91:6269–6271, 1987.

[19] K Toukan and A Rahman. Molecular-dynamics study of atomic motions in water. *Phys. Rev. B*, 31:2643–2648, 1985.

[20] J Hermans and H.J.C Berendsen. A consistent empirical potential for water–protein interactions. *Biopolymers*, 23:1513–1518, 1984.

[21] J.A McCammon, B.R Gelin, and M Karplus. Dynamics of folded proteins. *Nature*, 267:585–590, 1977.

[22] M Levitt and A Warshel. Computer simulation of protein folding. *Nature*, 253:694–698, 1975.

[23] M Levitt and R Sharon. Accurate simulation of protein dynamics in solution. *Proc. Nati. Acad. Sci. USA*, 85:7557, 1988.

[24] W.F Van Gunsteren and H.J.C Berendsen. Computer simulation as a tool for tracing the conformational differences between proteins in solution and in the crystalline state. *J. Mol. Biol*, 176:559–564, 1984.

[25] C.F Wong and J.A McCammon. Unknown. *Isr. J.Chem*, 27:211–215, 1987.

[26] R.M.J Cotterill. Computer simulation of model lipid membrane dynamics. *BBA - Biomembranes*, 433:264–270, 1976.

[27] A.J Kox, J Michels, and F Wiegel. Simulation of a lipid monolayer using molecular dynamics. *Nature*, 287:317–319, 1980.

[28] P Van der Ploeg and H.J.C Berendsen. Molecular dynamics of a bilayer membrane. *J. Chem. Phys.*, 49:233–248, 1982.

[29] B Jonsson, O Edholm, and O Teleman. Molecular dynamics simulations of a sodium octanoate micelle in aqueous solution. *J. Chem. Phys.*, 85:1–14, 1986.

[30] E Egberts and H.J.C Berendsen. Molecular dynamics simulation of a smectic liquid crystal with atomic detail. *J. Chem. Phys.*, 89:3718, 1988.

[31] E Egberts, S.J Marrink, and H.J.C Berendsen. Molecular dynamics simulation of a phospholipid membrane. *Eur. Biophys. J.*, 22:423–436, 1994.

[32] S.J Marrink, M. Berkowitz, and H.J.C Berendsen. Molecular dynamics simulation of a membrane/water interface: the ordering of water and its relation to the hydration force. *Langmuir*, (9):3122–3131, 1993.

[33] S.J Marrink, E Lindahl, O Edholm, and A.E Mark. Simulation of the spontaneous aggregation of phospholipids into bilayers. *Journal of the ...*, 123:8638–8639, 2001.

[34] O Edholm, O Berger, and F Jähnig. Structure and fluctuations of bacteriorhodopsin in the purple membrane: a molecular dynamics study. *J. Mol. Biol.*, 250:94–111, 1995.

[35] D.P Tieleman, S.J Marrink, and H.J.C Berendsen. A computer perspective of membranes: molecular dynamics studies of lipid bilayer systems. *BBA-Rev. Biomembranes*, 1331:235–270, 1997.

[36] P.J Stansfeld and M.S.P Sansom. Molecular simulation approaches to membrane proteins. *Struc. Fold. Des.*, 19:1562–1572, 2011.

[37] X Daura, B Jaun, D Seebach, W.F van Gunsteren, and A.E Mark. Reversible peptide folding in solution by molecular dynamics simulation1. *J. Mol. Biol.*, 280:925–932, 1998.

[38] X Daura, K Gademann, B Jaun, D Seebach, W.F van Gunsteren, and A.E Mark. Peptide folding: when simulation meets experiment. *Angew. Chem. Int. Edit.*, 38:236–240, 1999.

[39] Y Duan. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282:740–744, 1998.

[40] C.D Snow, H Nguyen, V.S Pande, and M Gruebele. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature*, 420:102–106, 2002.

[41] M Shirts and V.S Pande. Screen savers of the world unite! *Science*, 290:1903–1904, 2000.

[42] D.E Shaw, P Maragakis, K Lindorff-Larsen, S Piana, R.O Dror, M.P Eastwood, J.A Bank, J.M Jumper, J.K Salmon, and Y Shan. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330:341, 2010.

[43] D.E Shaw, R.O Dror, J.K Salmon, JP Grossman, K.M Mackenzie, J.A Bank, C Young, M.M Deneroff, B Batson, and K.J Bowers. Millisecond-scale molecular dynamics simulations on anton. *P. C. High Perf. Comp. Netw. Stor. Anal.*, page 39, 2009.

[44] T Narumi, R Susukita, T Ebisuzaki, G Mcniven, and B Elmegreen. Molecular dynamics machine: Special-purpose computer for molecular dynamics simulations. *Mol. Simulat.*, 21:401–415, 1999.

[45] H.J.C Berendsen, D van der Spoel, and R van Drunen. Gromacs: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.*, 91:43–56, 1995.

[46] G.R Bowman, V.A Voelz, and V.S Pande. Atomistic folding simulations of the five-helix bundle protein σ685. *J. Am. Chem. Soc.*, 133:664–667, 2011.

[47] K Lindorff-Larsen, S Piana, R.O Dror, and D.E Shaw. How fast-folding proteins fold. *Science's STKE*, 334:517, 2011.

[48] T.E Cheatham III. Simulation and modeling of nucleic acid structure, dynamics and interactions. *Curr. Opin. Struc. Biol.*, 14:360–367, 2004.

[49] E Fadda and R.J Woods. Molecular simulations of carbohydrates and proteinâ"carbohydrate interactions: motivation, issues and prospects. *Drug Discov. Today*, 15:596–609, 2010.

[50] O Guvench and A.D MacKerell. Comparison of protein force fields for molecular dynamics simulations. *Method. Mol. Biol.*, 443:63–88, 2008.

[51] H Berendsen. *Simulating the physical world: hierarchical modeling from quantum mechanics to fluid dynamics*. Cambridge University Press, 2007.

[52] B Hess, C Kutzner, D van der Spoel, and E Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput*, 4:435–447, 2008.

[53] R.W Zwanzig. Hightemperature equation of state by a perturbation method. i. nonpolar gases. *J. Chem. Phys.*, 22:1420–1428, 1954.

[54] P.G Bolhuis, D Chandler, C Dellago, and P.L Geissler. T ransitionp aths ampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.*, 53:291–318, 2002.

[55] T Huber, A.E Torda, and W.F Gunsteren. Local elevation: a method for improving the searching properties of molecular dynamics simulation. *J Comput Aid Mol Des*, 8:695–708, 1994.

[56] H.S Hansen and P.H Hünenberger. Using the local elevation method to construct optimized umbrella sampling potentials: Calculation of the relative free energies and interconversion barriers of glucopyranose ring conformers in water. *J. Comput. Chem.*, 31:1–23, 2009.

[57] E.M Moore. Cramming more components onto integrated circuits. *Electronics*, 38:1–4, 1965.

[58] W.F Van Gunsteren and D Bakowies. Biomolecular modeling: goals, problems, perspectives. *Angew. Chem. Int. Edit*, 2006.

[59] M Deserno. Mesoscopic membrane physics: Concepts, simulations, and selected applications. *Macromol. Rapid Comm.*, 30:752–771, 2009.

[60] G.A Voth. *Coarse graining of condensed phase and biomolecular systems.* 2009.

[61] A Soper. Empirical potential monte carlo simulation of fluid structure. *Chem. phys.*, 1996.

[62] S Izvekov, M Parrinello, C.J Bernham, and G.A Voth. Effective force fields for condensed phase systems from ab initio molecular dynamics simulation: A new method for force-matching. *J. Chem. Phys.*, 120:1–19, 2004.

[63] W.G Noid, J Chu, G.S Ayton, V Krishna, S Izvekov, G.A Voth, A Das, and H.C Andersen. The multiscale coarse-graining method. i. a rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.*, 128:244114, 2008.

[64] I.R Cooke, K Kremer, and M Deserno. Tunable generic model for fluid bilayer membranes. *Phys. Rev. E*, 72, 2005.

[65] S Miyazawa and R.L Jernigan. *J. Mol. Biol.*, 256:623–644, 1996.

[66] W. Shinoda, R DeVane, and M.L Klein. Multi-property fitting and parameterization of a coarse grained model for aqueous surfactants. *Mol. Simulat.*, 33:27–36, 2007.

[67] S.J Marrink, H.J Risselada, S Yefimov, D.P Tieleman, and A.H de Vries. The martini force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 111:7812–7824, 2007.

[68] J.E Jones. On the determination of molecular fields. ii. from the equation of state of a gas. *P Roy Soc A*, 106:463–477, 1924.

[69] S.J Marrink, A.H de Vries, and A.E Mark. Coarse grained model for semiquantitative lipid simulations. *J. Phys. Chem. B*, 108:750–760, 2004.

[70] L Monticelli, S.K Kandasamy, X Periole, R.G Larson, D.P Tieleman, and S.J Marrink. The martini coarse-grained force field: extension to proteins. *J, Chem. Phys.*, 4:819–834, 2008.

[71] C.A López, A.J Rzepiela, A.H de Vries, L Dijkhuizen, P.H Hünenberger, and S.J Marrink. Martini coarse-grained force field: Extension to carbohydrates. *J. Chem. Phys.*, 5:3195–3210, 2009.

[72] H Lee and R.G Larson. Coarse-grained molecular dynamics studies of the concentration and size dependence of fifth-and seventh-generation pamam dendrimers on pore formation in dmpc bilayer. *J. Phys. Chem. B*, 112:7778–7784, 2008.

[73] J Wong-Ekkabut, S Baoukina, W Triampo, I.M Tang, D.P Tieleman, and L Monticelli. Computer simulation study of fullerene translocation through lipid membranes. *Nat. Nanotechnol.*, 3:363–368, 2008.

[74] R Baron, A.H de Vries, P.H Hünenberger, and W.F Van Gunsteren. Configurational entropies of lipids in pure and mixed bilayers from atomic-level and coarse-grained molecular dynamics simulations. *J. Phys. Chem. B*, 2006.

[75] S.O Yesylevskyy, L.V Schäfer, D Sengupta, and S.J Marrink. Polarizable water model for the coarse-grained martini force field. *PLoS Comput. Biol.*, 6:e1000810, 2010.

[76] X Periole, M Cavalli, S.J Marrink, and M.A Ceruso. Combining an elastic network with a coarse-grained molecular force field: Structure, dynamics, and intermolecular recognition. *J. Chem. Phys.*, 5:2531–2543, 2009.

[77] S.J Marrink, X Periole, D.P Tieleman, and A.H de Vries. Comment on "on using a too large integration time step in molecular dynamics simulations of coarse-grained molecular models" by m. winger, d. trzesniak, r. baron and w. f. van gunsteren, phys. chem. chem. phys., 2009, 11, 1934. *Phys. Chem. Chem. Phys.*, 12:2254, 2010.

[78] M Fuhrmans, B.P Sanders, S.J Marrink, and A.H de Vries. Effects of bundling on the properties of the spc water model. *Theor. Chem. Acc.*, 125:335–344, 2010.

[79] N ShakerGaafar, N Karger, S Wappmann, and H-D Lüdemann. p, tdependence of selfdiffusion in liquid ethanol and the propanols. *Berich. Bunsen Gesell.*, 97:805–811, 1993.

[80] N Kahya, D Scherfeld, K Bacia, B Poolman, and P Schwille. Probing lipid mobility of raft-exhibiting model membranes by fluorescence correlation spectroscopy. *J. Biol. Chem.*, 278:28109–28115, 2003.

[81] A Filippov, G Orädd, and G Lindblom. The effect of cholesterol on the lateral diffusion of phospholipids in oriented bilayers. *Biophys. J.*, 84:3079–3086, 2003.

[82] J Kusba, L Li, I Gryczynski, G Piszczek, M Johnson, and J.R Lakowicz. Lateral diffusion coefficients in membranes measured by resonance energy transfer and a new algorithm for diffusion in two dimensions. *Biophys. J.*, 82:1358–1372, 2002.

[83] S Ramadurai, A Holt, L.V Schäfer, V.V Krasnikov, D.T.S Rijkers, S.J Marrink, J.A Killian, and B Poolman. Influence of hydrophobic mismatch and amino acid composition on the lateral diffusion of transmembrane peptides. *Biophys. J.*, 2010.

[84] H.J.C Berendsen, J.P.M Postma, W.F Van Gunsteren, A DiNola, and J.R Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684, 1984.

[85] G Bussi, D Donadio, and M Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126:014101, 2007.

[86] P Depa, C Chen, and J.K Maranas. Why are coarse-grained force fields too fast? a look at dynamics of four coarse-grained polymers. *J. Chem. Phys.*, 134:014903, 2011.

2

Determining Equilibrium Constants for Dimerization Reactions from Molecular Dynamics Simulations

This chapter is based upon the manuscript:
Determining Equilibrium Constants for Dimerization Reactions from Molecular Dynamics Simulations by Djurre H. de Jong, Lars V. Schäfer, Alex H. de Vries, Siewert J. Marrink, Herman J. C. Berendsen, Helmut Grubmüller, J. Comput. Chem., **2011**, *32*(9), 1919-1928

Abstract

With today's available computer power, free energy calculations from equilibrium molecular dynamics simulations 'via counting' become feasible for an increasing number of reactions. An example is the dimerization reaction of transmembrane alpha-helices. If an extended simulation of the two helices covers sufficiently many dimerization and dissociation events, their binding free energy is readily derived from the fraction of time during which the two helices are observed in dimeric form. Exactly how the correct value for the free energy is to be calculated, however, is unclear, and indeed several different and contradictory approaches have been used. In particular, results obtained via Boltzmann statistics differ from those determined via the law of mass action. Here we develop a theory that resolves this discrepancy. We show that for simulation-systems containing two molecules, the dimerization free energy is given by a formula of the form $\Delta G \propto \ln(P_1/P_0)$. Our theory is also applicable to high concentrations that typically have to be used in molecular dynamics simulations to keep the simulation system small, where the textbook dilute approximations fail. It also covers simulations with an arbitrary number of monomers and dimers and provides rigorous error estimates. Comparison with test simulations of a simple Lennard Jones system with various particle numbers as well as with reference free energy values obtained from radial distribution functions show full agreement for both, binding free energies as well as dimerization statistics.

2.1 Introduction

The computation of free energy differences is the aim of many molecular simulation studies. As a thermodynamic state function, the free energy provides insights into the molecular driving forces for the studied process, and often enables a direct and quantitative comparison to experiments. However, in many cases, it is not trivial to obtain

free energy differences from simulations of large condensed-phase systems, because it requires proper and extensive sampling of the underlying thermodynamic ensemble, for example through Monte Carlo (MC) or molecular dynamics (MD) techniques.

A number of MD-based simulation protocols for calculating free energy differences have been devised. Thermodynamic integration and free energy perturbation approaches, based on an alchemical transformation of one group of atoms into another, are frequently used. [1, 2, 3, 4] Also non-equilibrium methods have been successfully applied to calculate free energies of molecular systems.[5, 6, 7, 8, 9, 10] To study the energetics of self-assembly processes, such as the binding of two (or more) molecules, the umbrella sampling technique[11] is often applied, in which harmonic (umbrella) potentials drive the system along a pre-defined reaction coordinate, for example the distance between the molecules.

Today, the increasing available computer power and ongoing development of efficient algorithms and coarse-grain force fields enables longer simulations of large systems, thus opening the way to a straightforward alternative: To carry out an extended equilibrium MD simulation and obtain the free energy difference from directly counting the fractions of simulation time spent in the respective states. Such an approach has been applied to, for example, the dimerization of chirally related organic molecules,[12] folding / unfolding of small peptides,[13, 14] dimerization of methane molecules[15] as well as of charged[16] and hydrophobic[17] amino acid pairs in water, and a dimer of transmembrane helices in a lipid bilayer[18].

However, contradicting approaches and formulae have been employed to calculate free energy differences. In particular, as further explained in the Theory section, directly using the ratio of the observed Boltzmann probabilities [13, 14, 16, 17, 19] yields different results compared to approaches adopting the law of mass action to simulations of two dimerizing molecules.[12, 15, 18] For example, if the system is found in a dimerized state during a fraction P_1 of the total simulation time, and in a monomeric state during a fraction $P_0 = 1 - P_1$, the free energy difference would in the first case be given by an equation of the form $\Delta G \propto \ln(P_1/P_0)$, whereas in the latter case, it is $\propto \ln(P_1/P_0^2)$. Which of the two approaches is correct? Furthermore, it is not trivial to provide a generalized formalism as well as reliable error estimates for simulations with more than two molecules, which may provide better sampling.

Here, we develop a rigorous theory for dimerization reactions involving an arbitrary number of molecules, including only two, and derive how dimerization free energies can be calculated from simulations by direct counting. First, we will use thermodynamic arguments to show that an equation of the form $\Delta G \propto \ln(P_1/P_0)$ is the correct formula for simulations of two dimerizing molecules. Second, we present a general statistical mechanical treatment of dimer association / dissociation reactions of any number of molecules, and demonstrate how the law of mass action is recovered. Third, we discuss how the counter-intuitive disagreement between the Boltzmann treatment and the naive application of the law of mass action is resolved by careful consideration of the respective ensembles. We finally test our theoretical results against MD simulations, and compare free energies obtained from direct counting with those from radial distribution functions.

2.2 Theory

As shown in Figure 2.1, we consider the dimerization of two molecules in solution, A and B,

$$A + B \rightleftharpoons AB \quad , \tag{2.1}$$

for which the law of mass action reads

$$K_a = \frac{[AB]c^{\varnothing}}{[A][B]} \quad , \tag{2.2}$$

with association constant K_a, concentrations $[X]$, and $c^{\varnothing}$ an agreed standard concentration, usually 1 mol l^{-1}. Eq. (2.2) assumes that the system is sufficiently diluted, such that concentrations can be used instead of activities.

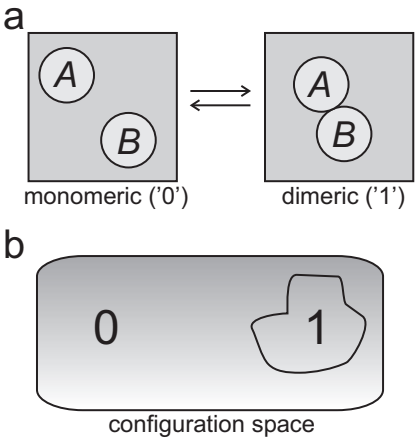

Figure 2.1: Dimerization of two molecules A and B within a given volume (a). The configuration space (b) is divided into two parts, monomeric (no dimer, '0') and dimeric (one dimer, '1').

Two particles

For a mixture of n_i mol of species i, the Gibbs free energy is related to the thermodynamic (chemical) potentials through

$$G = \sum_i n_i \mu_i \quad . \tag{2.3}$$

Thus, for two molecules,

$$G_1 = \frac{1}{N_{Av}}\mu_{AB} + \frac{N_s}{N_{Av}}\mu_s \tag{2.4}$$

and

$$G_0 = \frac{1}{N_{Av}}\mu_A + \frac{1}{N_{Av}}\mu_B + \frac{N_s}{N_{Av}}\mu_s \tag{2.5}$$

for the dimer and monomer states, 1 and 0, respectively. Here, N_s is the number of solvent molecules, μ_s the thermodynamic potential of the solvent, and N_{Av} is Avogadro's number. Thus,

$$N_{Av}G_1 = \mu^{\varnothing}_{AB} + RT \ln \frac{1}{c^{\varnothing} N_{Av} v} + N_s \cdot \left[\mu^{\varnothing}_s + RT \ln x_{s,1}\right] \tag{2.6}$$

and

$$N_{Av}G_0 = \mu^{\varnothing}_A + \mu^{\varnothing}_B + 2RT \ln \frac{1}{c^{\varnothing} N_{Av} v} + N_s \cdot \left[\mu^{\varnothing}_s + RT \ln x_{s,0}\right] \quad , \tag{2.7}$$

where $R = k_B N_{Av}$ is the gas constant, v is the total volume of the system, and $x_{s,1}$ and $x_{s,0}$ are the mole fractions of the solvent (assuming ideal solution) for the dimer and monomer, respectively. Taking the difference of Eqs. (2.6) and (2.7) and neglecting the small difference between $x_{s,1}$ and $x_{s,0}$ yields

$$N_{Av}(G_1 - G_0) = \mu^{\varnothing}_{AB} - \mu^{\varnothing}_A - \mu^{\varnothing}_B + RT \ln \left(c^{\varnothing} N_{Av} v\right) \quad . \tag{2.8}$$

We now assume an MD simulation of two molecules A and B, surrounded by solvent molecules in a simulation box. The monomers can form a dimer according to Eq. (2.1), defined by an unambiguous definition (for example a distance criterion). The system is observed to be in its dimeric state during a fraction P_1 ('one dimer') of the total simulation time and in its monomeric state ('zero dimers') during a fraction $P_0 = 1 - P_1$. We further assume that the simulation time is long enough for sufficiently many transitions to occur between the two states, such that the system can be considered to be in thermodynamic equilibrium. Our aim is to calculate the equilibrium constant K_a for the association reaction Eq. (2.1), or, equivalently, the (standard) association free energy

$$\Delta G^{\varnothing} = -k_B T \ln K_a \tag{2.9}$$

from the simulation, i.e., from P_0 and P_1.

To derive an expression for the free energy difference, the configuration space is divided into two parts 0 and 1, representing the monomeric and dimeric states, respectively (Figure 2.1b). For simulations at constant v, T, the probability to be in a defined state is proportional to $\exp(-A/k_B T)$, where A is the Helmholtz free energy of that state. Thus,

$$\frac{P_1}{P_0} = \exp\left[\frac{-(A_1 - A_0)}{k_B T}\right] \quad . \tag{2.10}$$

To obtain ΔG, the difference between the Gibbs functions,

$$G_1 - G_0 = -k_B T \ln \frac{P_1}{P_0} + (p_1 - p_0)v \tag{2.11}$$

is desired. The pv term is small for most reactions in solution and is therefore disregarded here; however, it can be determined from the pressures during the simulation if needed. For simpler notation, all partition functions further below refer to simulation ensembles at constant v, T.

Combining Eq. (2.8) and Eq. (2.11) yields the expected Boltzmann relation between the standard free energy change and the probability ratio observed in the simulation,

$$\Delta G^{\varnothing} = -RT \ln \frac{P_1}{P_0} - RT \ln \left(c^{\varnothing} N_{Av} v\right) = \Delta G - RT \ln \left(c^{\varnothing} N_{Av} v\right) \quad . \tag{2.12}$$

Alternatively, the equilibrium constant can be expressed using Eq. (2.9),

$$K_a = \frac{P_1}{P_0} c^{\varnothing} N_{Av} v \quad . \tag{2.13}$$

Subsequently, we will omit explicit reference to the standard state; it can easily be reintroduced via the above Eqs. (2.12) and (2.13).

Generalization for many particles

We now assume an equilibrium between N molecules, N_A of which are of type A, and $N_B = N - N_A$ of which are of type B. We further assume that particles of the

same type do not dimerize, and that polymerization does not occur. For example, for $N = 4$ and $N_A = N_B = 2$, the relevant states are

$$A + A + B + B \rightleftharpoons AB + A + B \rightleftharpoons AB + AB\,. \tag{2.14}$$

Generalizing Eqs. (2.12) and (2.13), we aim at relating the statistics of observed dimers AB in the MD simulation to the (macroscopic) association constant K_a and free energy ΔG for the dimerization reaction Eq. (2.1). As in Figure 2.1, the configuration space of the N molecules within the (3-dimensional) volume v is divided into ($3N$-dimensional) subvolumes V_0 (only monomers), V_1 (1 dimer, $N-2$ monomers), ..., V_m (m dimers, $N-2m$ monomers).

From the respective Helmholtz free energies, Eq. (2.10), the probability of finding the system in a fully monomeric state reads

$$P_0 = \frac{Z_0}{Z} = \frac{\int_{V_0} e^{-U/k_BT} dV}{\int_V e^{-U/k_BT} dV} \quad , \tag{2.15}$$

and that of finding exactly $m \leq \min(N_A, N_B)$ dimers is

$$P_m = \frac{Z_m}{Z} = \frac{\int_{V_m} e^{-U/k_BT} dV}{\int_V e^{-U/k_BT} dV} \quad , \tag{2.16}$$

with configurational partition functions Z_m, integrated over those regions of configuration space with m dimers and $N-2m$ monomers, interaction potential U, and partition function $Z = \sum_{m=0}^{\min(N_A,N_B)} Z_m$. We note that the kinetic part of the partition functions can always be factored out and therefore cancels in the above equations. From these probabilities, the average number $\langle m \rangle$ of dimers is readily obtained,

$$\langle m \rangle = \frac{1}{Z} \sum_{m=1}^{\min(N_A,N_B)} m Z_m \quad . \tag{2.17}$$

Neglecting the interaction energy of distant (unbound) particles, the above partition functions can be expressed in terms of an excess free energy per dimer (with respect to the ideal gas term),

$$G^* = k_BT \ln \langle e^{U/k_BT} \rangle_{V_1} \quad , \tag{2.18}$$

which is independent of the box volume. With this abbreviation, one obtains

$$Z_0 = \int_{V_0} e^{-U/k_BT} dV = V_0 \tag{2.19}$$

and

$$Z_1 = \int_{V_1} e^{-U/k_BT} dV = V_1 e^{-G^*/k_BT} \quad . \tag{2.20}$$

Assuming additive interaction potentials, the partition function for m dimers then reads

$$Z_m = \int_{V_m} e^{-U/k_BT} dV = V_m e^{-mG^*/k_BT} \quad , \tag{2.21}$$

which reduces the problem to estimating the configuration space volumes V_m.

To determine the configuration space volume V_0 for which the system consists of only monomers, note that the first molecule can be placed anywhere within the simulation box volume v. For the second molecule, only those positions are allowed for which it does not form a dimer with the first molecule, which yields the reduced volume $v - v_\mathrm{D}$. Here, v_D denotes the dimerization volume. Further, molecules A and B are assumed to occupy the same volume, i.e., the dimerization volume v_D equals the volume excluded by the repulsive interaction between AA or BB, respectively. For the third molecule, similarly, only a volume $v - 2v_\mathrm{D}$ remains, etc.

With $x = v_\mathrm{D}/v$, one thus obtains to second order in the particle concentrations $[A] + [B] = N/(N_A v)$,

$$V_0 = v(v - v_\mathrm{D})(v - 2v_\mathrm{D}) \cdot \ldots \cdot (v - (N-1)v_\mathrm{D}) = v^N \prod_{j=1}^{N-1} (1 - jx) \quad . \tag{2.22}$$

For the configuration space volume V_m of all states of m dimers and $N - 2m$ monomers, from similar but somewhat more involved reasoning

$$V_m = v^N m! \binom{N_A}{m} \binom{N_B}{m} x^m \prod_{j=1}^{m-1} (1 - jy) \prod_{j=0}^{N-2m-1} (1 - my - jx) \tag{2.23}$$

follows, where v_{AB} is the (average) volume excluded by each dimer AB (Figure 2.2b), and $y = v_{AB}/v$. To see why Eq. (2.23) holds true, first place m molecules to form a monomeric state, which yields the first product term as in Eq. (2.22). Next, place further m molecules to form m dimers, such that each of these molecules is restricted to a volume v_D, thus yielding the x^m-term. Finally, place $N - 2m$ further monomers within the remaining volume fraction $v^N(1 - my)$, with each monomer further reducing the available volume fraction by x. To verify the combinatorics note that this procedure yields only one out of all possible ways to select m molecules from the N_A molecules of type A and from the N_B molecules B, hence the two binomials. Finally, having selected m molecules of each type for dimerization, there are $m!$ ways of joining those into m dimers.

Note that with the convention that products for which the final value of the running index is smaller that the starting value equal unity (i.e., $\prod_{j=1}^{0} = 1$), Eq. (2.23) reduces to Eq. (2.22) for $m = 0$.

Eqs. (2.22) and (2.23) are independent of the shapes of the volumes. However, v_D and v_{AB} (and thus x and y) may in general be hard to determine. For spherical particles, a useful estimate is obtained by assuming a constant (average) interaction energy between the particles (Figure 2.2). In this case, the distance distribution between overlapping spheres is $p(r) \propto r^2$, yielding an average overlap of $v_\mathrm{D}/8$ and, hence, $y = x \cdot 15/8$. For the general case, it is important to realize that by choosing a criterion to define the dimer state, for example a distance cut-off, one implicitly determines the dimer volume, v_D. It is thus also possible to determine V_m numerically, without prior knowledge of v_D or v_{AB} by placing N non-interacting particles in a volume through a Monte Carlo run, as is demonstrated in section 2.3.

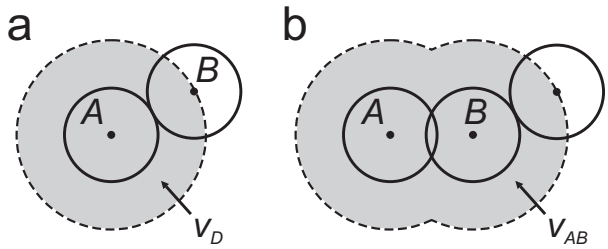

Figure 2.2: Definition of (a) dimerization volume v_{D} and (b) average dimer exclusion volume $v_{\mathrm{A}B}$ for the special case of spherical particles. If the center of particle B falls within the dimerization volume (grey) centered at particle A, the two particles are considered a dimer. The average volume of two combined overlapping dimerization volumes defines the dimer exclusion volume, within which a third particle is not considered monomeric.

Limiting cases

It is noteworthy to consider the case of moderately large N and low concentrations (i.e., $Nv_{\mathrm{D}} \ll v$ and, therefore, $x < y \ll 1$). In this case, using $\binom{N}{m} \approx N^m/m!$ and expanding the logarithm of the result to first order, Eq. (23) simplifies to, Eq. (2.23) simplifies to first order to

$$V_m \approx v^N \frac{(N_A N_B v_D/v)^m}{m!} e^{-\frac{1}{2}N^2 v_D/v} \quad . \tag{2.24}$$

Combining Eq. (2.24) with Eq. (2.16) and (2.21), after proper normalization , a Poisson distribution follows for the probabilities of finding m dimers:

$$P_m = \frac{\lambda^m e^{-\lambda}}{m!} \quad , \tag{2.25}$$

with

$$\lambda = N_A N_B \frac{v_D}{v} e^{-G^*/k_B T} \quad . \tag{2.26}$$

For very large N, Eq. (2.23) can be approximated by a Gaussian function in m of width $m^{1/2}$ and with a maximum at $m = m_{max}$ given by

$$\frac{m_{\mathrm{max}}}{(N_A - m_{\mathrm{max}})(N_B - m_{max})} = \frac{v_{\mathrm{D}}}{v} e^{-G^*/k_B T} \quad . \tag{2.27}$$

Because the relative width of this function tends to zero for large m, and with $[A] = (N_A - m)/(N_{Av}v)$, $[B] = (N_B - m)/(N_{Av}v)$, and $[AB] = m/(N_{Av}v)$, the law of mass action is readily recovered,

$$K_a = \frac{[AB]}{[A][B]} = v_{\mathrm{D}} N_{Av} e^{-G^*/k_B T} \quad . \tag{2.28}$$

K_a from counting and error estimate

The above framework enables to determine K_a from the number of dimers and monomers observed during a simulation.

For $N = 2$, Eqs. (2.17) and (2.21), and using Eqs. (2.22) and (2.23), yield

$$\frac{n_1}{n_0} \approx \frac{P_1}{P_0} = \frac{V_1}{V_0} e^{-G^*/k_B T} = \frac{1}{v/v_D - 1} e^{-G^*/k_B T} \quad , \tag{2.29}$$

or

$$G^* = -k_B T \ln \left[\frac{n_1}{n_0} \left(\frac{v}{v_D} - 1 \right) \right] \quad , \tag{2.30}$$

where n_1 and n_0 are the number of snapshots from the trajectory containing one dimer or two monomers, respectively. As can be seen, for simulation volumes v that are small compared to the molecular volumes v_D, an estimate for the latter is required. The association constant is then obtained by combining Eqs. (2.28) and (2.30),

$$K_a = N_{Av} \frac{n_1}{n_0} (v - v_\mathrm{D}) \quad . \tag{2.31}$$

For the general case of N particles, proceeding along similar lines, Eq. (2.17) serves to relate the average number of dimers $\langle m \rangle$ to G^*. Hence, G^* can be obtained from the dimer frequencies observed in a simulation of few particles — either via the second order approximations Eqs. (2.22) and (2.23), or via numerical integration, e.g., through Monte Carlo approaches, as demonstrated further below.

Moreover, simulations with $N > 3$ provide an independent approach to calculate G^*: For the number n_m of snapshots from the trajectory containing m dimers, Eqs. (2.17) and (2.21) yield

$$\frac{n_m}{n_0} \approx \frac{P_m}{P_0} = \frac{V_m}{V_0} e^{-mG^*/k_B T} \quad . \tag{2.32}$$

Therefore, the quantity

$$g_m = k_B T \ln \frac{n_0 V_m}{n_m V_0} \tag{2.33}$$

should satisfy $g_m = mG^*$, i.e., be proportional to m. Plotting g_m as a function of m thus provides the excess free energy per dimer G^* as the slope, and the plot yields a straight line only for properly chosen v_D.

Finally, error estimates are readily obtained from a Bayesian approach by considering the conditional probability

$$P(G^{*\prime}|n_0, n_1, n_2, \ldots) \propto P(n_0, n_1, n_2, \ldots|G^{*\prime}) \cdot P(G^{*\prime}) \tag{2.34}$$

$$\approx P(n_0, n_1, n_2, \ldots|G^{*\prime}) \cdot 1 = \prod_{m=0}^{\min(N_A, N_B)} \left[\frac{V_m e^{-mG^{*\prime}/k_B T}}{\sum_k V_k e^{-kG^{*\prime}/k_B T}} \right]^{\tilde{n}_m} \quad . \tag{2.35}$$

Here, a uniform *a priori* probability $P(G^{*\prime})$ for the dimer interaction free energy is assumed, and $\tilde{n}_m = n_m \cdot \Delta t / t_\mathrm{c}$ is the effective (i.e., statistically independent) number

of snapshots containing $m = 0, 1, 2, \ldots$ dimers. Several methods to determine this number are available, which critically determines the obtained error estimate, e.g., correlation analysis,[20, 21, 22] block averaging,[23, 24] and bootstrap analysis.[25, 26] For the test simulations presented below, the statistically independent number of snapshots is estimated from the average time t_c between collisions relative to the time spacing Δt of snapshots in the trajectory.

This probability distribution serves to calculate G^* and an estimate of its statistical uncertainty σ_{G^*} via

$$G^* = \int_{-\infty}^{\infty} G^{*\prime} P(G^{*\prime}|n_0, n_1, n_2, \ldots) dG^{*\prime} \tag{2.36}$$

and

$$\sigma_{G^*}^2 = \int_{-\infty}^{\infty} (G^{*\prime} - G^*)^2 P(G^{*\prime}|n_0, n_1, n_2, \ldots) dG^{*\prime} \quad . \tag{2.37}$$

An example python program is provided in the Appendix.

Resolving the seeming contradiction

We have shown above that for dilute systems of two dimerizing molecules, the Boltzmann approach, $\Delta G \propto \ln P_1/P_0$ [Eq. (2.12) and (2.13)], yields correct free energies, whereas direct application of the law of mass action provides wrong results. Nevertheless, as the above results show, the law of mass action can be derived from the Boltzmann approach and is in this sense compatible — as must be. What, then, is wrong with the expression $\Delta G \propto \ln P_1/P_0^2$ suggested by the law of mass action? As we will show in the following, *both* approaches are in fact correct; however, to apply the law of mass action to simulations of, e.g., only two dimerizing molecules requires a careful consideration of the relevant thermodynamic ensembles.

To demonstrate this, we again consider a two-particle MD system, and define the concentrations [A], [B], and [AB] from the respective probabilities,

$$[A] = \frac{N_A P_0}{v^{eff} N_{Av}} \quad , \tag{2.38}$$

$$[B] = \frac{N_B P_0}{v^{eff} N_{Av}} \quad , \tag{2.39}$$

and

$$[AB] = \frac{N_{AB} P_1}{v^{eff} N_{Av}} \quad , \tag{2.40}$$

where, for our two-particle system, $N = 2$, $N_A = N_B = N_{AB} = 1$, and v^{eff} is an appropriate effective volume. Inserting these concentrations into the law of mass action yields the puzzling result

$$K_a = \frac{[\mathrm{A}B] c^{\varnothing}}{[\mathrm{A}][\mathrm{B}]} = \frac{P_1}{P_0^2} c^{\varnothing} N_{Av} v^{eff} \quad . \tag{2.41}$$

Note, however, that it has not yet been defined how v^{eff} relates to the volume v of the simulation box. In contrast to what is saliently assumed by the above application of the law of mass action in Eq. (2.41), the MD ensemble does *not* represent

a (macroscopic) volume $v^{eff} = v \cdot N/2$ containing N interacting molecules (Figure 2.3b). Rather, it is an ensemble of $N/2$ simulation systems (Figure 2.3a), i.e., $N/2$ separate (periodic) boxes, each of volume v and containing two molecules. The crucial difference is that molecules from different boxes can never dimerize and, thus, the association rate, k^+, differs for the two ensembles. Since the dissociation rate, k^-, is unaffected, and $K_a \propto k^+/k^-$, the equilibrium concentrations also differ for the two ensembles. In particular, the time-averaged fraction P_0 of molecules in the monomeric state obtained from the MD simulation is, generally, not equal to the ensemble fraction expected for the macroscopic volume.

The two different ensembles can be reconciled by compensating for the fact that for each molecule within the $N/2$ MD boxes, a dimerization partner is available only during a fraction P_0 of the time. This is achieved by decreasing the effective volume by this factor, i.e., $v^{eff} := P_0 v N/2$ (note that properly correcting P_0 instead of v^{eff} yields the same results). Inserting this expression into Eq. (2.41) recovers Eq. (2.13), thus resolving the apparent contradiction.

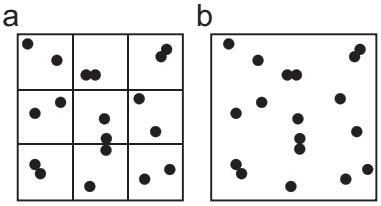

Figure 2.3: Two different ensembles are considered. Using dimerization frequencies obtained from an MD simulation of two molecules in a periodic box implicitly assumes an ensemble (a) consisting of replicas of the (microscopic) simulation system. Accordingly, the molecules cannot interact across the boxes (the configurations in a) represent snapshots at different time points, not periodic boxes). Straightforward application of the law of mass action, in contrast, refers to a macroscopic system of molecules (b), which can all form dimers with each other. As a consequence, the two ensembles generate different monomer / dimer ratios.

2.3 Molecular Dynamics Simulations

We carried out equilibrium MD simulations to demonstrate how the above framework can be applied to obtain association constants from simulations. Our simulation systems comprised of N van der Waals particles in a box, with N ranging from 2 to 64. $N/2$ of the particles were considered to be of type A and $N/2$ of type B, respectively (for uneven N, there was one excess A particle). The systems were simulated within periodic boundary conditions at constant volume using the Gromacs (v. 4.0.5) simulation package.[27] The temperature was kept constant using stochastic temperature coupling with an inverse friction coefficient of 5 ps. The neighbor list was updated at every integration time step, which was set to 50 fs. Test simulations with 10 fs and 20 fs integration time steps yielded identical results within the statistical errors. The particles had a mass of 72 amu and were interacting through a Lennard-Jones 6-12 potential

$$V_{LJ}(r) = 4\epsilon \left(\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right) \quad , \tag{2.42}$$

with $\sigma = 0.47$ nm and $\epsilon = 4$ kJ mol^{-1}. The potential was smoothly shifted to zero between 0.9 and 1.2 nm. For each N, we studied three different concentrations, corresponding to a volume per particle of $v/N = 27$, 54, and 108 nm^3, respectively.

Using these volumes, and at the simulation temperature of 298 K, the systems are in the gaseous state.

To obtain comparable statistics for a given computational cost for the different systems, each simulation was carried out for $64/N$ μs of simulation time. For example, the simulation system with $N = 2$ particles was simulated for 32 μs, whereas the simulation time for the system with $N = 64$ particles was 1 μs. These simulation times are long enough to have sufficiently many (i.e., thousands) of dimer association / dissociation events.

To obtain K_a from the simulations, the number of particles A within a distance r_c of any particle B was counted along the trajectory using the g_mindist tool of Gromacs. Subsequently, the obtained set of contacts was filtered for higher-order oligomers, which were discarded in order to not erroneously count them as dimers (the average number of particles in higher order oligomers was less than 1%, even at the highest concentration used). The dimer cut-off distance $r_c = 0.7$ nm was chosen such that the dimer peak observed in the radial distribution function is included. We did not investigate the dependence of K_a on the chosen dimer cut-off r_c,[28, 29] as our aim here was to compare different approaches for the calculation of K_a using the same cut-off.

Using the final (filtered) set of dimer contacts, G^* and its statistical uncertainty σ_{G^*} were calculated using the above Bayesian approach (Eq. (2.36) and Eq. (2.37); see Appendix for an example python program). To estimate the configuration space volumes, we used both the analytic approximation (Eq. 2.23) and a Monte Carlo approach (see below), which is numerically exact. Finally, the equilibrium constant was calculated from G^* using the standard state form of Eq. (2.28),

$$K_a = \frac{v_{\mathrm{D}}}{v^{\varnothing}} e^{-G^*/k_B T} \quad , \tag{2.43}$$

where $v^{\varnothing} = (c^{\varnothing} N_{Av})^{-1}$ is the (molecular) standard volume (1.66 nm^3) and $v_{\mathrm{D}} = 4/3\ \pi\ r_c^3$ (Figure 2.2). A uniform distribution $P(G^{*\prime})$ was used as a prior. To estimate the number of statistically independent snapshots, we calculated the average time between particle collisions according to $t_{\mathrm{c}} = \sigma_{\mathrm{c}}^{-1}\ \bar{c}^{-1/2}\ v/N$, with collision cross section $\sigma_{\mathrm{c}} = \pi\ (2\sigma_{\mathrm{LJ}})^2$ and mean velocity $\bar{c}$. We obtained $t_{\mathrm{c}} \approx 25$, 50, and 100 ps for the systems with $v/N = 27$, 54, and 108 nm^3, respectively.

2.4 Results and Discussion

Figure 2.4 summarizes the results obtained for the three different particle concentrations studied. The association constant K_a should be independent of the number of particles and of the volume of the simulation box. Figure 2.4 shows that this is indeed the case: Averaged over all particle numbers N, the obtained association constants ($\pm$ std. dev.) are 1.730 $\pm$ 0.014, 1.748 $\pm$ 0.009, and 1.752 $\pm$ 0.011 for $v/N = 27$, 54, and 108 nm^3, respectively. Furthermore, the statistical errors (bars in Figure 2.4) turn out to be independent of N, due to the comparable simulation times of 64/N μs per system.

As a check for the calculated K_a, Figure 2.5a shows a plot of g_m versus m. The linear dependence suggested by Eq. (2.33) holds, and the fit yields $G^* = -0.6951$ k$_B$T ($K_a = 1.734$), in excellent agreement with $K_a = 1.733 \pm 0.008$ obtained from counting the number of dimers in the simulation with $N = 64$, $v/N =$ 27 nm^3 (Figure 2.4).

Figure 2.5b confirms that for the simulated systems with moderately large N, the probabilities of finding m dimers follow a Poisson distribution, as derived above, Eq. (2.25).

As another, independent check for K_a, we integrated over the bound and unbound parts of the radial distribution function $g(r)$, as obtained from the simulation with $N = 2$ particles, according to

$$K_a = \frac{4\pi R^3 \int_0^{r_c} r^2 g(r) dr}{3v^{\varnothing} \int_{r_c}^{R} r^2 g(r) dr} \quad . \qquad (2.44)$$

The thus obtained equilibrium constant of $K_a = 1.74 \pm 0.02$ (solid line in Figure 2.4) also agrees with the result from counting.

To assess the accuracy of the analytical second-order approximation for the configuration space volumes V_m, Eq. (2.23), we placed $N = N_A + N_B$ non-interacting particles in a periodic 3-dimensional volume through Monte Carlo (MC) sampling, and counted the number of snapshots containing m dimers and $N - 2m$ monomers. As above, a dimer was defined by a distance criterion between any two particles A and B. Trimers and higher-order oligomers were discarded. The volume v and particle number N were systematically varied.

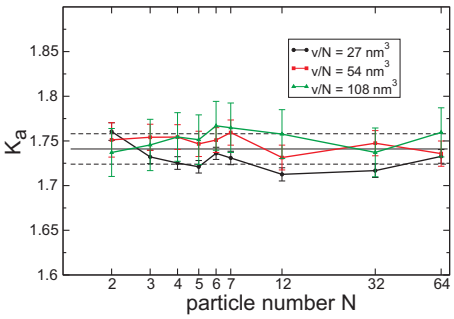

Figure 2.4: Association constant K_a obtained from counting the snapshots containing m dimers and $N-2m$ monomers in equilibrium MD simulations of a van der Waals gas. Systems with $N = 2$, 3, 4, 5, 6, 7, 12, 32, and 64 particles were simulated at volumes $v/N = 27$, 54, and 108 nm^3, respectively (black, red, and green curves, respectively); statistical errors were estimated using Bayesian statistics. The solid horizontal line gives K_a as obtained from integrating over the bound and unbound parts of the radial distribution function (Eq. 2.44); here, the statistical error (dashed lines) is the difference between the K_a's obtained from separately analyzing the two halves of the trajectory.

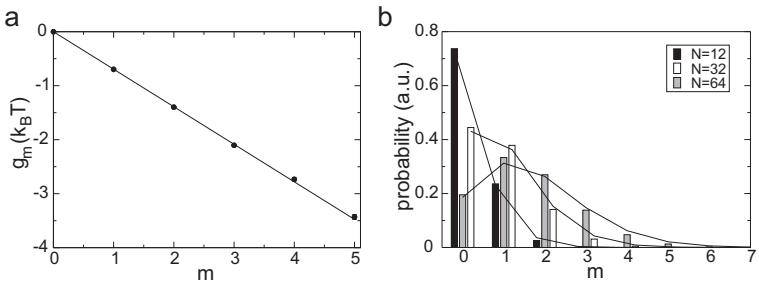

Figure 2.5: a) Plot of g_m over m, obtained from the simulation with $N = 64$ particles in a box with volume $v/N = 27$ nm^3. The slope of the linear fit (solid line) yields the same G^* as obtained from counting. b) The bars show the distributions of snapshots containing m dimers for the simulation systems with 12, 32, and 64 particles, respectively ($v/N = 27$ nm^3). They follow the corresponding Poisson distributions, Eq. (2.25), plotted as lines. The individual distributions are slightly shifted along the m-axis for clarity.

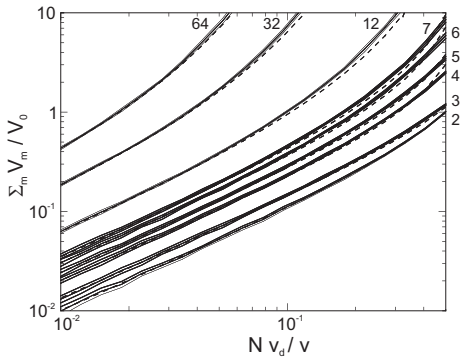

Figure 2.6: Comparison of dimer/monomer frequencies as a function of particle concentration, obtained from the second-order analytical approximation Eq. 2.23 (dashed lines) and Monte Carlo sampling (solid lines). Results are shown for $N = 2$, 3, 4, 5, 6, 7, 12, 32, and 64 (bottom to top). Statistical errors are shown as side lines.

Figure 2.6 shows that at low concentrations ($Nv_D/v \leq 0.1$), Eq. (2.23) is a very good approximation for the ratio of dimer/monomer configuration space volumes, particularly for small N. For higher concentrations $0.1 < Nv_D/v < 0.4$, Eq. (2.23) predicts a slightly too low ratio, whereas for even higher concentrations, the analytical approximation might slightly overestimate V_m/V_0, at least for $N \leq 7$. For $N = 2$, the results obtained from Eq. (2.23) and from MC sampling agree along the entire concentration range, because the analytical formula is exact in this case: The dimer configuration space volume V_m only contains terms beyond the second order for $N > 2$ (the monomer configuration space volume V_0 lacks higher order terms). In summary, from Figure 2.6 we conclude that for the diluted systems studied here ($0.01 < Nv_D/v < 0.05$), the second order analytical approximation Eq. (2.23) yields sufficiently accurate estimates of the configuration space volumes, and no further concentration-dependent correction is required.

2.5 Summary and Conclusions

We presented a derivation of the thermodynamics of dimerization reactions, and lay out how to calculate equilibrium constants and corresponding free energies from simulations of a limited number of dimerizing molecules. Using thermodynamic arguments, we have shown that naive application of the law of mass action, Eq. (2.41), yields wrong results in particular for simulations of only few dimerizing molecules, and that correct results are obtained via the Boltzmann factor of the ratio of the observed frequencies, Eq. (2.13). The difference between the two approaches can be significant, in particular if $P_0 \ll 1$ of course, as is the case for the systems studied in [12, 18].

We further derived through statistical mechanics the thermodynamics of dimerization reactions of any number of particles, and a Bayesian statistics approach to estimate equilibrium constants and free energies with their statistical errors from simulations. Finally, we showed that the two approaches can be reconciled by carefully considering the different underlying thermodynamic ensembles. We applied our approach to extract equilibrium constants from molecular dynamics simulations of systems containing different numbers of dimerizing particles.

One may ask whether there is an optimal system size to obtain statistically accurate free energies of dimerization from MD simulations through direct counting, given a certain available amount of computer time. From our results, we would argue in fa-

vor of simulating systems with $N = 2$ dimerizing molecules, for the following reasons. First, Eq. (2.23) is exact for the two-particle case, thus no concentration-dependent correction needs to be applied. Second, no trimers (or higher order oligomers) can occur, thus simplifying the analysis. Third, the computer time grows at least linearly with N, while our results show that simply having a larger number of molecules in the simulation box does not *per se* improve the statistical accuracy as compared to a simulation with two dimerizing molecules and correspondingly longer sampling time. For example, it would be better to carry out, e.g., four independent simulations with $N = 2$ molecules instead one single simulation with $N = 8$, as the latter would suffer from non-optimal parallel scaling.

Another question of practical importance is the choice of the simulation volume. One might argue that a large volume is desirable, because in that case concentrations can be used instead of activities, and second order effects (and thus the dimerization volume v_D) can be neglected. However, such an approach would suffer from a low statistical accuracy due to the small number of associations / dissociations. In addition, for simulations with explicit solvent, one seeks to reduce the number of solvent molecules as much as possible, since their treatment is computationally usually the most expensive part of the simulation. This discussion also underscores the importance of including v_D within our theory for obtaining free energies from counting in MD simulations.

Finally, we would like to discuss the advantages of running extended equilibrium simulations over biased simulations, such as umbrella sampling. From the latter, K_a can be calculated from the obtained potential of mean force $V_{\mathrm{m}f}(r)$ according to Eq. (2.44), with $g(r) = \exp(-V_{\mathrm{m}f}(r)/k_BT)$. This approach may seem more straightforward. However, it has the disadvantage that the system needs to be driven along a pre-defined reaction coordinate. This may be, for example, distances, angles, dihedrals, or (linear) combinations of these — even curvilinear coordinates may be required in certain cases. In general, the definition of a proper reaction coordinate may not always be straightforward and, furthermore, involve the derivation of Jacobian corrections that can become cumbersome for more complicated reaction coordinates. In such cases, it appears more convenient to run an unbiased simulation and choose the parameters to define the different states in an *a posteriori* manner during the analysis.

2.6 Supplementary information

Example python program using Bayes statistics for calculating G^* from an MD simulation. The configuration space volumes are estimated using Eq. (2.23).

```
from numpy import *
from math import factorial

def VM(na,nb,m,v,vd,y=None):
    '''Returns VM as a function of #particles (na,nb),\
    #dimers(m), volume(v), dimer volume(vd), excl.volume(y)'''
    x = vd/(1.0*v)
    if y == None: y = 15.*x/8.
    n = na + nb
    prod=1.0*factorial(m)*v**(1.0*n)
```

```
    prod=prod*factorial(na)/(factorial(m)*factorial(na-m))
    prod=prod*factorial(nb)/(factorial(m)*factorial(nb-m))
    prod=prod*x**m
    for j in range(1,m):
        prod=prod*(1.0-y*j)
    for j in range(1,n-2*m):
        prod=prod*(1.0-m*y-x*j)
    return prod

def bayes(vmarr,nf,na,nb,v,vd):
    '''A probability distribution for Gd is calculated, \
    given vmarr (array containing the # monomers,\
    single dimers (m=1), double dimers (m=2), etc \
    for every sample set, e.g. obtained with g_mindist) \
    and nf, the effective nr of samples (dt/tcoll), and na/nb.'''
    # Define the range and resolution of Gd
    gdmax = 2.; res=2000
    gdarr=2*gdmax*.arange(res)/(1.0*res)-gdmax
    #Choose a flat prior
    parr=ones(res)/(1.0*res)
    for n in range(len(vmarr)):
        for i in range(0,res):
            norm =0.0
            max_M=len(vmarr[n,:])
            for k in range(0,max_M):
                norm = norm + VM(na,nb,k,v,vd) * exp(-k*gdarr[i])
            prod=1.0
            for m in range(0,max_M):
                VMc=VM(na,nb,m,v,vd)
                prod=prod*(VMc*exp(-m*gdarr[i])/norm)**(1.0*nf*vmarr[n,m])
            parr[i]=prod*parr[i]
        parr = parr/(sum(parr*gdarr))
        gd = sum(gdarr * (parr/sum(parr)));
        print gd
    gdsigma = sum((gdarr -gd)**2 * (parr/sum(parr)));
    print 'gd = %.8f +/- %.8f'%(gd,sqrt(gdsigma))
```

2.7 Bibliography

[1] M.R Shirts, D.L Mobley, and J.D Chodera. Alchemical free energy calculations: Ready for prime time? *Annu. Rep. Comput. Chem.*, 3:41–59, 2007.

[2] J.L Knight and C.L Brooks III. lambda-dynamics free energy simulation methods. *J. Comput. Chem.*, 30:1692–1700, 2009.

[3] Y.Q Deng and B Roux. Computations of standard binding free energies with molecular dynamics simulations. *J. Phys. Chem. B*, 113:2234–2246, 2009.

[4] C.D Christ, A.E Mark, and W.F van Gunsteren. Basic ingredients of free energy calculations: A review. *J. Comput. Chem.*, 31:1569–1582, 2010.

[5] C Jarzynski. Nonequilibrium equality for free energy difference. *Phys. Rev. Lett.*, 78:2690–2693, 1997.

[6] C. Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: a master-equation approach. *Phys. Rev. E*, 56:5018–5035, 1997.

[7] G.E Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible markovian systems. *J. Stat. Phys.*, 90:1481–1487, 1998.

[8] F.M Ytreberg, R.H Swendsen, and D.M Zuckerman. Comparison of free energy methods for molecular systems. *J. Chem. Phys.*, 125:184114, 2006.

[9] M. Goette and H. Grubmüller. Accuracy and convergence of free energy differences calculated from nonequilibrium switching processes. *J. Comput. Chem.*, 30:447–456, 2009.

[10] G Hummer. *Free energy calculations: theory and applications in chemistry and biology.* Springer, 2007.

[11] G.M Torrie and J.P Valleau. Non-physical sampling distributions in monte-carlo free-energy estimation – umbrella sampling. *J. Comput. Phys*, 23:187–199, 1977.

[12] P.H Hünenberger, J.K Granwehr, J.N Aebischer, N Ghoneim, E. Haselbach, and W.F Van Gunsteren. Experimental and theoretical approach to hydrogen-bonded diastereomeric interactions in a model complex. *J. Am. Chem. Soc.*, 119:7533–7544, 1997.

[13] X Daura, W.F Van Gunsteren, and A.E Mark. Folding-unfolding thermodynamics of a beta-heptapeptide from equilibrium simulations. *Proteins–Stucture Function and Genetics*, 34:269–280, 1999.

[14] B.L de Groot, X. Daura, A.E Mark, and H Grubmüller. Essential dynamics of reversible peptide folding: Memory-free conformational dynamics governed by internal hydrogen bonds. *J. Mol. Biol.*, 309:299–313, 2001.

[15] Y Zhang and J.A McCammon. Studying the affinity and kinetics of molecular association with molecular-dynamics simulation. *J. Chem. Phys.*, 118:1821–1827, 2003.

[16] A.S Thomas and A.H Elcock. Molecular simulations suggest protein salt bridges are uniquely suited to life at high temperatures. *J. Am. Chem. Soc.*, 126:2008–2214, 2004.

[17] H. Yang and A.H Elcock. Association lifetimes of hydrophobic amino acid pairs measured directly from molecular dynamics simulations. *J. Am. Chem. Soc.*, 125:13968–13969, 2003.

[18] E Psachoulia, P.W Fowler, P.J Bond, and M.S.P Sansom. Helix-helix interactions in membrane proteins: coarse-grained simulations of glycophorin a helix dimerization. *Biochemistry*, 47:10503–10512, 2008.

[19] G.J Moro and M.G Severin. Bimolecular kinetics according to a stochastic analysis of reactant dynamics. *J. Chem. Phys.*, 114:4565–4578, 2001.

[20] G.M Jenkins and D.G Watts. *Spectral analysis and its applications.* Holden-Day, San Francisco, 1968.

[21] S.K Schiferl and D.C Wallace. Statistical errors in molecular dynamics averages. *J. Chem. Phys.*, 83:5203–5209, 1985.

[22] T.P Straatsma, H.J.C Berendsen, and A.J Stam. Estimation of statistical errors in molecular simulation calculations. *Mol. Phys.*, 57:89–95, 1986.

[23] H Flyvbjerg and H.G Petersen. Error estimates on averages of correlated data. *J. Chem. Phys.*, 91:461–466, 1989.

[24] B Hess. Determining the shear viscosity of model liquids from molecular dynamics simulations. *J. Chem. Phys.*, 116:209–217, 2002.

[25] B. Efron. Bootstrap methods - another look at the jackknife. *Ann. Stat.*, 7:1–26, 1979.

[26] V Knecht and H Grubmüller. Mechanical coupling via the membrane fusion snare protein syntaxin 1a: a molecular dynamics study. *Biophys. J.*, 84:1527–1547, 2003.

[27] B Hess, C Kutzner, D van der Spoel, and E Lindahl. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4:435–447, 2008.

[28] D. Shoup and A. Szabo. Role of diffusion in ligand binding to macromolecules and cell-bound receptors. *Biophys. J.*, 40:33–39, 1982.

[29] W.L Jorgensen and D.L Severance. Aromatic-aromatic interactions: Free energy profiles for the benzene dimer in water, chloroform, and liquid benzene. *J. Am. Chem. Soc.*, 112:4768–4774, 1990.

3

Dimerization of amino acid side chains: lessons from the comparison of different force fields

This chapter is based upon the manuscript:
Dimerization of Amino Acid Side Chains: Lessons from the Comparison of Different Force Fields by Djurre H. de Jong, Xavier Periole, Siewert J. Marrink, J. Chem. Theory Comput., **2012**, *8*, 1003-1014

Abstract

The interactions between amino acid side chains govern protein secondary, tertiary as well as quaternary structure formation. For molecular modeling approaches to be able to realistically describe these phenomena, the underlying force fields have to represent these interactions as accurately as possible. Here we compare the side chain-side chain interactions for a number of commonly used force fields, namely the all-atom OPLS, the united-hydrogen Gromos, and the coarse grain Martini force field. We do so by calculating the dimerization free energies between selected pairs of side chains and structural characterization of their binding modes. To mimic both polar and non-polar environments, the simulations are performed in water, *n*-octanol, and decane. In general, reasonable correlations are found between all three force fields, with deviations of the order of 1 kT in aqueous solvent. In apolar solvent, however, significantly larger differences are found, especially for charged amino acid pairs between the OPLS and Gromos force fields, and for polar interactions in the Martini force field in comparison to the higher resolution models. Interestingly, even in cases were the dimerization free energies are similar, the binding mode may differ substantially between the force fields. This was found to be especially the case for aromatic residues. In addition to the inter force field comparison, we compared the various force fields to a knowledge based potential. The two independent approaches show good correlation in aqueous solvent with an exception of aromatic residues for which the interaction strength is lower in the knowledge based potentials.

3.1 Introduction

For the last couple of decades molecular dynamics (MD) computer simulations have become a valuable tool to study proteins, from their dynamics and folding into a functional structure to their assembly into complexes. [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] The correct formation of these structures critically depends on a delicate balance between attractive and repulsive forces of the constituent amino acids. Together, these interactions determine the secondary,

tertiary and quaternary structures formed, giving rise to biological function. For molecular modeling approaches to be successful, it is therefore essential to describe these interactions as accurately as possible. It is a continuous challenge to improve the molecular force fields, in order to capture more and more realistic behavior. Much efforts have recently been devoted to the improvement of the protein backbone conformational preferences, which has an important contribution to protein folding. [11, 12, 13] Tests on the performance of the side chain interactions, however, are less abundant. In this work we focus on the performance of side chains (SC) by looking at association constants and kinetics of SC amino acid analogues (SCA). Three different force fields (FF) are compared: two atomistic, the OPLS and Gromos 53, and one coarse-grain (CG), Martini 2.1.

The parametrization of the SC non-bonded interactions in MD force fields might follow different strategies leading to different accuracies and resolutions. The OPLS FF has been parameterized against conformational energies and properties of organic liquids. [14, 15] In this FF hydrogen atoms do not carry van der Waals interactions but are represented by a Coulomb interaction site. The Gromos 53A5/53A6 FF has been parameterized against free enthalpies of solvation of small molecular building blocks in water and cyclohexane. [16] Gromos is a united-atom (UA) FF where aliphatic hydrogens are not explicitly modeled, i.e. they have no non-bonded interactions. This reduces the number of interaction sites to compute and thus leads to a speed up of simulations. The CG Martini FF has even less interaction sites: on average four heavy atoms are mapped into one bead. This leads to a significant computational speed up, but at the cost of simulation details. The Martini FF has been parameterized primarily targeting the partitioning free energies of model compounds between polar and apolar phases. [17, 18]

To assess amino acid SC parameters, we report on the interaction strength of SCA as characterized by their dimerization (or binding, association) free energies, ΔG^{dimer}. Earlier studies looked at the dimerization free energies of a limited set of SC pairs in water, [19] or at all SC combinations in water in order to derive effective residue-based contact energies for the development of CG potentials. [20] Here we systematically calculated ΔG^{dimer} for pairs of SCA in polar (water) as well as apolar solvents (n-octanol and decane) in order to probe the SC in a mimic of the diversity of biological environments. While for soluble proteins the values of ΔG^{dimer} obtained in water will be most relevant, in lipid membranes the dielectric screening much more resembles that of apolar solvents and consequently the values obtained in octanol and decane are more relevant. For all solvents we analyzed the structural binding modes found in the different force fields and looked at the kinetics of the association process through calculation of the survival correlation times of the SCA pairs, in order to better explain possible differences in ΔG^{dimer}.

The comparison of different force fields can only address the consistency between the force fields. While a strong similarity between force fields that have been parameterized against different experimental data, would certainly increase the confidence in the force fields, only a comparison to experimental data can give a direct assessment of the quality of the force fields. To the best of our knowledge, experimental binding free energies of SCAs in solution are not available in the literature. The only data we found pertains to the backbone analogue. [21] However an indirect measure of the affinity of SCAs for each other can be obtained from a population analysis of their contacts in proteins, which has been used to estimate effective binding free energies. [22] Thus, as an independent check for the side chain interactions we compare the results obtained for the molecular dynamics FFs to a knowledge based potential.

The chapter is organized as follows. In section 3.2, a description of the FF parameters and methods to obtain ΔG^{dimer} is given. In section 3.3, ΔG^{dimer} for 21 different amino acid SCA pairs are reported and compared. This section is split into the comparison of the high resolution FFs, OPLS and Gromos (section 3.3), followed by the comparison of the Martini CG FF to both OPLS and Gromos FF (section 3.3), the analysis of structural

binding modes (section 3.3) and finally the comparison of all three FFs to a knowledge based potential (section 3.3). In section 3.3 we analyze the dimerization kinetics of the different FFs in water. A short concluding section ends this chapter.

3.2 Methods

System set-up

All systems simulated consisted of two amino acid SCA solvated either in water, *n*-octanol or decane. The systems contained 880, 113, and 80 solvent molecules for water, octanol and decane, respectively, in a cubic unit-cell of ≈ 3.0 nm. For the CG system, 880 water molecules corresponds to 220 water beads, of which 10% were replaced by antifreeze particles. [17]

The dimerization free energies, ΔG^{dimer}, were calculated for 21 selected amino acid SCA pairs (See tables 3.1 and 3.2) and were classified into five groups: charged (Lys-Glu, Lys-Lys, Arg-Asp), polar (Asn-Asn, Ser-Ser, Gln-Asn, Gln-Gln), apolar (Leu-Leu, Val-Val, Leu-Val), mixed (Leu-Asn, Leu-Gln, Leu-Lys, Asn-Lys, Met-Ser) and aromatic (Trp-Trp, Tyr-Tyr, Phe-Phe, His-Phe, Trp-Tyr). The SCAs were constructed by replacing the C_α in the C_β-C_α bond by a hydrogen and omitting the rest of the backbone. Backbone-backbone dimerization free energies were also calculated. A backbone (BB) analogue was defined by a Gly residue as found in a continuous protein backbone. In other words, the C-terminus only has a carbonyl oxygen attached, not a complete acid group. Although this is not a chemically realistic and stable molecule, it best mimics the interactions found in a protein.

Table 3.1: Three letter amino acid abbreviations and side chain analogs used.

Side chain	Analogue
Leu	*iso*-butane
Val	propane
Lys	butyl amine
Glu	propionic acid
Phe	toluene
Asn	acetamide
Trp	3-methylindole
Tyr	*p*-cresol
Ser	methanol
Met	methyl ethyl sulfide
Arg	propyl guanidine
Asp	acetic acid
Gln	propion amide
His	4-methyl imidazole
BB	*see text*

The SCAs were initially randomly placed in the solvent box. The systems were then minimized and equilibrated for 20000 steps, followed by a 240 ns production run during which frames were saved every 500 steps for analysis. All simulations were performed using version 4.0.x of the Gromacs [23] simulation package.

Simulation and force field parameters

All simulations were performed using three levels of description: the all-atom (AA), the united-atom (UA), and the coarse-grain (CG) level. For the AA simulations, the OPLS FF [14][15] was used, as implemented in the Gromacs 4.0.x package [23]. The SPC model [24] was used to represent water. Although the TIP3P model is a more common choice in combination with OPLS, we decided upon SPC, to ensure that we were comparing properties of the protein FFs and not variations of the solvents. Parameters for *n*-octanol and decane compatible with the OPLS FF were taken from Garrido *et al.*[25] The Lennard-Jones (LJ) and Coulombic interactions were cutoff at 1.4 nm with a neighbor list update every 5 steps. To correct for the truncation of electrostatic interactions beyond the cutoff a reaction-field

(RF) correction was applied. [26] The relative dielectric constant for the RF, ϵ_{rf}, was set to 54, 8.8 and 2 for water, *n*-octanol and decane, respectively.

For the UA simulations, the Gromos FF [16] was used, as implemented in the Gromacs 4.0.x package. [23] For version 53 of the Gromos FFs two parameter sets have been used. For the apolar solvents 53A5 was used, for which the amino acid parameters have been derived in cyclohexane. For polar solvents 53A6 is used, which is the same as 53A5, except for the partial charges, that have been adjusted to reproduce the hydration free enthalpies in water. The SPC model [24] was used to represent water. Parameters for *n*-octanol and decane compatible with the Gromos FF were taken from Garrido *et al.* [25] The LJ interactions and Coulombic interactions were handled as done in the AA simulations. The relative dielectric constant for the RF, ϵ_{rf}, was set to 54, 8.8 and 2 for water, *n*-octanol and decane, respectively.

For the CG simulations, the Martini FF [17] and its extension to proteins [18] was used. The parameters for the SCAs and the solvents were as reported in these publications. Simulations were also performed with the recently parameterized polarizable Martini water model. [27] For the backbone analogue, the 'P5' particle type was used, which represents the backbone in an unstructured peptide. Following the standard protocol for Martini, the LJ interactions were smoothly shifted to zero between 0.9 and 1.2 nm, using the switch potential implemented in Gromacs. The coulombic interactions were shifted between 0 and 1.2 nm, for both regular and polarizable Martini. The neighbor list was updated every 10 steps. Effective dielectric constants $\epsilon_r = 15$ for regular Martini and $\epsilon_r = 2.5$ for polarizable Martini water were used to screen the Coulombic interactions.

In all simulations the temperature was kept close to its target value, $298K$, using Berendsen's weak coupling algorithm ($\tau_{\mathrm{T}} = 0.1, 0.1$ and 0.3 ps for AA, UA, and CG, respectively). The SCAs and the solvents were coupled separately. The pressure was maintained close to 1 bar using an isotropic weak coupling scheme ($\tau_{\mathrm{P}} = 1.0, 1.0$ and 3.0 ps for AA, UA, and CG, respectively). [28] The integration time step was 2 fs for both AA and UA simulations and 30 fs for the CG systems. All bonds in the AA and UA systems, as well as the bonds of the aromatic residues in the CG systems, were constraint with the LINCS-algorithm. [29]

Calculation of dimerization free energies

The free energy of dimerization, ΔG^{dimer}, was calculated using the formula derived for two particle systems, as explained in chapter 2 of this thesis:

$$\Delta G^{bind} = \Delta G^{\varnothing} = -k_B T \ln \left(\frac{v_d n_1}{v^{\varnothing} n_0} \left[\frac{v}{v_d} - 1 \right] \right) \tag{3.1}$$

where k_B is Boltzmann's constant, T the temperature, v_d the dimer volume (calculated as the sum of the experimental SC volumes[30]), $v^{\varnothing}$ the standard volume of 1.66 nm^3 (equivalent to a concentration of 1 mol/l), v the volume of the system and n_0 and n_1 are the respective number of monomers and dimers counted in the trajectory. The discrimination between the bound and unbound states was based on the distance between the center-of-mass (COM) of the SCs. A SC pair was considered in a bound state if the distance between their COMs was less than a cutoff. The cutoff was different for each pair, chosen to equal the distance of the first minimum of the SC-SC radial distribution function (RDF). Most simulations showed hundreds to thousands of association and dissociation events. Errors are calculated using the Bayesian statistics approach as previously described (see chapter 2.

In order to calculate statistically reliable free energy of dimerization at least 50 association and dissociation events are needed. For the amino acid SC pairs that did not show sufficient spontaneous events, a potential of mean force (PMF) was calculated as a function of the SC's COM distance. Simulations were run for distances constrained in the range 0.24-1.50 nm with a 0.02 nm interval. At each distance the system was simulated for 2 ns from which

the first 500 ps were discarded as equilibration period. Over the remaining simulation time the mean constraining force was calculated using the constraint pulling code implemented in Gromacs and integrated as described by Hess *et al.* [31]

Analysis of structural binding modes

The structural modes of dimerization of the SCAs were analyzed on the basis of their relative orientation at short distances (¡ 0.75 nm). The SCAs relative orientation was described by a set of simple geometrical reaction coordinates described here. The distance between the center-of-masses (COM) of the SCAs was used in all cases. For aromatic side chains both the angle between the normals of the ring planes (*'angle'*) and the dihedral angle between two vectors through SCA rings (*'dihedral'*) were used. The ring plane was defined by the triangle of three atoms in the ring. The dihedrals were defined using two atoms on each ring. For non-aromatic SCAs we used the angle between vectors describing the long axis of the molecules, where the vector was defined by two atoms of the side chain. Both the angle and the dihedral where plotted against the distance between the SCAs. The probability distributions for each reaction coordinate were calculated, normalized and converted into free energy surfaces using:

$$\Delta G = -k_B T \ln(P) \tag{3.2}$$

where k_B is boltzman's constant, T is the temperature and P is the normalized probability. Analyses were performed using standard Gromacs analysis tools and local scripts. Visual analysis of the trajectories was performed using VMD [32].

Calculation of complex lifetimes

The lifetimes of the bound state of amino acid SCs pairs was calculated using the following time-correlation function also called survival correlation function [33]:

$$S_L(t) = \frac{1}{(t_{tot} - \Delta t)} \sum_{t=0}^{t_{tot}-\Delta t} p(t, t + \Delta t) \tag{3.3}$$

where $p(t, t + \Delta t)$ is set to 1 if the SCs are bound for a period of time Δt after their initial dimerization at time t and t_{tot} is the length of the trajectory analyzed. To speed up the analysis, the 240 ns trajectories were split into 50 blocks of 4.8 ns, which is still about two orders of magnitude longer than the observed lifetimes (see Results). A running average of 5 ps was applied to filter out fast association-dissociation events. The lifetime of a complex, τ_L, was obtained by a weighted (with the standard deviation) least squares fit of the mean of $S_L(t)$ over the 50 blocks using the following expression:

$$< S_L(t) > \approx \exp[-(t/\tau_L)] \tag{3.4}$$

3.3 Results and Discussion

We present the analysis of the dimerization behavior of a representative set of amino acid side chain pairs. Table 3.2 lists the complete set of dimerization free energies, ΔG^{dimer}, calculated for 21 pairs, in three different solvents (water, *n*-octanol, and decane) and for three different force fields (OPLS, Gromos, and Martini). For pairs that showed insufficient dimerization events, PMFs were determined and the depth of the well at contact distance of these PMFs is reported instead (Table 3.3). The results are presented and discussed in more detail in the following sections. First

	Water				Octanol			Decane		
Pair	Pol.Mart.	Martini	Gromos	OPLS	Martini	Gromos	OPLS	Martini	Gromos	OPLS
Charged										
Lys & Glu	-0.00 ± 0.13	-1.67 ± 0.05	-0.08 ± 0.08	-1.35 ± 0.06	-3.53 ± 0.06	$- \pm -$	$- \pm -$	-10.35 ± 1.71	$- \pm -$	$- \pm -$
Lys & Lys	0.37 ± 0.12	0.85 ± 0.18	0.88 ± 0.22	1.21 ± 0.31	0.20 ± 0.12	$- \pm -$	$- \pm -$	-5.14 ± 0.07	$- \pm -$	$- \pm -$
Arg & Asp	-0.16 ± 0.12	-1.66 ± 0.05	-0.13 ± 0.07	-2.78 ± 0.05	-4.16 ± 0.06	$- \pm -$	$- \pm -$	$- \pm -$	$- \pm -$	$- \pm -$
Polar										
Ser & Ser	-0.02 ± 0.10	0.21 ± 0.12	0.58 ± 0.15	0.54 ± 0.14	-0.24 ± 0.09	-0.17 ± 0.15	-0.33 ± 0.15	-0.90 ± 0.06	-4.61 ± 0.14	-2.33 ± 0.06
Gln & Asn	0.11 ± 0.12	0.13 ± 0.14	0.07 ± 0.11	-0.15 ± 0.10	-0.82 ± 0.06	-0.00 ± 0.13	0.53 ± 0.21	-1.65 ± 0.12	-7.32 ± 0.74	$- \pm -$
Asn & Asn	0.23 ± 0.13	0.35 ± 0.15	0.22 ± 0.11	0.06 ± 0.11	-0.81 ± 0.07	0.11 ± 0.14	-0.53 ± 0.09	-1.50 ± 0.10	-7.14 ± 1.13	$- \pm -$
Gln & Gln	0.25 ± 0.13	-0.02 ± 0.13	-0.14 ± 0.10	-0.53 ± 0.11	-0.71 ± 0.07	-0.47 ± 0.11	-0.55 ± 0.12	-1.34 ± 0.05	-6.96 ± 0.96	$- \pm -$
Mixed										
Leu & Asn	0.20 ± 0.11	0.85 ± 0.18	-0.49 ± 0.19	-0.46 ± 0.19	0.88 ± 0.20	0.51 ± 0.41	0.46 ± 0.38	-0.02 ± 0.10	0.25 ± 0.26	-0.32 ± 0.19
Leu & Gln	0.15 ± 0.10	0.41 ± 0.14	-0.29 ± 0.09	-0.38 ± 0.09	0.50 ± 0.14	0.33 ± 0.14	-0.18 ± 0.11	-0.02 ± 0.10	-0.18 ± 0.08	-0.24 ± 0.08
Leu & Lys	-0.33 ± 0.07	-0.12 ± 0.09	0.15 ± 0.11	0.17 ± 0.12	0.14 ± 0.12	0.48 ± 0.16	0.60 ± 0.19	-1.88 ± 0.05	-0.38 ± 0.07	-0.12 ± 0.08
Asn & Lys	0.23 ± 0.12	0.08 ± 0.12	-0.01 ± 0.11	0.07 ± 0.11	-1.21 ± 0.06	-1.49 ± 0.06	-1.58 ± 0.05	-3.80 ± 0.05	$- \pm -$	$- \pm -$
Met & Ser	-0.17 ± 0.09	0.05 ± 0.11	0.16 ± 0.11	0.14 ± 0.11	-0.02 ± 0.10	-0.02 ± 0.11	0.72 ± 0.21	-0.37 ± 0.08	-0.32 ± 0.07	-0.59 ± 0.08
Apolar										
Leu & Leu	-1.19 ± 0.04	-0.88 ± 0.03	-0.37 ± 0.09	-0.34 ± 0.09	-0.19 ± 0.08	-0.37 ± 0.08	-0.52 ± 0.08	0.11 ± 0.11	-0.29 ± 0.07	-0.24 ± 0.10
Val & Val	-1.02 ± 0.05	-0.50 ± 0.07	-0.10 ± 0.10	-0.11 ± 0.09	-0.07 ± 0.09	-0.64 ± 0.06	-0.18 ± 0.10	0.11 ± 0.11	-0.06 ± 0.08	-0.11 ± 0.10
Leu & Val	-1.04 ± 0.05	-0.72 ± 0.05	-0.16 ± 0.10	-0.35 ± 0.09	-0.04 ± 0.09	-0.17 ± 0.09	0.00 ± 0.11	0.11 ± 0.11	-0.03 ± 0.08	-0.29 ± 0.10
Aromatic										
Trp & Trp	-1.66 ± 0.06	-1.73 ± 0.06	-1.40 ± 0.05	-1.28 ± 0.05	-0.40 ± 0.11	-1.09 ± 0.06	0.56 ± 0.18	-0.81 ± 0.10	-1.21 ± 0.05	-1.49 ± 0.06
Tyr & Tyr	-1.07 ± 0.08	-1.25 ± 0.07	-0.92 ± 0.07	-1.08 ± 0.06	-0.13 ± 0.13	-1.42 ± 0.06	0.41 ± 0.18	-0.95 ± 0.09	-3.93 ± 0.07	-2.27 ± 0.06
Phe & Phe	-2.06 ± 0.05	-1.81 ± 0.06	-0.49 ± 0.08	-0.79 ± 0.07	-0.08 ± 0.14	0.35 ± 0.18	-0.05 ± 0.14	-0.52 ± 0.59	-1.84 ± 0.05	-0.53 ± 0.08
His & Phe	-0.99 ± 0.08	-0.88 ± 0.09	-0.66 ± 0.07	-0.71 ± 0.07	0.07 ± 0.16	-0.19 ± 0.16	-0.36 ± 0.09	-0.83 ± 0.09	-6.16 ± 0.18	-1.25 ± 0.07
Trp & Tyr	-1.56 ± 0.06	-1.32 ± 0.07	-1.05 ± 0.06	-1.11 ± 0.06	-0.33 ± 0.11	-0.63 ± 0.10	-0.43 ± 0.10	-1.02 ± 0.09	-2.38 ± 0.04	-1.47 ± 0.07
Backbone										
BB & BB	0.23 ± 0.13	0.35 ± 0.15	-0.37 ± 0.08	0.33 ± 0.13	-0.81 ± 0.07	-1.38 ± 0.07	-0.23 ± 0.20	-1.50 ± 0.10	-4.04 ± 0.08	-1.66 ± 0.07

Table 3.2: Binding free energies (k_BT) for pairs of amino acids in water, octanol and decane. Pairs have been grouped according to chemical properties: charged, apolar, polar, aromatic, mixed, and backbone. Statistical errors were obtained using Bayesian statistics. Only dimerization free energies obtained from free simulations are shown.

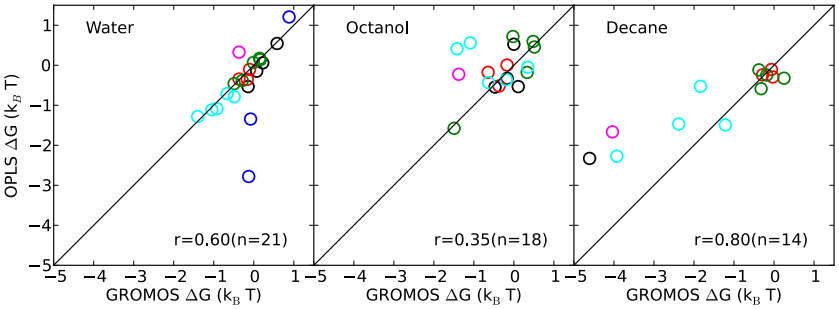

Figure 3.1: Comparison of dimerization free energies for the OPLS and Gromos FFs in water (left panel), *n*-octanol (middle panel) and decane (right panel). Blue circles represent charged SC pairs, black circles polar SC pairs, red circles apolar SC pairs, green circles mixed SC pairs, cyan circles aromatic SC pairs and magenta circles the backbone-backbone pair. The black line represents a regression of one. r is the Pearson correlation coefficient, calculated over n data points.

the AA OPLS FF is compared to the UA Gromos FF, followed by a comparison of both atomistic FFs to the CG Martini FF. The modes of binding are structurally characterized. The ΔG^{dimer} obtained with the three FFs are then compared to contact energies obtained from a knowledge-based potential. Finally the lifetimes of the bound SCA pairs in water for the different FF are discussed.

OPLS versus Gromos

The correlation between the SCAs ΔG^{dimer} obtained using the Gromos 53A5/53A6 and OPLS FF in the three different solvents is shown in Figure 3.1. The raw data together with error estimates are listed in tablee 3.2.

In water, the ΔG^{dimer} averaged over the 21 pairs is -0.22 ± 0.51 and -0.42 ± 0.81 k_BT for the Gromos and OPLS FFs, respectively. The root mean square difference (RMSD) between the two FFs is 0.68 k_BT. The oppositely charged amino acid (Arg-Asp and Lys-Glu, blue circles) show much stronger interaction (more negative ΔG^{dimer}) for the OPLS

Table 3.3: Deepest well points obtained from potentials of mean force, given in k_BT. PMFs were only calculated for amino acid pairs that showed insufficient association-dissociation events in the free simulations.

	Octanol			Decane		
Pair	Martini	Gromos	OPLS	Martini	Gromos	OPLS
Asn & Asn	–	–	–	-2.7 ± 0.1	-7.5 ± 0.5	-11.7 ± 0.1
Gln & Gln	–	–	–	-2.7 ± 0.1	-8.0 ± 0.2	-9.2 ± 0.2
Gln & Asn	–	–	–	-2.8 ± 0.1	-7.8 ± 0.2	-10.1 ± 0.6
Asn & Lys	–	–	–	-4.5 ± 0.1	-24.6 ± 0.2	-29.3 ± 0.3
Arg & Asp	-5.0 ± 0.1	-19.2 ± 1.1	-18.8 ± 2.2	-11.7 ± 0.1	-118.7 ± 0.2	-134.3 ± 0.2
Lys & Glu	-4.5 ± 0.1	-18.5 ± 1.7	-17.0 ± 1.9	-11.2 ± 0.1	-127.6 ± 0.2	-149.8 ± 0.3

FF. Omitting these two pairs, the RMSD decreases to 0.23 k_BT. The RMSDs for the polar, apolar, aromatic and mixed SCA groups are 0.24, 0.11, 0.17 and 0.06 k_BT, respectively. The correlation coefficient, r, is 0.60 over 21 pairs and 0.92 when omitting the charged pairs (Figure 3.1).

In n-octanol, for the 18 pairs that showed sufficient association-dissociation events, there is a considerable, inter FF spread of the ΔG^{dimer}. The averaged ΔG^{dimer} are -0.35 ± 0.63 and -0.09 ± 0.56 k_BT for the Gromos and OPLS FF, respectively. RMSDs are 0.72 k_BT for all 18 pairs and 0.42, 0.30, 1.12 and 0.41 k_BT for the polar, apolar, aromatic and charged SC groups. The correlation coefficient over the 18 pairs was $r = 0.35$ (Figure 3.1). For pairs that did not show sufficient association-dissociation events, the PMF was calculated to characterize the interaction strength. This was the case for Lys-Lys, Arg-Asp and Lys-Glu pairs. The well's depth in the PMF for the oppositely charged SCA pairs (Arg-Asp and Lys-Glu) is similar for the OPLS and Gromos FFs (cf. Table 3.3). For the equally charged pair (Lys-Lys) no well depth could be determined since the PMF does not show a minimum. Both PMFs show a good agreement (data not shown).

Note that simulations of n-octanol have shown that it may adopt local structure. [34] These structures are more pronounced in hydrated octanol as compared to dry octanol, which we choose for our simulations. Careful examination of the trajectories showed that no macroscopic structure (micelles or inverse micelles) formed. The local structure was characterized by calculating the oxygen-oxygen and oxygen-hydrogen RDFs, which described similar structure as shown in the figure 4 of MacCallum *et al.* [34] (data not shown). As far as we could determine the local structure observed did not impair the sampling of the SCA-SCA interactions.

In decane, only 14 pairs showed sufficient association-dissociation events and their averaged ΔG^{dimer} are -1.80 ± 2.01 and -0.92 ± 0.77 k_BT for Gromos and OPLS, respectively. The RMSDs are 1.71 k_BT for the 14 pairs and 2.28, 0.16, 2.43 and 0.34 k_BT for the polar, apolar, aromatic and charged SC groups. The correlation coefficient over the 14 pairs is $r = 0.80$ (Figure 3.1). The depth of the well in the PMF for the pairs that showed insufficient association-dissociation events in decane (i.e. Lys-Lys, Arg-Asp, Asn-Lys, Gln-Gln, Asn-Asn, Gln-Asn, Arg-Asp and Lys-Glu) are indicated in table 3.3. Two PMFs (Lys-Glu and Lys-Lys) are shown in Figure 3.3. In the case of the polar pairs (Asn-Asn, Gln-Gln and Asn-Gln) the wells are considerably deeper for the OPLS FF ($\sim 1 - 4$ k_BT), consistent with the lack of association-dissociation events in the free simulations of these pairs using the OPLS FF. In the case of the oppositely charged and mixed pairs (Arg-Asp, Lys-Glu and Asn-Lys) a similar trend is observed although the differences are larger ($\sim 5 - 22$ k_BT). For the equally charged Lys-Lys pair the interaction is virtually the same in both the Gromos and the OPLS FF (Figure 3.3).

The RMSD between the FFs over all pairs that showed sufficient association-dissociation events are 0.68, 0.72, and 1.71 k_BT for water, n-octanol and decane, respectively and 1.10 k_BT averaged over all three solvents. The fact that their performance is most alike in water (smallest RMSD) is to be expected, given that amino acids parameters have been primarily derived for aqueous solutions. Larger differences between the two FFs are found in n-octanol and decane. Especially the difference in dimerization free energies for aromatic and charged side chains is worrisome, with a RMSD exceeding 2 k_BT. The discrepancy may partly result from the difference in solvent models (absence of aliphatic hydrogens and corresponding partial charges in the Gromos FF). The interactions between charged SCAs are stronger in the OPLS FF in all solvents. We attribute this to the difference in the distribution of partial charges on charged groups between both FFs. The OPLS FF has larger partial charges when compared to the Gromos FF, leading to larger dipoles in the molecules and therefore stronger interactions, especially in solvents of low polarity. The aromatic SCs correlate well in water, yet some stronger interaction is observed with the Gromos FF in

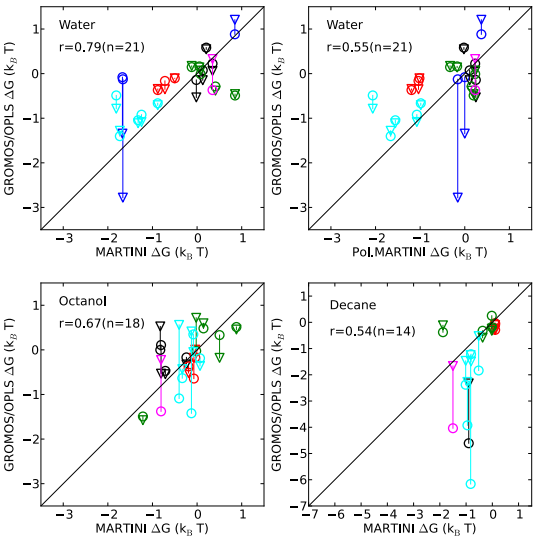

Figure 3.2: Comparison of dimerization free energies between Martini and OPLS or Martini and Gromos FFs in water (top left panel), polarizable water (top right panel), octanol (bottom left panel), and decane (bottom right panel). Circles (○) represent Gromos-Martini comparison, triangles (▽) represent OPLS-Martini comparison. Lines connect the same pair in each of the two atomistic models. Blue are charged, black are polar, red are apolar, green are mixed and cyan are aromatic SC pairs and magenta is the backbone-backbone pair. The black line represents a regression of one. r is the Pearson correlation coefficient, calculated over n SC pairs.

both n-octanol (Trp-Trp, Tyr-Tyr) and decane (all except Trp-Trp). In these cases the distribution of charge difference between the ring carbons and protons is larger in the Gromos FF than in the OPLS FF, leading to larger local dipoles.

Backbone-backbone interactions are also stronger for the Gromos FF. This is true in all solvents, and again most pronounced in decane. Experimental data on the dimerization free energy of the backbone analogue molecule n-methylacetamide (NMA) point to an even larger difference between dimerization in water versus organic solvent. Values in the range of 7-9 k_BT are listed by Ben-Tal *et al.* [21], compared to ΔG^{dimer}(water-decane) of 3.7 and 2.0 k_BT for Gromos and OPLS, respectively. However, the authors also note that the interpretation of the experimental data is far from trivial.

Martini versus atomistic

Figure 3.2 shows the correlation plots between the ΔG^{dimer} obtained for amino acid SCAs using the Martini CG FF and the joint data obtained for the atomistic FFs discussed above. The raw data and associated error estimates are given in table 3.2. To ease the comparison, the results for the atomistic FFs were averaged and are referred to as $\overline{AT}$ in the following.

In water, the average ΔG^{dimer} over the 21 pairs is -0.44 ± 0.85, -0.42 ± 0.81 and -0.22 ± 0.51 for the Martini, OPLS and Gromos FFs, respectively. The RMSD over these pairs and between the Martini FF and $\overline{AT}$ is 0.54 k_BT. It should be noted that for the charged residues there is a large spread between OPLS and Gromos (cf. the connecting lines in Figure 3.2). Omitting the charged SC yields a RMSD value of 0.52 k_BT. The small difference reflects a cancellation between the Gromos and OPLS FFs, with the average dimerization free energy close to the value obtained for the Martini FF. The groups of polar, apolar, aromatic, and mixed SCAs have RMSDs of 0.27, 0.46, 0.58 and 0.69 k_BT, respectively. The Pearson correlation coefficient over all pairs is $r = 0.79$ (Figure 3.2). The average ΔG^{dimer} over the 21 pairs is not affected when switching to the polarizable Martini water model: -0.45 ± 0.74. The RMSDs change to 0.66 k_BT for all pairs and 0.42, 0.84, 0.70, and 0.46 k_BT for polar,

apolar, aromatic, and mixed SCAs, respectively. The correlation of the dimerization free energies with respect to the AA models becomes slightly worse (Pearson coefficient drops to 0.55, Figure 3.2), however, this is largely attributed to the effect of the charged pairs. Whereas with standard Martini water the $\overline{AT}$ value is closely matched, the polarizable model matches Gromos very closely, but not OPLS.

In *n*-octanol, the averaged ΔG^{dimer} over the 18 pairs showing sufficient association-dissociation events are -0.22 ± 0.49, -0.09 ± 0.56 and -0.35 ± 0.63 k_BT for Martini, OPLS and Gromos FFs, respectively. The RMSDs between Martini and $\overline{AT}$ are 0.40 k_BT for all pairs and 0.63, 0.24, 0.27, and 0.38 k_BT for the polar, apolar, aromatic, and mixed SCAs, respectively. The correlation coefficient over these 18 pairs is $r = 0.67$ (Figure 3.2). For the charged pairs, insufficient sampling was achieved to determine ΔG^{dimer}. The PMFs, however, show that the Martini FF systematically underestimates the attraction between unlike-charges when compared to both atomistic FFs, as evidenced by the well depths of Lys-Glu and Arg-Asp pairs (Table 3.3). Similarly, the repulsion between like-charges is underestimated (Lys-Lys pair, data not shown).

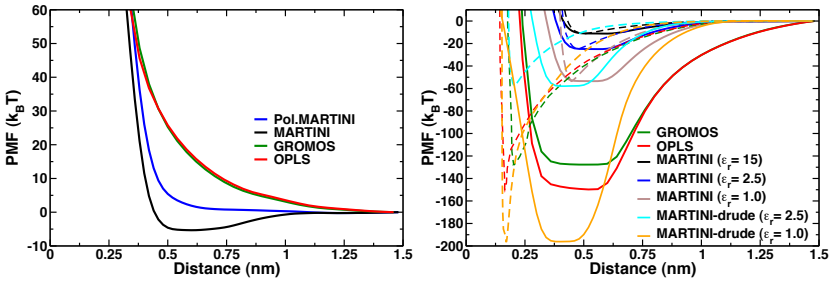

Figure 3.3: Potential of mean force for Lys-Lys (left) and Lys-Glu (right) SC pairs in decane. Solid lines are PMF plotted against the center of mass distance, dashed lines are PMFs plotted against charge group minimum distance. Black lines are Martini, green lines are Gromos, red lines are OPLS, blue lines are polarizable water Martini ($\epsilon_r = 2.5$), brown lines are Martini without dielectric screening ($\epsilon_r = 1.0$), cyan lines are an alternative model for charged residues in Martini (drude-like, see text) with $\epsilon_r = 2.5$, orange lines are the same model with $\epsilon_r = 1.0$. Errors do not exceed the thickness of the lines.

In decane, for the 14 pairs that showed sufficient association-dissociation events the averaged ΔG^{dimer} are -0.61 ± 0.61, -0.92 ± 0.77, and -1.80 ± 2.01 k_BT for the Martini, OPLS and Gromos FFs, respectively. The RMSDs are 1.36 k_BT for all pairs and 2.57, 0.29, 1.70, and 0.82 k_BT for the polar, apolar, aromatic, and mixed SCAs, respectively. Over these 14 pairs the correlation coefficient is $r = 0.54$ (Figure 3.2). The PMFs calculated for the charged SCA pairs indicate much weaker Coulombic interactions in the Martini FF than in the atomistic ones (table 3.3). The examples of the Lys-Glu and Lys-Lys pairs are shown in Figure 3.3. The Lys-Lys pair exhibits a slightly attractive interaction with the Martini FF, while the atomistic FFs are purely repulsive. The behavior of the charged SCA interactions in the Martini FF is improved with the polarizable version of the water model, also shown in Figure 3.3. The depth of the well of the Lys-Glu pair PMF decreases from -11 to -25 k_BT, albeit still much smaller than the $-120/-140$ k_BT value when using the atomistic models. In the case of the Lys-Lys pair, the attractive range observed with the standard Martini water disappears with the polarizable model.

The RMSDs of ΔG^{dimer} between the Martini FF and $\overline{AT}$ are 0.54, 0.40 and 1.36 k_BT

for water, n-octanol and decane, respectively and 0.86 k_BT averaged over the three solvents. Comparing the Martini FF used with polarizable water to $\overline{AT}$ in water gives a RMSD of 0.65 k_BT. Thus, the difference between Martini and Gromos/OPLS is of the order of k_BT, comparable to the difference between the two atomistic FFs. Main outliers to this last observation are the interaction between Phe side chains in water, which is more attractive in the Martini FF, and interactions involving charged and polar SCAs in the apolar media.

In the case of polar SCs this discrepancy is due to the lack of an explicit representation of the polarity. The introduction of partial charges, in line with the polarizable water model [27], may improve this behavior.

For the charged residues, the large difference observed between the atomistic and CG apporaches results from two factors. First, in Martini a high implicit dielectric screening ($\epsilon_r = 15$) is used, while in the atomistic models no dielectric screening is used ($\epsilon_r = 1$). Reducing the dielectric screening used in Martini simulations does increase the interaction strength between the charged SCs (See Figure 3.3): i) in the polarizable Martini model ($\epsilon_r = 2.5$) the interaction strength is increased by a factor of 2; ii) when no dielectric screening is used ($\epsilon_r = 1.0$), the interaction strength increases further by another factor of two (not shown). Second, the size of the CG beads (defined by the van der Waals potential), which limits the approach of the charged beads to $\sim$ 0.5 nm, whereas the charges in the atomistic models can come much closer. Because the coulombic interactions fall off as $1/r$, the total interaction strength between the charged atomistic side chains is also affected and is much more attractive. This effect becomes evident when the PMFs are plotted against the minimum distance between the charges instead of the COM distance of the SCAs (dashed lines in Figure 3.3). To increase the binding strength in the Martini force field, we designed an alternative model for charged side chains in which the electrostatic and van der Waals interactions are carried by two different particles connected by a constrained bond of length 0.11 nm (drude-like). The van der Waals center and the charge both have a mass of 36 amu. Using this setup, the charges may approach each other much closer and the interaction increases by approximately a factor of 5 for the lysine/glutamate pair as compared to normal Martini (cf. Figure 3.3). Combining this appraoch with an absence of implicit dielectric screening ($\epsilon_r = 1$) increases the interaction strength by a factor 18 compared to the normal Martini topology. The latter setup yields SCA pair interactions stronger than when using the Gromos and OPLS force fields. This is due to the spread of the charge over multiple atoms in the atomistic force field. In chapter 7 a new model for charged side-chains in the Martini force field is developed that fixes this problem.

Structural analysis

To further characterize the binding of SCAs in different force fields and solvents, we analyzed the structural binding modes of the SCA pairs. To do so we build 2D probability distributions correlating different reaction coordinates. For the aromatic pairs, both the angle between the ring normals (*'angle'*) and the dihedral angle between two axes within the ring planes (*'dihedral'*) are correlated with the COM distance. For the non-aromatic pairs the angle between the SCAs long axes was correlated with the COM distance. In the Martini FF only the aromatic pairs have sufficient beads to define the angle and dihedral. Therefore for the other pairs only OPLS and Gromos could be compared. The analysis of the angles was combined with a careful visual inspection of the trajectories. The results are shown in Figure 3.4 and Figure 3.5 for the case of the Phe-Phe pair; plots for the other pairs can be found in the supporting information of [35], figure 1 to 22.

Phenyl rings in general and phenylalanines residues in particular are known to bind in mainly three different modes: T-stacking, straight-stacking and shifted-stacking. T-stacking and shifted-stacking are most abundant in proteins [36]. For water and octanol Figure 3.4 and

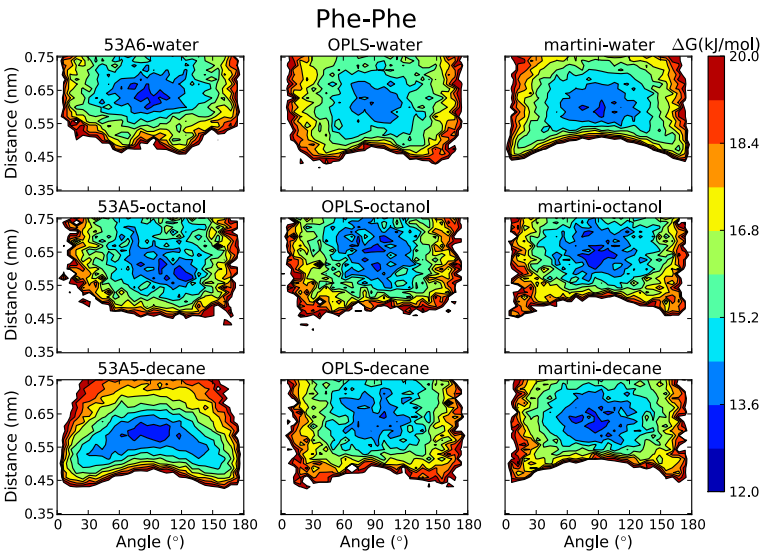

Figure 3.4: Two dimensional histograms showing the free energy profiles of the interactions between two phenylalanine side chains in different solvents and different force fields. The angle between the normal of the ring plains is plotted against the center of mass distance. The normal of the ring plains is defined by C_γ, $C_{\epsilon 1}$ and $C_{\epsilon 2}$ for Gromos and OPLS and all three ring beads for Martini. Colors indicate free energy levels in kJ/mol.

Figure 3.5 show some typical characteristics, that are seen in all of the aromatic pairs. First, the lower free energy around a 90° angle is due to entropic reasons: more conformations are possible for a perpendicular orientation as compared to parallel orientations. Second, the position of the free energy minimum in the distance dimension corresponds to the first minimum of the radial distribution function. And third the lower free energy at low distance and angles around 0 and 180° is due to the flat stacking of the rings that is preferred at small distances. Except for those general characteristics, in water and octanol no specific binding mode is found in either of the force fields. In decane, Gromos shows a strong preference for the shifted stacking, where the methyl groups are located straight above the ring of the other side chain. In Figure 3.5 this orientation corresponds to the the lower free energy at 0°. A cluster analysis showed that this conformation accounted for $\sim 40\%$ of the bound state. For OPLS in decane no strongly specific orientation was observed. For Martini, orientations in all three solvents are governed by optimal packing of the LJ spheres, leading to a preferred binding in the ridges between the coarse grain beads.

For Gromos and OPLS, two tyrosine rings in water are often in a stacked conformation, where there is no preferred orientation for the ring substituents (Supplementary Information of de Jong *et al.*[35], figures 1 and 2). In decane, the interaction is mainly guided by the hydrogen bond between the alcohol groups for OPLS and Gromos. Gromos has a second conformation at closer distance in which the alcohol group points to the middle of the ring of the second molecule. In octanol, this latter conformation is observed for both Gromos and OPLS, however less pronounced. Using Martini, in decane and octanol the rings are

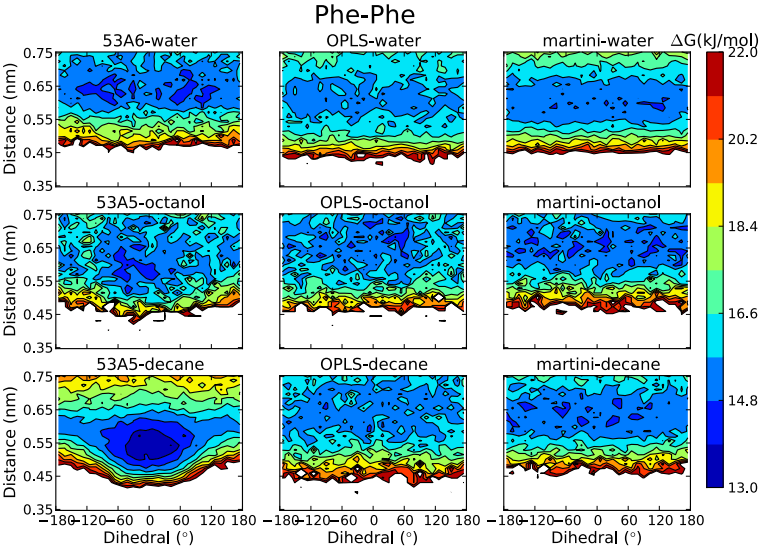

Figure 3.5: Two dimensional histograms showing the free energy profiles of the interactions between two phenylalanine side chains in different solvents and different force fields. The dihedral angle between the vectors in the plane of the ring is plotted against distance between the centers of mass of the SCAs. The vectors are defined by C_γ and C_ζ for Gromos and OPLS and the first two ring beads for Martini. Colors indicate free energy levels in kJ/mol.

mainly found in a stacked conformation. In water, orientations are again governed by sphere packing.

The histidine-phenylalanine pair has no specific binding mode in water for any of the force fields ([35], SI, figures 3 and 4). In decane, OPLS and Gromos favor a conformation where the ϵ_2 hydrogen is pointing to the middle of phenyl ring, leading to a 90° angle between the rings. This orientation very strongly preferred for Gromos compared to OPLS. The difference is less pronounced in octanol. For Martini, orientations in all three solvents are governed by the packing of the CG beads.

The tryptophan-tryptophan pair preferably binds with the amine hydrogen of one molecule pointing to the six ring of the second molecule in both Gromos and OPLS FFs ([35], SI, figures 5 and 6). This orientation is strongly present in decane and, for OPLS, in octanol. In water no specific binding mode is observed. In Martini there is a specific binding mode where the rings stack parallel, in that way maximizing the total contact surface.

Using Gromos and OPLS, the Trp-Tyr pair has two main conformations: either the Tyr alcohol group is pointing towards the middle of the Trp 5-ring or the Trp amine group is pointing towards the middle of Tyr 6-ring ([35], SI, figures 7 and 8). In OPLS the former conformation is more abundant (equivalent to the 90° dihedral angle, see supplementary information), while in Gromos the latter is (equivalent to the ±180° dihedral angle). In water there is no preferred orientation, except for small distances (< 0.5nm) where the rings are mainly found in a stacked orientation. This orientation is preferred for Martini in all three solvents.

For the four polar pairs (Ser-Ser, Asn-Asn, Gln-Gln and Gln-Asn) OPLS and Gromos behave very similar in all solvents ([35], SI, figures 9, 10, 11, and 12). Interactions in octanol and decane are governed by hydrogen bonds . In decane this effect is the most pronounced, leading to only a small number of structural arrangements being sampled and a strong binding energy (table 3.2 and table 3.3). For the Ser-Ser pair the hydrogen bond is between the alcohol groups, which leads to an average angle between the molecules around 90°. For the other polar pairs a molecule angle around 180 degrees, corresponding to the hydrogen bond between the oxygen and either of the amine hydrogens, is most prevalent. In octanol the partial charges are more strongly screened and the interactions are consequently weaker and less specific, although the hydrogen bond is still most prevalent. The pairs involving Asn or Gln also sample the hydrogen bond between both amine groups. In water the screening of charges is even stronger and no specific binding is observed.

All three pairs of apolar molecules as well as the mixed pairs (Leu-Asn, Leu-Gln, Leu-Lys and Met-Ser) show no specific binding mode in any of the solvents when using a high resolution force field ([35], SI, figures 13 to 19). The angles wobble around 90 degrees due to entropic reasons. Valine-valine has a slightly smaller minimum distance when in contact.

The charged-polar mixed pair Asn-Lys has a strongly preferred binding mode where one of the (charged) amine protons forms a hydrogen bond with the Asn oxygen ([35], SI, figure 20). The two minima found in the plot result from the presence of two conformations of the lysine carbon tail (either stretched or U-shaped).

The charged Lys-Glu pair in decane and octanol always bind via their charged groups. A large range of angles and associated change of COM-COM distance is sampled and results from the change of conformation of the carbon tails of both molecules ([35], SI, figure 21). In water, using the OPLS force field the side chains show the same mode of binding, although less pronounced. In Gromos, potentially due to the down scaling of the charges in the 53A6 compared to the 53A5 version there is no specific binding mode in water.

For the charged Arg-Asp pair, there are strong differences between force fields and solvents ([35], SI, figure 22). In decane, using OPLS the charged Asp oxygens are strongly coordinated by the η_1 and η_2 hydrogens of Arg, while in octanol they are coordinated by η_2 and ϵ hydrogens. In water both conformations are found. In decane, using 53A5, Asp binds in between the η_2 and ϵ carbons of Arg, with the planes of the molecules being perpendicular. In octanol, Asp binds Arg at both η_1/η_2 and η_2/ϵ positions, with the molecule planes primarily in a perpendicular orientation. In water, using 53A6, no specific binding mode is observed.

Comparison to knowledge-based potentials

Unfortunately there is no experimental data that could be used to directly compare the calculated ΔG^{dimer} against. There is, however, a long history of extracting side chain-side chain contact energies for amino acid pairs from their occurrence in protein structures. [37][38][39] These contact energies have been used to derive energy functions for CG FFs, the so called knowledge-based potentials (KBP), [40] as well as to test existing FFs. [41] Although the environment of an amino acid buried in a protein is different than free in solution (e.g. different orientational freedom and screening of electrostatics) and the methods used to extract such KBPs are not flawless, [20] these contact energies provide us with a useful independent benchmark. We use the set of contact energies obtained by Miyazawa and Jernigan [22] (MJ). It was extracted from a set of 20000 SC contacts from which a delicate balance of terms was carefully derived and corrected for potential biases.

The comparison of the dimerization free energies obtained from our simulations to the contact energies obtained by MJ is shown in Figure 3.6. The correlation coefficients, r, between the data from MJ and the OPLS, Gromos, regular and polarizable Martini FFs and over 20 SCA pairs are $0.48, 0.67, 0.72$, and 0.87, respectively. Note that in MJ the backbone

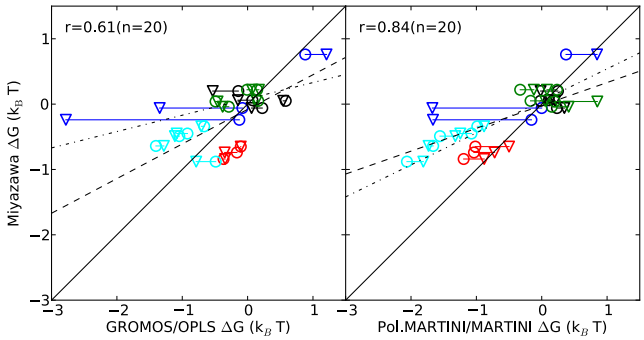

Figure 3.6: Comparison of calculated SC dimerization free energies to knowledge-based potentials (KBP) derived by Miyazawa and Jernigan [22]. In the left panel, KPB is compared to the atomistic FFs Gromos (circles (○)) and OPLS (triangles (▽)). In the right panel, KPB is compared to the CG FFs Martini (circles (○)) and polarizable Martini (triangles (▽)). Lines connect equal SC pairs. Blue are charged, black are polar, red are apolar, green are mixed, and cyan are aromatic SC pairs. The black line represents a regression of 1. Dashed and dashed-dotted lines represent regression fits for Gromos and OPLS (left panel) and Martini and polarizable Martini (right panel). r is the Pearson correlation coefficient, calculated over n data points of the average of Gromos and OPLS (left panel) and of Martini and polarizable Martini (right panel).

was not considered. The slope, a, of a linear regression fit of the MJ data versus the Gromos, OPLS, regular and polarizable Martini FFs data was $0.53, 0.25, 0.35$ and 0.49, respectively. The correlation and regression coefficients are in line with those found by Betancourt *et al.* [20] in their comparison of the Gromos 43A1 FF to the data from MJ ($r = 0.55$ and $a = 0.30$, fitted including all possible amino acid pairs). It is also notable that the correlation with an earlier set of contact energies derived by MJ [40] is much lower (data not shown), suggesting that the additional corrections were valuable.

Overall the two Martini FFs compare better to the MJ SC-SC contact energies than the atomistic ones. Notably the OPLS and the polarizable Martini FFs have stronger unlike charge-charge interactions in comparison to the data from MJ. The correlation coefficients become 0.70 and 0.88, respectively, when these interactions are omitted from the analysis.

In all four FFs the interactions between aromatic SCs are stronger than those obtained by MJ, in line with the observation that values of interaction energies below $-1\ k_BT$ (the case for the rings) estimated by the method used by MJ might be underestimated (too weak). [20] The deviation of these side chains primarily leads to regression coefficients smaller than one for all FFs, as was also found by Betancourt *et al.* It is interesting to note that no significant configurational preference for ring-ring interactions and persistent over FFs was found in our simulations. This is in great contrast with first principle (ab initio) calculations and protein structure analysis. This suggests that ring-ring interactions in an aqueous solvent might be less specific than buried in a protein and/or in a non-polar environment. It was also notable that in most cases simulated the SC-SC interactions experienced an increase of conformational sampling with the increase of the polarity of the solvent (from decane to water).

Considering that the KBP was obtained from structures of soluble proteins, [22] the choice to compare to ΔG^{dimer} obtained from simulations in water seems reasonable. However, the

Table 3.4: Complex lifetimes of amino acid side chain complexes in ps. Pairs have been grouped according to chemical properties: charged, apolar, polar, aromatic, mixed, and backbone. Statistical errors were obtained as the fit error in the least squares fitting procedure.

Pair	Pol.Mart.	Martini	Gromos	OPLS
Charged				
Lys & Glu	24.3 ± 0.7	68 ± 2	12.9 ± 0.3	$(1.5 \pm 0.2) \times 10^2$
Lys & Lys	12.8 ± 0.3	9.9 ± 0.8	20 ± 3	18 ± 2
Arg & Asp	38 ± 2	68 ± 1	12.2 ± 0.2	$(7 \pm 5) \times 10^2$
Polar				
Ser & Ser	20.9 ± 0.6	16 ± 1	12.8 ± 0.5	9.6 ± 0.2
Gln & Asn	19.5 ± 0.3	24 ± 2	16 ± 4	16 ± 3
Asn & Asn	22 ± 1	16.1 ± 0.6	16.0 ± 0.7	20 ± 1
Gln & Gln	18.7 ± 0.6	25 ± 2	23.1 ± 0.9	28 ± 3
Mixed				
Leu & Asn	14.3 ± 0.5	8.4 ± 0.4	17 ± 1	16 ± 2
Leu & Gln	14.2 ± 0.2	13.4 ± 0.7	23.6 ± 0.7	31 ± 2
Leu & Lys	18.5 ± 0.4	16.1 ± 0.8	18.3 ± 0.6	21.6 ± 0.6
Asn & Lys	17.0 ± 0.3	20 ± 1	20.9 ± 0.7	17.4 ± 0.4
Met & Ser	17.1 ± 0.2	13.3 ± 0.3	15.3 ± 0.3	15.8 ± 0.7
Apolar				
Leu & Leu	25.7 ± 0.1	20.3 ± 0.1	27.5 ± 0.8	27.3 ± 0.8
Val & Val	29.5 ± 0.5	17.6 ± 0.5	21.2 ± 0.4	19.7 ± 0.8
Leu & Val	28.7 ± 0.5	16.5 ± 0.5	24 ± 1	27.8 ± 0.8
Aromatic				
Trp & Trp	124 ± 7	110 ± 9	65 ± 2	58 ± 2
Tyr & Tyr	71 ± 4	67 ± 3	56 ± 4	52 ± 2
Phe & Phe	102 ± 2	84 ± 3	36 ± 2	40 ± 2
His & Phe	67 ± 4	56 ± 4	34 ± 1	33.0 ± 0.8
Trp & Tyr	119 ± 8	64 ± 3	53 ± 2	41.0 ± 0.6
Backbone				
BB & BB	22 ± 1	16.1 ± 0.6	24.6 ± 0.5	16.9 ± 0.8

dielectric environment inside a protein is not the same as in aqueous solution. We therefore also calculated the correlation coefficients between the KBP and the ΔG^{dimer}calculated in n-octanol and decane for the different FFs. The correlation coefficients were systematically lower than those obtained in water and always below 0.2 (data not shown).

Kinetics of side chain association

To assess the variations between the FFs with respect to the association kinetics of the various SCAs, the lifetimes of the SCA bound complex, τ_L, were computed using Eqs. 3.1 and 3.2. Only the aqueous solvent case was considered, for which the results are listed in table 3.4. The data show that, for most SCA pairs, the lifetime of the complex is of the order of 10-50 ps, irrespective of the nature of the side chains. Since the lifetimes are expected to be proportional to $\exp(-\Delta G^{dimer})$, the differences between the FFs are largely reflecting differences in dimerization free energy. To appreciate if there are additional effects, we calculated a residual lifetime defined as $\tau^* = \tau_L / \exp(-\Delta G^{dimer})$. Since the lifetimes

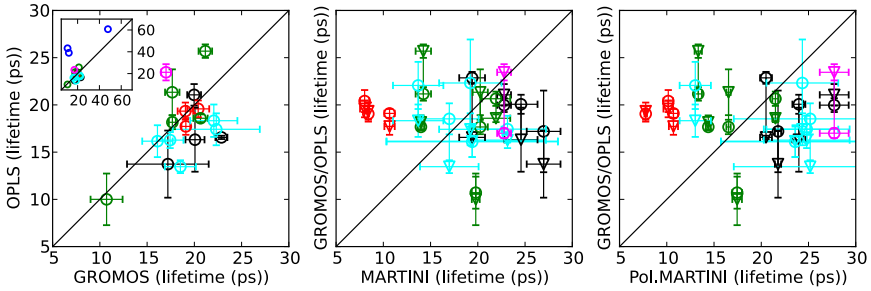

Figure 3.7: Comparison of residual lifetimes of amino acid SCAs in water between different force fields. In the first pane, circles indicate a SCA pair. In the second and third pane circles compare Gromos to Martini and triangles compare OPLS to Martini. Black represents polar SC pairs, red represents apolar SC pairs, green represent mixed pairs, cyan represents aromatic SC pairs and magenta represents the backbone-backbone pair. The inset in the first pane shows the charged pairs (blue circles). Charged pairs have been omitted for the second and third plot. Errors in both dimensions are shown as error bars. The black line represents a regression of one.

are inversely proportional to the dissociation constant, $\tau_L \propto k_{off}^{-1} \propto \exp(-\Delta G^{dimer}) k_{on}^{-1}$, the residual lifetimes will be inversely proportional to k_{on}. Figure 3.7 shows the correlation of these residual lifetimes between the FFs.

The comparison between the residual lifetimes obtained using the OPLS and Gromos FFs shows a clear correlation (Figure 3.7). Within the error bounds, for most pairs the residual lifetimes are the same. The average residual lifetimes for all but the charged pairs are 19 ± 3ps and 18 ± 4ps for Gromos and OPLS respectively. Thus, for the OPLS and Gromos FFs the variations in ΔG^{dimer} are mainly affecting k_{off}. Given the small variations in the residual lifetimes, both within and between force fields k_{on} is the same for different residues in both force fields.

Comparing residual lifetimes obtained using OPLS and Gromos with those obtained using Martini and polarizable Martini show a larger residual spread in the values obtained with the CG FFs (Figure 3.7). There is a strong influence of the type of amino acid: the apolar pairs have low residual lifetimes (9 ± 1ps for Martini and 10 ± 1ps for polarizable Martini), the polar pairs have the highest residual lifetimes (23 ± 3ps for Martini and 23 ± 3ps for polarizable Martini), and the mixed and aromatic pairs are intermediate (18 ± 3ps and 18 ± 3ps for Martini and 17 ± 3ps and 22 ± 5ps for polarizable Martini).

The residual spread in the data points indicates that there is a difference in k_{on}. Assuming that k_{on} will be mainly influenced by diffusion of the molecules in the solvent, our results can be explained by the faster diffusion of apolar residues with respect to polar ones. (Compare for example the diffusion coefficient of $1.4 \times 10^{-5}\text{cm}^2\text{s}^{-1}$ for glutamine to $1.9 \times 10^{-5}\text{cm}^2\text{s}^{-1}$ for leucine, data not shown). Why diffusion of apolar beads in Martini is somewhat faster than polar beads is not entirely clear, however.

3.4 Conclusions and Outlook

To assess the performance of both atomistic and coarse grain force fields, we have determined the dimerization free energies, ΔG^{dimer}, of 21 amino acid side chain analogue (SCA) pairs in solvents with different degrees of hydrophobicity (water, *n*-octanol and decane). Overall good correlations were found between the all-atom OPLS and united-atom Gromos FFs, and notably also with the Martini FF which uses a lower resolution. The charged SCAs proved to be the most variable over the FFs, especially in the low dielectric solvents. This was the case not only when comparing the CG Martini FF to the atomistic FFs, but also when comparing OPLS and Gromos. The use of the polarizable Martini water model improved the agreement of the charge-charge interactions between the Martini FF with the atomistic FFs, while not affecting the other interactions. Concerning the difference between the atomistic FFs, the OPLS FF has stronger Coulomb interactions than the Gromos FF, leading to larger dimerization free energies in apolar solvents.

The high correlation and deviations found in the FFs contribute to define the areas where one may be confident in the force fields and the points that need be taken with care and improved. For instance, the structural analysis of the binding modes of the respective SCA pairs points to significant differences in the packing of, especially, aromatic residues between the force fields. One would expect that this will also affect the way aromatic residues pack inside a protein. Furthermore, the differences in interaction energies between charged and polar residues in apolar solvents is a serious point of worry. The choice of force field will have consequences for processes such as membrane poration by charged aggregates and peptides (e.g dendrimers and antimicrobial peptides), the cooperativity of charged amino acids penetrating in membranes, binding of ligands in hydrophobic pockets inside proteins and membrane protein complexation. In order for a force field to better cover those kind of phenomena, membrane proteins should be used in addition to soluble proteins for the validation of the force field.

On the basis of the deviations of the MARTINI FF with respect to the atomistic FFs, a few recommendations could be made for future improvements. The phenylalanine residue pair seems to bind too strongly in the MARTINI FF. The Phe SC should be made slightly more attractive toward water than it is currently. A similar conclusion was recently reached by Singh and Tieleman[42] comparing the binding of WimleyWhite pentapeptides to lipid bilayers. The oppositely charged and polar SCs were found to bind too weakly in the apolar solvents. For the charged SCs, this is slightly improved by using the polarizable MARTINI model, which reduces the screening of Coulombic interactions, and improved further by adding an extra, charge carrying bead to the model. In the case of the polar SCs, it could be advantageous to introduce polarizable particles in the line of the polarizable water model. In chapter 7 these solutions are further developed into a new set of parameters for the Martini forcefield.

The side chain contact energy pairs obtained by Miyazawa and Jernigan from the occurrence of the SC pairs in a large set of protein structures correlated remarkably well with the ΔG^{dimer}in water obtained with the various FFs. This is even more striking considering the completely different approaches used to derive the potentials. It is worth mentioning that the comparison with contact energies derived earlier by a different method [40] did not correlate as well as with the current [22] set. The larger deviations observed for the aromatic SCs suggest that they may involve more effects difficult to account for, or that the FFs are not parameterized to the same level of accuracy as the other SCs.

The current work demonstrates the capability of modern large scale computational power to systematically assess the quality of side chain interactions in molecular dynamics force fields and identifies some possible improvements to be made with respect to these interactions.

3.5 Bibliography

[1] J.A McCammon, B.R Gelin, and M Karplus. Dynamics of folded proteins. *Nature*, 267:585–590, 1977.

[2] M Vendruscolo and C.M Dobson. Protein dynamics: Moore's law in molecular biology. *Curr. Biol.*, 21:R68–R70, 2011.

[3] C Clementi. Coarse-grained models of protein folding: toy models or predictive tools? *Curr. Opin. Struc. Biol.*, 18:10–15, 2008.

[4] D.E Shaw, P Maragakis, K Lindorff-Larsen, S Piana, R.O Dror, M.P Eastwood, J.A Bank, J.M Jumper, J.K Salmon, and Y Shan. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330:341, 2010.

[5] G.R Bowman, V.A Voelz, and V.S Pande. Taming the complexity of protein folding. *Curr. Opin. in Struc. Biol.*, 21:4–11, 2010.

[6] P.L Freddolino, C.B Harrison, Y Liu, and K Schulten. Challenges in protein-folding simulations. *Nat. Phys.*, 6:751–758, 2010.

[7] H.A Scheraga, M Khalili, and A Liwo. Protein-folding dynamics: overview of molecular simulation techniques. *Annu. Rev. Phys. Chem.*, 58:57–83, 2007.

[8] K.A Dill, S.B Ozkan, M.S Shell, and T.R Weikl. The protein folding problem. *Annu. Rev. Biophys.*, 37:289–316, 2008.

[9] T Cellmer, D Bratko, J M Prausnitz, and H W Blanch. Protein aggregation in silico. *Trends Biotechnol.*, 25:254–261, 2007.

[10] X Periole, T Huber, S.J Marrink, and T.P Sakmar. G protein-coupled receptors self-assemble in dynamics simulations of model bilayers. *J. Am. Chem. Soc.*, 129:10126–10132, 2007.

[11] A.D MacKerell, M Feig, and C.L Brooks. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.*, 25:1400–1415, 2004.

[12] V Hornak, R Abel, A Okur, B Strockbine, A Roitberg, and C Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Struct., Funct., Bioinf.*, 65:712–725, 2006.

[13] N Schmid, A.P Eichenberger, A Choutko, S Riniker, M Winger, A.E Mark, and W.F Gunsteren. Definition and testing of the gromos force-field versions 54a7 and 54b7. *Eur. Biophys. J.*, 40:843–856, 2011.

[14] W.L Jorgensen, D.S Maxwell, and J Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118:11225–11236, 1996.

[15] G.A Kaminski, R.A Friesner, J Tirado-Rives, and W.L Jorgensen. Evaluation and reparametrization of the opls-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B*, 105:6474–6487, 2001.

[16] C Oostenbrink, A Villa, A.E Mark, and W.F Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos forcefield parameter sets 53a5 and 53a6. *J. Comput. Chem.*, 25:1656–1676, 2004.

[17] S.J Marrink, H.J Risselada, S Yefimov, D.P Tieleman, and A.H de Vries. The martini force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 111:7812–7824, 2007.

[18] L Monticelli, S.K Kandasamy, X Periole, R.G Larson, D.P Tieleman, and S.J Marrink. The martini coarse-grained force field: extension to proteins. *J. Chem. Theory Comput.*, 4:819–834, 2008.

[19] H Yang and A Elcock. Association lifetimes of hydrophobic amino acid pairs measured directly from molecular dynamics simulations. *J. Am. Chem. Soc.*, 125:13968–13969, 2003.

[20] M Betancourt. Another look at the conditions for the extraction of protein knowledge-based potentials. *Proteins*, 76:72–85, 2009.

[21] N Ben-Tal, D Sitkoff, I.A Topol, A.S Yang, S.K Burt, and B Honig. Free energy of amide hydrogen bond formation in vacuum, in water, and in liquid alkane solution. *J. Phys. Chem. B*, 101:450–457, 1997.

[22] S Miyazawa and R.L Jernigan. Selfconsistent estimation of interresidue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins*, 34:49–68, 1999.

[23] B Hess, C Kutzner, D van der Spoel, and E Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4:435–447, 2008.

[24] H.J.C Berendsen, J.P.M Postma, W.F Van Gunsteren, and J Hermans. *Interaction models for water in relation to protein hydration*. Reidel, Dordrecht, The Netherlands, 1981.

[25] N.M Garrido, A.J Queimada, M Jorge, E.A Macedo, and I.G Economou. 1-octanol/water partition coefficients of n-alkanes from molecular simulations of absolute solvation free energies. *J. Chem. Theory Comput.*, 5:2436–2446, 2009.

[26] I.G Tironi, R Sperb, P.E Smith, and W.F van Gunsteren. A generalized reaction field method for molecular dynamics simulations. *J. Chem. Phys.*, 102:5451–5459, 1995.

[27] S.O Yesylevskyy, L.V Schäfer, D Sengupta, and S.J Marrink. Polarizable water model for the coarse-grained martini force field. *PLoS Comput. Biol.*, 6:1–17, 2010.

[28] H.J.C Berendsen, J.P.M Postma, W.F Van Gunsteren, A DiNola, and J.R Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[29] B Hess, H Bekker, H.J.C Berendsen, and J.G.E.M Fraaije. Lincs: a linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18:1463–1472, 1997.

[30] M Sandberg, L Eriksson, J Jonsson, M Sjöström, and S Wold. New chemical descriptors relevant for the design of biologically active peptides. a multivariate characterization of 87 amino acids. *J. Med. Chem.*, 41:2481–2491, 1998.

[31] B Hess, C Holm, and N Van der Vegt. Osmotic coefficients of atomistic nacl (aq) force fields. *J. Chem. Phys.*, 124:164509, 2006.

[32] W Humphrey, A Dalke, and K Schulten. Vmd: visual molecular dynamics. *J. Mol. Graphics*, 14:33–38, 1996.

[33] A.E García and L Stiller. Computation of the mean residence time of water in the hydration shells of biomolecules. *J. Comput. Chem.*, 14:1396–1406, 1993.

[34] J.L MacCallum and D.P Tieleman. Structures of neat and hydrated 1-octanol from computer simulations. *J. Am. Chem. Soc.*, 124:15085–15093, 2002.

[35] D.H de Jong, X Periole, and S.J Marrink. Dimerization of amino acid side chains: lessons from the comparison of different forcefields. *J. Chem. Theory Comput.*, 8:1003–1014, 2012.

[36] R Castellano and F Diederich. Interactions with aromatic rings in chemical and biological recognition. *Angew. Chem. Int. Edit.*, 42:1210–1250, 2003.

[37] S Tanaka and H.A Scheraga. Medium-and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, 9:945–950, 1976.

[38] S Miyazawa and R.L Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18:534–552, 1985.

[39] M Sippl. Calculation of conformational ensembles from potentials of mena force*:: An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, 213:859–883, 1990.

[40] S Miyazawa and R.L Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, 256:623–644, 1996.

[41] M.R Betancourt and S.J Omovie. Pairwise energies for polypeptide coarse-grained models derived from atomic force fields. *J. Chem. Phys.*, 130:195103, 2009.

[42] G Singh and D.P Tieleman. Using the wimleyi-white hydrophobicity scale as a direct quantitative test of force fields: The martini coarse-grained model. *J. Chem. Theory Comput.*, 7:2316–2324, 2011.

4

Lipid packing drives the segregation of transmembrane helices into disordered lipid domains in model membranes

This chapter is based upon the manuscript:
Lipid packing drives the segregation of transmembrane helices into disordered lipid domains in model membranes by Lars V. Schäfer, Djurre H. de Jong, Andrea Holt, Andrzej J. Rzepiela, Alex H. de Vries, Bert Poolman, J. Antoinette Killian, Siewert J. Marrink, P. Natl. Acad. Sci. USA, **2011**, *108*(4), 1343-1348

Abstract

Cell membranes are comprised of multicomponent lipid and protein mixtures that exhibit a complex partitioning behavior. Regions of structural and compositional heterogeneity play a major role in the sorting and self-assembly of proteins, and their clustering into higher-order oligomers. Here, we use computer simulations and optical microscopy to study the sorting of transmembrane helices into the liquid-disordered domains of phase-separated model membranes, irrespective of peptide-lipid hydrophobic mismatch. Free energy calculations show that the enthalpic contribution due to the packing of the lipids drives the lateral sorting of the helices. Hydrophobic mismatch regulates the clustering into either small dynamic or large static aggregates. These results reveal important molecular driving forces for the lateral organization and self-assembly of transmembrane helices in heterogeneous model membranes, with implications for the formation of functional protein complexes in real cells.

4.1 Introduction

The heterogeneity of biological membranes plays an important role in cellular function [1] [2]. Despite experimental progress in recent years [3] , the characterization of lateral organization in biological membranes, however, remains challenging due to the lack of tools to study fluctuating nanoscale assemblies in living cells [4, 5, 6]. Model membranes [7, 8, 9, 10] as well as isolated plasma membranes [11, 12] are more frequently studied, because large-scale phase separation can be observed in these systems. In particular, at cholesterol concentrations reminiscent of biological membranes (10-30 mol% cholesterol), ternary mixtures of saturated and unsaturated lipids can segregate into coexisting lipid domains of different fluidity, a liquid-ordered (Lo) and a liquid-disordered (Ld) phase. Probing the structural and dynamical properties of these fluid domains has received a lot of attention, as it is presumably linked to the formation of lipid nanodomains ("rafts") in real cells [13, 14].

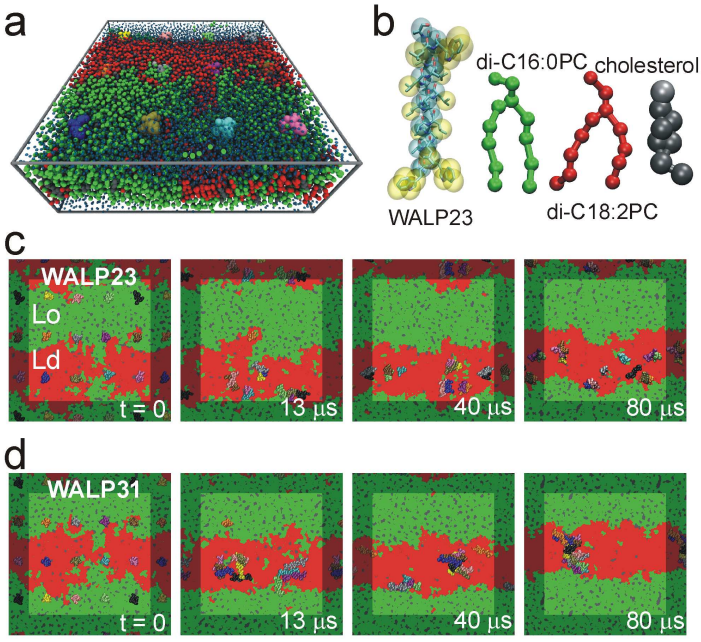

Figure 4.1: Simulation of the sorting and clustering of WALP peptides in a model bilayer with coexisting fluid domains. (A) WALP peptides (colored spheres) embedded in a ternary mixture of di-C16:0PC (green), di-C18:2PC (red), and cholesterol (gray), solvated by water (blue). (B) Coarse grain representation of WALP23 (cyan and yellow), di-C16:0PC (green), di-C18:2PC (red), and cholesterol (gray), shown as spheres and sticks. The atomistic structure of the peptide is also shown (sticks). (C and D) Sorting and clustering of WALP23 (C) and WALP31 peptides (D) in the disordered lipid domain.

The structure and function of membrane proteins is intimately connected with their lipid environment [15, 16]. Because of the heterogeneity of the cell membrane, proteins partition between different lipid domains, are recruited to specific locations, and form functional complexes [17, 18, 19]. This lateral organization is very important for various cellular processes, such as membrane fusion [20, 21], protein trafficking [22], and signal transduction [23, 24, 25]. Although lipids and integral membrane proteins are well studied by themselves, the molecular properties that determine the specific interactions between them remain poorly understood. Interactions important for protein sorting and self-assembly are (specific) protein-protein and protein-lipid forces, and indirect lipid-mediated forces. The latter category includes, for instance, forces due to entropic confinement of lipid chains and forces arising from the mismatch between the hydrophobic parts of the protein and the bilayer. Recent computational studies highlight the importance of lipid-mediated interactions for membrane protein clustering [26, 27, 28, 29].

Integral membrane proteins have a large structural variety, but all contain hydrophobic

regions that span the bilayer; these regions are often single α-helices or α-helical bundles. Many properties of membrane proteins are thus determined by the interactions between these transmembrane (TM) helices and the surrounding lipids. Here, we investigate the molecular interactions between TM helices, specifically WALP peptides, and heterogeneous model membranes with coexisting Lo and Ld lipid domains. Synthetic WALP peptides were designed as probes to study the behavior of TM domains in model systems, such as their mode of insertion, aggregation, and partitioning [30, 31].

WALPs form α-helical TM domains comprised of alternating alanine and leucine residues, flanked by tryptophanes to anchor the peptide at the membrane interface. Varying the length of the peptide's hydrophobic stretch allows for tailoring of desired properties, such as the hydrophobic mismatch. WALPs and other TM peptides were observed to be excluded from Lo domains in model membranes, irrespective of the hydrophobic mismatch, see, e.g., refs. [32] and [33]. It was speculated that the TM peptides disturb the tight packing of lipids and cholesterol in Lo [14]; however, the underlying thermodynamic driving forces are not well understood. Our study combines both large-scale molecular dynamics (MD) computer simulations of planar membranes and confocal fluorescence microscopy of giant unilamellar vesicles (GUVs) to investigate peptide sorting under different hydrophobic mismatch conditions. In addition, the simulations yield insight into the clustering of the peptides and the thermodynamic driving forces for lateral sorting.

4.2 Results

Simulations Show Partitioning of TM Helices into Ld Domain.

Figure 4.1 A and B shows the simulation system, consisting of 12 WALP peptides inserted into a lipid bilayer patch and solvated by explicit water. All molecules are represented at a coarse grain (CG) level, retaining near-atomic resolution. The bilayer consists of a ternary mixture of dipalmitoylphosphatidylcholine (di-C16:0PC), dilinoleoylphosphatidylcholine (di-C18:2PC), and cholesterol in a 0.42:0.28:0.30 molar ratio, about 2,000 lipids in total. The membrane is separated into two coexisting fluid domains that can spontaneously form from a random lipid distribution, as evidenced in previous simulations [34]. Characterization of these domains revealed that the Lo membrane domain is mainly composed of the fully saturated di-C16:0PC lipid and cholesterol, whereas the Ld domain is enriched in the doubly unsaturated di-C18r:2PC lipid and has a reduced cholesterol content. The Lo domain is thicker (PC headgroup-headgroup distance, $d_0 \approx 4.6$ nm) than the Ld domain ($d_0 \approx 3.8$ nm). In the current chapter, we studied the partitioning of two different peptides, namely WALP23 [AW_2L-$(AL)_8$-W_2A] and WALP31 [AW_2L-$(AL)_{12}$- W_2A] in these domains. The shorter WALP23 (peptide length $l_p \approx 3.4$ nm) is expected to fit well into the Ld domain, whereas the longer WALP31 ($l_p \approx 4.6$ nm) could span the Lo domain and thus would display a rather large positive hydrophobic mismatch in the Ld domain. As shown in Figure 4.1A, monomeric peptides were initially inserted into the membrane. CG-MD simulations using the Martini force field [35, 36] were then carried out for 80 μs for the WALP23 and WALP31 systems, respectively.

Figure 4.1 C and D shows that the peptides adopt a distinct non-homogeneous distribution during the simulation. Both the WALP23 and WALP31 peptides are expelled from the Lo domain and partition into the Ld domain, irrespective of their length. For the membrane studied here, in which the Ld domain spans about 10 nm, this sorting process is completed within 2 μs, after which most peptides are incorporated into the Ld phase. We never observed that peptides (transiently) returned to the Lo phase during the time span of our simulations, suggesting a rather large driving force for peptide sorting into the Ld phase. Attaching two

palmitoyl anchors to the N-terminus of WALP23 does not alter the partitioning, in agreement with experiments[32]. The addition of eight palmitoyl anchors, however, leads to preferential binding to the domain boundary (see section 4.5). The lateral segregation of the peptides leads to an increase of the effective peptide/lipid ratio to about 150 in the disordered domain, where dynamic self-assembly of the peptides into oligomeric aggregates is observed. The size and stability of the aggregates depends on the mismatch, as is addressed below.

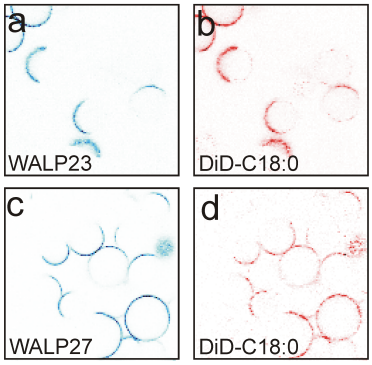

Figure 4.2: Confocal images of GUVs containing Alexa Fluor 488-labeled WALP peptides. (A and B) The vesicles were composed of di-C18:1PC/ cholesterol/egg-sphingomyelin (1:1:1) and WALP23 (blue fluorescence channel) and DiD-C18:0 (red). (C and D) The vesicles contained di-C14:1PC/cholesterol/egg-sphingomyelin (1:1:1) and WALP27 (blue) and DiD-C18:0 (red).

Confocal Microscopy Verifies That TM Helices Prefer the Ld Domain.

Next, we experimentally tested the predictions of our simulations by optical fluorescence microscopy. We incorporated Alexa Fluor 488-labeled WALP peptides into GUVs composed of cholesterol, di-C181PC, and egg-sphingomyelin (1:1:1). The fluorescent lipid dye DiD-C180 was added as a marker for the Ld phase[8, 11]. First, we incorporated labeled WALP23 into the GUVs and recorded fluorescence images (Figure 4.2A, blue channel). The confocal image in Figure 4.2B shows the fluorescence of the Ld marker DiD-C18:0 (red channel), confirming the phase separation. The colocalization of the WALP23 and DiD-C18:0 fluorescence indicates that the peptides partition into the Ld domains (containing mainly di-C18:1PC) and are expelled from the Lo domains (composed of mainly cholesterol and egg-sphingomyelin). In a next step, we repeated the experiments with a membrane in which di-C18:1PC was replaced by di-C14:1PC, thereby decreasing the hydrophobic thickness of the Ld phase by about 0.6 nm. To maximize the hydrophobic mismatch, we incorporated the WALP27 peptide, which is 0.6 nm longer than WALP23. Thus, taken together, the difference is 1.2 nm, identical to the length difference between the WALP23 and WALP31 peptides used in our MD simulations. Also here, the Lo domains match the hydrophobic length of the longer peptide much better than the thinner Ld domains. The confocal images shown in Figure 4.2C and D reveal that Alexa Fluor 488-labeled peptides and DiD-C18:0 still colocalize, demonstrating that WALP peptides prefer the Ld phase over the Lo phase irrespective of the hydrophobic mismatch. These confocal images confirm our MD simulations and agree with previous detergent extraction experiments[32].

Free Energy Calculations.

To roughly understand the underlying molecular driving forces for the observed lateral sorting of the peptides, we calculated, by means of thermodynamic integration (TI), the partitioning free energies of a single WALP peptide between the Lo and Ld phase (for details, see section 4.5). The results are summarized in Table 4.1. The free energies of transferring a peptide from the Lo to the Ld domain are $\Delta G_{Lo/Ld} = 63 \pm 8$ kJ mol^{-1} and 58±8 kJ mol^{-1} for WALP23 and WALP31, respectively. The strong preference of both peptides for the Ld phase explains why during the free simulations, peptides once incorporated into the Ld domain never returned to the Lo domain. The transfer free energies $\Delta G_{Lo/Ld}$ result from large counteracting enthalpic and entropic contributions (Table 4.1). The enthalpy favors partitioning into the Ld phase, whereas the entropy would favor partitioning into Lo. This can be rationalized by the disorder introduced when a TM peptide is accommodated in the well-ordered Lo phase. However, lipid-lipid and lipid-cholesterol packing is optimal in the Lo phase, and any foreign inclusion comes with a large enthalpic penalty that is not outweighed by this increase in entropy. By contrast, the lipids in the Ld domain are already strongly disordered, and incorporation of a TM helix thus does not yield a significant increase in entropy in this case. At the same time, due to the less densely packed lipids in Ld, the inclusion of the TM helix is enthalpically more favorable in Ld than in Lo, again outplaying the entropic contribution.

Table 4.1: Free energies, enthalpies, and entropies at 300 K of partitioning of single peptides between the Lo and Ld phases. All values are in kJ mol^{-1}. [a] $\Delta G_{\mathrm{Lo/Ld}}$ was obtained from TI simulations. [b] $\Delta H_{\mathrm{Lo/Ld}}$ was estimated from the system's potential energy, and $-T\Delta S_{\mathrm{Lo/Ld}} = \Delta G_{\mathrm{Lo/Ld}} - \Delta H_{\mathrm{Lo/Ld}}$. Standard errors are given (see section 4.5). Note, the decomposition of the free energy into enthalpy and entropy, and thus the absolute values of $\Delta H_{\mathrm{Lo/Ld}}$ and $\Delta S_{\mathrm{Lo/Ld}}$ should be interpreted at a qualitative level (see section 4.5).

	$\Delta G^{\mathrm{a}}_{\mathrm{Lo/Ld}}$	$\Delta H^{\mathrm{b}}_{\mathrm{Lo/Ld}}$	$T\Delta S^{\mathrm{b}}_{\mathrm{Lo/Ld}}$
WALP23	-63±8	-594±47	531±48
WALP31	-58±8	-276±33	531±34

The length of the peptide, and hence mismatch, does also play a role. The enthalpic gain of moving the shorter WALP23 peptide from the Lo to the Ld domain is significantly higher compared to that of the longer WALP31 peptide, which shows no hydrophobic mismatch with respect to the Lo phase (594 kJ mol^{-1} and 276 kJ mol^{-1} for WALP23 and WALP31, respectively; see Table 4.1). These results are backed up by the calculation of lipid tail order parameters from CG and atomistic simulations; see section 4.5. Taken together, for both peptides, the free energy gain upon moving from Lo to Ld results from the strong enthalpic contribution due to peptide-induced breaking up of the tight lipid packing in the Lo domain, which cannot be overcome by the entropic preference for the Lo phase.

Clustering of Peptides Depends on Mismatch.

Figure 4.3A and B quantifies the cluster formation of the WALP23 and WALP31 peptides, respectively, revealing the distinct aggregation behavior of the two peptides. The shorter WALP23 peptides display a very dynamic clustering behavior. After an initial equilibration phase of ca. 15 μs, WALP23 peptides are mostly present as monomers, dimers, and trimers, and clusters continuously reorganize by exchanging peptides, with average cluster lifetimes of a few microseconds. Calculating the average number of monomers, dimers, and trimers over the last 60 μs of the trajectory yields n_{mon}:n_{dim}:n_{trim} =2.8:2.0:1.0. Thus, the populations of monomer and dimer are similar at the effective peptide/lipid ratio of about 150 in the Ld phase. For the dimers as the most abundant aggregate, and assuming that the law of mass action can be applied, we estimate the equilibrium association constant according to [37]: $K = (n_{dim}/n_{tot})(n_{mon}n^2_{tot})$, where n_i is the average number of moles of species i in the Ld

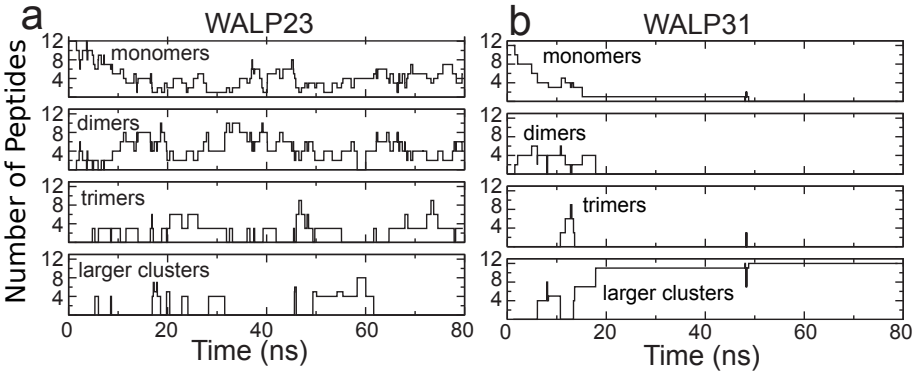

Figure 4.3: Clustering of WALP peptides during simulations. The number of WALP23 (A) or WALP31 (B) peptides that are present as a monomer, or part of a dimer, trimer, or a larger cluster are plotted over time. Two WALPs were considered bound if the distance between any two particles of the peptides was smaller than 0.7 nm.

domain. Using $n_{tot} = 540$ (number of lipids in the Ld domain), we obtain $K = 140 \pm 60$. The dimerization free energy is then $\Delta G = RT \ln K = 12 \pm 2$ kJ mol^{-1}, in agreement with experiments [38] and other simulations [39]; see section 4.5 for details. Under experimental conditions, WALP peptides are found to be monomeric at low peptide concentrations [31]. However, at peptide/lipid ratios of about 125 to 150, clustering (dimerization) was observed, a process that is additionally fostered by the presence of cholesterol [40]. Thus, the presence of weakly bound WALP23 dimers as found in our simulations is in agreement with the available experimental data.

Most of the WALP23 dimers observed in the simulation adopt a right-handed configuration, with a negative helix crossing angle (see section 4.5). Although left-handed packing of helices is more abundant in known membrane protein structures [41, 42], right-handedness is also found, for example, in the glycophorin A (GpA) dimer [43]. Similar to WALP, GpA is a single transmembrane α-helix, and its dimer has a rather small interface that does not involve hydrogen bonds or salt bridges. In the WALP23 trimers formed during the simulation, we mostly observed a linear arrangement of the peptides (i.e., the two outermost peptides within a trimer are not in direct contact with each other), again with a preference for a right-handed configuration. Left-handed helix-helix contacts are also occasionally present during the simulation; however, such aggregates are rather short-lived and rapidly dissociate again.

In contrast to WALP23, the longer WALP31 tends to form large clusters. In addition, the clustering behavior is much less dynamic as compared to WALP23, with fewer association / dissociation events observed. After 13 μs, a heptameric and a tetrameric cluster are formed, and a single WALP31 is not yet incorporated into the Ld domain (orange peptide, Figure 4.1D); this peptide joins the heptameric cluster after 50 μs (Figure 4.3B). Because the growing size of the clusters slows down their lateral diffusion, it takes until $t = 70\mu$s for the encounter of the two clusters to occur, yielding a 12-mer that is stable for the rest of the simulation time. The large WALP31 cluster is rather unstructured; that is, it has no clear tertiary structure or long-range order. As shown in Figure 4.1, the clusters occasionally approach the domain boundary, but get reincorporated into the core of the disordered domain,

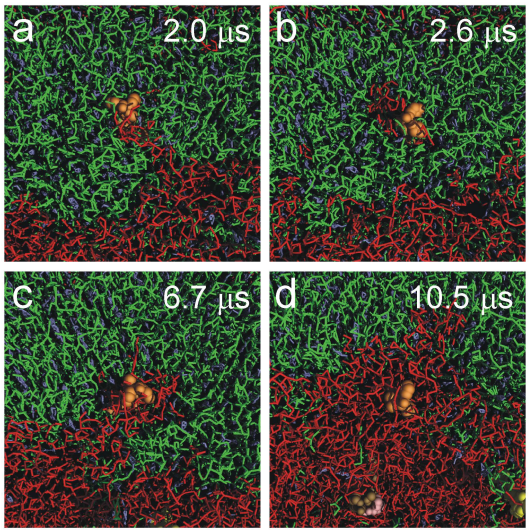

Figure 4.4: Proteins moving lipids moving proteins. A WALP23 (orange spheres), initially located in the Lo phase (green), gets incorporated into the Ld phase (red) through concerted peptide and lipid motion. (A) A peninsula of di-C18:2PC lipids bridges the peptide to the core of the Ld domain. (B) The peptide gathers an island of di-C18:2PC lipids (red) in its environment. (C and D) The di-C18:2PC island merges with the Ld domain, and the peptide subsequently diffuses away from the domain boundary.

demonstrating that the clusters, like the monomers, also prefer the Ld over the Lo domain.

To check and validate our coarse grain simulations, we transformed a snapshot (after $t = 40\mu$s) from the CG-MD simulation of the WALP23 system to its underlying atomistic representation [44] and carried out a 100-ns atomistic MD simulation (see section 4.5). During this simulation, the majority of the WALP23 clusters stay bound (a single dimer dissociates), consistent with the behavior shown in Figure 4.3. The atomistic simulations thus support the clustering observed in the coarse grain simulations.

4.3 Discussion

Our computational modeling data as well as the confocal microscopy results reveal that model transmembrane helices, such as the WALP peptides, do not like to be embedded in the liquid-ordered phase. Rather, they are observed to spontaneously partition into the liquid-disordered phase. The near-atomic resolution of our simulation approach allows us to study the sorting process in molecular detail. This is further illustrated in Figure 4.4, showing a time series of the incorporation of a particular WALP23 peptide into the Ld domain. Although most peptides are rapidly expelled from the Lo domain, this particular peptide gets temporarily trapped and enters the Ld domain only after 10 μs. The incorporation of this peptide is facilitated by the transient formation of an island of unsaturated lipids pervading the ordered domain, emphasizing that it is the collective motion of both lipids and peptides that plays a role in the lateral sorting process.

Thus far, no TM peptides or proteins that partition into Lo domains in model membranes were reported [33, 45, 46, 47, 48, 49], irrespective of the amino acid sequence. However, the underlying molecular mechanisms and thermodynamic driving forces were poorly understood. Our simulations reveal that the driving force for this observed sorting is the enthalpic cost associated with the presence of a cylindrical object (the TM helix) inside the ordered lipid phase. Thus, although we used synthetic WALP peptides as generic models, our proposed

mechanism is general and likely to be relevant for protein sorting, also *in vivo*. To test the generality, we additionally studied the partitioning of the α-helical TM domain of the syntaxin 1A protein (TM-Sx1A, residues 259-288) and of the linker for activation of T cells (TM-LAT, residues 2-32); see section 4.5. LAT is frequently used as a raft marker, but differential partitioning behavior in model membranes and in plasma membranes was reported [48, 50]. In our simulations, TM-Sx1A partitions into the Ld domain, in agreement with experimental findings [21, 45, 51, 52, 53]. TM-LAT also prefers Ld, in agreement with experiments on model membranes [48]. However, unlike TM-Sx1A, we find the larger TM-LAT clusters to be located close to the domain boundary interface.

In living cells, the differences between Lo and Ld domains are less pronounced than they can be in model membranes, due to the more complex lipid composition and high protein concentration in the plasma membrane, which may hamper the tight lipid packing in Lo [49, 50]. In particular, in the model membrane used in our simulations, the almost complete depletion of unsaturated lipids from Lo yields a very ordered domain. Furthermore, the Ld domain is rather strongly disordered due to the doubly unsaturated di-C18:2PC, explaining the observed large free energy differences (Table 1). McIntosh and coworkers [54] measured through detergent extraction experiments much smaller transfer free energies of TM peptides between lipid domains in bilayers comprised of dioleoylphosphatidylcholine, cholesterol, and sphingomyelin (1:1:1), which are closer to the critical point and might thus more closely resemble the situation in cell membranes. They report transfer free energies of only about 25 kJ mol^{-1} for the peptides P-23, P-29, and P-31 (α-helical peptides with a hydrophobic TM part and charged residues at the termini). However, such a quantitative analysis of detergent extraction experiments has to be taken with care [55], and larger free energy differences were estimated based on elastic theories [56].

For integral membrane proteins, a widely accepted assumption is that sorting is controlled by the length of the TM domains [18]. This mechanism is supported by the finding that the TM parts of proteins in the Golgi membrane (enriched in unsaturated lipids) are on average spanned by 15 amino acids, whereas TM domains of plasma membrane proteins comprise on average 20 amino acids [22]. We observed a preference for the Ld phase regardless of the length of the peptide and hence the degree of hydrophobic mismatch. One key difference between single TM helices, such as WALP peptides, and large integral membrane proteins is that the former can alleviate the energetic penalty due to positive hydrophobic mismatch through adopting a large tilt angle (according to our simulations, the average tilt angle of WALP31 in the Ld domain is indeed substantial, around 45°; see section 4.5). The tilting of large integral membrane proteins as a whole is energetically unfavorable. Rather, they may induce stronger local bilayer deformations than single α-helices or adapt their hydrophobic thickness through structural rearrangements, such as changing the tilt of individual α-helices relative to each other. The associated free energy costs influence partitioning. To probe the effect of helix tilt, an additional simulation was carried out in which we restricted tilting of WALP31. The peptide preferentially adsorbs at the Lo/Ld domain interface (see section 4.5). This result confirms that hydrophobic mismatch also plays an important role for the sorting of integral membrane proteins, and we predict that proteins that do not fit well into the Ld domain may display "linactant" [57, 58] behavior at the Lo/Ld interface. Also TM peptides with several palmitoyl anchors can accumulate at the domain interface (see section 4.5). However, specific lipid factors, such as raft gangliosides, can also be important for sequestering into the Lo phase [12]. Finally, protein-protein interactions can further alter the driving forces for LoLd partitioning [59, 60] by modulating the number of lipidprotein contacts. Clustering of transmembrane proteins and peptides may occur as a response to hydrophobic mismatch, as it reduces the number of lipids bound to the nonmatching membrane inclusion. This was confirmed by experiments on various transmembrane proteins and peptides [31]. The increased tendency to cluster observed for WALP31 supports this idea of

mismatch driven self-aggregation.

By combining molecular dynamics simulations and confocal microscopy, we show why transmembrane helices partition into disordered lipid domains in heterogeneous model membranes, irrespective of helix-lipid mismatch. Our study demonstrates that computer simulations, in combination with experiments, can not only contribute to unravel the molecular forces driving complex phase behavior in biological systems, they are also poised to become a powerful tool for predicting the localization of proteins and lipids in heterogeneous cellular environments and foster the design of bioinspired molecules with tailored properties.

4.4 Materials and Methods

Molecular Dynamics Simulations.

All simulations were carried out with the Gromacs package, version 4.0.5 [61]. In the coarse grain simulations, the Martini force field [35, 36] was used. An integration time step of 20 fs was applied, together with the standard settings for the nonbonded interactions. Constant particle number, pressure, temperature ensembles were simulated within periodic boundary conditions, with peptides, lipids, and water coupled separately to a heat bath at 295 K (τ_T = 0.3 ps). The system was semi-isotropically coupled to a pressure bath at 1 bar (τ_p = 3 ps). As the starting structure for the membrane platform, we took the phase-separated bilayer obtained after 20 μs of coarse grain simulation from ref. [34]. Subsequently, TM peptides were inserted, regularly distributed on a lateral grid. For the WALP simulations, 12 peptides were inserted (Figure 4.1A), with neighboring peptides inserted in an alternating manner (i.e., 6 peptides were inserted "head up," and 6 peptides "head down"). For TM-Sx1A and TM-LAT, 9 peptides were inserted in a parallel manner (see section 4.5 for details). The final systems comprised of TM helices, 828 di-C16:0PC, 540 di-C18:2PC, and 576 cholesterol molecules, solvated by 12,600 CG water beads (one water bead represents four real water molecules). Prior to the free MD simulations, the systems were energy-minimized (1,000 steps steepest descent), followed by 1 ns of MD with positional restraints on the peptide beads (force constant = 1,000 kJ mol^{-1} nm^{-2}). For details about the resolution transformation and atomistic simulations, see section 4.5. To convert simulation times to real times, a factor of 4 was used, obtained by comparing the lateral diffusion coefficient of a single WALP23 peptide in a bilayer to experimental data; see section 4.5 and ref. [62]. All times reported in the text are real times.

Confocal Microscopy.

Cys-WALP23/27 [Acetyl-CGWW(LA)$_{810}$LWWA-amide] were synthesized using Fmoc/tBu peptide solid-phase synthesis as described previously for related KALP peptides [63]. The peptides were labeled with Alexa Fluor 488 C5-maleimide (Molecular Probes, Invitrogen Corp.) using procedures similar to the ones for another thiol-reactive fluorophore [64]. A lipid mixture of cholesterol, sphingomyelin (extracted from egg-yolk), and 1,2-dimyristoleoyl-sn-glycero-3-phosphocholine (di-C14:1PC) or 1,2-dioleoyl-sn-glycero-3-phosphocholine (di-C18:1PC) was prepared from lipid stock solution in chloroform with molar ratio 1:1:1 (lipids were purchased from Avanti Polar Lipids). The labeled peptides dissolved in trifluoroethanol and the fluorescent lipid marker DiD-C180 (solid, 1,1'-di-octadecyl-3,3,3',3'-tetra- methyl-indodicarbocyanine, 4-chlorobenzene-sulfonate salt, Molecular Probes, Invitrogen Corp.), dissolved in methanol, were added to the lipid mixture at a ratio of 1:100,000. The peptide/lipid mixture was applied to indiumtin oxide-coated microscope glasses (thickness 0.13 − 0.16 mm; 15-30 Ω, SPI Supplies), and solvents were evaporated. GUVs were produced using

a homemade electroformation chamber assembled of a pair of indiumtin oxide-coated glasses filled with buffer (10 mM KPi, pH 7.0), to which an alternating current with 10 Hz and 2 V (corresponding to ca. 400 V/m) was applied for at least 2 h. For imaging, the electroformation chamber was directly mounted on the sample stage of a laser scanning confocal microscope equipped with two lasers. A blue argon ion laser (488 nm) and a red He-Ne laser (633 nm) were used to excite the Alexa 488 fluorophore and the lipid marker DiD-C18:0, respectively.

4.5 Supplementary Information.

Lateral diffusion of WALPs: Comparing experimental and simulation time scales

Simulation times from coarse grain molecular dynamics (CG-MD) simulations can be compared to real times by accounting for the smoother energy landscape of CG models as compared to atomistic models: Since CG models lack (part of) the friction due to motion of the atoms, the dynamics observed in CG-MD simulations is faster than at the atomistic level. For the Martini CG model, a time conversion factor of 4 was found adequate to match the simulated and experimental diffusion coefficients of both bulk water and DPPC lipids in a bilayer. However, this time conversion factor cannot *a priori* be expected to be applicable to other processes as well, since it depends on the energy barriers and the relevant mass and length scales involved in the process under study. Here, to convert the simulation times from our CG-MD simulations of membrane-embedded WALP peptides to experimental time scales, we simulated a WALP23 peptide in dioleoylphosphatidylcholine (DOPC) bilayers, and compared the obtained lateral diffusion coefficient to experiments[62, 65].

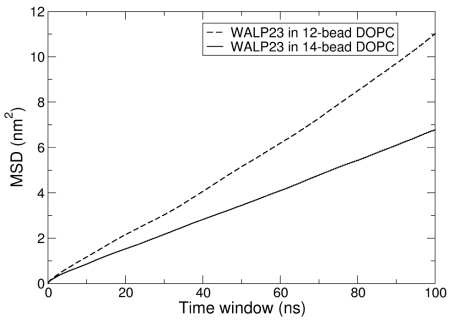

Figure 4.5: Mean square displacement of a single WALP23 in a 12-bead (dashed line) and a 14-bead (solid line) DOPC bilayer. The lateral diffusion coefficient was obtained from a linear fit to the msd curve in the time interval 10 to 100 ns (simulation time).

MD simulations. Two CG-MD simulations using the Martini model were carried out, in which a single WALP23 peptide was embedded into a pre-equilibrated bilayer of 476 DOPC molecules solvated by 6816 CG water beads (one CG water bead represents 4 water molecules). In the first simulation, a 14-bead Martini model was used for DOPC, in which each lipid chain comprised 5 beads. In the second simulation, DOPC comprised 12 beads, i.e., only 4 beads per chain. In both cases, the lipid chains contained a C3-type bead (in the third position), which together with the appropriate bonded parameters models the double bond. These two different models were used, because the measured thicknesses of DOPC bilayers can vary substantially (depending on the experimental conditions), with a typical range between 3.7 and 4.3 nm. The 12- and 14-bead DOPC models yield a bilayer thickness of about 3.90 and 4.35 nm, respectively, a range comparable to that observed in experiments. The systems were equilibrated for 200 ns prior to data collection. NpT ensembles

were simulated at 300 K, with the simulation parameters as described in Methods. The simulation time was 4 μs for each system.

Lateral diffusion. The lateral diffusion coefficient of the peptide was calculated from the long time mean square displacement (msd) of the peptide over time according to

$$D_{lat} = lim_{t\to\infty} \frac{\langle |r(t+t_0) - r(t_0)|^2 \rangle}{4t} \tag{4.1}$$

where r is the center of mass vector of the peptide in the membrane plane. The time window averaging was done over all initial time origins t_0. The overall center of mass motion was removed. The statistical error was obtained by separately analyzing the two halves of the trajectory. The obtained msd curves, shown in Figure 4.5, are to a good approximation linear over the time interval 10 to 100 ns. Linear fits yield $D_{lat} = 15.6 \pm 2.4$ $\mu m^2 s^{-1}$ and $23.2 \pm 2.0 \mu m^2 s^{-1}$ for WALP23 in the 14-bead and 12-bead DOPC bilayers, respectively, which are about 3 to 4.4 times larger as compared to the value of $D_{lat} = 5.3 \mu m^2 s^{-1}$ measured in giant unilamellar vesicles composed of pure DOPC[62] or of a 3:1 mixture of DOPC and DOPG[65]). Thus, we conclude that a factor of about 4 is appropriate to convert the peptide diffusion time observed in our CG-MD simulations to the experimental time scale. Unless stated otherwise, all times reported in section 4.2 are effective times, that is, simulation times multiplied by a factor 4.

TI calculations of partitioning free energy

Here we describe the details of the free energy calculations presented in section 4.2. The full results of the thermodynamic integration (TI) simulations are summarized in Table 4.5.

System setup. For the TI simulations, we used two bilayer patches with lipid compositions corresponding to the Lo and Ld domains, respectively. The Lo bilayer patch comprised 368 di-C16:0PC and 224 cholesterol molecules, corresponding to a 0.62:0.38 molar ratio. The Ld bilayer patch comprised 336 di-C18:2PC, 32 di-C16:0PC, and 72 cholesterol molecules, corresponding to a 0.76:0.08:0.16 molar ratio. These mixtures closely match the equilibrium compositions in the core of the Lo and Ld domains, respectively[34]. Initially, the PC and cholesterol molecules were randomly distributed in the membrane plane (with equal amounts of each molecule type in both monolayers), and the bilayers were solvated by 5328 CG water beads (one CG water bead represents 4 atomistic water molecules). Then, into both bilayers, a single WALP peptide was inserted in its transmembrane orientation. To generate starting structures for the subsequent simulations, the systems were energy minimized (steepest descent, 1000 steps) and equilibrated for 4 μs in free MD simulations, using the Martini force field and simulation parameters as described in Methods.

TI procedure. As shown in Figure 4.6, the free energies of extracting a peptide from the membrane bilayer to the gas phase were determined by thermodynamic integration with a coupling parameter λ. This was done in two steps: First, all non-bonded interactions of the peptide with the surrounding bilayer and water as well as within the peptide were turned off; second, the interactions were turned back on in vacuo. Thus, $\lambda = 0$ and $\lambda = 1$ correspond to the fully coupled and uncoupled states, respectively. Note, the helicity of the WALP peptides is imposed by the dihedral (bonded) interactions, and is thus not affected during the TI. From the free energies of extracting a peptide from the bilayer to the gas phase, the transfer free energy between the Lo and Ld domains was calculated as $\Delta G_{Lo/Ld} = \Delta G_{Lo/vac} - \Delta G_{Ld/vac}$. In the TI simulations, NpT ensembles were simulated using stochastic temperature coupling to T = 300K and an inverse friction coefficient of 5 ps. This friction coefficient yields a similar lipid lateral diffusion coefficients as observed in the MD simulations of the large 2000-lipid bilayer. The center of mass of the peptide was kept close to the center of the bilayer by means of a harmonic position restraint along the bilayer normal (force constant

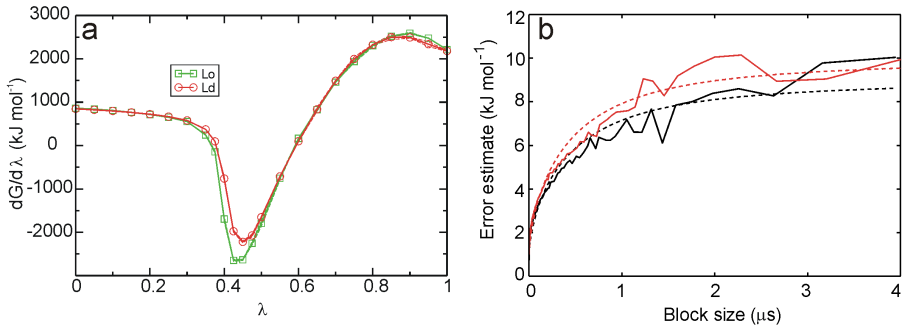

Figure 4.6: a) Results from thermodynamic integration simulations for WALP23. The green and red curves show the TI in the Lo and Ld bilayers, respectively. Statistical errors are plotted as dashed sidelines. Similar curves were obtained for WALP31. b) Convergence of the errors of the potential energy (used to estimate the enthalpy, see text) as a function of block size (solid lines), together with the analytical estimate (bi-exponential function, dashed lines) for the Ld bilayer without (black) and with WALP23 embedded (red).

1000 kJ mol^{-1} nm^{-2}). This restraint prevents the peptide from slipping out of the bilayer at λ close to 1, while still allowing it to diffuse, reorient, and tilt in the bilayer. A soft-core potential was used for the nonbonded Lennard-Jones interactions (the WALP peptides bear no (partial) charges in the Martini force field), with potential height $\alpha = 1.3$, λ-power = 1, and interaction-range $\sigma = 0.47$ nm[66]. We used a basic spacing between the λ-points of 0.05; additional λ-points were added at $\lambda = 0.375$, $\lambda = 0.425$, and $\lambda = 0.475$, where the curvature of $dG/d\lambda$ is maximal (see Figure 4.6a), yielding a total of 24 λ-points. At each λ-point, the system was simulated for 500 ns (which, using the time conversion factor of 4 observed for DOPC bilayers, see above, may correspond to about 2 μs of atomistic simulation time). The first 20% of each simulation were considered equilibration and excluded from the analysis. Error estimates were obtained by fitting a bi-exponential function to the block average of $dG/d\lambda$ as a function of block size, as implemented in the g_analyze tool of Gromacs and described in Ref.[67].

Enthalpy estimation. To obtain the enthalpies, we carried out long MD simulations of the Lo and Ld bilayer patches with and without the peptides, and of the individual peptides in vacuo. The systems were simulated for 40 μs and 16 μs for the Lo and Ld bilayer patches, respectively, in addition to an initial equilibration phase of 4 μs. The vacuum systems were simulated for 100 ns. The enthalpies were calculated by subtracting the potential energy of the pure bilayer and the peptide in vacuo from the potential energy of the membrane-embedded peptide. We neglected the contribution pΔV, as it is small for condensed liquid phase systems at ambient pressures (here, $p\Delta V \approx 0.3$ kJ mol^{-1} for WALP31). We also neglected the contribution due to the liberation of the solute[68] , as it is (i) smaller than the statistical error, and (ii) cancels out in the calculation of $\Delta H_{Lo/Ld}$. Error estimates (shown in Figure 4.6b) were determined by block averaging, see above.

Enthalpy - entropy compensation. As stressed in Table 1, the absolute numbers of the enthalpy and the entropy obtained with a coarse grain model should be interpreted with care. However, the Lo/Ld partitioning free energy differences should be accurate. This may in particular be expected for the Martini model, because it was parameterized against ex-

perimental partitioning free energies. In general, coarse grain models have a smaller entropy than atomistic models due to the reduced number of degrees of freedom. Thus, to arrive at the desired correct free energy, this smaller entropy has to be compensated for by a corresponding change in enthalpy For more details on this topic, we refer to the work of Baron *et al.*[69, 70] . For the dimerization of WALP23 in DOPC, a direct comparison of enthalpies and entropies obtained from experiment and simulation is possible: Yano *et al.*[38] measured values of $\Delta H = 31$ kJ mol^{-1} and $T\Delta S = +19$ kJ mol^{-1}, which can be compared to the respective values of 30 kJ mol^{-1} and +15 kJ mol^{-1} obtained by Ash and coworkers using the Martini force field[39]. Thus, not only is the dimerization free energy difference ΔG obtained with the Martini model in good agreement with experiment, but also the enthalpic and entropic parts.

Our main conclusion here is that the peptide sorting is enthalpy-driven. This conclusion is of qualitative nature, independent of the exact values. It is based on the observation that the inclusion of a TM helix introduces substantial disorder (increases entropy) in the Lo domain, but not in Ld (see Table 4.5). Thus, solely in terms of entropy, the peptides would segregate into Lo, which is not observed. Our free energy calculations show that in Lo, the entropy gain is outplayed by an enthalpy loss, which is thus the driving force for the overall process. Also at the atomistic level, the degree of disorder introduced by inserting a TM peptide into the Lo domain can be expected to be larger than in the Ld domain, and again the entropic driving force towards Lo will have to be over-compensated by enthalpy. Of course, this only holds if the resulting free energy upon moving a peptide from Lo to Ld is negative not only at the CG but also at the atomistic level. This can be safely assumed to be the case, because it is also observed in the experiments.

Table 4.2: Free energies, enthalpies, and entropies obtained from TI simulations. All values in kJ mol^{-1}. Results are obtained at T=300 K. $\Delta G_{Lo/vac}$ and $\Delta G_{Ld/vac}$ are the free energies of transferring a single peptide from the Lo and Ld bilayer patches, respectively, to the vacuum. $\Delta G_{Lo/Ld}$ was obtained as $\Delta G_{Lo/vac}\Delta G_{Ld/vac}$. Standard errors are given. The statistical errors in Lo/Ld were obtained by assuming that the values in Lo/vac and Ld/vac are independent.

	$\Delta G_{Lo/Ld}$	$\Delta G_{Lo/vac}$	$\Delta G_{Ld/vac}$
WALP23	-63 ± 8	678 ± 5	741 ± 5
WALP31	-58 ± 8	855 ± 5	913 ± 6
	$\Delta H_{Lo/Ld}$	$\Delta H_{Lo/vac}$	$\Delta H_{Ld/vac}$
WALP23	-594 ± 47	71 ± 45	665 ± 14
WALP31	-276 ± 33	500 ± 32	776 ± 10
	$-T\Delta S_{Lo/Ld}$	$-T\Delta S_{Lo/vac}$	$-T\Delta S_{Ld/vac}$
WALP23	531 ± 48	607 ± 46	76 ± 15
WALP31	218 ± 34	354 ± 32	137 ± 12

To test our hypothesis, we calculated lipid chain order parameters using both CG and atomistic MD simulations. The results, given in the following section, support the idea that the peptide sorting is enthalpy-driven.

Lipid chain order parameters

We calculated lipid chain order parameters to substantiate the interpretation of our TI calculations, and further investigate the configurations of the lipids around the peptides, both using CG and atomistic simulations. Peptide-bound lipids are compared to bulk lipids.

$$P_2 = \frac{1}{2}(3\langle cos^2\theta\rangle - 1) \tag{4.2}$$

Method. The second-rank order parameter, was computed for consecutive bonds in the CG lipids, with the angle between the bond vector and the bilayer normal. The square brackets denote an ensemble average. Perfect alignment with the bilayer normal is indicated by $P_2 = 1$, perfect anti-alignment by $P_2 = -0.5$, and a random orientation by $P_2 = 0$. Order parameters were calculated by averaging over 100 evenly spaced snapshots taken from the last μs of the free simulations of the bilayer systems used in the TI simulations (see above). Figure 4.5 compares the order parameters of the lipids that are bound to the peptide to those of unbound lipids. A lipid was considered bound to the peptide if the distance between its phosphate head group bead and any bead of the peptide was smaller than 0.8 nm. On average, 6 lipids were bound to a peptide. For the unbound lipids, the order parameters were averaged over all 368 and 336 di-C16:0PC and di-C18:2PC lipids present in the Lo and Ld peptide-free bilayers, respectively. In Ld, di-C18:2PC preferentially binds to the peptides, and thus the di-C16:0PC molecules in Ld were excluded from the analysis.

To obtain order parameters from atomistic simulations, two snapshots of WALP23 in the Lo and Ld bilayers, respectively (taken after 1000 ns of CG simulation) were back-mapped to their underlying atomistic representations using the reverse transformation procedure detailed below. For both systems, a 60 ns atomistic MD simulation was carried out; the first 30 ns of each simulation were discarded as equilibration time (for simulation details, see below). Prior to analysis, the atomistic trajectories were converted to the Martini CG representation to enable a comparison to the order parameters directly obtained from the CG simulations (Figure 4.5).

Results. Figure 4.5 shows that in the Lo domain, peptide-bound lipids have a lower order parameter as compared to lipids in the bulk. This effect is more pronounced for CG di-C16:0PC lipids in the vicinity of WALP23 (Figure 4.5a, black line) than for those close to WALP31 (Figure 4.5a, blue line): The average chain order parameter (averaged over bonds 4-7) drops from 0.793 for CG di-C16:0PC in the bulk Lo phase to 0.645 for CG lipids bound to WALP23, a decrease of 19%. CG lipids bound to WALP31 have an average chain order parameter of 0.728, which is only 8% lower than in the bulk. In addition, lipids bound to WALP23 show a decreased P_2 for bond 2 in the PC head group, which is not the case for lipids bound to WALP31. Although P_2 order parameters cannot be straightforwardly translated into entropies, the observed decrease of P_2, which is about two times stronger for WALP23 than for WALP31, agrees very well with the entropy differences obtained from the TI calculations, see Table 4.5. In contrast to the ordered domain, inclusion of WALP23 or WALP31 into the Ld domain does not substantially influence the order parameters as compared to the peptide-free case (Figure 4.5b). Thus, the high degree of disorder already intrinsic to Ld is not further decreased by the peptide, in line with the small entropy gain upon peptide inclusion (Table 4.5).

The order parameters obtained from the atomistic simulations are shown in Figure 4.5c,d. In the Lo domain, bonds in peptide-bound di-C16:0PC chains have lower order parameters than in bulk lipids: The average chain order parameters (averaged over bonds 4-7) are 0.704 and 0.814 for peptide-bound and bulk lipids, respectively, a difference of 0.11. For di-C18:2PC lipids in the Ld domain, the respective values are 0.200 and 0.240, a difference of only 0.04. Thus, in agreement with the CG simulations, also in the atomistic simulations WALP23 does introduce substantial disorder in the Lo, but not in the Ld domain. Hence, also the atomistic simulations show that there is an entropic driving force for peptide sorting into Lo that has to be overcompensated by an enthalpic preference for Ld.

Crossing and tilt angles of WALP23 dimers and trimers

Here we present details about the clusters formed by the WALP23 peptide.

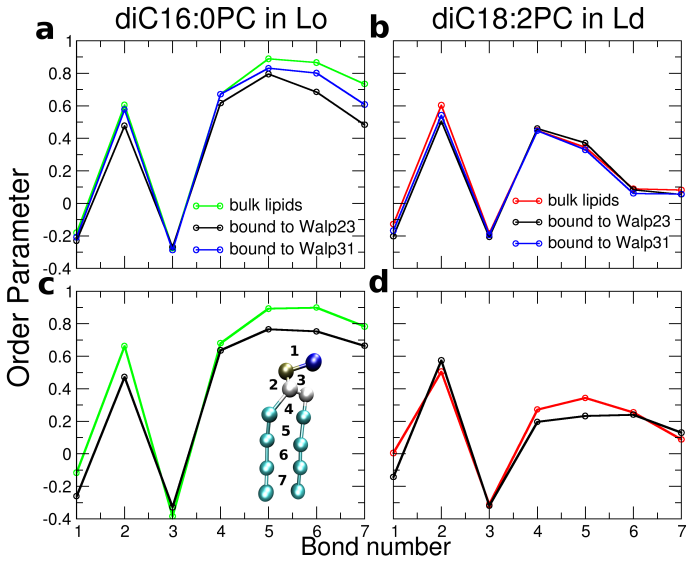

Figure 4.7: P_2 order parameters for consecutive bonds in PC lipids with respect to the bilayer normal. The data for bonds 4-7 are averaged over both lipid chains. The bond numbers are shown in c, inset. P_2 of CG di-C16:0PC in Lo (a) and CG di-C18:2PC in Ld (b) bound to WALP23 or WALP31 (black and blue lines, respectively), are compared to P_2 of bulk lipids in (peptide-free) Lo and Ld bilayers (green and red lines, respectively). Panels (c) and (d) show the respective order parameters obtained from atomistic simulations.

Crossing angles. Figure 4.8 shows that the crossing angle distributions for the right-handed dimers are rather broad, with a mean value at $\theta = -26.7°$ and a standard deviation of 6.6° (solid line). For trimers, $\theta = -24.4^{\theta}$ and a standard deviation of 6.5° was found (dashed line). These crossing angles are slightly smaller than the −40° observed in the right-handed glycophorin A dimer[43]. The inset of Figure 4.8b shows a representative crossing angle time trace for a WALP23 dimer (stable between 8 and 24 μs). The crossing angle displays transitions to values around −55° and −5° on the μs time scale, emphasizing the large conformational flexibility. To observe these rare events requires long time scales, as enabled by the use of CG models.

Table 4.3: Tilt angles of monomeric WALP peptides in Lo and Ld. Statistical errors (obtained from block averaging, see above) are below 1°.

Peptide	**Tilt in Lo (± std. dev.)**	**Tilt in Ld (± std. dev.)**
WALP23	8.8° ± 4.7°	20.9° ± 9.3°
WALP23	13.7° ± 5.3°	45.3° ± 7.8°

Tilt angles. The tilt angles (Figure 4.8) of single peptides were obtained from the extended simulations of the small bilayer patches used in the TI calculations and are summarized in Table 4.3. As expected, the WALP23 peptide in Lo (negative mismatch) adopts the

smallest tilt angle (8.8°). This tilt is very small, as it falls almost within the tilt due to thermal fluctuations. Under hydrophobic matching conditions, intermediate tilt angles of about 15° to 20° are observed, in very good agreement with data from experiments and other simulations[71, 72]. WALP31 in Ld reacts on the extreme mismatch conditions by adopting a very large tilt angle of about 45°.

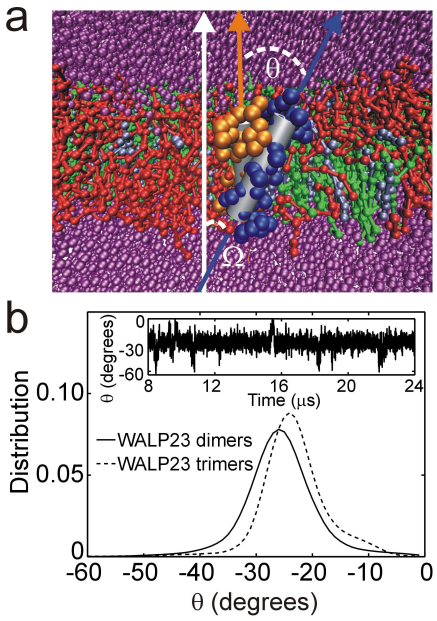

Figure 4.8: Crossing-angles of WALP23 dimers and trimers. a) Snapshot from CG-MD simulation showing a right-handed dimer of WALP23 molecules (side view). The two peptides are shown as orange and blue spheres, respectively; grey cylinders represent α-helices with an arrow along the helix axis. The white arrow shows the membrane normal. The inter-helix crossing-angle, θ and helix tilt-angle, Ω are indicated by dashed lines. di-C16:0PC and di-C18:2PC lipids are shown in green and red, respectively; cholesterol is colored grey; water is shown in purple. b) Distributions of crossing angles for WALP23 dimers (solid line) and trimers (dashed line). Inset: Representative crossing angle time- trace for individual WALP23 dimer.

Parallel vs. anti-parallel WALP23 dimers. In FRET experiments of Sparr and coworkers, a clear preference for an antiparallel arrangement of the WALP helices in a dimer was observed[40]. In our simulations, we found a 4:1 ratio of antiparallel (AP) and parallel (P) WALP23 dimers, in qualitative agreement with the experiments (Table 4.3). It was speculated that AP is preferred over P due to the favorable interaction between antiparallel helix-macrodipoles. Although the Martini model does not have explicit dipolar interactions between the helix backbones, the model differentiates between backbone beads close to the C- and N-termini of the helix in such a way that cross-interactions between C- and N-terminal backbone beads are more favorable than those between the same termini.

Partitioning of WALP31 with restricted tilt

Here we show that a WALP31 peptide that is restrained to an upright orientation prefers to reside at the domain boundary.

System setup. To study the partitioning of WALP31 in the absence of tilting, a single WALP31 transmembrane peptide was energy-minimized in vacuo (100 steps, steepest descent), and subsequently incorporated into the core of the Lo domain, such that it was not tilted with respect to the membrane normal. After 1000 steps of steepest descent energy minimization, a 20 μs MD simulation (Martini force field) was carried out, during which the motion of the backbone beads of residues 2 and 30 along the membrane normal (z-axis) was restricted through harmonic potentials (force constant 1000 kJ mol^{-1} nm^{-2}). This procedure yielded a reduced tilting of the peptide (mean tilt angle of about 10°).

Results. Within the first 4 μs the WALP31 peptide diffuses to the domain boundary, where it stays for the rest of the simulation time.

Atomistic simulation

To test whether the peptides clustering behavior might have been affected by the coarse grain representation of the molecules, we also performed an MD simulation using an atomistic force field.

Table 4.4: Peptide clusters in WALP23 system after t = 40 μs of CG-MD simulation. P = parallel; AP = antiparallel; RH = right-handed; LH = left-handed. C_α-rmsd is the value after 100 ns of atomistic MD.

Cluster	P/AP	RH/LH	C_α-rmsd
dimer1	AP	RH	0.24 nm
dimer2	P	RH	0.41 nm
dimer3	AP	RH	0.29 nm
dimer4	P	LH	0.29 nm
trimer	AP,P	LH,RH	0.23 nm
monomer	-	-	0.22 nm

Methods. In the atomistic simulation, the peptides were described with the 53A6 parameter set of the Gromos force field[73] , with the latest update on peptide backbone dihedral angle parameters that enhance the stability of α-helices[74]. For the phosphatidylcholine lipids we used an in-house version of a new Gromos53A6-based lipid force field; parameters and additional information are obtainable from the authors upon request. Parameters for cholesterol were taken from Marrink *et al.*[75] and the SPC water model[76] was used. To obtain the starting structure for the atomistic simulations, a snapshot from the CG simulation of the WALP23 system was transformed to the underlying atomistic representation using our recent resolution transformation algorithm[44]. During the resolution transformation, the system was cooled down from an initial temperature of 1300 K to the desired target temperature of 295 K by 100 ps of simulated annealing, during which the atomistic particles were coupled to their corresponding CG beads through harmonic restraints. Subsequently, the coupling was gradually removed within a time span of 10 ps. These annealing simulations were carried out in the NVT ensemble. No constraints were applied, and an integration time step of 1 fs was used. To control the temperature, stochastic coupling with an inverse friction constant $\tau_T = 0.1$ ps was applied. The other parameters for the resolution transformation were set to the standard values[44]. During the subsequent free atomistic MD simulations, application of the Lincs[77, 78] and SETTLE[79] algorithms allowed for an integration time step of 2 fs. Temperature and pressure coupling was applied similar to the CG simulations, with time constants $\tau_T = 0.1$ ps and $\tau_p = 1$ ps, respectively. Non-bonded interactions within 0.9 nm were updated at every time step, and interactions between 0.9 and 1.4 nm every 10 steps. A reaction field contribution[80] was added to the electrostatic interactions beyond this long-range cutoff, with $\epsilon_{rf} = 54$. *System details.* The snapshot after t = 40 μs from the CG-MD simulation of the WALP23 system was transformed to the atomistic representation, followed by a free 100 ns atomistic MD simulation. Including peptides, lipids, and water molecules, the system comprised of about 241,000 particles. As summarized in Table 4.4, at the beginning of the atomistic simulation, there were four WALP23 dimers, one trimer, and one monomer. Two of the dimers contain antiparallel helices, whereas the other two dimers consist of parallel peptides (averaged over the CG simulation, the AP/P dimer ratio is 4:1, see above). There is one left-handed dimer (dimer4), whereas the other three dimers adopt the right-handed configuration. The peptides in the trimer form the typical linear arrangement: The central helix forms an antiparallel left-handed pair with one of the outer helices, and a parallel right-handed pair with the other.

Results. Figure 4.9a shows a zoom-in on the antiparallel right-handed dimer3 (Table 4.4) after 100 ns of atomistic simulation; the C_α-rmsd time traces of the peptides during the atomistic simulation are shown in Figure 4.9b. The final rmsd of the oligomers with respect to their starting configuration (the back-mapped CG configuration) is summarized in Table

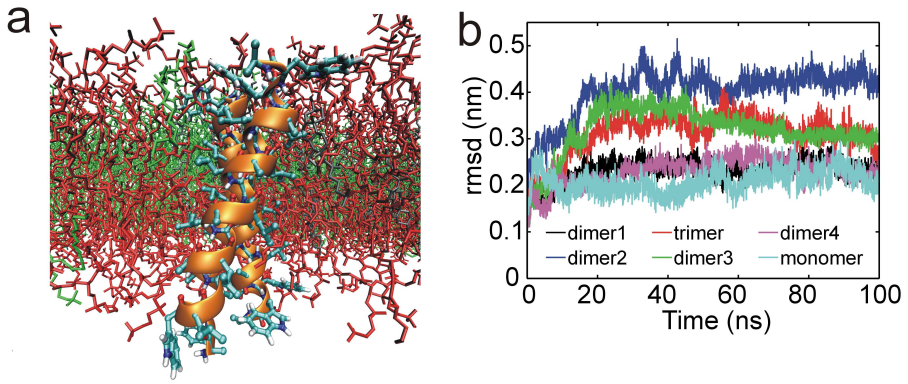

Figure 4.9: Atomistic molecular dynamics simulation of WALP23 system. a) Snapshot (side view) showing a zoom-in on an antiparallel right-handed WALP23 dimer (dimer3, Table 4.4) after 100 ns of simulation. The peptides are shown in ball-and-sticks representation with α-helices colored orange. Lipid molecules are shown as thin sticks, water molecules are not shown for clarity. The color code is the same as in Figure 4.1b) Root-mean-square deviation (rmsd) of peptide C_α-atoms with respect to starting configuration.

4.4. All peptides stay α-helical during the simulation. Three of the dimers (dimer1, dimer3, and dimer4) as well as the trimer stay bound, as indicated by an rmsd < 0.3 nm. Such low rmsd values are typical for atomistic MD simulations of small proteins (i.e., of comparable size to the WALP peptides) in the absence of large-scale conformational changes. The parallel dimer2, however, dissociates, and the two peptides become separated by lipid molecules. This single dissociation event shows that the simulation time of 100 ns, although considered short, can in principle be sufficient to probe the stability of the oligomers. This conclusion is also supported by the rmsd curves in Figure 4.9, which do not show a steady rise over the simulation time. Although the atomistic simulations are computationally too demanding to be able to yield a fully equilibrated ensemble, the results are consistent with the conclusions drawn from our CG simulations: WALP23 peptides in the Ld phase form stable dimers and trimers, with occasional dissociation (and, on longer time scales, re-association) events.

Coarse grain MD simulation of TM domain of syntaxin 1A

We carried out an additional CG-MD simulation of the α-helical transmembrane region of the syntaxin 1A (Sx1A) protein. Sx1A partitions into the Ld domain and forms weakly bound clusters.

System Setup. An atomistic α-helical structure of the TM domain of murine syntaxin 1A (TM-Sx1A, residues 259-288, sequence SKARRKKIMIIICCVILGIIIASTIGGIFG) was constructed using the pepbuild webserver (`www.imtech.res.in/bvs/pepbuild`) and converted into its CG representation. The subsequent MD simulation protocol used was very similar to the one applied for the WALP peptides: Initially, nine TM-Sx1A proteins were inserted into the same phase-separated lipid bilayer that was used for the WALP simulations; the TM-Sx1A helices were regularly distributed on a lateral grid. All nine proteins were inserted in a parallel manner, such that the

basic amino acids at the N-terminus are located at the same side of the membrane. After energy minimization and a short MD with positional restraints on the protein beads, the system was simulated for 8 μs (simulation time) using the Martini force field (simulation parameters identical to those applied in the WALP simulations). The C-terminus was capped with a negative charge, whereas the N-terminus was capped with an uncharged P4-type bead. The helicity of TM-Sx1A[80] was kept stable by dihedral potentials, similar to the WALPs.

Results. The observed partitioning and clustering of TM-Sx1A closely resembles that of WALP23 in that TM-Sx1A segregates into the Ld domain and forms small clusters, with frequent association / dissociation events on the μs time scale. Figure 4.10 shows a typical configuration, taken after 5 μs of simulation time. All proteins are located in the Ld domain, where they form (mostly) dimers and trimers that are in equilibrium with monomeric proteins.

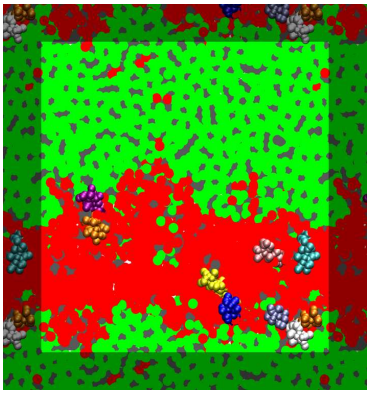

Figure 4.10: Representative snapshot after 5 μs of simulation time for Syntaxin 1A. Proteins are shown as colored spheres. Color code the same as in Figure 4.1.

Coarse grain MD simulations of TM domain of the linker for activation of T cells (LAT)

Additional CG-MD simulations of the transmembrane domain of the linker for activation of T cells (LAT) protein[81] were carried out, using both palmitoylated and non-palmitoylated forms.

System Setup. An atomistic α-helical structure of the TM domain of LAT (EADWLSPVGLGLLLLPFLVTLLAALCVRCRE, residues 232 of murine LAT, with substitution Trp for Ala4) was constructed using the pepbuild webserver (`www.imtech.res.in/bvs/pepbuild`) and converted into its CG representation. Two MD simulations were carried out. In the first simulation, palmitoyl anchors were attached to both terminal Cys residues[82], whereas in the second simulation, non-palmitoylated LAT was used. The simulation protocol was identical to the one applied for TM-Sx1A (see above): Initially, nine LAT proteins (regularly distributed on a lateral grid) were inserted into the bilayer, in a parallel manner. After energy minimization and a short MD with positional restraints on the protein beads, both the palmitoylated and non-palmitoylated LAT systems were simulated for 11 μs (simulation time), using the Martini force field (same simulation parameters as used above). Both termini were capped with uncharged P4-type beads; the helicity of residues 3-28 was kept stable through dihedral potentials.

Results. In agreement with experiments on model membranes[48], LAT TM domains partition into the Ld domain, irrespective of bound palmitoyl anchors. Due to the initial placement of monomeric proteins in the membrane, the LATs first sequester as monomers into Ld, where they form clusters. The clustering behavior is different for palmitoylated and non-palmitoylated LAT: For the former, occasional dissociations of clusters are observed, whereas non-palmitoylated LAT clusters do not break up again within the simulation time. The stronger clustering of LAT is similar to WALP31, due to the comparably long hydrophobic stretches. The monomeric LATs sample both the Lo/Ld interface and the bulk Ld domain. The LAT clusters also prefer Ld, but remain close to the interface (see Figure 4.11) and do not return to the core of the Ld domain on the time scale of our simulations.

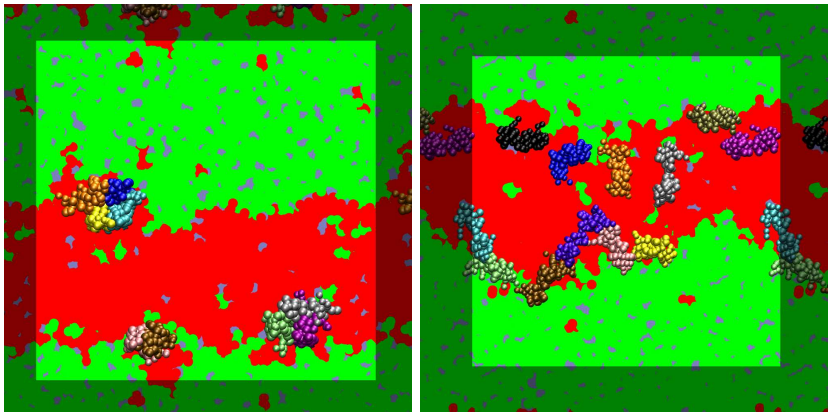

Figure 4.11: Representative snapshots (after 9 and 11 μs of simulation time) from CG-MD simulation of palmitoylated LAT TM domain (left) and octa-palmitoylated WALP23 (right). Proteins are shown as colored spheres. Color code the same as in Figure 4.1.

Coarse grain MD simulations of palmitoylated WALP23 peptides

Two additional CG-MD simulations were carried out. In the first simulation, two palmitoyl anchors were attached to WALP23, whereas in the second simulation, eight palmitoyl anchors per peptide were used. The doubly palmitoylated peptides partition into the Ld domain, as was also observed in experiments[32], whereas the octa-palmitoylated WALP23 peptides are located at the domain boundary interface. In chapter 5 the influence of adding lipid anchors to proteins is further investigated.

System Setup. The system setup and simulation parameters were the same as for the non-palmitoylated WALPs described above. For the doubly palmitoylated WALP, both palmitoyl anchors were attached to the side chain of an additional Glu residue attached to the N-terminus of the peptide. To generate the octa-palmitoylated peptides, four palmitoyl tails were attached to the N- and C-terminal residues, respectively. The simulation times were 9 μs for both simulations.

Results. The partitioning and clustering of the doubly-palmitoylated WALP23 peptides does not substantially differ from that of their non-palmitoylated counterparts. Peptides carrying eight palmitoyl anchors also preferentially partition into the Ld domain, however, they are bound to the domain boundary interface (see Figure 4.11). The fluctuations of the interface are enhanced, suggesting a reduced line tension. Such poly-palmitoylated peptides are thus potential line-active molecules, a prediction that can be tested experimentally.

Dimerization free energy of WALP23: Statistical error, comparison to literature

As described in section 4.2, we estimated the association constant K of WALP23 in the Ld domain form the average number of monomers and dimers observed during the simulation. For this analysis, the first 20 μs of the 80 μs simulation were discarded as equilibration time. To obtain the statistical error σK, we calculated the time autocorrelation function of monomer-monomer contacts, which displays a single-exponential decay with $\tau \approx 4\mu$s.

We then applied a block averaging procedure, using 7 independent blocks of length 2τ, to calculate the standard error in K. The error in the dimerization free energy $\Delta G = -RTlnK$ is then obtained using $\sigma_{\Delta G} = RT/K\sigma_K$.

The dimerization free energy of -12 ± 2 kJ mol^{-1}we obtained for WALP23 in Ld can be compared to Martini simulations of W.L Ash and coworkers[39]. From potential of mean force calculations of WALP23 dimers in DOPC, a dimerization free energy of -15 kJ mol^{-1} was obtained, in good agreement with our value. In addition, both values also agree with the value of $\Delta G = -12.7 \pm 0.4$ kJ mol^{-1}obtained fromFRETexperimentsofverysimilarpeptidesinDOPC[38]. The peptides used in the experiments were X-(AALALAA)3-Y, with X = 7-nitro-2-1, 3-benzoxadiazol-4-yl (NBD) and Y = NH2 (I), or X = Ac and Y = NHCH2CH2-S-N-[4-[[4-(dimethylamino)phenyl]azo]-phenyl]maleimide (DABMI) (II), and FRET from I to II was measured.

4.6 Bibliography

[1] D Lingwood and K Simons. Lipid rafts as a membrane-organizing principle. *Science*, 327:46–50, 2010.

[2] K Simons and E Ikonen. Functional rafts in cell membranes. *Nature*, 387:569–572, 1997.

[3] C Eggeling, C Ringemann, R Medda, G Schwarzmann, K Sandhoff, S Polyakova, V.N Belov, B Hein, C Von Middendorff, A Schönle, and S.W Hell. Direct observation of the nanoscale dynamics of membrane lipids in a living cell. *Nature*, 457:1159–1162, 2009.

[4] J.F Hancock. Lipid rafts: contentious only from simplistic standpoints. *Nat. Rev. Mol. Cell Biol.*, 7:456–462, 2006.

[5] M Edidin. The state of lipid rafts: from model membranes to cells. *Ann. Rev. Biophys. Biomol. Struc.*, 32:257–283, 2003.

[6] K Jacobson, O.G Mouritsen, and R.G.W Anderson. Lipid rafts: at a crossroad between cell biology and physics. *Nat. Cell Biol.*, 9:7–14, 2007.

[7] T Baumgart, S.T Hess, and W.W Webb. Imaging coexisting fluid domains in biomembrane models coupling curvature and line tension. *Nature*, 425:821–824, 2003.

[8] N Kahya, D Scherfeld, K Bacia, B Poolman, and P Schwille. Probing lipid mobility of raft-exhibiting model membranes by fluorescence correlation spectroscopy. *J. Biol. Chem.*, 28109–28115:273, 2003.

[9] A.V Samsonov, I Mihalyov, and F.S Cohen. Characterization of cholesterol-sphingomyelin domains and their dynamics in bilayer membranes. *Biophys. J.*, 81:1486–1500, 2001.

[10] S.L Veatch, I.V Polozov, K Gawrisch, and S.L Keller. Liquid domains in vesicles investigated by nmr and fluorescence microscopy. *Biophys. J.*, 86:2910–2922, 2004.

[11] T Baumgart, A.T Hammond, P Sengupta, S.T Hess, D.A Holowka, B.A Baird, and W.W Webb. Large-scale fluid/fluid phase separation of proteins and lipids in giant plasma membrane vesicles. *Proc. Nat. Acad. Sci. USA*, 104:3165, 2007.

[12] D Lingwood, J Ries, P Schwille, and K Simons. Plasma membranes are poised for activation of raft phase coalescence at physiological temperature. *Proc. Nat. Acad. Sci.*, 105:10005–10010, 2008.

[13] D.A Brown and E London. Structure and origin of ordered lipid domains in biological membranes. *J. Membr. Biol.*, 164:103–114, 1998.

[14] K Simons and W.L.C Vaz. Model systems, lipid rafts, and cell membranes1. *Annu. Rev. Biophys. Biomol. Struct.*, 33:269–295, 2004.

[15] R Phillips, T Ursell, P Wiggins, and P Sens. Emerging roles for lipids in shaping membrane-protein function. *Nature*, 459:379–85, 2009.

[16] J.N Sachs and D.M Engelman. Introduction to the membrane protein reviews: the interplay of structure, dynamics, and environment in membrane protein function. *Biochemistry*, 75:707–712, 2006.

[17] M.F Hanzal-Bayer and J.F Hancock. Lipid rafts and membrane traffic. *FEBS Lett.*, 581:2098–2104, 2007.

[18] H Sprong, P van der Sluijs, and G van Meer. How proteins move lipids and lipids move proteins. *Nat. Rev. Mol. Cell Biol.*, 2:504–513, 2001.

[19] G Van Meer, D.R Voelker, and G.W Feigenson. Membrane lipids: where they are and how they behave. *Nat. Rev. Mol. Cell Biol.*, 9:112–124, 2008.

[20] L.H Chamberlain, R.D Burgoyne, and G.W Gould. Snare proteins are highly enriched in lipid rafts in pc12 cells: implications for the spatial control of exocytosis. *Proc. Nat. Acad. Sci. USA*, 98:5619–5624, 2001.

[21] T Lang, D Bruns, D Wenzel, D Riedel, P Holroyd, C Thiele, and R Jahn. Snares are concentrated in cholesterol-dependent clusters that define docking and fusion sites for exocytosis. *EMBO J.*, 20:2202–2213, 2001.

[22] M Bretscher and S Munro. Cholesterol and the golgi apparatus. *Science*, 261:1280–1281, 1993.

[23] M Kawabuchi, Y Satomi, T Takao, Y Shimonishi, S Nada, K Nagal, A Tarakhovsky, and M Okada. Transmembrane phosphoprotein cbp regulates the activities of src-family tyrosine kinases. *Nature*, 404:990–1003, 2000.

[24] S Moffett, D.A Brown, and M.E Linder. Lipid-dependent targeting of g proteins into rafts. *J. Biol. Chem.*, 275:2191–2198, 2000.

[25] K Simons and D Toomre. Lipid rafts and signal transduction. *Nat. Rev. Mol. Cell Biol.*, 1:31–41, 2000.

[26] X Periole, T Huber, S.J Marrink, and T.P Sakmar. G protein-coupled receptors self-assemble in dynamics simulations of model bilayers. *J. Am. Chem. Soc*, 129:10126–10132, 2007.

[27] U Schmidt, G Guigas, and M Weiss. Cluster formation of transmembrane proteins due to hydrophobic mismatching. *Phys. Rev. Lett.*, 101:128104, 2008.

[28] B Reynwar and M Deserno. Membrane composition-mediated protein-protein interactions. *Biointerphases*, 3:FA117–FA124, 2008.

[29] F.J.M de Meyer, M Venturoli, and B Smit. Molecular simulations of lipid-mediated protein-protein interactions. *Biophys. J.*, 95:1851–1865, 2008.

[30] M.R.R de Planque and J.A Killian. Protein-lipid interactions studied with designed transmembrane peptides: role of hydrophobic matching and interfacial anchoring (review). *Mol. Membr. Biol.*, 20:271–284, 2003.

[31] T.K.M Nyholm, S Özdirekcan, and J.A Killian. How protein transmembrane segments sense the lipid environment. *Biochemistry*, 46:1457–1465, 2007.

[32] B.Y van Duyl, D.T.S Rijkers, B de Kruijff, and J.A Killian. Influence of hydrophobic mismatch and palmitoylation on the association of transmembrane [alpha]-helical peptides with detergent-resistant membranes. *FEBS Lett.*, 523:79–84, 2002.

[33] M.E Fastenberg, H Shogomori, X Xu, D.A Brown, and E London. Exclusion of a transmembrane-type peptide from ordered-lipid domains (rafts) detected by fluorescence quenching: Extension of quenching analysis to account for the effects of domain size and domain boundaries. *Biochemistry*, 42:12376–12390, 2003.

[34] H.J Risselada and S.J Marrink. The molecular face of lipid rafts in model membranes. *Proc. Nat. Acad. Sci. USA*, 105:17367, 2008.

[35] L Monticelli, S.K Kandasamy, X Periole, R.G Larson, D.P Tieleman, and S.J Marrink. The martini coarse-grained force field: extension to proteins. *J. Chem. Theory Comput.*, 4:819–834, 2008.

[36] S.J Marrink, H.J Risselada, S Yefimov, D.P Tieleman, and A.H De Vries. The martini force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 111:7812–7824, 2007.

[37] K Pleming. Standardizing the free energy change of transmembrane helix-helix interactions. *J. Mol. Biol.*, 323:563–571, 2002.

[38] Y Yano, M Ogura, and K Matsuzaki. Measurement of thermodynamic parameters for hydrophobic mismatch 2: Intermembrane transfer of a transmembrane helix. *Biochemistry*, 45:3379–3385, 2006.

[39] W Ash. *Helix-helix interactions in membrane proteins probed with computer simulations.* PhD thesis, University of Calgary, 2009.

[40] E Sparr, W.L Ash, P.V Nazarov, D.T.S Rijkers, M.A Hemminga, D.P Tieleman, and J.A Killian. Self-association of transmembrane -helices in model membranes. *J. Biol. Chem.*, 280:39324–39331, 2005.

[41] J.U Bowie. Helix packing in membrane proteins1. *J. Mol. Biol.*, 272:280–289, 1997.

[42] A.R Curran and D Engelman. Sequence motifs, polar interactions and conformational changes in helical membrane proteins. *Curr. Opin. Struc. Biol.*, 13:412–417, 2003.

[43] K.R MacKenzie, J.H Prestegard, and D.M Engelman. A transmembrane helix dimer: structure and implications. *Science*, 276:131–133, 1997.

[44] A.J Rzepiela, L.V Schäfer, N Goga, H.J Risselada, A.H De Vries, and S.J Marrink. Reconstruction of atomistic details from coarsegrained structures. *J. Comput. Chem.*, 31:1333–1343, 2010.

[45] K Bacia, C Schuette, N Kahya, R Jahn, and P Schwille. Snares prefer liquid-disordered over "raft"(liquid-ordered) domains when reconstituted into giant unilamellar vesicles. *J. Biol. Chem.*, 279:37951–37955, 2004.

[46] A.T Hammond, F.A Heberle, T Baumgart, D Holowka, B Baird, and G.W Feigenson. Crosslinking a lipid raft component triggers liquid ordered-liquid disordered phase separation in model plasma membranes. *Proc. Nat. Acad. Sci. USA*, 102:6320–6325, 2005.

[47] A Vidal and T.J McIntosh. Transbilayer peptide sorting between raft and nonraft bilayers: comparisons of detergent extraction and confocal microscopy. *Biophys. J.*, 89:1102–1108, 2005.

[48] H Shogomori, A Hammond, A.G Ostermeyer-Fay, D.J Barr, G.W Feigenson, E London, and D.A Brown. Palmitoylation and intracellular domain interactions both contribute to raft targeting of linker for activation of t cells. *J Biol. Chem.*, 280:18931–18942, 2005.

[49] J Nikolaus, S Scolari, E Bayraktarov, N Jungnick, S Engel, A.P Plazzo, M Stökl, R Volkmer, M Veit, and A Hermann. Hemagglutinin of influenza virus partitions into the nonraft domain of model membranes. *Biophys. J.*, 99:489–498, 2010.

[50] H.J Kaiser, D Lingwood, I Levental, J.L Sampaio, L Kalvodova, L Rajendran, and K Simons. Order of lipid phases in model and plasma membranes. *Proc. Nat. Acad. Sci. USA*, 106:16645, 2009.

[51] R Laage, J Rohde, B Brosig, and D Langosch. A conserved membrane-spanning amino acid motif drives homomeric and supports heteromeric assembly of presynaptic snare proteins. *J. Biol. Chem.*, 275:17481, 2000.

[52] J.J Sieber, K.I Willig, C Kutzner, C Gerding-Reimers, B Harke, G Donnert, B Rammer, C Eggeling, S.W Hell, H Grubmüller, and T Lang. The snare motif is essential for the formation of syntaxin clusters in the plasma membrane. *Biophys. J.*, 317:1072–1076, 2006.

[53] J.J Sieber, K.I Willig, C Kutzner, C Gerding-Reimers, B Harke, G Donnert, B Rammner, C Eggeling, S.W Hell, and H Grubmüller. Anatomy and dynamics of a supramolecular membrane protein cluster. *Science*, 317:1072, 2007.

[54] T McIntosh, A Vidal, and S.A Simon. Sorting of lipids and transmembrane peptides between detergent-soluble bilayers and detergent-resistant rafts. *Biophys. J.*, 85:1656–1666, 2003.

[55] S Munro. Lipid rafts: Elusive or illusive? *Cell*, 115:377–388, 2003.

[56] J.A Lundbaek, O.S Andersen, T Werge, and C Nielsen. Cholesterol-induced protein sorting: an analysis of energetic feasibility. *Biophys. J.*, 84:2080–2089, 2003.

[57] S Trabelsi, S Zhang, T.R Lee, and D.K Schwartz. Linactants: surfactant analogues in two dimensions. *Phys. Rev Lett.*, 100:1–4, 2008.

[58] L.V Schäfer and S.J Marrink. Partitioning of lipids at domain boundaries in model membranes. *Biophys. J.*, 99:L91–L93, 2010.

[59] J Tong, M Briggs, D Mlaver, A Vidal, and T.J McIntosh. Sorting of lens aquaporins and connexins into raft and nonraft bilayers: role of protein homo-oligomerization. *Biophys. J.*, 97:2493–2502, 2009.

[60] J.F Girmens N Buzhynskyy, W Faigle, and S Scheuring. Human cataract lens membrane at subnanometer resolution. *J. Mol. Biol.*, 374:162–169, 2007.

[61] B Hess, C Kutzner, D van der Spoel, and E Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4:435–447, 2008.

[62] S Ramadurai, A Holt, L.V Schäfer, V Krasnikov, and B Poolman. Influence of hydrophobic mismatch and amino acid composition on the lateral diffusion of transmembrane peptides. *Biophys. J.*, 99:1482–1489, 2010.

[63] M.R.R de Planque, J.A.W Kruijtzer, R.M.J Liskamp, D Marsh, D.V Greathouse, R.E Koeppe II, B de Kruijff, and J.A Killian. Different membrane anchoring positions of tryptophan and lysine in synthetic transmembrane -helical peptides. *J. Biol. Chem.*, 274:20839–20846, 1999.

[64] A Holt, R.M Koehorst, T Rutters-Meijneke, M.H Gelb, D.T.S Rijkers, M.A Hemminga, and J.A Killian. Tilt and rotation angles of a transmembrane model peptide as studied by fluorescence spectroscopy. *Biophys. J.*, 97:2258–2266, 2009.

[65] S Ramadurai, A Holt, V Krasnikov, G van den Bogaart, J.A Killian, and B Poolman. Lateral diffusion of membrane proteins. *J. Am. Chem. Soc.*, 131:12650–12656, 2009.

[66] J Wong-Ekkabut, S Baoukina, W Triampo, I.M Tang, D.P Tieleman, and L Monticelli. Computer simulation study of fullerene translocation through lipid membranes. *Nat. Nanotechnol.*, 3:363–368, 2008.

[67] B Hess. Determining the shear viscosity of model liquids from molecular dynamics simulations. *J. Chem. Phys.*, 116:209, 2002.

[68] B Hess and N.F.A van der Vegt. Hydration thermodynamic properties of amino acid analogues: a systematic comparison of biomolecular force fields and water models. *J. Phys. Chem. B*, 110:17616–17626, 2006.

[69] R Baron, A.H de Vries, P Hünenberger, and W.F van Gusteren. Configurational entropies of lipids in pure and mixed bilayers from atomic-level and coarse-grained molecular dynamics simulations. *J. Phys. Chem. B*, 110:15602–15614, 2006.

[70] R Baron, D Trzesniak, A.H de Vries, A Elsener, S.J Marrink, and W.F van Gunsteren. Comparison of thermodynamic properties of coarsegrained and atomiclevel simulation models. *ChemPhysChem*, 8:452–461, 2007.

[71] A Holt, L Rogier, V Reat, F Jolibois, O Saurel, J Czaplicki, J.A Killian, and A Milon. Order parameters of a transmembrane helix in a fluid bilayer: case study of a walp peptide. *Biophys. J.*, 2258-2266:97, 2010.

[72] J Salgado. The dynamic orientation of membrane-bound peptides: bridging simulations and experiments. *Biophys. J.*, 131:15194–15202, 2007.

[73] C Oostenbrink, A Villa, A.E Mark, and W.F Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos forcefield parameter sets 53a5 and 53a6. *J. Comput. Chem.*, 25:1656–1676, 2004.

[74] A Choutko, S Riniker, M Winger, A.E Mark, W.F Van Gunsteren, N Schmid, and A.P Eigenberger. Definition and testing of the gromos force-field versions 54a7 and 54b7. *Eur. Biophys. J.*, 40:843856, 2011.

[75] S.J Marrink, A.H De Vries, T.A Harroun, J Katsaras, and S.R Wassall. Cholesterol shows preference for the interior of polyunsaturated lipid membranes. *J. Am. Chem. Soc*, 130:10–11, 2008.

[76] H.J.C Berendsen, J.P.M Postma, W.F Van Gunsteren, and J Hermans. *Intermolecular forces*, chapter Interaction models for water in relation to protein hydration, pages 331–342. D. Reidel Publishing Company, 1981.

[77] B Hess, H Bekker, H.J.C Berendsen, and J.G.E.M Fraaije. Lincs: a linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18:1463–1472, 1997.

[78] B Hess. P-lincs: A parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.*, 4:116–122, 2008.

[79] S Miyamoto and P.A Kollman. Settle: an analytical version of the shake and rattle algorithm for rigid water models. *J. Comput. Chem.*, 13:952–962, 1992.

[80] I.G Tironi, R Sperb, P.E Smith, and W.F Van Gunsteren. A generalized reaction field method for molecular dynamics simulations. *J. Chem. Phys.*, 105:5451–5461, 1995.

[81] W Zhang, J Sloan-Lancaster, J Kitchen, R.P Trible, and L.E Samelson. Lat: the zap-70 tyrosine kinase substrate that links t cell receptor to cellular activation. *Cell*, 93:83–92, 1998.

[82] W Zhang, R Trible, and L.E Samelson. Lat palmitoylation: Its essential role in membrane microdomain targeting and tyrosine phosphorylation during t cell activation. *Immunity*, 9:239–246, 1998.

5

Molecular view on protein sorting into liquid-ordered membrane domains mediated by gangliosides and lipid anchors

This chapter is based upon the manuscript:
Molecular view on protein sorting into liquid- ordered membrane domains mediated by gangliosides and lipid anchors by Djurre H. de Jong, Cesar A. Lopez, Siewert J. Marrink, Faraday Discussions, **2012**, *In Press*, DOI: 10.1039/C2FD20086D

Abstract

We present results from coarse grain molecular dynamics simulations of mixed model membranes consisting of saturated and unsaturated lipids together with cholesterol, in which lipid-anchored membrane proteins are embedded. Membrane proteins studied are the peripherally bound H-Ras, N-Ras, and Hedgehog, and the transmembrane peptides WALP and LAT. We provide a molecular view on how the presence and nature of these lipid anchors affects partitioning of the proteins between liquid-ordered and liquid-disordered domains. In addition, we probed the role of the ganglioside lipid GM1 on the protein sorting, showing formation of GM1-protein nano- domains that act as shuttles between the differently ordered membrane regions.

5.1 Introduction

The organisational principles of the cell membrane are amongst the great open questions in biology[1, 2, 3]. The cell membrane (or plasma membrane) consists of a mixture of different lipids, transmembrane proteins and membrane anchored soluble proteins. In order to bring together functional components in such a complex mixture the plasma membrane is believed to be compartmentalized, thus restricting the conformational search problem. These membrane compartments, or domains are typically known as lipid rafts[4] and are enriched in sphingolipids, cholesterol and specific proteins. The current view describes these rafts *in vivo* as nanoscale assemblies that may condense into larger platforms under relevant conditions[2]. *In vitro*, the segregation of ternary lipid mixtures into a liquid ordered (Lo) and liquid disordered (Ld) membrane is considered to be a good model for membrane compartmentalization [5, 6, 7]. In such model systems, transmembrane proteins are typically found segregated into the Ld phase, even in the case of raft-associated proteins. This leaves

us with a fundamental question how cells are able to regulate protein sorting into more ordered regions of the membrane. In order to target proteins towards rafts, lipids are believed to play an important role.

The lipid-protein interaction could be chemically enforced, i.e. via lipid anchors, or occur through either specific or non-specific binding. Different types of lipid anchors include glycophosphatidylinositol (GPI), palmitoyl, isoprenyl or sterol anchors, and are found for both peripheral and transmembrane proteins[8]. These anchors have a natural affinity for the raft domains and may drag the protein along. Binding of proteins to raft lipids could be another mechanism by which proteins are being sorted. Specific lipid binding sites are found, for example, for the p24 transmembrane protein recognizing certain sphingolipids[9] or the influenza virus M2 protein binding cholesterol[10]. Besides sphingolipids and cholesterol, ganglioside lipids are an important component of rafts[11, 12, 13, 14, 15]. Gangliosides are ceramide based lipids with an oligosaccharide head group. GM1 and GM3, two well known ganglioside lipid species, bind different proteins (both soluble[16, 17] and transmembrane[18, 19]) and promote formation of extended raft domains upon crosslinking with antibodies[20]. Despite the vast amount of research effort in the last decade, the underlying organising principles of cell membranes are not yet fully understood. To gain molecular insight in the driving forces of membrane protein segregation, coarse grain (CG) molecular dynamics simulations are a promising technique[21, 22, 23, 24]. The CG Martini model[25, 26] has proven to work well in this respect. Applying the Martini model, we recently were able to correctly predict the partitioning of transmembrane (TM) helices between domains in Lo-Ld phase separated model membranes consisting of ternary mixtures of saturated lipids, unsaturated lipids, and cholesterol[27, 28]. These simulations allowed us to evaluate the underlying mechanism by which small proteins are sorted into the Ld phase. The driving force for partitioning into the Ld phase was shown to originate from lipid packing effects. Here we extend these studies in order to investigate the role of lipid anchors in steering the sorting of membrane proteins. The proteins considered are the peripheral proteins H-Ras and N-Ras, which have a farnesyl anchor together with two or one palmitoyl anchors, respectively, and Hedgehog, a cholesterol anchored peripheral protein. We also studied the TM peptides WALP and linker for activation of T-cells (LAT) with two palmitoyl anchors attached. Additionally, we simulated the ability of small amounts of GM1 gangliosides to affect the sorting process in case of the TM peptides.

5.2 Methods

System setup

All systems are based upon a spontaneously demixed bilayer consisting of 828 dipalmitoyl-phosphatidylcholine (di-C16:0PC, DPPC), 540 dilinoeoyl-phosphatidylcholine (di-C18:2PC, DLiPC), and 144 cholesterol CG molecules, solvated by 12600 CG water beads representing 50,400 real waters. The simulation cell measures approximately 21.7 by 21.7 nm in the lateral (x,y) dimensions and 7.5nm in the perpendicular z-direction. Domanski *et al.*[28] found this mixture to be weakly phase separating into Ld and Lo domains with a line tension of 2 ± 2 pN. Additionally, the systems contained either four copies of anchored peripheral proteins (H-Ras, N-Ras, or Hedgehog) or twelve lipid anchored transmembrane peptides (WALP23 or LAT). In some systems 32 ganglioside lipids (GM1) were added. An overview of the systems simulated is given in Table 5.1 and a graphical overview of the membrane constituents is shown in Figure 5.1.

In systems containing the anchored peripheral proteins H-Ras, N-Ras, or Hedgehog, the four proteins were initially positioned on a regular grid. The globular domains were put

Table 5.1: Overview of simulated bilayer systems.

System	Composition	Simulation time (μs)	Number of simulations
Reference			
A	DPPC:DLiPC:Chol (828:540:144)	10	1
Peripheral proteins			
B	A + 4 H-Ras	12	1
C	A + 4 H-Ras (no Pal.)	12	1
D	A + 4 N-Ras	12	1
E	A + 4 Hedgehog	12	1
Transmembrane peptides			
F	A + 12 WALP	12	1
G	A + 12 WALP (no Pal.)	12	1
H	A + 12 LAT	12	1
Gangliosides			
F	A + 4% GM1	12	1
G	F + 4% GM1	12	2
H	H + 4% GM1	12	2

approximately 2 nm above the membrane, all on the same face of the membrane. To accommodate the lipid anchors, one of the membrane lipids was replaced by either one (N-Ras) or two (H-Ras) of the palmitoyl anchors (effectively lowering the amount of lipids in a monolayer by $\sim 1\%$ for the four proteins in total). The farnesyl tail was inserted in the membrane separately. For Hedgehog, a membrane cholesterol was substituted. The system size in the z-direction was enlarged by 6 nm to ensure that periodic images would not interact and subsequently 24043 additional water beads were added to solvate the system. In case of the Ras proteins, 24 sodium counter ions were added to neutralize the system, equivalent to four times the minus 6 charge of a single protein. For the systems containing LAT-peptides, 12 peptides were inserted in a parallel manner on a regular grid. The parallel orientation is consistent with *in vivo* systems where membrane proteins are oriented with positively charges residues towards the cytoplasm, the so called positive inside rule[29]. The WALP peptides were inserted in a similar manner, except they were positioned in an alternating parallel-antiparallel fashion, mimicking the random up-down orientation found in experiments[30]. Both LAT and WALP peptides were initially placed overlapping with the membrane, and overlap was subsequently removed by energy minimization (see below). Addition of 12 sodium counter ions assured neutralizing the charge of the 12 LAT peptides.

In systems containing GM1 lipids, 16 lipids were added to both monolayers, equivalent to ~ 4 mol% lipids. The GM1 lipids were initially positioned on a regular grid. In systems containing both peptides and GM1 lipids the peptide and ganglioside grids were shifted with respect to each other, in order to have no initial overlap between the GM1 lipids and peptides. One additional sodium ion per GM1 lipid was added as counter ion.

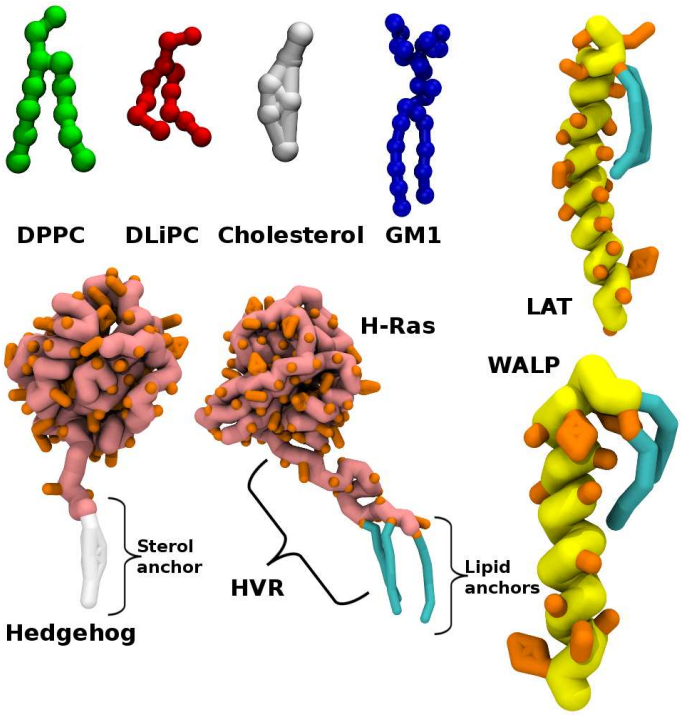

Figure 5.1: Coarse grain representation of membrane constituents considered in this work. The Martini model is used, which maps on average four heavy atoms into a single CG bead. Secondary structure of the proteins is constraint in this model. Molecules are depicted at different scales. The color scheme is corresponding to the one used in the remaining figures. The N-Ras protein is similar to H-Ras, but lacks one of the palmitoyl anchors. HVR stands for hyper variable region, an unstructured part of the Ras protein.

Molecular parameters

The Martini CG force field[25, 26] was used for all systems. The parameters for DPPC, DLiPC and cholesterol, as well as those for water and ions, can be found in previous publications[25, 31]. The CG representation of all membrane constituents is shown in Figure 5.1. All parameters used in this study can be downloaded from `http://cgmartini.nl`.

For H-Ras and N-Ras, structures were taken from the PDB-database (121P[32] & 3CON[33]). H-Ras is crystallized in its activated state, bound to the GTP mimicking inhibitor GCP, and N-Ras is crystallized in its inactive state, bound to GDP. The soluble domain is connected to the membrane via lipid anchors that reside in an unstructured stretch of amino acids known as the hyper variable region, HVR (cf. Figure 5.1). The C-terminal HVR, residues 166-186, was modelled as a random coil using Pymol[34]. The missing residues in the structure of N-Ras were modelled based upon homology to H-Ras. The complete protein was converted

to CG based on the Martini protein force field[26] and using the martinize-script[35]. Ligands were left out and the ternary structure was kept stable using an elastic network36. The C_α-root mean square deviation (RMSD) of the soluble protein domain (excluding the flexible HVR) never exceeds 0.3 nm for both H-Ras and N-Ras due to the elastic network. Palmitoyl anchors, using the standard Martini mapping for palmitoyl chains Na-C1-C1-C1-C1, were connected to cysteine 181 and 184 (H-Ras) or 184 (N-Ras) side chains using a harmonic bond with equilibrium distance r_0=0.39 nm and force constant f_c=5000 kJ mol^{-1} nm^{-2}. The C-terminal cysteine residue 186 was farnesilated. Our model of the farnesyl tail consisted of a linear sequence of three C3 beads, all connected via harmonic bonds with $r_0 = 0.49$ nm and f_c=8000 kJ mol^{-1} nm^{-2}. The angle bending is restricted by harmonic potentials with a 140° equilibrium angle and f_c=200 kJ mol^{-1}. The first C3 bead was connected to the cysteine side chain bead using a harmonic bond with $r_0 = 0.39$ nm, f_c=5000 kJ mol^{-1} nm^{-2}.

The Hedgehog protein structure was taken from the PDB-database (1VHH[36]). The three missing residues (Ser196, Gly197, Gly198) were modelled as a random coil. The atomistic structure was converted to CG based on the Martini protein force field[26] and using the martinize-script[35]. For the cholesterol anchor the standard Martini cholesterol parameters were used. The hydroxyl bead of the cholesterol was connected to the C-terminal glycine residue by a harmonic bond with $r_0 = 0.35$ thus creating two carboxylic acid groups, one on the side chain and one on the backbone C-terminus.

Topologies for the WALP23 (GW_2L-$(AL)_8$-W_2A) peptides[37] were also generated using the Martini protein parameters[26], thus constraining the α-helicity for the whole peptide. A glutamic acid is connected to the N-terminal glycine via a succimide moiety[38]. This reverses the N-to-C terminus direction of the glutamic acid backbone, thus creating two carboxylic acid groups, one on the side chain and one on the backbone C-terminus. Palmitoyl anchors are connected to both carboxylic acids (see Figure 5.1). Succinimide is modelled as a polar P5 particle and connected to the glycine and glutamate backbone beads by relatively weak harmonic bonds with $r_0 = 0.47$ nm, f_c=200 kJ mol^{-1} nm^{-2}. The palmitoyl tails are connected to the glutamate backbone and side chain bead by a harmonic bond with $r_0 = 0.34$ nm, and f_c=1250 kJ mol^{-1} nm^{-2}. No angle or dihedral potentials involving either the succinimide or glutamate beads were defined.

The LAT peptide (EADWLSPVGLGLLLLPFLVTLLAALCVRCRE, residues 232 of Murine LAT, with substitution Trp for Ala[4]) was modelled in a similar way as the WALP peptide. Palmitoyl anchors were connected to both cysteine residues side chains using harmonic bonds with $r_0 = 0.39$ nm, f_c=5000 kJ mol^{-1} nm^{-2}.

The GM1 ganglioside lipids were parameterized based upon the extension of the Martini model towards carbohydrates[39]. A detailed description of the parameterization of GM1 can be found in the supplementary information of [40].

Simulation parameters

All systems were simulated using the Gromacs MD package[41] (version 4.5.3). After minimizing the system energy by 500 steps steepest descent, the systems were simulated using a leap-frog integrator with a time step of 20 fs. A constant particle number, pressure and temperature (NpT) ensemble was applied. Pressure in the lateral (xy) and normal (z) dimensions was coupled separately to a 1 bar external bath with coupling time constant, $\tau_p = 3.0$ ps and compressibility, $\chi = 3.0 \cdot 10^{-5}$ bar^{-1}. Temperature is kept constant at 295 K by coupling to an external temperature bath with a coupling time constant, $\tau_T = 1.0$ ps. Three groups of molecules were coupled separately to avoid heat flow: water and ions, lipids and proteins, and cholesterol. Most systems were run for more than 10 μs, see Table 5.1 for total simulation times. Note that we report actual simulations times; due to the smoothing of the potential energy surface in CG models, the effective time is longer. For lipids and proteins

in the Martini model, the speed-up factor is about fourfold[42], i.e. 10 μs simulation time would correspond to 40 μs real time.

Analysis

The preferential partitioning of membrane components is calculated as the relative number of contacts of a lipid species with each of the other lipid constituents, corrected for the total number of lipids in the system:

$$p_A = \frac{c_A/n_A}{\sum_x c_x/n_x} \tag{5.1}$$

where p_A is the preferential partitioning with membrane component A, c_A the number of contacts with component A and n_A the number of molecules of component A. Using this formula random mixing would give an equal preferential partitioning with all components (e.g. 0.50 for a two component comparison). Note that Eq. 5.1 does not take into account the number of contacts per molecule of the different membrane constituents. Since in this chapter we only compare lipids of equivalent size, this effect will be negligible. Contacts were defined with respect to the GL1 and GL2 beads for lipids, all anchor beads for the soluble proteins and either the peptide with the anchor or just the anchor for the TM peptides. Two molecules were counted in contact if they were within 1.1 nm, roughly corresponding to the second solvation shell. The second solvation shell was chosen to obtain better statistics. Using only the first solvation shell (< 0.8 nm) gave comparable results. The Gromacs analysis tool g_mindist was used to calculate the number of contacts, analysing a trajectory frame every 1 ns. The first part of each of the simulations was omitted as equilibration phase. The system was considered equilibrated if the number of DPPC-DPPC contacts became constant, typically requiring between 2 and 5μs. Comparing the number of contacts between various components showed the number of DPPC-DPPC contacts to be in general a good indicator for the state of equilibration of the system. Density profiles were calculated by first centering the bilayer in the box based upon the last tail bead of all lipids over the course of the simulation. Subsequently the number density profiles were calculated using the Gromacs analysis tool g_density and normalized afterwards.

5.3 Results and discussion

We study the preferential partitioning of several bilayer constituents in a phase separated model membrane. Our reference system (mixture A) is a weakly phase separated membrane containing DPPC, DLiPC, and cholesterol in a 828:540:144 molar ratio, approximating 6:4:1. It consists of a liquid ordered (Lo) domain enriched in DPPC and cholesterol and a liquid disordered (Ld) domain enriched in DLiPC. The highly dynamic domain behavior and a line tension of 2 ± 2 pN of this system resemble the characteristics of *in vivo* systems[2, 20] and thus our system is very suitable to study the preferential partitioning of different bilayer constituents. The remainder of the results section is split into three parts. First, we describe simulations of the partitioning of peripheral proteins, which are anchored to the membrane by either palmitoyl and farnesyl anchors or a cholesterol anchor. Next we consider single helix transmembrane peptides, with two palmitoyl anchors. Finally, we look at the partitioning behavior of gangliosides and their ability to influence the partitioning of membrane lipids and transmembrane peptides. Table 5.1 gives an overview of the contents and total simulation time of the simulated systems. To quantify the preferential partitioning we calculated the normalized number of contacts with DPPC and DLiPC lipids. We compare against DPPC and DLiPC as they are the main constituents of the Lo and Ld phase, respectively, and serve as a marker for these phases.

Table 5.2: Normalized number of contacts p_A of bilayer systems containing peripheral proteins. The values for a system in one row are normalized (add up to one). Errors in p_A, based on block-averaging and assuming independent blocks over 500 ns, are of the order of 0.03.

	Reference		H-Ras		N-Ras	
	DPPC	DLiPC	DPPC	DLiPC	DPPC	DLiPC
DPPC	0.79	0.21	0.78	0.22	0.78	0.22
DLiPC	0.19	0.81	0.20	0.80	0.20	0.80
anchor			0.68	0.32	0.41	0.59

	H-Ras (no Pal.)		Hedgehog	
	DPPC	DLiPC	DPPC	DLiPC
DPPC	0.79	0.21	0.80	0.20
DLiPC	0.19	0.81	0.18	0.82
anchor	0.22	0.78	0.74	0.26

Partitioning of peripheral proteins into Lo domain mediated by lipid anchors

To investigate the partitioning of membrane anchored soluble proteins we added four copies of a Ras protein to the reference mixture. The Ras proteins are peripheral membrane proteins belonging to the family of GTPases and are involved in signal transduction pathways that control cell growth and proliferation[43]. Here we study both N-Ras and H-Ras (mixture B and D), which have a high sequence identity (>90%) over the 165 N-terminal residues (G-domain) and much lower sequence identity (10-15%) in the C-terminal[24] residues forming the HVR[44]. The membrane anchors are found in the HVR, and consist of a farnesyl tail at the terminal cysteine residue (Cys 186), and either one or two palmitoyl chains in N-Ras (Cys 181) and H- Ras (Cys 181 and 184), respectively[44], see Figure 5.1. We also simulated a system containing depalmitoylated H-Ras (mixture C), i.e. H-Ras with one farnesyl but no palmitoyl tails. Starting from a regular distribution over the membrane, the three different Ras proteins show distinct behavior. H-Ras is observed to partition into the Lo phase over a time scale of several μs, whereas N-Ras and depalmitoylated H-Ras have a preference for the Ld phase, as shown in Figure 5.3. The soluble G-domains form an aggregate within 5 μs of simulation time for all three proteins. Once the cluster has formed only minor reorientations of the monomers with respect to each other are observed. Table 5.2 shows the normalized number of contacts of the Ras proteins with DPPC and DLiPC. The addition of the anchored protein does not affect the (de)mixing of the DPPC and DLiPC lipids as visible from the normalized number of contacts which stay close to the corresponding values in the reference system. The H-Ras protein anchors have more contacts with the DPPC lipids, in other words they preferentially partition to the Lo phase consistent with the images shown in Figure 5.2. N-Ras and de-palmitoylated H-Ras have more contacts with DLiPC, most pronounced for the latter. Especially N-Ras can be considered as line-active, i.e. spending most time at the border between Lo and Ld domains (bottom row Figure 5.2).

Our data on partitioning of Ras proteins are in agreement with recent *in vitro* studies, showing a preference for N-Ras to accumulate at the domain boundaries45. K-Ras, which only has a farnesyl anchor in addition to a polybasic membrane unit, is found[45] to partition into the Ld domain consistent with our results for depalmitoylated H-Ras. Recent results from CG simulations by Janosi *et al.* [46] who studied the partitioning of the isolated Ras anchors, are consistent with our data, and confirm the antagonistic action of farnesyl and palmitoyl tails.

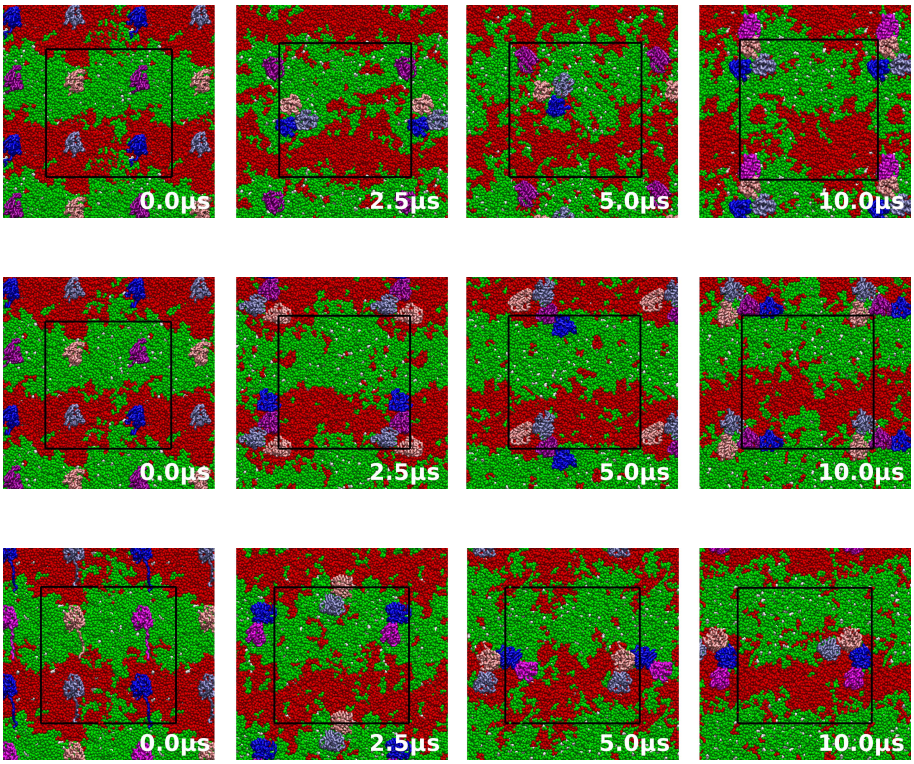

Figure 5.2: Anchor driven partitioning of Ras proteins into membrane domains. Top view snapshots from a simulation of four copies of the H-Ras (top), de-palmitoylated H-Ras (middle) and N-Ras (bottom) peripheral proteins in a DPPC:DLiPC:cholesterol lipid mixture. The different copies of the proteins are colored pink, purple, blue and light blue. DPPC, DLiPC and cholesterol are colored red, green and white, respectively. The black line shows the simulation box.

Interestingly, *in vivo* data[47, 48] show the partitioning of H-Ras to either the raft or non-raft phase to be regulated by additional factors. In particular, GTP loading (i.e. activation) segregates H-Ras from the raft phase to the non-raft phase while the GDP bound form (i.e. inactive) segregates to the Lo domain. This different behavior upon GTP binding is hypothesized to be caused by a structural change of the soluble N-terminal domain, coupling to a different orientation or conformation of the HVR to which the anchors are attached[49]. A change in membrane insertion of the Lo-phase-preferring palmitoyl anchors versus the, Ld-phase-preferring farnesyl anchors could thus be achieved, changing the relative propensity for either phase. Our observation that the GCP bound (activated) H-Ras partitions to the Lo phase, appears in contradiction with this. Apparently, in our simulations, the driving force provided by the lipid anchors is insufficiently modulated by the structural differences

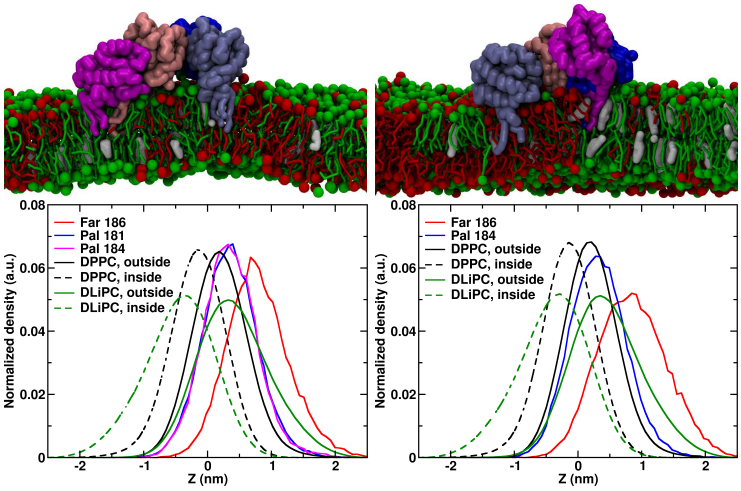

Figure 5.3: Positioning of palmitoyl and farnesyl anchors of Ras proteins inside the membrane. Top: sideview of the H-Ras (left) and N-Ras (right) protein complexes after 10 μs. The different copies of the proteins are colored pink, purple, blue and light blue. DPPC, DLiPC and cholesterol are colored red, green and white, respectively. Bottom: Normalized density profiles along the z-axis for H-Ras (left) and N-Ras (right) lipid anchors. Density profiles are plotted for the last tail beads of the farnesyl (red) and palmitoyl (blue and pink) anchors and DPPC (black) and DLiPC (green) lipids. Solid and dashed lines indicate the inner and outer monolayer, respectively.

due to activation or deactivation. This might be explained by several factors: first, the H-Ras structure used here was determined using GCP as a (inhibiting) ligand instead of the natural GTP. Possible small structural changes as a result of this might influence the partitioning behavior. Second, the HVR region is modelled as a random coil, which might be inappropriate for the activated state of Ras. Third, the relative partitioning of the palmitoyl versus farnesyl anchors might be depending on the overall lipid composition of the membrane.

To study the role of the lipid anchors and the HVR in more detail, we characterized the lateral position of the lipid anchors and fluctuations of the HVR region. For the anchors we analyzed their particle density along the bilayer normal. The result is shown in Figure 5.3 together with a graphical snapshot of the final configuration of the systems. When comparing the position of the last tail bead of the membrane lipids, it can be seen that the palmitoyl and farnesyl anchors are not inserted as deeply into the bilayer as DPPC lipids. This is true for both H-Ras and N-Ras. In other words the attachment of lipid anchors to the protein backbone slightly pulls the anchors out of the membrane. The farnesyl tail is inserted less deeply into the membrane compared to the palmitoyl anchors which can be attributed to the shorter length of the farnesyl unit. In all-atom simulations, Gorfe *et al.*[49] found the farnesyl tail of H-Ras to be inserted deeper into the membrane as compared to the palmitoyl tails, however in their model the farnesyl tail is modelled by a longer (saturated) hexadecyl unit, which affects the membrane insertion. The HVR does not adopt a stable, well defined structure, comparable to the work of Gorfe *et al.*[49] and experimental results[50]. The

fluctuations, measured by the pairwise RMSD (mean: 0.14, 0.13 and 0.18 nm for H-Ras, depalmitoylated H-Ras and N-Ras, respectively), are similar to those reported by Gorfe *et al.*[49] for the HVR with G-domain connected.

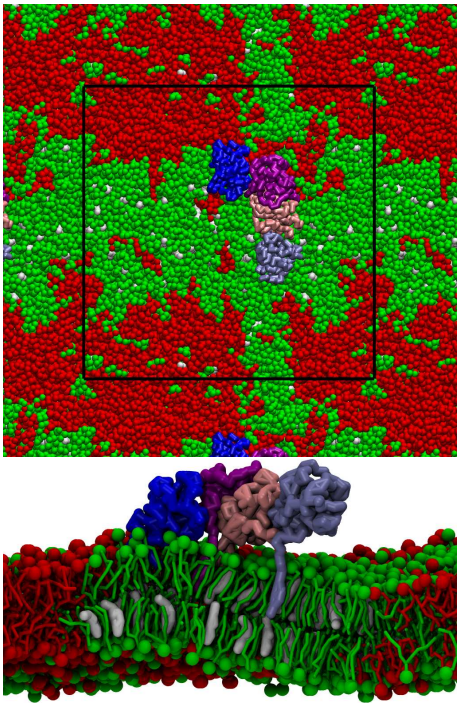

Figure 5.4: Anchor driven partitioning of Hedgehog proteins into the Lo domain. Top (left) and side (right) view of a simulation of 4 Hedgehog peripheral proteins in a DPPC:DLiPC:cholesterol lipid mixture after 12 μs. The different copies of the proteins are colored pink, purple, blue and light blue. DPPC, DLiPC and cholesterol are colored red, green and white, respectively. The black line shows the simulation box.

A much less abundant membrane anchoring mechanism is via the attachment of a sterol moiety, found in the Hedgehog protein family[36]. Although the complete form of the human Hedgehog protein is both C-terminally sterolated and N-terminally palmitoylated[51], it was found that the N-terminal sterol anchor is sufficient for the protein to partition to detergent-resistant membrane patches[52]. Here we study the partitioning of the N-terminally sterol anchored Hedgehog protein, by placing four protein copies in the reference mixture (mixture E, Table 5.1). Similar to the palmitoyl anchors of the Ras proteins, also the cholesterol anchor is able to drive the peripheral protein into the Lo domain. Figure 5.4 shows the final snapshots from the simulation, with the Hedgehog protein clearly residing in the Lo domain. Once in the Lo phase the proteins form a cluster and never enter the Ld phase again.

Analysis of the normalized number of contacts of the four cholesterol anchored Hedgehog proteins with the other membrane components, shown in Table 5.2, underlines the strong preference for saturated lipids. Partitioning of cholesterol-anchored Hedgehog into the Lo phase is in agreement with the experimentally known behavior[52]. Together, our data on peripheral membrane proteins show that lipid anchors provide a strong driving mechanism for selective partitioning. The anchors considered, either palmitoyl or sterol anchors, show a strong preference for the Lo phase in line with expectations. The ability of the farnesyl anchor of the Ras proteins to provide a counter force toward the Ld phase is also apparent from our simulations.

Lipid anchors also modulate partitioning behavior of TM peptides

Both the synthetic WALP transmembrane peptide and the linker for activation of T-cells (LAT) have been extensively studied with respect to domain partitioning. WALP peptides were specifically designed to study the behavior of TM proteins in lipid bilayers[53] and have been shown to partition into the Ld phase in model membranes, both with[38] and without[27] palmitoyl anchors. Palmitoylated LAT has been both reported to partition to

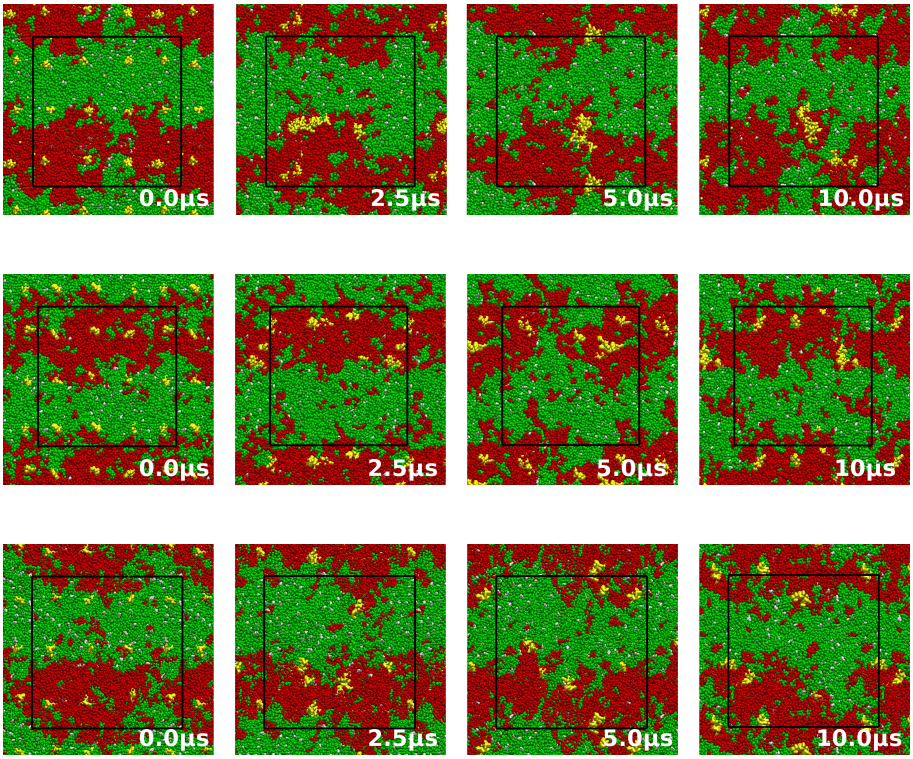

Figure 5.5: Anchor driven partitioning of TM peptides toward the Lo-Ld phase boundary. Top view of bilayer systems containing 12 LAT peptides (top) or 12 WALP peptides (middle) with anchors and 12 WALP peptides without anchors (bottom). Green is the saturated DPPC lipids, red the unsaturated DLiPC lipids, white is cholesterol, and yellow the TM peptides.

the Lo and Ld domain, depending on the method of preparation of the membrane system. In synthetically devised lipid mixtures palmitoylated LAT prefers the Ld phase[54], whereas in plasma membrane spheres, obtained from real cells, it prefers the Lo domain[8]. It has been hypothesized that this contrasting behavior in the differently prepared membranes systems can be attributed to a degree of order difference between the Lo and the Ld phase in the two systems. To study the behavior of these lipid anchored TM peptides, we added either 12 WALP23 or 12 LAT peptides with two palmitoyl chains each to the reference mixture (mixtures F,H, Table 5.1). For comparison, a system containing WALP without palmitoyl anchors was also simulated (mixture G). Figure 5.5 shows snapshots from these simulations. Both LAT and WALP peptides are segregated out of the bulk of the Lo phase within 1 μs. Once the peptides have segregated from the Lo phase, they show very different behavior. For the WALP peptides the behavior is very dynamic. First, they form peptide-peptide

clusters that constantly form and break up again, consistent with earlier observations[27]. Monomers, dimers and trimers are observed throughout the simulation. Second, the WALP peptides do not remain in the bulk of the Ld phase but often reside close to the domain boundary and in transiently formed lipid peninsulas or lipid islands. The addition of the transmembrane peptides does not notably change the mixing of the lipids (Table 5.3). The normalized number of contacts indicates a strong preference of the WALP peptide for the Ld phase, but not as strong as in the case of WALP without anchors. To study the orientation of the peptides we calculated the number of contacts of the anchor to DPPC and DLiPC. The preference for the Ld phase remains, although less strong, indicative of an orientation, and consecutive dragging force, of the (saturated) lipid anchors towards the Lo phase. In experiments[38] on artificial lipid mixtures, single or double palmitoylation is not sufficient to bring WALP into the Lo phase, matching our current results.

Table 5.3: Normalized number of contacts p_A of bilayer systems containing TM peptides. The values for a system in one row are normalized (add up to one). Errors in p_A, based on block-averaging and assuming independent blocks over 500 ns, are of the order of 0.03.

	Reference		WALP	
	DPPC	DLiPC	DPPC	DLiPC
DPPC	0.79	0.21	0.80	0.20
DLiPC	0.19	0.81	0.19	0.81
anchor			0.40	0.60
peptide			0.22	0.78
	LAT		WALP (no Pal.)	
	DPPC	DLiPC	DPPC	DLiPC
DPPC	0.79	0.21	0.80	0.20
DLiPC	0.19	0.81	0.19	0.81
anchor	0.67	0.33		
peptide	0.27	0.73	0.09	0.91

In contrast to WALP, the LAT peptides form a constantly growing cluster: once a peptide binds to a partner, they never break up again (Figure 5.5). The lipid anchors of the clustered peptides are oriented towards the same side of the cluster and towards the Lo domain, but the peptides still reside in the Ld domain. This leads to a higher normalized number of contacts of the anchors to DPPC compared to the combination of peptides and anchors (Table 5.3). The LAT peptides show a slightly higher normalized number of contacts to DPPC lipids compared to WALP, but the preferred contact is with the unsaturated lipids found in the Ld region. The partitioning of palmitoylated LAT into the Ld phase agrees with the experimental results using artificial membranes[54], and not those in plasma membrane spheres[8]. This indicates that the difference in order between the Lo and Ld domains in the current mixture is larger than that of real membranes, assuming that the plasma spheres more closely resemble the *in vivo* situation. We conclude that the palmitoylation of TM peptides indeed provides a driving force toward the more ordered membrane domains. Two saturated lipid anchors, however, are not sufficient to allow the peptides to dissolve into the Lo domains. The orientation of the lipid anchors towards the Lo domains endorses the important role of lipid anchors; in fact the anchored peptides show linactant behaviori[55], residing most of the time at the domain boundaries.

Gangliosides are able to shuttle proteins into the Lo phase

Ganglioside lipids are strongly amphiphilic lipids due to their large carbohydrate head group and their ceramide based lipid tails. They are enriched in detergent resistant membrane domains and are thought to play an important role in the formation of those domains[12, 13, 20]. Here we investigate the effect of GM1 gangliosides (see Figure 5.1) on membrane domain formation and their interaction with TM peptides. We first looked at the partitioning of GM1 lipids between the Lo and Ld domains of our ternary lipid system. In order to do so

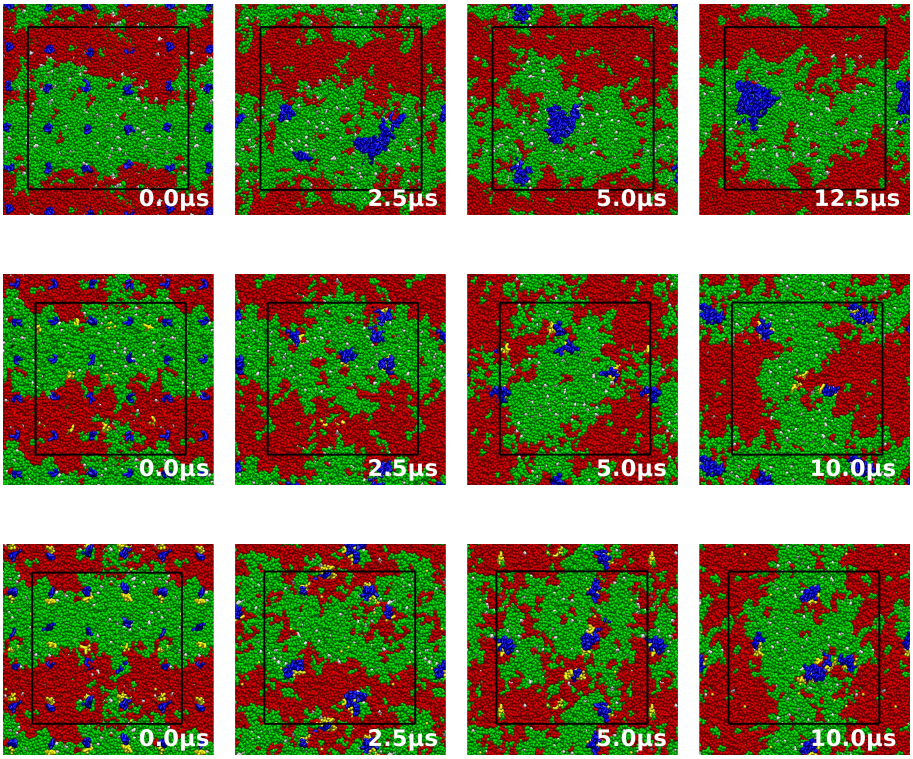

Figure 5.6: Ganglioside mediated shuttling of TM peptides toward the Lo domains. Top view of bilayer systems containing 4 mol% GM1 lipids (top), 4 mol% GM1 lipids and 12 WALP peptides (middle) or 4mol% GM1 and 12 LAT peptides (bottom). Green is used to depict the saturated DPPC lipids, red the unsaturated DLiPC lipids, white the cholesterols, blue the GM1 and yellow the TM peptides.

we added 4 mol% GM1 lipids to our reference mixture (mixture I). The top row of Figure 5.6 illustrates the time evolution of this system. Placed initially on a homogeneous grid, the GM1 lipids are observed to aggregate into a single domain in about 10 μs. These nanodomains are registered over both monolayers and dynamic, i.e., the gangliosides remain in a fluid state. To quantify the extent of mixing in this quarternary membrane mixture, the normalized number of contacts of the different system components with respect to DPPC (Lo) and DLiPC (Ld) were calculated. The data are gathered in Table 5.4, from which a number of conclusions can be drawn. First, the extent of demixing of DPPC and DLiPC does not change significantly upon addition of the small amount of GM1 (compare to reference mixture in Table 5.2). Second, GM1 shows a very strong preference for DPPC lipids over DLiPC lipids, implying GM1 associates with the Lo lipids; this is also evident from the snapshots in Figure 5.6. Third, when preferential partitioning parameters are calculated

Table 5.4: Normalized number of contacts p_A of lipids in GM1 containing bilayers. The values for a system in one row are normalized (add up to one). Errors in p_A, based on block-averaging and assuming independent blocks over 500 ns (No peptide) or based on the differences between two simulations (WALP and LAT) are of the order of 0.03.

	No Peptides				
	DPPC	DLiPC	DPPC	DLiPC	GM1
DPPC	0.78	0.22	0.49	0.14	0.37
DLiPC	0.19	0.81	0.18	0.78	0.04
GM1	0.94	0.06	0.06	0.0	0.94
cholesterol			0.30	0.05	0.65
peptide					
anchor					
	WALP				
GM1	DPPC	DLiPC	DPPC	DLiPC	GM1
DPPC	0.78	0.22	0.57	0.16	0.27
DLiPC	0.20	0.80	0.17	0.68	0.15
GM1	0.66	0.34	0.07	0.03	0.90
cholesterol			0.50	0.09	0.41
peptide	0.44	0.56	0.05	0.05	0.90
anchor	0.65	0.35	0.05	0.05	0.90
	LAT				
	DPPC	DLiPC	DPPC	DLiPC	GM1
DPPC	0.77	0.23	0.54	0.17	0.29
DLiPC	0.21	0.79	0.19	0.68	0.13
GM1	0.69	0.31	0.06	0.03	0.91
cholesterol			0.39	0.09	0.52
peptide	0.34	0.66	0.03	0.07	0.90
anchor	0.61	0.39	0.07	0.04	0.90

with respect to three groups, DPPC, DLiPC and GM1, GM1 shows the strongest preference for self-association, consistent with the formation of a nano-domain. Finally, the number of contacts for cholesterol indicates an enrichment of the GM1 domain in cholesterol with respect to the surrounding Lo domain. Experimental data on the phase behavior of GM1 containing membranes are limited (see recent reviews[12, 13]). Besides, interpretation of these data is not straightforward due to the fact that GM1 lipids can have either saturated and unsaturated tails next to the sphingosine moiety. In our simulations, we use a fully saturated tail. Experimentally, fully saturated GM1 shows a high affinity for Lo phases, in line with our results. Moreover, there is growing evidence[12, 13] that GM1 can form laterally separated Lo nano-domains, and this corroborates our findings. Based on measurements on mixtures of GM1, sphingomyelin (SM), and cholesterol, a depletion of cholesterol from the ganglioside domain was concluded[56]. This is in contrast to our observation, which indicates a small but significant cholesterol enrichment of the GM1 domain. The difference might be explained by the difference in lipid mixtures. SM lipids are known to have a higher affinity for cholesterol than DPPC lipids do, therefore cholesterol depletion of GM1 domains is expected with respect to SM but not DPPC. Realizing that GM1 lipids also have a ceramide backbone, the condensation of GM1 and cholesterol into a single domain seems rather plausible.

Next we address the question to which extent GM1 can have an effect on the partitioning of TM peptides. To do so, we added 4 mol% GM1 to the membranes containing 12 doubly palmitoylated WALP or LAT peptides (mixtures J,K, Table 5.1). Snapshots from these

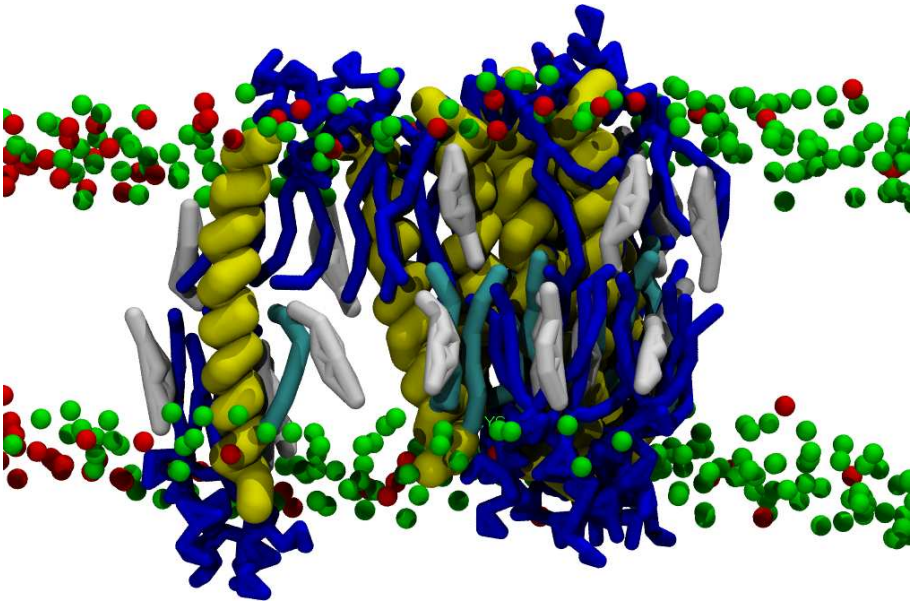

Figure 5.7: Molecular view of a raft-embedded nano-domain composed of gangliosides and LAT peptides. The snapshot shows a cluster of GM1 lipid (blue) and LAT peptides (yellow) with lipid anchors (cyan). Cholesterol molecules are shown in white. Of the DPPC and DLiPC lipids only the head groups are shown as green and red spheres, respectively.

simulations are shown in Figure 5.6, together with the GM1 containing membrane without peptides discussed above. Initially, the peptides and GM1 lipids were added on overlapping grids. Due to the larger number of GM1 lipids (16) compared to peptides (12), this placement results in some close peptide-GM1 contacts right from the beginning, as well as some isolated peptides and gangliosides. Within a few μs, however, almost all of the peptides and GM1 lipids are binding to each other, forming small, mixed clusters which merge to form larger clusters. The mixtures with WALP or LAT peptides show comparable behavior. At the end of the simulations, no clusters containing only GM1 lipids or TM peptides are observed. It is likely that the simulations have not yet reached equilibrium at this stage, and that eventually the few remaining clusters at 10 μs will coalesce into one nano-domain. In contrast to the GM1 clusters in the systems containing no peptides, the clusters are depleted of cholesterol (with respect to the Lo domain), but do contain DLiPC, as can be inferred from the increased normalized number of contacts for GM1 and cholesterol with the other constituents (Table 5.4). Closer inspection of the organization of the GM1/peptide clusters reveals a tendency to remain at the Lo/Ld domain boundary, which explains the apparent substitution of cholesterol for unsaturated lipids. The presence of the GM1/peptide clusters also leads to slightly more mixing of DPPC and DLiPC, which again would point to linactant behavior. Comparing systems that contain 4 mol% GM1 to systems without GM1 (cf. Tables 5.2 and 5.3), both WALP and LAT peptides show a higher affinity for the DPPC rich

Lo domain in presence of the gangliosides. This is in support of the hypothesized role of GM1 as helping proteins to sort into Lo domains[11, 12, 13].

Taken together, our results underline the important role that gangliosides in general, and GM1 in particular, may have in the overall organization of the cell membrane. GM1 seems to have the remarkable combination of properties that i) makes them readily partition into Lo domains, and ii) allows them to bind to TM peptides. These features provide them with the ability to act as nano-shuttles for sorting of membrane embedded proteins into ordered membrane regions. Although we did not attempt to further unravel the driving forces for the GM1-peptide association, it appears that the large oligosaccharide head group plays an important role in this process the membrane embedded WALP or LAT peptides leave room for the GM1 head group, if it were, to embrace the peptides. The organization of the GM1 induced nano-domain is illustrated in Figure 5.7 by a snapshot from the simulation involving LAT peptides, showing the ganglioside coat around the TM helices. Whether or not a similar embracement can take place with larger proteins remains to be seen. It would also be interesting to investigate what effect a higher concentration of cholesterol has on the GM1 conformation in our model, and consequently on the binding of GM1 to membrane proteins, since it is known that cholesterol effects both the head group orientation and binding of soluble proteins[57]. However, this is outside the scope of the current study.

Limitations

For a proper interpretation of the results described in the preceding sections, it is important to discuss some of the limitations of our model. An important simplification is the fixation of secondary protein structure in the Martini protein model[26]. Due to the limited resolution of the protein backbone, realistic folding is out of reach and the secondary structure is restrained to that of the initial (e.g. crystal) structure. In the current application, this limitation might have affected the partitioning behavior of the Ras proteins, which are anchored in the membrane via a flexible loop (the HVR, see Figure 5.1). As the structure of this loop is not resolved we modelled it as a random coil, a questionable assumption. Besides, activation of the Ras protein may affect the secondary structure of this loop *in vivo*, but this we obviously cannot take into account. A similar restriction applies to the conformational flexibility of the GM1 oligosaccharide head group, which was restrained using an elastic network as explained in the supporting material. A further point of concern is the oligomerization we observe in case of the soluble proteins. Both Ras variants as well as the Hedgehog protein form linear aggregates that do not dissociate once formed. Although it is not clear from experimental data to what extent this clustering is realistic or not, it could be that the Martini model overestimates the binding affinity of soluble proteins. We recently showed that, on the level of individual amino acids, dimerization free energies of amino acid side chain pairs are well reproduced in comparison to atomistic data[58]. This would point to a collective effect arising between protein surfaces, possibly related to the dewetting of these surfaces by the CG water beads which are large compared to real water molecules (recalling that a CG water represents four real waters). Eventually, some of the results obtained here with the Martini model will have to be checked using atomistic models. Considering the three orders of magnitude speed-up of the CG model with respect to atomistic models, this is not yet possible due to computational limitations.

Another point worth discussing is our choice of membrane composition. We face two challenges: on the one hand we need a lipid mixture that undergoes strong enough phase separation to result in well distinguishable Lo and Ld domains on a length scale of $\sim$20 nm, the size of our simulation box. On the other hand we like to resemble the *in vivo* situation, which is closer to a near critical mixture characterized by fluctuating nano-domains rather than macroscopic phase separation[2]. The mixture of DPPC, DLiPC, and cholesterol at

approximately 6:4:1 ratio used in this study fulfils these criteria it is weakly phase separating resulting in distinct regions of Lo and Ld domains that are yet very dynamic and able to rearrange on the sub-microsecond time scale of our simulations. Still, the partitioning of lipid anchored proteins in such a mixture will be determined by the difference in chemical potential of these proteins in either the Lo or Ld phase, and this may depend critically on the composition of the respective domains. Systematic exploration of the partitioning behavior of membrane proteins as a function of lipid composition is clearly needed. To arrive at a more realistic description of compartmentalization in real membranes, a few more steps are required. The current model membranes are yet lacking the complex composition of *in vivo* plasma membranes. Some important differences are: i) the asymmetry between the inner and outer monolayers is lacking, ii) the use of DPPC as saturated lipid, instead of sphingomyelin found *in vivo*, iii) the lack of minor membrane constituents, both lipids and proteins, and iv) the lack of cytoskeleton interacting with the membrane. We are currently working on a complex, more realistic lipid-protein mixture to further study the organizational principles in cell membranes.

5.4 Conclusion

We have simulated the partitioning behavior of several membrane constituents that are thought to be involved, or even play an important role, in the domain formation of the plasma membrane. We find that peripheral Ras proteins partition to the Lo or Ld phase depending on the type of anchor. Double palmitoylated H-Ras prefers the Lo phase, whereas single palmitoylated N-Ras resides at the domain boundaries. Depalmitoylated H-Ras, with only a farnesyl anchor remaining, partitions into the Ld phase. Our results for the Hedgehog protein show that a single cholesterol anchor is sufficient to bring the protein in the Lo phase. Transmembrane peptides partition to the Ld phase, but saturated lipid anchors drive them toward the Lo-Ld domain boundary. Addition of GM1 ganglioside lipids does decrease the preference of the transmembrane peptides for the Ld even further. Interestingly, we find that the GM1 and peptides are capable of forming small nano-domains with high affinity for the Lo phase. This indicates that GM1 might play an important role in the recruiting of transmembrane proteins to membrane rafts.

5.5 Bibliography

[1] L.A Bagatolli, J.H Ipsen, A.C Simonsen, and O.G Mouritsen. An outlook on organization of lipids in membranes: searching for a realistic connection with the organization of biological membranes. *Prog. Lipid Res.*, 49:378–89, 2010.

[2] D Lingwood and K Simons. Lipid rafts as a membrane-organizing principle. *Science*, 327:46–50, 2010.

[3] G Van Meer, D.R Voelker, and G.W Feigenson. Membrane lipids: where they are and how they behave. *Nat. Rev. Mol. Cell biol.*, 9:112–124, 2008.

[4] K Simons and D Toomre. Lipid rafts and signal transduction. *Nat. Rev. Mol. Cell Biol.*, 1:31–41, 2000.

[5] T Baumgart, A.T Hammond, P Sengupta, S.T Hess, D.A Holowka, B.A Baird, and W.W Webb. Large-scale fluid/fluid phase separation of proteins and lipids in giant plasma membrane vesicles. *P. Natl. Acad. Sci. USA*, 104:3165, 2007.

[6] N Kahya, D Scherfeld, K Bacia, B Poolman, and P Schwille. Probing lipid mobility of raft-exhibiting model membranes by fluorescence correlation spectroscopy. *J. Biol. Chem.*, 278:28109–28115, 2003.

[7] S.L Veatch, I.V Polozov, K Gawrisch, and S.L Keller. Liquid domains in vesicles investigated by nmr and fluorescence microscopy. *Biophys. J.*, 86:2910–2922, 2004.

[8] I Levental, D Lingwood, M Grzybek, Ü Coskun, and K Simons. Palmitoylation regulates raft affinity for the majority of integral raft proteins. *P. Natl. Acad. Sci. USA*, 107:22050–22054, 2010.

[9] F Contreras, A Ernst, P Haberkant, P Björkholm, E Lindahl, B Gönen, C Tischer, A Elofsson, G Von Heijne, C Thiele, R Pepperkok, F Wieland, and B Brügger. Molecular recognition of a single sphingolipid species by a protein/'s transmembrane domain. *Nature*, 481:525–529, 2012.

[10] B Thaa, I Levental, A Hermann, and M Veit. Intrinsic membrane association of the cytoplasmic tail of influenza virus m2 protein and lateral membrane sorting regulated by cholesterol binding and palmitoylation. *Biochem. J.*, 437:389–397, 2011.

[11] C Yuan and L.J Johnston. Distribution of ganglioside gm1 in l-a-dipalmitoylphosphatidylcholine/cholesterol monolayers: A model for lipid rafts1. *Biophys. J.*, 79:2768–2781, 2008.

[12] A Prinetti, N Loberto, V Chigorno, and S Sonnino. Glycosphingolipid behaviour in complex membranes. *Biochim. Biophys. Acta*, 1788:184193, 2009.

[13] B Westerlund and J Slotte. Differential ability of cholesterol-enriched and gel phase domains to resist benzyl alcohol-induced fluidization in multilamellar lipid vesicles. *Biochim. Biophys. Acta*, 24542461:1788, 2009.

[14] S Sonnino, L Mauri, V Chigorno, and A Prinetti. Gangliosides as components of lipid membrane domains. *Glycobiology*, 17:1R–13R, 2006.

[15] R.K Yu, Y.T Tsai, T Ariga, and M Yanagisawa. Structures, biosynthesis, and functions of gangliosides-an overview. *J. Oleo Sci.*, 60:537–544, 2011.

[16] N Ichikawa, K Iwabuchi, H Kurihara, K Ishii, T Kobayashi, T Sasaki, N Hattori, Y Mizuno, K Hozumi, Y Yamada, and E Arikawa-Hirasawa. Binding of laminin-1 to monosialoganglioside gm1 in lipid rafts is crucial for neurite outgrowth. *J. Cell Sci.*, 122:289–299, 2009.

[17] R.G Zhang, D.L Scott, M.L Westbrook, S Nance, B.D Spangler, G.G Shipley, and E.M Westbrook. The three-dimensional crystal structure of cholera toxin. *J. Mol. Biol.*, 251:563–573, 1995.

[18] T Mitsuda, K Furukawa, S Fukumoto, H Miyazaki, T Urano, and K Furukawa. Overexpression of ganglioside gm1 results in the dispersion of platelet-derived growth factor receptor from glycolipid-enriched microdomains and in the suppression of cell growth signals. *J. Biol. Chem.*, 277:11239–11246, 2002.

[19] N Kawashima, S Yoon, K Itoh, and K.I Nakayama. Tyrosine kinase activity of epidermal growth factor receptor is regulated by gm3 binding through carbohydrate to carbohydrate interactions. *J. Biol. Chem.*, 284:6147–6155, 2009.

[20] H.J Kaiser, D Lingwood, I Levental, J.L Sampaio, L Kalvodova, L Rajendran, and K Simons. Order of lipid phases in model and plasma membranes. *P. Natl. Acad. Sci. USA*, 106:16645, 2009.

[21] B Reynwar and M Deserno. Membrane composition-mediated protein-protein interactions. *Biointerphases*, 3:FA117–FA124, 2008.

[22] F.J.M de Meyer, M Venturoli, and B Smit. Molecular simulations of lipid-mediated protein-protein interactions. *Biophys. J.*, 95:1851–1865, 2008.

[23] X Periole, T Huber, S.J Marrink, and T.P Sakmar. G protein-coupled receptors self-assemble in dynamics simulations of model bilayers. *J. Am. Chem. Soc*, 129:10126–10132, 2007.

[24] D.L Parton, J.W Klingelhoefer, and M.S.P Sansom. Aggregation of model membrane proteins, modulated by hydrophobic mismatch, membrane curvature, and protein class. *Biophys. J.*, 101:691–699, 2011.

[25] S.J Marrink, H.J Risselada, S Yefimov, D.P Tieleman, and A.H De Vries. The martini force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 111:7812–7824, 2007.

[26] L Monticelli, S.K Kandasamy, X Periole, R.G Larson, D.P Tieleman, and S.J Marrink. The martini coarse-grained force field: extension to proteins. *J. Chem. Theory Comput.*, 4:819–834, 2008.

[27] L.V Schäfer, D.H de Jong, A Holt, A.J Rzepiela, A.H de Vries, B Poolman, and J.A Killian S.J Marrink. Lipid packing drives the segregation of transmembrane helices into disordered lipid domains in model membranes. *P. Natl. Acad. Sci. USA*, 108:1343–1348, 2011.

[28] J Domański, S.J Marrink, and L.V Schäfer. Transmembrane helices can induce domain formation in crowded model membranes. *BBA - Biomembranes*, 1818:984–994, 2012.

[29] G von Heijne. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J.*, 5:3021–3027, 1986.

[30] E Sparr, W.L Ash, P.V Nazarov, D.T.S Rijkers, M.A Hemminga, D.P Tieleman, and J.A Killian. self-association of transmembrane -helices in model membranes. *J. Biol. Chem.*, 280:39324–39331, 2005.

[31] H.J Risselada and SJ Marrink. The molecular face of lipid rafts in model membranes. *P. Natl. Acad. Sci. USA*, 105:17367, 2008.

[32] U Krengel. *Struktur und Guanosintriphosphat-Hydrolysemechanismus des C-terminal verkrzten menschlichen Krebsproteins p21H-ras.* PhD thesis, University of Heidelberg, 1991.

[33] A Wernimont and A Edwards. In situ proteolysis to generate crystals for structure determination: An update. *PLoS ONE*, 4:e5094, 2009.

[34] LLC Schrödinger. The pymol molecular graphics system, version 1.2r1.

[35] D.H de Jong, W.F.D Bennet, G Singh, C Arnarez, T.A Wassenaar, L.V Schäfer, X Periole, D.P Tieleman, and S.J Marrink. Improved protein side-chain parameters for the martini coarse-grained force field. *J. Chem. Theory Comput.*

[36] J.A Porter, K.E Young, and P.A Beachy. Cholesterol modification of hedgehog signaling proteins in animal development. *Science*, 274:255–259, 1996.

[37] J.A Killian, I Salemink, M.R.R de Planque, G Lindblom, R.E Koeppe II, and D.V Greathouse. Induction of nonbilayer structures in diacylphosphatidylcholine model membranes by transmembrane -helical peptides: importance of hydrophobic mismatch and proposed role of tryptophans. *Biochemistry*, 35:1037–1045, 1996.

[38] B.Y van Duyl, D.T.S Rijkers, B de Kruijff, and J.A Killian. Influence of hydrophobic mismatch and palmitoylation on the association of transmembrane [alpha]-helical peptides with detergent-resistant membranes. *FEBS Lett.*, 523:79–84, 2002.

[39] C.A López, A.J Rzepiela, A.H De Vries, L Dijkhuizen, P.H Hünenberger, and S.J Marrink. Martini coarse-grained force field: Extension to carbohydrates. *J. Chem. Theory Comput.*, 5:3195–3210, 2009.

[40] D.H de Jong, C.A Lopez, and S.J Marrink. Molecular view on protein sorting into liquid-ordered membrane domains mediated by gangliosides and lipid anchors. *Faraday Discuss.*, 2013.

[41] B Hess, C Kutzner, D van der Spoel, and E Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4:435–447, 2008.

[42] S Ramadurai, A Holt, L.V Schäfer, V Krasnikov, and B Poolman. Influence of hydrophobic mismatch and amino acid composition on the lateral diffusion of transmembrane peptides. *Biophys. J.*, 99:1482–1489, 2010.

[43] K Wennerberg, K.L Rossman, and C.J Der. The ras superfamily at a glance. *J. Cell Sci.*, 118:843–846, 2005.

[44] I Prior and J.F Hancock. Ras trafficking, localization and compartmentalized signalling. *Semin. Cell Dev. Biol.*, 23:145–153, 2011.

[45] K Weise, S Kapoor, C Denter, J Nikolaus, N Opitz, S Koch, G Triola, A Hermann, H Waldmann, and R Winter. Membrane-mediated induction and sorting of k-ras microdomain signaling platforms. *J. Am. Chem. Soc.*, 113:880–887, 2011.

[46] L Janosi, Z Li, J Hancock, and A.A Gorfe. Organization, dynamics, and segregation of ras nanoclusters in membrane domains. *P. Natl. Acad. Sci. USA*, 109:8097–8102, 2012.

[47] I.A Prior, A Harding, J Yan, J Sluimer, R.G Parton, and J.F Hancock. Gtp-dependent segregation of h-ras from lipid rafts is required for biological activity. *Nat. Cell Biol.*, 3:368–375, 2001.

[48] H Niv, O Gutman, Y Kloog, and Y.I Henis. Activated k-ras and h-ras display different interactions with saturable nonraft sites at the surface of live cells. *J. Cell Biol.*, 157:865–872, 2002.

[49] A.A Gorfe, M Hanzal-Bayer, D Abankwa, J.F Hancock, and J.A McCammon. Structure and dynamics of the full-length lipid-modified h-ras protein in a 1, 2-dimyristoylglycero-3-phosphocholine bilayer. *J. Med. Chem.*, 50:674–684, 2007.

[50] R Thapar, J Williams, and S.L Campell. Nmr characterization of full-length farnesylated and non-farnesylated h-ras and its implications for raf activation. *J. Mol. Biol.*, 343:1391–1408, 2004.

[51] R.B Pepinsky, C Zeng, D Wen, P Rayhorn, D.P Baker, K.P Williams, S.A Bixler, C.M Ambrose, E.A Garber, K Miatkowski, F.R Taylor, E.A Wang, and A Galdes. Identification of a palmitic acid-modified form of human sonic hedgehog. *J. Biol. Chem.*, 273:14037–14045, 1998.

[52] A Rietveld, S Neutz, K Simons, and S Eaton. Association of sterol-and glycosylphosphatidylinositol-linked proteins with drosophila raft lipid microdomains. *J. Biol. Chem.*, 274:12049–12054, 1999.

[53] M.R.R de Planque and J.A Killian. Protein-lipid interactions studied with designed transmembrane peptides: role of hydrophobic matching and interfacial anchoring (review). *Mol. Mem. Biol.*, 20:271–284, 2003.

[54] H Shogomori, A Hammond, A.G Ostermeyer-Fay, D.J Barr, G.W Feigenson, E London, and D.A Brown. Palmitoylation and intracellular domain interactions both contribute to raft targeting of linker for activation of t cells. *J. Biol. Chem.*, 280:18931–18942, 2005.

[55] S Trabelsi, S Zhang, T.R Lee, and D.K Schwartz. Linactants: surfactant analogues in two dimensions. *Phys. Rev. Lett.*, 100:1–4, 2008.

[56] A Ferraretto, M Pitto, P Palestini, and M Masserini. Lipid domains in the membrane: thermotropic properties of sphingomyelin vesicles containing gm1 ganglioside and cholesterol. *Biochemistry*, 26:9232–9236, 1997.

[57] D Lingwood, B Binnington, T RóG, I Vattulainen, M Grzybek, U Coskun, C.A Lingwood, and K Simons. Cholesterol modulates glycolipid conformation and receptor activity. *Nat. Chem. Biol.*, 7:260–262, 2011.

[58] D.H de Jong, X Periole, and S.J Marrink. Dimerization of amino acid side chains: lessons from the comparison of different forcefields. *J. Chem. Theory Comput.*, 8:1003–1014, 2012.

6

Parameterization of thylakoid cofactors at atomistic and coarse grain resolution

Abstract

The thylakoid membrane, responsible for the conversion of sunlight to chemical energy in green plants, algae and cyano bacteria, is a complex mixture of special lipids, proteins and a number of different cofactors. Many of the processes of interest in the thylakoid membrane are on length scales that are difficult to study experimentally, but accessible using (coarse grain) molecular dynamics simulations. In this chapter we parameterize the most important cofactors, namely plastoquinone, plastoquinol, heme b, chlorophyll A , pheophytin and β-carotene. The parameters obtained at atomistic resolution are consistent with the Gromos force field and at coarse grain resolution with the Martini force field. The atomistic parameters are based upon the building blocks available for the Gromos force field and *ab initio* calculations. Simulations using the atomistic topologies serve as the basis for the coarse grain model. Thermodynamic properties of the molecules using the topologies are compared to experimental values when available. Finally, the cofactor topologies are tested, either in a DPPC or thylakoid membranes, or in complex with the heme binding protein bacterioferritin.

6.1 Introduction

Most life on earth is dependent on sunlight for its energy. The conversion of sunlight into chemical energy happens largely in the chloroplasts of green plants, algae and in cyano bacteria. In these organisms the conversion is taking place in the thylakoid membrane: a densely packed membrane mostly consisting of thylakoid lipids[1] and four large protein complexes. The complexes, Photosystem I and II (PSI and PSII), cytochrome b_6f and F-ATPase perform the light conversion. PSII absorbs photons to oxidize water and reduce the cofactor plastoquinone into plastoquinol. The plastoquinol is reoxidized by cytochrome b_6f, which uses the released energy to reduce plastocyanin. The reduced plastocyanin is reoxidized by PSI. The released energy, together with the energy from the uptake of another photon by PSI, is used to reduce ferredoxin, which is released to the cell stroma. During these steps a proton gradient is created over the membrane, which is ultimately used by F-ATPase to create ATP[2].

The whole process is a complex interplay between many proteins, the thylakoid membrane and a large number of cofactors. In the remainder of the chapter we will focus on PSII and

more specifically the cofactors that are involved in its function. In recent years several high resolution structures of PSII have been published[3, 4], which opened the way for molecular dynamics (MD) simulation studies to investigate its inner dynamics[5, 6, 7, 8]. However, given the current computational power the large size of the complex makes the study of the system at time scales relevant for most processes well beyond the reach of standard atomistic computer simulations. A possible way to reduce the computational cost for molecular simulations is the use of coarse grain (CG) models. In those models the resolution is reduced as compared to atomistic models by grouping together several atoms in CG interaction sites. A reduced number of interaction sites leads to a lower number of calculations performed and hence faster simulations (see also Chapter 1).

In this study the Martini CG force field is used[9]. The Martini force field applies a mapping of four heavy atoms to one coarse grain bead and is mainly parametrized against thermodynamic data such as partitioning free energies between polar and apolar phases. For the Martini force field, parameters are already available for proteins[10], simple lipids[9] and specific thylakoid lipids[11]. Here we aim to obtain parameters for a number of cofactors found in PSII and the thylakoid membrane: plastoquinone (PL9one), plastoquinol (PL9ol), heme (HEM), chlorophyll A (CLA), pheophytin (PHO) and β-carotene (BCR) (see Figure 6.1 and 6.2).

Atomistic parameters for some of the cofactor molecules compatible with the OPLS or AMBER force fields have been published before [12, 13, 14, 15, 16, 17]. Here we derive a set of parameters consistent with the Gromos force field[18], for the same cofactors except for heme; parameters for heme are already available for Gromos[18]. The Gromos parameters form a basis for the parametrization of the CG parameters and can be used in atomistic and multiscale simulations[19].

In the next section the protocols used to obtain and test both the united atom Gromos and the coarse grain Martini models are described. Next, the final topologies are described and critical steps in the parameterization process are detailed. Results derived with the new topologies are discussed in section 6.3. Finally, a summary concludes this chapter.

6.2 Methods

The parametrization was done in a two step process: first, atomistic topologies consistent with the Gromos force field were obtained. Second, coarse grain topologies for the Martini force field were obtained, partially based on the atomistic parameters. Both steps will be described separately.

Simulation parameters

All simulations were run using the Gromacs simulation package[20], version 4.5.x. For the atomistic simulations a timestep of 2 fs was applied and the coordinates and system energies were written out every 10 ps. A twin-range cut-off scheme was used, where the neighborlist extended to 1.4 nm and was updated every 5 steps. Electrostatic interactions were treated using the reaction-field algorithm[21] with ϵ_{rf}=2, van der Waals interactions were cut-off after 1.4 nm. Temperature and pressure were kept constant at 303 K and 1.0 bar by coupling to an external bath[22] with coupling parameters τ_{t} and τ_{p} being 0.1 and 0.5 ps^{-1}, respectively. All bonds were constraint using the LINCS algorithm[23, 24]. Atomistic trajectories were mapped to CG trajectories using the reverse transformation tool[25].

For the CG simulations similar parameters as for the atomistic simulations were used, except for the van der Waals interactions which were smoothly shifted to zero between 0.9 and 1.2 nm, the electrostatic interactions were explicitly screened by a screening parameter

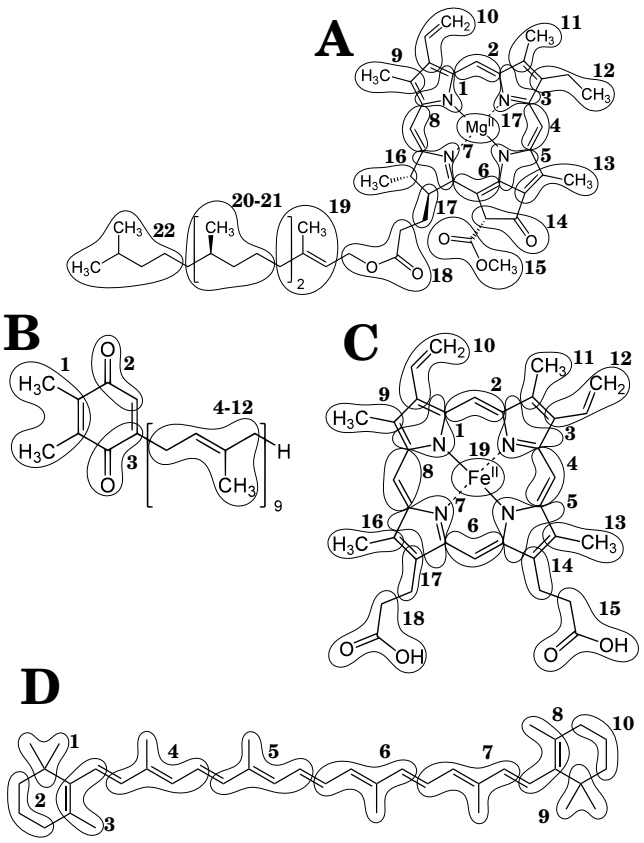

Figure 6.1: Molecular structure of the PSII cofactors. Rounded shapes encircle atoms that are mapped together to one CG bead. A) Chlorophyll A (CLA). The CG eight bead ring topology is shown. Pheophytin (PHO) has the same topology, lacking the central magnesium. B) Plastoquinone (PL9one). Plastoquinol (PL9ol) has the same topology, except for the carbonyl ring substituents, which are replaced by alcohol groups. C) Heme b (HEM). The CG eight bead ring topology is shown. D) β-Carotene (BCR).

(ϵ_r) of 15 and were smoothly shifted to zero between 0.0 and 1.2 nm. The neighbor list of 1.4 nm was updated every 10 steps. Pressure and temperature coupling parameters, τ_t and τ_p were set to 1.0 and 3.0 ps^{-1}. Bonds were not constrained, unless explicitly set in the topology.

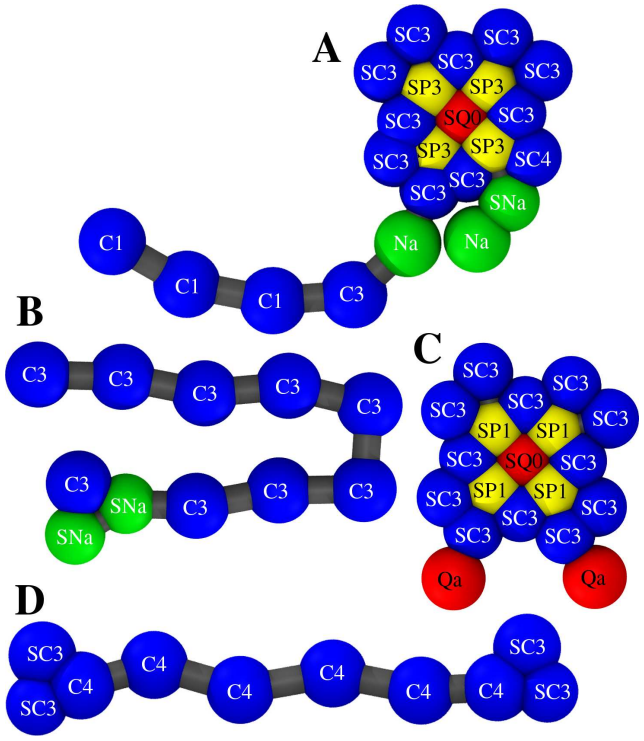

Figure 6.2: CG structure of the PSII cofactors. The colors are indicative of the bead type: Red, charged; Yellow, polar; Green, intermediate; Blue, apolar. A) Chlorophyll A (CLA). The CG eight bead ring topology is shown. Pheophytin (PHO) has the same topology, lacking the central magnesium. B) Plastoquinone (PL9one). Plastoquinol (PL9ol) has the same topology, except for the SNa ring beads, which are replaced by more polar SP2 beads. C) Heme b (HEM). The CG eight bead ring topology is shown. D) β-Carotene (BCR).

Parametrization of atomistic models

An initial guess for the topology of all six molecules was obtained using either the PRODRG-server[26] or the Automated Topology Builder[27]. The atom types (defining the non-bonded interactions) and bonded interactions in these topologies were checked and adjusted if necessary based upon the Gromos force field building blocks[18] and the force field definitions[28]. To obtain partial charges, first the optimal geometry of an all-atom structure was obtained using the Restricted Hartree-Fock (HF) method implemented in GAMESS-US[29] with a 6-31G* basis set. Second, using the optimized structure, atom-centered charges were calculated using the same method and fitting the electric potential to a Connolly surface with a density of 71.4 points per $Å^2$ while constraining the total monopole and dipole of the molecules obtained in the HF calculations.

For the elongated shape of CLA this protocol resulted in unrealistic high partial charges in the (aliphatic) tail. In order to avoid this issue, we calculated the partial charges for CLA using an in-house implementation of the Dipole Preserving Analysis[30] (DPA) in GAMESS-UK[31].

The sets of partial charges were subsequently adjusted as follows: charges for CH_x groups that are not part of an aromatic system were summed and assigned to the carbon atom, since the Gromos force field has no explicit hydrogens; charges were adjusted to respect the symmetry of a molecule; finally, all charges were rounded to two decimal places.

Parameterization of coarse grain models

The mapping of atoms to CG interactions sites and the choice of bead types was based on the building blocks described in Table 3 in Marrink *et al.*[9] and following the recipe described in [32]. Although no absolute rules can be given for the choices made, we followed a few general guidelines: 1) Atoms were grouped in order to include chemical groups in one bead; 2) The number of atoms in a bead should be close to four for normal type beads, and close to three for S-type beads. An exception was made for the much heavier metal centers of CLA and HEM, which are represented by a single S-type bead; 3) Ring motifs should consist of at least three beads; 4) Repeated patterns should be modeled using a repeated topology and symmetric parts of the molecules should have a symmetric topology; 5) Similar molecules should share similar topologies and therefore differences between molecules, for example the magnesium bead in CLA that is not present in PHO, should be concentrated at the last beads of the topology for simplicity. In this way the numbering of beads and bonded interactions are the same for most of the molecule.

In general, coarse grain interaction sites contain no partial charges. Exceptions are the central bead (iron and magnesium) in HEM (+0.4 e) and CLA (+1.0 e) and the four surrounding beads (-0.1 e and -0.25 e for HEM and CLA respectively). This strategy was adopted in order to mimic the strong polar nature of the metal center. Note however that the polarizing effect of these (partial) charges is limited in Martini 2.1, since other polar groups do not have any partial charges. It will effect interactions with charged beads such as ions and charged amino acids, but its effect will be stronger when combined with polarizable water[33] and the recently released parameters for Martini with explicit polar amino acids (Chapter 7).

Bonded interactions were obtained by mapping atomistic systems to CG resolution. In order to do so an atomistic simulation of a single cofactor molecule was simulated using the Gromos topologies for 10 ns in a system containing 133 decane molecules. In general, the conformations sampled will, to some extend, depend on the solvent used and it might not be possible to find a CG topology that optimally mimics the conformations sampled in all solvents. Here we choose to base the CG topology on a simulation in (apolar) decane, since the cofactors are rather hydrophobic and will spend most time in apolar environments.

Atomistic parameters obtained in the first step of the parametrization were used for the solute, while the model of decane was taken from [34]. Distributions for all bonds, angles and dihedrals were extracted from atomistic trajectories mapped to CG resolution and from these distributions initial guesses for the bonded parameters were obtained. Next, in an iterative procedure the parameters were adjusted to obtain optimal distributions from a CG simulation, while maintaining numerical stability. For every iteration a 10 ns simulations of the CG cofactor molecule in 133 CG decane molecules was run, using the currently obtained parameters. Distributions for all bonded interactions were obtained and compared to distributions from the mapped atomistic simulations. These steps were repeated until satisfactory distributions were obtained.

Setup of test systems

Heme b. HEM was simulated bound to the protein bacterioferritin. A structure was obtained from the model with PDB entry 3E2C and converted to a CG model using the *martinize.py* script[35]. The Martini parameters for proteins [10] are used in combination with an Elnedyn elastic network[36]. The interactions between the proteins and the heme b are purely non-bonded, e.g. no bonds are used to keep the heme b in place. The system was solvated by 5007 CG water beads (equivalent to 20,028 water molecules) and neutralized by adding 22 sodium ions. It was energy minimized for 500 steps using the steepest descent algorithm and simulated for 10000 steps (100 ps) with position restraints on the protein backbone and a smaller (10 fs) time step. Finally, the system was simulated for 1.5 μs, using the settings as describe above for Martini simulations.

Plastoquinone and plastoquinol. Plastoquinone and plastoquinol were simulated in a 128 lipid DPPC bilayer, using both coarse grain and atomistic models and at a coarse grain resolution in a thylakoid membrane. The CG simulations of a DPPC bilayer were started from a pre-equilibrated bilayer, obtained from `www.cgmartini.nl`. The bilayer consisted of 128 lipids solvated in 2000 CG water beads, equivalent to 8000 water molecules. Four plastoquinone molecules were placed in the water phase. The systems were energy minimized for 500 steps using the steepest descent algorithm and simulated for 10000 steps at a small time step (5 fs) to properly relax possible overlap between water and plastoquinone. Since the two molecules have the same number of beads, plastoquinone is converted plastoquinol by simply changing the molecular topology.

The CG simulations in a thylakoid membrane with a composition representative of a bacterial thylakoid membrane[37] were started from a preformed bilayer. The bilayer consisted of 72 phosphatidylglycerol (PG) lipids, 180 digalactosyldiacylglycerol (DGDG) lipids, 288 monogalactosyldiacylglycerol (MGDG) lipids, 180 sulfoquinovosyl diacylglycerols (SQDG) lipids. The lipids have an oleoyl tail at the sn-1 position and a palmitoyl tail at sn-2 position, except for 72 of the SQDG lipids that had two palmitoyl tails. The bilayer was solvated in 21258 CG water beads (equivalent to 85032 water molecules) and neutralized by 252 sodium ions. Parameters for the CG thylakoid lipids have been developed by Lopez *et al.*[11]. Plastoquinones were included in the system in the same way as for the DPPC systems.

The CG simulations were performed with simulation parameters as described above for CG systems, except that the temperature was set to 310 and 324 K for the thylakoid and DPPC membranes, respectively and the pressure coupling, was switched to a semi-isotropic scheme.

The atomistic systems were obtained from a CG structure in which the plastoquinones had already moved into the bilayer, using the reverse transformation tool[25]. The simulation used the PL9one parameters obtained in this study in combination with the Berger lipids[38] incorporated in the Gromos53A6 force field. Simulations were performed with simulation parameters as described above for atomistic systems, except that the temperature was set to 324 K (10 degrees above the phase transition temperature of DPPC) and the pressure coupling, which was switched to a semi-isotropic scheme in order to decouple the pressure in the bilayer plane and the pressure perpendicular to it.

β-carotene BCR was simulated in a 128 DPPC lipid bilayer using both CG and UA models. The systems contained four BCR molecules and were setup in the same way and using the same simulation parameters as the PL9 simulations.

Analysis The conformations sampled by cofactors simulated in a bilayer system were analyzed using a cluster analysis algorithm described by Daura *et al.*[39] and implemented in the Gromacs tool g_cluster. An RMSD cut-off of 0.15 nm was used in the analysis. The electron densities were calculated along the Z-axis of the box using the Gromacs analysis

tool g_analyze. The electrons of the atoms mapped to one CG bead were added together. To properly compare the densities obtained from UA and CG simulations, the UA trajectories were mapped to CG using the reverse transformation tool[25] prior to the calculations of the electron densities. Diffusion rates were calculated from the mean square displacements (MSD) averaged over all molecules in the system. The linear regime of MSD curve in one dimension was fitted to:

$$MSD = 2 \cdot D \cdot t + c \tag{6.1}$$

The diffusion in the X and Y direction (D_x and D_y) diffusion are calculated separately to check for anisotropy. The final value is the mean of the D_x and D_y, the error is the standard error calculated over all molecules.

6.3 Results and Discussion

Figure 6.1 shows the atomistic structures and coarse grain mappings of PL9one, HEM, CLA and BCR. Figure 6.2 shows the CG beads types and molecule geometries.

Parameterization of the atomistic force field

For all molecules only modest changes were made with respect to the initial topologies generated by the Automated Topology Builder[27]. All final topologies were stable in simulations using an integration time step of 2 fs, with all bonds constrained. The full topologies in Gromacs format can be found in section 6.5.

Chlorophyll A and pheophytin. The bonded and non-bonded interactions for both molecules are equal, except for atoms bound to the central magnesium. The topology contains three explicit hydrogens connected to the porphyrin ring. These hydrogens carry a +0.1 *e* charge and no van der Waals potential. The magnesium ion in chlorophyll A carries a +1.0 *e* charge, which is compensated for by partial positive charges in the porphyrin ring, yielding a neutral molecule. In pheophytin, the magnesium is replaced by two hydrogen atoms, each with a +0.1 *e* partial charge. Partial charges in the porphyrin ring are lower in PHO compared to CLA. Overall PHO is a neutral molecule as well.

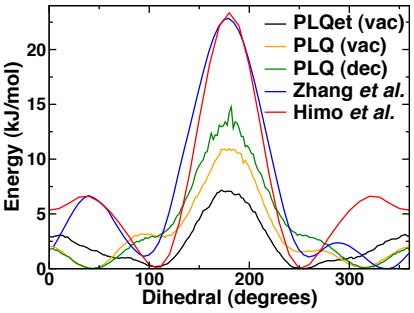

Figure 6.3: Energy distribution of the dihedral connecting the plastoquinone head and tail. The energy distribution has been calculated for PL9et (solid line) where the poly-prenyl tail has been replaced by an ethyl group, PL9one in vacuum (dotted line) and PL9one in decane (dashed line).

β-Carotene. The topology for BCR was largely based on the existing topology for retinol[18]. Retinol consists of a head group equal to BCR with a tail of two repeated units and an alcohol group connected to the last tail unit. The model for BCR was created by removing the alcohol group and repeating the topology in reverse direction. The two retinol based parts of the topology are connected by a double bond. The topology contains no explicit hydrogens and no partial charges.

Plastoquinone and plastoquinol. The bonded and non-bonded interactions for both molecules are equal, except for the ester or alcohol substituents of the ring. The distribution of the

partial charges are taken from the charge calculations (see 6.2) for both molecules separately. The prenyl repeats all have the same atom types, bonded interactions and partial charges. The hydrogen connected to the carbon involved in the prenyl double bond is explicitly represented, and carries a +0.185 e partial charge to mimic the strong polarization of the double bond. In addition these hydrogens carry a small repulsive van der Waals term to avoid clashes with other atoms.

We calculated the potential energy distribution of the dihedral connecting the PL9one head group with the first prenyl tail unit (Figure 6.3). Zhang *et al.* [17] and Himo *et al.* [40] calculated this energy distribution using a quantum mechanical model for plastoquinone in vacuum and semi-plastoquinone radical in vacuum where the first tail unit was replaced by a ethyl group ($PLQet^-$), respectively. When comparing the distribution obtained here for PLQet to the one obtained by Himo *et al.* for $PLQet^-$ it can be seen that the position of the global maximum is correct, although too low. The positions of the secondary maxima are slightly off at 10 and 350 degrees (against 50 and 310 by Himo *et al*) and too low. The distributions obtained for PL9one in vacuum reproduces the main features of the distribution found by Zhang *et al.*, however the height of all barriers is off by approximately a factor 2 and the position of the secondary minima or shifted. When simulated in decane the global maximum becomes higher whereas most other features are tuned down. Although for both molecules the energy profiles do no perfectly match the QM energy profiles, the overall feature are sufficiently well reproduced to give a proper united atom topology.

Parametrization of CG force fields

The following CG topologies give stable simulations with a time step up to 20 fs. Full Martini topologies in Gromacs format can be found in section 6.6.

Table 6.1: LogP values for partitioning of CG cofactors between water (W) and octanol (Oct) or water and cyclohexane (Chx). Subscript numbers indicate the tail lengths of plastoquinone and plastoquinol: either 1 or 9 prenyl tail units. **a.** Eight bead ring topology. **b.** Four bead ring topology. **c.** Ref.[41] **d.** Ref.[42] **e.** Ref.[43] **f.** Octanol-water partitioning. **g.** Cyclohexane-water partitioning. **h.** Ref.[44]

	$\mathbf{Log_{10}P}_{W\rightarrow Oct/Chx}$		
Molecule	**CG**	**UA**	**Literature**
Chlorophyll A	15.8[a],12.1[b]	27.0	2.12[c]
Pheophytin	16.9[a]	-	-
Heme b	6.7[a]	-	0.95[c]
β-carotene	17.2	-	17.62[d], 2.88[e]
$Plastoquinone_9$	20.2[f], 27.9[g]	-	-
$Plastoquinone_1$	5.1[f], 4.7[g]	-	> 3[f,h], > 3[g,h]
$Plastoquinol_1$	2.6[f], 0.56[g]	-	> 3[f,h], 0.23[g,h]

Chlorophyll A and pheophytin. Topologies using four, six or eight beads were tested to model the CG topology of the conjugated ring structure. The 6-bead variant was not preferred because it does not maintain the fourfold symmetry of the ring in contrast to the 4 and 8 bead ring variants. Figures 6.1 and 6.2 only show the 8 bead mapping scheme. The water-octanol partitioning of both topologies with a 4 and 8 bead ring was calculated and compared to the experimental value[41](See Table 6.1). The LogP values for CLA (15.8) and PHO (16.9) indicate that both topologies are much too hydrophobic. To investigate the origin of the difference, we also calculated the water-octanol partitioning using the atomistic topology obtained in this work. We obtained $\Delta G = -159 \pm 11\text{kJ mol}^{-1}$, LogP = 27, which is also too hydrophobic when compared to the experimental value (LogP

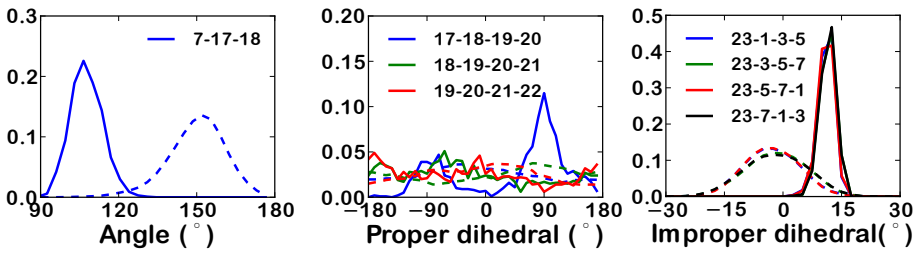

Figure 6.4: Distribution of bonded interactions obtained from a simulation of CG chlorophyll (solid lines) and a UA simulation of chlorophyll A mapped to CG (dashed line). The three plots show bond angles (left), proper dihedrals (middle) and improper dihedrals (right). The numbers in the legend define for which beads the distributions are measured, following the bead numbering in Figure 6.1.

= 2.12). Additionally, for pheophytin we calculated the LogP using a number of on-line available LogP predictors[45, 46, 47, 48]. The mean value over four different predictors is 10.24±1.85, which is lower than the LogP obtained from the simulations, but considerably higher than the experimental value.

It is very well possible that both the atomistic and CG force fields and the predictor algorithms do not properly reproduce the partitioning of chlorophyll A and pheophytin. Both the ring conjugation and the magnesium center are challenging features. The partitioning free energy for pheophytin might also not be representative of chlorophyll A. However, the large differences observed between the partitioning free energies we obtained shed enough doubt on the experimental values to settle with the CG topologies. We favor the 8-bead ring variant since it is thinner due to the use of S-beads, thus better representing the flat nature of the porphyrin ring.

This topology consists of 23 beads for chlorophyll A and 22 for pheophytin, mainly S-type particles in the ring and normal particles in the tail. In order to optimally maintain the symmetry of the molecule eight atoms have been doubly mapped: they are used in the determination of the center of mass (COM) of two adjacent beads. The masses of these atoms have to be halved for the COM calculation. The ring structure is very rigid and closely packed, giving rise to a high amount of stress. This stress is partially removed by defining mutual exclusions (from the non-bonded interactions) between all of the ring beads. The central magnesium is modeled by a single SQ0 bead and has a +1.0 e charge. The charge is equal to the charge on the Mg in the atomistic topology and mimics the strong polar nature of this atom. The four beads surrounding magnesium have a -0.25 e charge, thus neutralizing the structure. The magnesium ion is the last bead in the topology, allowing it to be removed without renumbering any interactions. Upon removal of the magnesium and all partial charges the topology represents pheophytin.

All CG bonded interactions were obtained by adjusting the parameters to give a satisfying reproduction of the distributions obtained from the atomistic simulations. The mean and standard deviations of distributions for all bonded interaction are given in the CG topologies in section 6.6. In order to obtain a synoptic measure for the bonded interactions in the different cofactors, the root mean square deviation (RMSD) of both the mean and standard deviation of the probability distributions averaged over all bonds, angles, proper and improper dihedrals have been calculated. Over 37 bonds, 26 angles, 3 proper and 13 im-

proper dihedrals the mean and standard deviation RMSDs were 0.006±0.003 nm, 9.3±3.8°, 37.3±11.7° and 9.4±4.0°, respectively. These RMSDs are relatively high for the angles and dihedrals due to a few interactions that strongly deviate from target distributions. For those interactions the distributions are shown in Figure 6.4. The deviation of a single angle (7-17-18) is most likely caused by the steric repulsion between beads later on in the tail and the ring (for example bead 19 with 17). Defining an angle potential over beads 7-17-18 does not improve the poorly reproduced distribution as shown in Figure 6.4. The deviation in the means of the four proper dihedrals is large, however their distributions are virtually flat, which makes a mean rather arbitrary. The four deviating improper dihedrals all involve the central magnesium. The large stress in the porphyrin ring makes it impossible to influence those distributions by defining a dihedral potential.

Heme b. The ring structure of heme is very comparable to that of chlorophyll A. Therefore a similar topology was used for both molecules, using eight beads to model the porphyrin ring. Heme is modeled to carry two negative charges, as both propionate sidechains (pKa=5.0) are deprotonated at physiological pH∼7.0. Like for chlorophyll the CG topology models a compound (LogP = 6.7) that is too hydrophobic compared to experiment (LogP = 0.95) (Table 6.1)[41]. The HEM is more hydrophilic than CLA, and the ratio of their LogP values closely matches the ratio of the experimental LogP values ($\text{LogP}_{\text{HEM}}/\text{LogP}_{\text{CLA}} = \sim 2.4$).

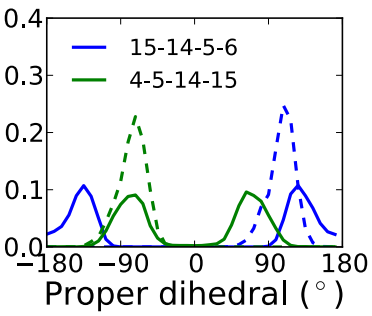

Figure 6.5: Distribution of bonded interactions obtained from a simulation of CG heme b (solid lines) and a UA simulation of heme b mapped to CG (dashed lines). The plot shows proper dihedrals. The numbers in the legend define for which beads the distributions are measured, following the bead numbering in Figure 6.1. Dihedral 4-5-14-15 is symmetrical with 8-7-17-18, 15-14-5-6 symmetrical with 18-17-7-6.

The topology of heme b contains 19 beads, of which all but two are small, S-type beads (see Figure 6.1). In heme double mapping and bead exclusions are applied as is the case for CLA to obtain the correct symmetry and partially relax strain in the ring.

All CG bonded interactions were obtained by adjusting the parameters to give a satisfying reproduction of the distributions obtained from the atomistic simulations. The mean and standard deviations of distributions for all bonded interaction are given in the CG topologies in section 6.6. Over 40 bonds, 20 angles, 4 proper and 12 improper dihedrals the mean and standard deviation RMSDs were 0.005±0.004nm, 2±2°, 93±96° and 1±4°, respectively. In the CG model, the dihedrals concerning the propionate sidechain beads (15-14-5-6, 4-5-14-15 and 18-17-7-6, 8-7-17-18) do not flip between both symmetric orientations visited by these dihedrals in the mapped UA trajectory (see Figure 6.5). Although the individual lobes of the bi-modal target distribution are reasonably well reproduced, the total RMSD is still high. Defining a dihedral potential over those atoms did not resolve this issue.

β-Carotene. The molecule has been modeled by 10 beads (see Figures 6.1 and 6.2. The two rings and its constituents are modeled by three beads, of which two are a S-type. The linker between the beads is divided in five equal parts, each modeled by a C4 bead with a five-to-one mapping. Non-bonded interactions between all beads in the ring and the first bead in the linker are excluded.

The LogP values for BCR found in literature diverge greatly. Cooper *et al.*[42] reported a value of 17.62 based on calculations and Liao *et al.*[43] found LogP = 2.88 based upon

UV-measurements. The average value for a series of on-line available LogP predictors[45, 46, 47, 48]was 6±1. The value obtained for our model (LogP=17.23) is close to the value found by Cooper *et al.* but high compared to most other reported values.

Over 11 bonds, 8 angles, 7 proper and 2 improper dihedrals the mean and standard deviation RMSDs were 0.007±0.02 nm, 12.6±3.5°, 94.0±57.5° and 21.1±12.3 °. Figure 6.6 shows the CG proper and improper dihedrals that do not give a well matching distribution. For symmetric pairs of dihedrals only one is shown. For the proper dihedrals this is due to the extended nature of the linker: if one or both of the angles contained in the dihedral is close to 180° this gives rise to a large force being exerted on the dihedral (if the angle is exactly 180° the dihedral is not calculated in Gromacs), leading to numerical instabilities. A solution for this problem has been devised[49], but has not been implemented in the main release of Gromacs yet. Testing with an in-house version of Gromacs (based upon version 4.0.7) gave dihedral distributions better matching the target distribution. However, in what follows, we will only consider a topology that can be simulated in the current releases of Gromacs. The CG improper dihedrals (3-4-1-2 and 8-7-9-10) do deviate from mapped UA distributions, however given the width of both distributions this is reasonable.

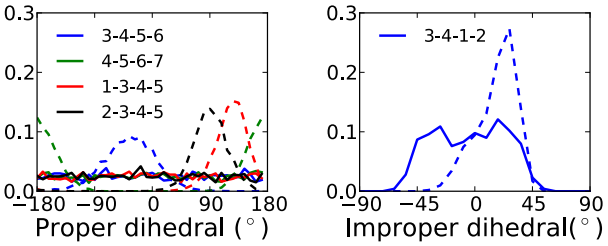

Figure 6.6: Distribution of bonded interactions obtained from a simulation of CG β-carotene (solid line) and a UA simulation of β-carotene mapped to CG (dashed line). The two plots show proper dihedrals (left) and improper dihedrals (right). The numbers in the legend define for which beads the distributions are measured, following the bead numbering in Figure 6.1.

Plastoquinone and plastoquinol. Both molecules have the same mapping (see Figure 6.1). The ring is represented by one C3 bead and two SNa particles in the case of plastoquinone. In the case of plastoquinol the SNa beads are replaced by more polar SP2 particles. The tail is modeled by 9 C3 beads, approximately a 1:5 mapping. The C3 bead was chosen based upon the water-octanol partitioning of *iso*-prenyl (-13 kJ mol^{-1})[50], similar to the partitioning of C3 beads (-14 kJ mol^{-1})[9]. Note that the *iso*-prenyl unit differs from the repeated unit in β-carotene by one double bond, giving *iso*-prenyl a slightly more apolar nature.

Partitioning free energies have been calculated for PL9one, PL1one, and PL1ol between water and cyclohexane, and water and octanol. Experimentally, the partitioning free energy for PL1ol between water and cyclohexane has been accurately determined and rough estimates for PL1ol water-octanol partitioning and PL1one water-cyclohexane and water-octanol partitioning are available[44]. The current model reproduces these values fairly well: LogP_{sim}=0.56 vs LogP_{exp}=0.23 for PL1ol water-cyclohexane partitioning. Between water and octanol PL1ol is slightly too hydrophilic (LogP_{sim} = 2.6 vs LogP_{exp} > 3.0). PL1one correctly has a LogP higher than 3.0 for the partitioning between water and both apolar solvents.

Non-bonded interactions between all beads in the ring and the first bead in the tail are excluded. Over 12 bonds, 10 angles, eight proper and two improper dihedrals the mean and standard deviation RMSDs were 0.007±0.007 nm, 6.1±5.1°, 7.9±26.1° and 0.18± 34.9°. All distributions are properly reproduced.

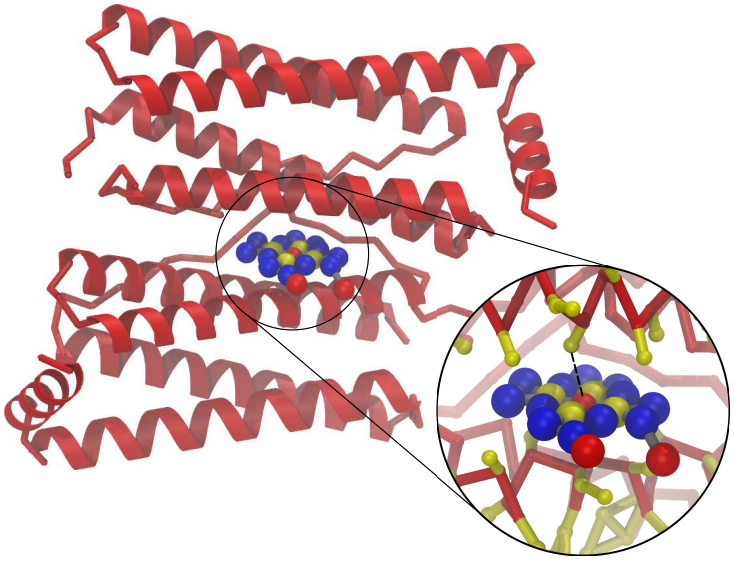

Figure 6.7: Snapshot of a heme molecule in between two bacterioferritin subunits after 1.5 μs of simulation. Proteins are shown in red, the heme uses the same coloring scheme as used in Figure 6.2. The inset gives a detailed view of the heme being coordinated by the methionine side chain.

Test system simulations

In order to investigate the behavior of the parametrized CG cofactors in a physiological environment we simulated the cofactors bound in a protein complex (HEM) or inside lipid bilayers (BCR, PL9one and PL9ol). The results were compared to the crystal structure or results from atomistic simulations, respectively.

Heme b. Inside the PS-II complex, heme b is bound in between the α and β subunits of cytochrome B559 or inside a pocket in cytochrome C550[4]. The central iron is coordinated by amino acid side chains of these proteins. To test if the CG heme b fits inside a protein binding pocket and the interactions with protein ligands are correctly modeled we simulated it in between two bacterioferritin proteins. Bacterioferritin is a homodimer of two 158 amino acid proteins that together encapsulate one heme (PDB entry: 3E2C). The heme iron is coordinated on both sides by a methionine. Simulating heme in between two seperate proteins means it is easier to escape, and thus a harder test, compare to when it would be simulated in complex with a cytochrome, where Heme b is largely surrounded by the protein. Although this test does not proof that the cofactor has been properly parametrized, it does give a strong indication that the molecule can be successfully simulated while ligated to a protein.

Over 1.5 μs of simulation the complex remains stable and the heme stays doubly coordinated. Omitting the first 100 ns of the simulation as equilibration period, the C_α RMSD of protein subunits A and B and the complex including the heme were 0.13±0.01, 0.15±0.02 and 0.24±0.02 nm, respectively. The distances between the heme iron and the coordinating methionines in the coarse grained crystal structure are 0.27 and 0.29 nm respectively. How-

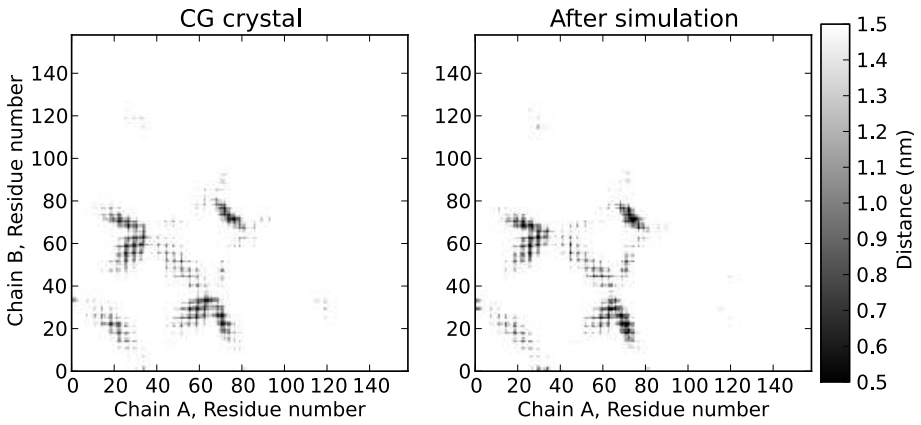

Figure 6.8: Comparison of residue-residue contact maps between chain A and B in the crystal structure (left) and the mean over the last 500 ns of the simulation(right).

ever, due to the large van der Waals radius of the CG beads, those small distances can not be maintained during the simulation. The average distace over the simulations are 0.48±0.03 and 0.53±0.03 nm, which is close to the minimum distance between two CG beads. The contact maps of the coarse grained crystal structure and the average over the last 500 ns of simulation (see Figure 6.8), show very similar patterns. Overall the analysis shows that the complex is very stable.

Plastoquinone and plastoquinol. The organization and dynamics of the CG PL9one and PL9ol in the thylakoid membrane (the "plastoquinone-pool") is of great importance for the transport between different protein complexes in the photosynthetic pathway. To test if our CG model properly models the behavior of PL9one and PL9ol in the membrane, we simulated PL9one and PL9ol inside a DPPC bilayer and inside a more complex lipid mixture, representative of the cyano bacterial thylakoid membrane. The location along the membrane normal, the conformations sampled by the molecule, the flip-flop rates and the lateral diffusion rates in DPPC are compared to atomistic simulations.

The systems containing four CG PL9one or PL9ol molecules were simulated for 4.0 μs each. In the analysis the first 200 ns were omitted as equilibration. The cofactors were initially placed in the water phase, but segregated into the lipid bilayer within tens of nanoseconds. At atomistic resolution a DPPC bilayer containing four PL9one molecules was simulated for 250 ns, of which the first 10 ns were omitted as equilibration. At the start of the simulation two PL9one molecules were placed inside each monolayer in order to shorten the slow equilibration phase.

Figure 6.9 shows the electron densities of UA PL9one and CG PL9one and PL9ol in a DPPC bilayer. In order to properly compare the CG and atomistic densities, the atomistic trajectory was mapped to CG before the analysis. For reference also the electron density distributions of the lipid phosphate bead are shown. The thickness of the coarse grain and atomistic bilayers is equal as reported by the distance between these PO_4 peaks (red lines in Figure 6.9). At both resolutions, the head groups reside just below the head group region of the bilayer (green lines), however the (mapped) atomistic head groups are buried ~ 0.3 nm deeper into the bilayer. This difference indicates that the CG PL9one head groups are slightly

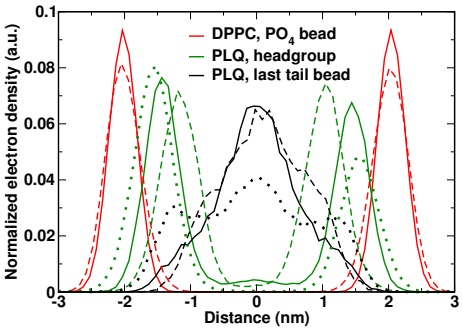

Figure 6.9: Normalized electron density plots for plastoquinone (PL9one) and plastoquinol (PL9ol) in a DPPC bilayer. Densities obtained with the coarse grain model (PL9one, solid lines; PL9ol, dotted lines) are compared to those obtained for the atomistic model mapped to coarse grain resolution (PL9one, dashed lines). Densities for the DPPC phosphate bead (red) are compared to those for the plastoquinone head group (green) and plastoquinone tail bead (black). Phosphate bead densities from the CG PL9ol simulation are similar as those obtained from the CG PL9one simulation and omitted for clarity.

too hydrophilic.

Using a SN0 particle in the head group gave slightly lower water-octanol and water-cyclohexane LogP values were smaller, but still larger than the lower bound given by experiments ($\mathrm{LogP_{exp}}$ >3). In addition the headgroups were slightly (∼0.1 nm) deeper buried, in better agreement with atomistic model, however the headgroups spend a considerable amount of the simulations time at the bilayer midplane, in contrast to atomistic model. Considering this, together with the fact that the SN0 particle does not mimic the hydrogen acceptor properties of the molecule by stronger binding to hydrogen donors and positively charged molecules, we decide to settle with the SNa. The end of the tail of both UA and CG PL0one resides mostly at the bilayer midplane. The CG PL9ol head group is on average found slightly closer to the bilayer interface, as would be expected from the more hydrophilic head group. The secondary maxima in the density distribution of the last tail bead are due to the higher occurrence of tail back folding (see below). Both PL9one and PL9ol are in contact relatively a large part of the time due to the high solute to lipid ratio (1:32 for both CG and UA), however the contact is mostly via the long tails. There is no particular preference for the head groups to stack.

A cluster analysis of the molecular conformations sampled, showed three preferred orientations in both the CG and UA simulations: 1) the PL9one head group in the lipid head group region of one monolayer with the tail in a L-shape with the last part of the tail at the bilayer midplane (*L*-conformation), 2) the tail being extended with the tail ends in the opposite monolayer, sometimes with slight S-turn at the bilayer midplane (*I*-conformation) and 3) the tail curling back up to the head group region (*U*-conformation). Figure 6.10 shows examples of the these conformation for UA (left) and CG (right) systems. In the CG and UA simulations containing 4 PL9ones the ratios L:I:U were 53(1):16(2):13(2)% and 46(14):18(3):13(3)%. The values between brackets indicate the standard deviation calculated over the four solutes. The L,I and U conformation together are found in 82% of the time; the remaining 28% and 23% consist of less well defined, often less extended, conformations. For CG PL9ol the L:I:U ratio is 47(1):10(1):19(1)%, a significantly higher occurrence of the U conformation. This stronger tendency of the tail to fold towards the lipid head group is consistent with the electron density of the last tail bead of PL9ol (see above) and is most likely due to the position closer to the water phase, due to the more hydrophilic head group as compared to PL9one.

During the simulation the PL9one head group was observed to flip from one monolayer to the other, a so called flip-flop event. In the CG simulation 164 flip-flop events were observed,

giving a flipflop rate of $1.1 \times 10^7 s^{-1}$. In the UA simulation the head group of one PL9one molecule was observed to flip twice from one monolayer to the other. Although two events are insufficient to obtain a statistically reliable number, it indicates a flipflop rate in the order of 2×10^6 s^{-1}. This gives speed-up of a factor of five for the CG model, in accordance with the speed-up typically found in Martini systems[51]. For CG PL9ol eight flops are observed, a flip-flop rate of 0.05×10^7 s^{-1}. This much lower flip-flop rate as compared to PL9one is due to the more hydrophilic headgroup of PL9ol.

The systems containing four CG PL9ones or PL9ols in the cyano bacterial thylakoid membrane model were simulated for 3 and 3.5 μs, respectively. The ratios of L, I and U conformations are very similar to those observed in DPPC: 58(1):14(3):15(4)% and 61(1):10(3):17(2)% for the systems containing PL9ones and PL9ols respectively. For the system containing four PL9ones 18 flipflop events were counted, resulting in a flipflop rate of 1.5 $\times 10^6$ s^{-1}, respectively, considerably slower than in a DPPC bilayer. For the simulations containing PL9ol no flipflop events were observed. Although it is not immediatly clear why the flip-flop rates in the thylakoid membrane are lower, a possible explanation might be a stronger binding to the less hydrophylic headgroups of the lipids in the thylakoid membrane as compared to DPPC.

β-carotene. In the PSII, BCR protects nearby chlorophyll molecules against excess light energy[52]. Although it is mostly buried inside the protein complex, its properties have been studied before in bilayer environments[53]. In order to compare the behavior of our CG model to these previous results, we also simulated BCR in a DPPC bilayer. For comparison we also simulated a system using the UA parameters obtained in this work.

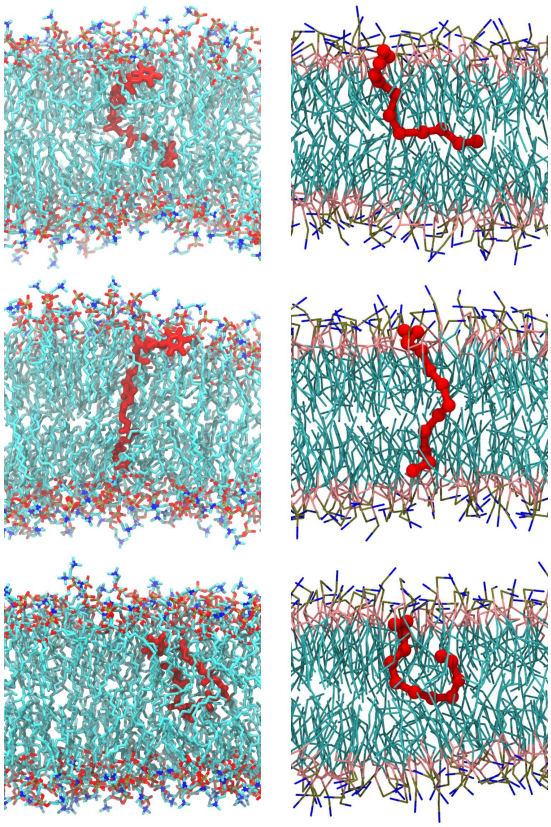

Figure 6.10: Side-by-side comparison of atomistic (left) and CG (right) plastoquinone in different conformations inside a DPPC lipid bilayer. Shown are the L-conformation (top), the I-conformation (middle) and the U-conformation (bottom).

The four CG BCR molecules were initially placed in the water phase. The system was simulated for 4 μs. One cofactor enters the bilayer in less than 10 ns. The other molecules cluster together and enter the bilayer together as a trimer. Inside the bilayer the trimer breaks up and the monomers

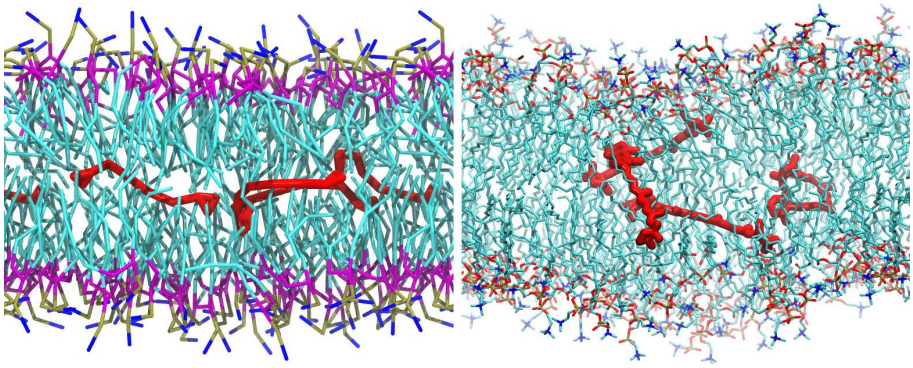

Figure 6.11: Side-by-side comparison of CG of four (left) and atomistic (right) β-carotenes in a flat orientation in the (DPPC) bilayer midplane.

do no cluster. They adopt a straight conformation and stay at the bilayer midplane (see Figure 6.11). The atomistic starting structure was reverse transformed from the last CG snapshot, in order to shorten the long equilibration phase. The system was simulated for 200 ns. The molecules remain in a elongated conformation on the bilayer midplane and show no specific clustering. The CG and UA BCR molecules reside approximately in the same region of the bilayer, as measured by the normalized head group electron density, shown in Figure 6.12. The difference between density distribution obtained for both head groups obtained for the UA model is indicative that the simulation has not yet reached equilibrium.

Due to their flat orientation in between the two monolayers, the BCR molecules are expected to diffuse faster than the surrounding DPPC lipids. The diffusion rates of CG BCR and DPPC were 0.15±0.015 and 0.078±0.002×10^{-5} cm^2 s^{-1}, respectively. The errors have been calculated as the standard error over all molecules. In the atomistic systems the diffusion rates were 0.032±0.006 and 0.014± 0.003 × 10^{-5} cm^2 s^{-1}. At both resolutions BCR diffuses a factor ∼2 faster. Both BCR and the lipids diffuse faster at CG resolution (4.7 and 5.6 times faster, respectively), in agreement with the speed-up generally observed for the Martini force field[51]. Experimentally, no diffusion rates for BCR in lipid bilayers are known, however Jemioa-Rzemiińska *et al.*[53] estimated the diffusion of BCR in a POPC bilayer to be 0.052·10^{-5} cm^2 s^{-1} from atomistic simulations using the OPLS force field. This diffusion rate is slightly higher than the one obtained for Gromos, which could be due to the different type of lipids used (the authors estimate the diffusion of POPC to be between 0.040 and 0.045 ×10^{-5} cm^2 s^{-1} based on values that had previously been reported in literature).

6.4 Conclusions

Topologies at atomistic and coarse grain resolution have been obtained for plastoquinone, plastoquinol, heme b, chlorophyll A, pheophytin and β-carotene. The behavior of these molecules at both resolutions has been tested in different systems. The topologies showed reasonable to good behavior, however all topologies have some weaker points.

CLA and HEM could be modeled by four or eight beads in the ring. Irrespective of the number of ring beads the molecules were too hydrophobic compared to values from literature.

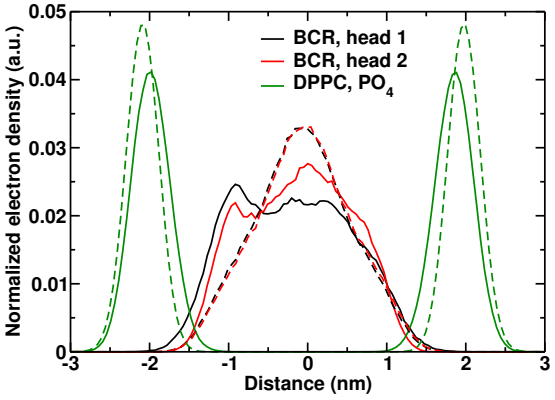

Figure 6.12: Normalized electron density plots for β-carotene in a DPPC bilayer. Densities obtained with the CG model (dashed lines) are compared to those obtained with the atomistic model mapped to coarse grain resolution (solid lines). Densities for the DPPC phosphate bead (green) are compared to those for both BCR head groups (red and black lines).

The eight bead ring model is slightly flatter due to the use of more S-type particles. Therefor we choose the eight bead model. In CG CLA all but one angle and four improper dihedrals reproduced the bonded interactions of the mapped atomistic very well. In HEM only two dihedrals showed less good behavior. In a simulation in a test system HEM was found to remain bound in the crystal structure binding site of bacterioferritin.

CG BCR reproduces the water-octanol partitioning behavior found in literature. Most bonded interactions show excellent behavior compared to the reference system. Dihedrals in the elongated tail were difficult to model using the currently in Gromacs implemented dihedral functions. Nonetheless does the CG BCR properly mimic the behavior of the UA BCR inside a DPPC bilayer.

Atomistic PLQ qualitatively mimics energy profiles for the head tail dihedral obtained by QM calculations. CG plastoquinone and plastoquinol with different tail lengths show good correspondence to experimental partitioning behavior and excellent agreement between UA and CG bonded interactions. In a DPPC bilayer CG PLQ behaves similar as UA PLQ.

The parametrized molecules are the most important cofactors present in the thylakoid membrane. Together with recently published parameters for lipids found in the thylakoid membrane[11] and high resolution crystal structures of the thylakoid proteins[4], these parameters open the way to detailed simulations of the thylakoid membrane.

6.5 Gromos topologies

Plastoquinone-9

```
[ moleculetype ]
; Name   nrexcl
plq      3

[ atoms ]
;  nr  type resnr resid atom cgnr charge  mass
; Headgroup
   1   C    1     PLQ   C     1   -0.417  12.0110
   2   HC   1     PLQ   H     2    0.242   1.0080
   3   C    1     PLQ   C     3    0.675  12.0110
   4   O    1     PLQ   O     4   -0.491  15.9994
   5   C    1     PLQ   C     5   -0.109  12.0110
```

```
   6   C    1     PLQ   C     6    0.604  12.0110
   7   O    1     PLQ   O     7   -0.504  15.9994
   8   C    1     PLQ   C     8    0.000  12.0110
   9   CH3  1     PLQ   CH3   9    0.000  15.0350
  10   C    1     PLQ   C    10    0.000  12.0110
  11   CH3  1     PLQ   CH3  11    0.000  15.0350
; U1: First tail unit
  12   CH2  1     PLQ   CH2  12    0.292  14.0270
  13   C    1     PLQ   C    13   -0.594  12.0110
  14   HC   1     PLQ   H    14    0.185   1.0080
  15   C    1     PLQ   C    15    0.117  12.0110
  16   CH3  1     PLQ   CH3  16    0.000  15.0350
  17   CH2  1     PLQ   CH2  17    0.000  14.0270
; U2: First tail unit
  18   CH2  1     PLQ   CH2  18    0.292  14.0270
  19   C    1     PLQ   C    19   -0.594  12.0110
  20   HC   1     PLQ   H    20    0.185   1.0080
  21   C    1     PLQ   C    21    0.117  12.0110
  22   CH3  1     PLQ   CH3  22    0.000  15.0350
  23   CH2  1     PLQ   CH2  23    0.000  14.0270
; U3: First tail unit
  24   CH2  1     PLQ   CH2  24    0.292  14.0270
  25   C    1     PLQ   C    25   -0.594  12.0110
  26   HC   1     PLQ   H    26    0.185   1.0080
  27   C    1     PLQ   C    27    0.117  12.0110
  28   CH3  1     PLQ   CH3  28    0.000  15.0350
  29   CH2  1     PLQ   CH2  29    0.000  14.0270
; U4: First tail unit
  30   CH2  1     PLQ   CH2  30    0.292  14.0270
  31   C    1     PLQ   C    31   -0.594  12.0110
  32   HC   1     PLQ   H    32    0.185   1.0080
  33   C    1     PLQ   C    33    0.117  12.0110
  34   CH3  1     PLQ   CH3  34    0.000  15.0350
  35   CH2  1     PLQ   CH2  35    0.000  14.0270
; U5: First tail unit
  36   CH2  1     PLQ   CH2  36    0.292  14.0270
  37   C    1     PLQ   C    37   -0.594  12.0110
  38   HC   1     PLQ   H    38    0.185   1.0080
  39   C    1     PLQ   C    39    0.117  12.0110
  40   CH3  1     PLQ   CH3  40    0.000  15.0350
  41   CH2  1     PLQ   CH2  41    0.000  14.0270
; U6: First tail unit
  42   CH2  1     PLQ   CH2  42    0.292  14.0270
  43   C    1     PLQ   C    43   -0.594  12.0110
  44   HC   1     PLQ   H    44    0.185   1.0080
  45   C    1     PLQ   C    45    0.117  12.0110
  46   CH3  1     PLQ   CH3  46    0.000  15.0350
  47   CH2  1     PLQ   CH2  47    0.000  14.0270
; U7: First tail unit
  48   CH2  1     PLQ   CH2  48    0.292  14.0270
  49   C    1     PLQ   C    49   -0.594  12.0110
  50   HC   1     PLQ   H    50    0.185   1.0080
  51   C    1     PLQ   C    51    0.117  12.0110
  52   CH3  1     PLQ   CH3  52    0.000  15.0350
  53   CH2  1     PLQ   CH2  53    0.000  14.0270
; U8: First tail unit
  54   CH2  1     PLQ   CH2  54    0.292  14.0270
  55   C    1     PLQ   C    55   -0.594  12.0110
  56   HC   1     PLQ   H    56    0.185   1.0080
  57   C    1     PLQ   C    57    0.117  12.0110
  58   CH3  1     PLQ   CH3  58    0.000  15.0350
```

```
  59   CH2  1     PLQ   CH2  59    0.000  14.0270
; U9: First tail unit
  60   CH2  1     PLQ   CH2  60    0.292  14.0270
  61   C    1     PLQ   C    61   -0.594  12.0110
  62   HC   1     PLQ   H    62    0.185   1.0080
  63   C    1     PLQ   C    63    0.117  12.0110
  64   CH3  1     PLQ   CH3  64    0.000  15.0350
  65   CH3  1     PLQ   CH3  65    0.000  15.0350

[ bonds ]
;  i    j funct
   1   2  2  gb_3
   1   3  2  gb_22
   1   5  2  gb_10
   3   4  2  gb_5
   3   8  2  gb_26
   5   6  2  gb_26
   6   7  2  gb_5
   6  10  2  gb_26
   8   9  2  gb_26
   8  10  2  gb_12
  10  11  2  gb_26
; Connection head group-> U1
   5  12  2  gb_26
; U1
   1  14  2  gb_3
   1  15  2  gb_10
  13  12  2  gb_26
  15  17  2  gb_26
  15  16  2  gb_26
; Connection U1-> U2
  17  18  2  gb_26
; U2
  19  20  2  gb_3
  19  21  2  gb_10
  19  18  2  gb_26
  21  23  2  gb_26
  21  22  2  gb_26
; Connection U2-> U3
  23  24  2  gb_26
; U3
  25  26  2  gb_3
  25  27  2  gb_10
  25  24  2  gb_26
  27  29  2  gb_26
  27  28  2  gb_26
; Connection U3-> U4
  29  30  2  gb_26
; U4
  31  32  2  gb_3
  31  33  2  gb_10
  31  30  2  gb_26
  33  35  2  gb_26
  33  34  2  gb_26
; Connection U4-> U5
  35  36  2  gb_26
; U5
  37  38  2  gb_3
  37  39  2  gb_10
  37  36  2  gb_26
  39  41  2  gb_26
```

```
  39  40  2  gb_26
; Connection U5-> U6
  41  42  2  gb_26
; U6
  43  44  2  gb_3
  43  45  2  gb_10
  43  42  2  gb_26
  45  47  2  gb_26
  45  46  2  gb_26
; Connection U6-> U7
  47  48  2  gb_26
; U7
  49  50  2  gb_3
  49  51  2  gb_10
  49  48  2  gb_26
  51  53  2  gb_26
  51  52  2  gb_26
; Connection U7-> U8
  53  54  2  gb_26
; U8
  55  56  2  gb_3
  55  57  2  gb_10
  55  54  2  gb_26
  57  59  2  gb_26
  57  58  2  gb_26
; Connection U8-> U9
  59  60  2  gb_26
; U9
  61  62  2  gb_3
  61  63  2  gb_10
  61  60  2  gb_26
  63  65  2  gb_26
  63  64  2  gb_26

[ pairs ]
;  i   j   funct
; Head
   1   7   1
   1   9   1
   1  10   1
   2   4   1
   2   6   1
   2   8   1
   3   6   1
   3  11   1
   4   5   1
   4   9   1
   4  10   1
   5   8   1
   5  11   1
   6   9   1
   7   8   1
   7  11   1
   9  11   1
; Connection head group-> U1
   1  13   1
   2  12   1
   3  12   1
   5  14   1
   5  15   1
   6  13   1
```

```
   7  12   1
  10  12   1
; U1
  12  16   1
  12  17   1
  14  16   1
  14  17   1
; Connection U1-> U2
  13  18   1
  15  19   1
  16  18   1
  17  20   1
  17  21   1
; U2
  18  22   1
  18  23   1
  20  22   1
  20  23   1
; Connection U2-> U3
  19  24   1
  21  25   1
  22  24   1
  23  26   1
  23  27   1
; U3
  24  28   1
  24  29   1
  26  28   1
  26  29   1
; Connection U3-> U4
  25  30   1
  27  31   1
  28  30   1
  29  32   1
  29  33   1
; U4
  30  34   1
  30  35   1
  32  34   1
  32  35   1
; Connection U4-> U5
  31  36   1
  33  37   1
  34  36   1
  35  38   1
  35  39   1
; U5
  36  40   1
  36  41   1
  38  40   1
  38  41   1
; Connection U5-> U6
  37  42   1
  39  43   1
  40  42   1
  41  44   1
  41  45   1
; U6
  42  46   1
  42  47   1
  44  46   1
```

```
  44  47   1
; Connection U6-> U7
  43  48   1
  45  49   1
  46  48   1
  47  50   1
  47  51   1
; U7
  48  52   1
  48  53   1
  50  52   1
  50  53   1
; Connection U7-> U8
  49  54   1
  51  55   1
  52  54   1
  53  56   1
  53  57   1
; U8
  54  58   1
  54  59   1
  56  58   1
  56  59   1
; Connection U8-> U9
  55  60   1
  57  61   1
  58  60   1
  59  62   1
  59  63   1
; U9
  60  64   1
  60  65   1
  62  64   1
  62  65   1

[ angles ]
;  i   j   k funct
   1   3   4 2     ga_30
   1   3   8 2     ga_27
   1   5   6 2     ga_27
   2   1   3 2     ga_25
   2   1   5 2     ga_25
   3   1   5 2     ga_27
   3   8   9 2     ga_27
   3   8  10 2     ga_27
   4   3   8 2     ga_30
   5   6   7 2     ga_30
   5   6  10 2     ga_27
   6  10   8 2     ga_27
   6  10  11 2     ga_27
   7   6  10 2     ga_30
   8  10  11 2     ga_27
   9   8  10 2     ga_35
  11  10   8 2     ga_35
; Connection head group-> U1
   1   5  12 2     ga_27
   6   5  12 2     ga_27
   5  12  13 2     ga_15
; U1
  13  15  17 2     ga_27
  13  15  16 2     ga_27
```

```
  14  13  15 2     ga_25
  14  13  12 2     ga_25
  15  13  12 2     ga_27
  17  15  16 2     ga_27
; Connection U1 -> U2
  15  17  18 2     ga_15
  19  18  17 2     ga_15
; U2
  19  21  23 2     ga_27
  19  21  22 2     ga_27
  20  19  21 2     ga_25
  20  19  18 2     ga_25
  21  19  18 2     ga_27
  23  21  22 2     ga_27
; Connection U2 -> U3
  21  23  24 2     ga_15
  25  24  23 2     ga_15
; U3
  25  27  29 2     ga_27
  25  27  28 2     ga_27
  26  25  27 2     ga_25
  26  25  24 2     ga_25
  27  25  24 2     ga_27
  29  27  28 2     ga_27
; Connection U3 -> U4
  27  29  30 2     ga_15
  31  30  29 2     ga_15
; U4
  31  33  35 2     ga_27
  31  33  34 2     ga_27
  32  31  33 2     ga_25
  32  31  30 2     ga_25
  33  31  30 2     ga_27
  35  33  34 2     ga_27
; Connection U4 -> U5
  33  35  36 2     ga_15
  37  36  35 2     ga_15
; U5
  37  39  41 2     ga_27
  37  39  40 2     ga_27
  38  37  39 2     ga_25
  38  37  36 2     ga_25
  39  37  36 2     ga_27
  41  39  40 2     ga_27
; Connection U5 -> U6
  39  41  42 2     ga_15
  43  42  41 2     ga_15
; U6
  43  45  47 2     ga_27
  43  45  46 2     ga_27
  44  43  45 2     ga_25
  44  43  42 2     ga_25
  45  43  42 2     ga_27
  47  45  46 2     ga_27
; Connection U6 -> U7
  45  47  48 2     ga_15
  49  48  47 2     ga_15
; U7
  49  51  53 2     ga_27
  49  51  52 2     ga_27
  50  49  51 2     ga_25
```

```
  50  49  48 2     ga_25
  51  49  48 2     ga_27
  53  51  52 2     ga_27
; Connection U7 -> U8
  51  53  54 2     ga_15
  55  54  53 2     ga_15
; U8
  55  57  59 2     ga_27
  55  57  58 2     ga_27
  56  55  57 2     ga_25
  56  55  54 2     ga_25
  57  55  54 2     ga_27
  59  57  58 2     ga_27
; Connection U8 -> U9
  57  59  60 2     ga_15
  61  60  59 2     ga_15
; U9
  61  63  65 2     ga_27
  61  63  64 2     ga_27
  62  61  63 2     ga_25
  62  61  60 2     ga_25
  63  61  60 2     ga_27
  65  63  64 2     ga_27

[ dihedrals ]
;  i   j   k   l funct
; Head group
   1   2   3   5  2  gi_1
   1   5   6  10  2  gi_1
   3   1   5   6  2  gi_1
   3   1   4   8  2  gi_1
   5   6  10   8  2  gi_1
   6  10   8   3  2  gi_1
   6   5   7  10  2  gi_1
   8   3   9  10  2  gi_1
   8   3   1   5  2  gi_1
  10   8   3   1  2  gi_1
  10   6   8  11  2  gi_1
; Connection head group-> U1
   5   1  12   6  2  gi_1
; U1
  13  14  15  12  2  gi_1
  15  13  17  16  2  gi_1
; U2
  19  20  21  18  2  gi_1
  21  19  23  22  2  gi_1
; U3
  25  26  27  24  2  gi_1
  27  25  29  28  2  gi_1
; U4
  31  32  33  30  2  gi_1
  33  31  35  34  2  gi_1
; U5
  37  38  39  36  2  gi_1
  39  37  41  40  2  gi_1
; U6
  43  44  45  42  2  gi_1
  45  43  47  46  2  gi_1
; U7
  49  50  51  48  2  gi_1
  51  49  53  52  2  gi_1
```

```
; U8
  55  56  57  54  2  gi_1
  57  55  59  58  2  gi_1
; U9
  61  62  63  60  2  gi_1
  63  61  65  64  2  gi_1

[ dihedrals ]
;  ai   aj   ak   al  funct    ph0      cp     mult
; Connection head group-> U1
   1   5  12  13  1  gd_10  ; -CH2-CR1-
  15  13  12   5  1  gd_10  ; -C-CH2-
;U1
  12  13  15  17  1  gd_15  ; -C-C-, double bond
  13  15  17  18  1  gd_34  ; -CH2-CH2-
; Connection U1-> U2
  15  17  18  19  1  gd_10  ; -CH2-CR1-
  21  19  18  17  1  gd_10  ; -C-CH2-
;U2
  18  19  21  23  1  gd_15  ; -C-C-, double bond
  19  21  23  24  1  gd_34  ; -CH2-CH2-
; Connection U2-> U3
  21  23  24  25  1  gd_10  ; -CH2-CR1-
  27  25  24  23  1  gd_10  ; -C-CH2-
;U3
  24  25  27  29  1  gd_15  ; -C-C-, double bond
  25  27  29  30  1  gd_34  ; -CH2-CH2-
; Connection U3-> U10
  27  29  30  31  1  gd_10  ; -CH2-CR1-
  33  31  30  29  1  gd_10  ; -C-CH2-
;U4
  30  31  33  35  1  gd_15  ; -C-C-, double bond
  31  33  35  36  1  gd_34  ; -CH2-CH2-
; Connection U4-> U5
  33  35  36  37  1  gd_10  ; -CH2-CR1-
  39  37  36  35  1  gd_10  ; -C-CH2-
;U5
  36  37  39  41  1  gd_15  ; -C-C-, double bond
  37  39  41  42  1  gd_34  ; -CH2-CH2-
; Connection U5-> U6
  39  41  42  43  1  gd_10  ; -CH2-CR1-
  45  43  42  41  1  gd_10  ; -C-CH2-
;U6
  42  43  45  47  1  gd_15  ; -C-C-, double bond
  43  45  47  48  1  gd_34  ; -CH2-CH2-
; Connection U6-> U7
  45  47  48  49  1  gd_10  ; -CH2-CR1-
  51  49  48  47  1  gd_10  ; -C-CH2-
;U7
  48  49  51  53  1  gd_15  ; -C-C-, double bond
  49  51  53  54  1  gd_34  ; -CH2-CH2-
; Connection U7-> U8
  51  53  54  55  1  gd_10  ; -CH2-CR1-
  57  55  54  53  1  gd_10  ; -C-CH2-
;U8
  54  55  57  59  1  gd_15  ; -C-C-, double bond
  55  57  59  60  1  gd_34  ; -CH2-CH2-
; Connection U8-> U9
  57  59  60  61  1  gd_10  ; -CH2-CR1-
  63  61  60  59  1  gd_10  ; -C-CH2-
;U9
```

```
  60  61  63  65  1  gd_15  ; -C-C-, double bond

[ mapping ]
 1     8  9 10 11
 2     1  2  3  4
 3     5  6  7
 4    12 13 14 15 16 17
 5    18 19 20 21 22 23
 6    24 25 26 27 28 29
 7    30 31 32 33 34 35
 8    36 37 38 39 40 41
 9    42 43 44 45 46 47
10    48 49 50 51 52 53
11    54 55 56 57 58 59
12    60 61 62 63 64 65
```

Plastoquinol-9

```
[ moleculetype ]
; Name   nrexcl
plqol      3

[ atoms ]
;  nr type  resnr resid atom cgnr charge  mass
; Headgroup
   1    C   1     PLQ   C     1   -0.390  12.0110
   2    HC  1     PLQ   H     2    0.260   1.0080
   3    C   1     PLQ   C     3    0.370  12.0110
   4    O   1     PLQ   O     4   -0.610  15.9994
   5    H   1     PLQ   H     5    0.430   1.0080
   6    C   1     PLQ   C     6   -0.110  12.0110
   7    C   1     PLQ   C     7    0.190  12.0110
   8    O   1     PLQ   O     8   -0.540  15.9994
   9    H   1     PLQ   H     9    0.380   1.008
  10    C   1     PLQ   C    10    0.000  12.0110
  11    CH3 1     PLQ   C    11    0.000  15.0350
  12    C   1     PLQ   C    12    0.000  12.0110
  13    CH3 1     PLQ   C    13    0.000  15.0350
; U1: First tail unit
  14    CH2 1     PLQ   C    14    0.292  14.0270
  15    C   1     PLQ   C    15    -0.59  12.0110
  16    HC  1     PLQ   H    16    0.185  1.0080
  17    C   1     PLQ   C    17    0.117  12.0110
  18    CH3 1     PLQ   C    18    0.000  15.0350
  19    CH2 1     PLQ   C    19    0.000  14.0270
; U2: First tail unit
  20    CH2 1     PLQ   C    20    0.292  14.0270
  21    C   1     PLQ   C    21    -0.59  12.0110
  22    HC  1     PLQ   H    22    0.185  1.0080
  23    C   1     PLQ   C    23    0.117  12.0110
  24    CH3 1     PLQ   C    24    0.000  15.0350
  25    CH2 1     PLQ   C    25    0.000  14.0270
; U3: First tail unit
  26    CH2 1     PLQ   C    26    0.292  14.0270
  27    C   1     PLQ   C    27    -0.59  12.0110
  28    HC  1     PLQ   H    28    0.185  1.0080
  29    C   1     PLQ   C    29    0.117  12.0110
  30    CH3 1     PLQ   C    30    0.000  15.0350
```

```
  31    CH2 1     PLQ   C    31    0.000  14.0270
; U4: First tail unit
  32    CH2 1     PLQ   C    32    0.292  14.0270
  33    C   1     PLQ   C    33    -0.59  12.0110
  34    HC  1     PLQ   H    34    0.185  1.0080
  35    C   1     PLQ   C    35    0.117  12.0110
  36    CH3 1     PLQ   C    36    0.000  15.0350
  37    CH2 1     PLQ   C    37    0.000  14.0270
; U5: First tail unit
  38    CH2 1     PLQ   C    38    0.292  14.0270
  39    C   1     PLQ   C    39    -0.59  12.0110
  40    HC  1     PLQ   H    40    0.185  1.0080
  41    C   1     PLQ   C    41    0.117  12.0110
  42    CH3 1     PLQ   C    42    0.000  15.0350
  43    CH2 1     PLQ   C    43    0.000  14.0270
; U6: First tail unit
  44    CH2 1     PLQ   C    44    0.292  14.0270
  45    C   1     PLQ   C    45    -0.59  12.0110
  46    HC  1     PLQ   H    46    0.185  1.0080
  47    C   1     PLQ   C    47    0.117  12.0110
  48    CH3 1     PLQ   C    48    0.000  15.0350
  49    CH2 1     PLQ   C    49    0.000  14.0270
; U7: First tail unit
  50    CH2 1     PLQ   C    50    0.292  14.0270
  51    C   1     PLQ   C    51    -0.59  12.0110
  52    HC  1     PLQ   H    52    0.185  1.0080
  53    C   1     PLQ   C    53    0.117  12.0110
  54    CH3 1     PLQ   C    54    0.000  15.0350
  55    CH2 1     PLQ   C    55    0.000  14.0270
; U8: First tail unit
  56    CH2 1     PLQ   C    56    0.292  14.0270
  57    C   1     PLQ   C    57    -0.59  12.0110
  58    HC  1     PLQ   H    58    0.185  1.0080
  59    C   1     PLQ   C    59    0.117  12.0110
  60    CH3 1     PLQ   C    60    0.000  15.0350
  61    CH2 1     PLQ   C    61    0.000  14.0270
; U9: First tail unit
  62    CH2 1     PLQ   C    62    0.292  14.0270
  63    C   1     PLQ   C    63    -0.59  12.0110
  64    HC  1     PLQ   H    64    0.185  1.0080
  65    C   1     PLQ   C    65    0.117  12.0110
  66    CH3 1     PLQ   C    66    0.000  15.0350
  67    CH3 1     PLQ   C    67    0.000  15.0350

[ bonds ]
;  i   j  funct
   1   2  2     gb_3
   1   3  2     gb_22
   1   6  2     gb_10
   3   4  2     gb_13
   3  10  2     gb_26
   4   5  2     gb_1
   6   7  2     gb_26
   7   8  2     gb_13
   7  12  2     gb_26
   8   9  2     gb_1
  10  11  2     gb_26
  10  12  2     gb_12
  12  13  2     gb_26
; Connection head group-> U1
   6  14  2     gb_26
```

```
; U1
  15  16  2     gb_3
  15  17  2     gb_10
  15  14  2     gb_26
  17  19  2     gb_26
  17  18  2     gb_26
; Connection U1-> U2
  19  20  2     gb_26
; U2
  21  22  2     gb_3
  21  23  2     gb_10
  21  20  2     gb_26
  23  25  2     gb_26
  23  24  2     gb_26
; Connection U2-> U3
  25  26  2     gb_26
; U3
  27  28  2     gb_3
  27  29  2     gb_10
  27  26  2     gb_26
  29  31  2     gb_26
  29  30  2     gb_26
; Connection U3-> U4
  31  32  2     gb_26
; U4
  33  34  2     gb_3
  33  35  2     gb_10
  33  32  2     gb_26
  35  37  2     gb_26
  35  36  2     gb_26
; Connection U4-> U5
  37  38  2     gb_26
; U5
  39  40  2     gb_3
  39  41  2     gb_10
  39  38  2     gb_26
  41  43  2     gb_26
  41  42  2     gb_26
; Connection U5-> U6
  43  44  2     gb_26
; U6
  45  46  2     gb_3
  45  47  2     gb_10
  45  44  2     gb_26
  47  49  2     gb_26
  47  48  2     gb_26
; Connection U6-> U7
  49  50  2     gb_26
; U7
  51  52  2     gb_3
  51  53  2     gb_10
  51  50  2     gb_26
  53  55  2     gb_26
  53  54  2     gb_26
; Connection U7-> U8
  55  56  2     gb_26
; U8
  57  58  2     gb_3
  57  59  2     gb_10
  57  56  2     gb_26
  59  61  2     gb_26
```

```
  59  60  2     gb_26
; Connection U8-> U9
  61  62  2     gb_26
; U9
  63  64  2     gb_3
  63  65  2     gb_10
  63  62  2     gb_26
  65  67  2     gb_26
  65  66  2     gb_26

[ pairs ]
;  i   j  funct
; Head
    1   8  1
    1  11  1
    1  12  1
    2   4  1
    2   7  1
    2  10  1
    3   7  1
    3  13  1
    4   6  1
    4  11  1
    4  12  1
    6  10  1
    6  13  1
    7  11  1
    8  10  1
    8  13  1
   11  13  1
; Connection head group-> U1
    1  15  1
    2  14  1
    3  14  1
    6  16  1
    6  17  1
    7  15  1
    8  14  1
   12  14  1
; U1
   14  18  1
   14  19  1
   16  18  1
   16  19  1
; Connection U1-> U2
   15  20  1
   17  21  1
   18  20  1
   19  22  1
   19  23  1
; U2
   20  24  1
   20  25  1
   22  24  1
   22  25  1
; Connection U2-> U3
   21  26  1
   23  27  1
   24  26  1
   25  28  1
   25  29  1
```

```
; U3
   26  30  1
   26  31  1
   28  30  1
   28  31  1
; Connection U3-> U4
   27  32  1
   29  33  1
   30  32  1
   31  34  1
   31  35  1
; U4
   32  36  1
   32  37  1
   34  36  1
   34  37  1
; Connection U4-> U5
   33  38  1
   35  39  1
   36  38  1
   37  40  1
   37  41  1
; U5
   38  42  1
   38  43  1
   40  42  1
   40  43  1
; Connection U5-> U6
   39  44  1
   41  45  1
   42  44  1
   43  46  1
   43  47  1
; U6
   44  48  1
   44  49  1
   46  48  1
   46  49  1
; Connection U6-> U7
   45  50  1
   47  51  1
   48  50  1
   49  52  1
   49  53  1
; U7
   50  54  1
   50  55  1
   52  54  1
   52  55  1
; Connection U7-> U8
   51  56  1
   53  57  1
   54  56  1
   55  58  1
   55  59  1
; U8
   56  60  1
   56  61  1
   58  60  1
   58  61  1
; Connection U8-> U9
```

```
   57  62  1
   59  63  1
   60  62  1
   61  64  1
   61  65  1
; U9
   62  66  1
   62  67  1
   64  66  1
   64  67  1

[ angles ]
;  i   j   k funct
   1   3   4  2    ga_30
   1   3  10  2    ga_27
   1   6   7  2    ga_27
   2   1   3  2    ga_25
   2   1   6  2    ga_25
   3   4   5  2    ga_42
   3   1   6  2    ga_27
   3  10  11  2    ga_27
   3  10  12  2    ga_27
   4   3  10  2    ga_30
   6   7   8  2    ga_30
   6   7  12  2    ga_27
   7   8   9  2    ga_42
   7  12  10  2    ga_27
   7  12  13  2    ga_27
   8   7  12  2    ga_30
  10  12  13  2    ga_27
  11  10  12  2    ga_35
  13  12  10  2    ga_35
; Connection head group-> U1
   1   6  14  2    ga_27
   7   6  14  2    ga_27
   6  14  15  2    ga_15
; U1
  15  17  19  2    ga_27
  15  17  18  2    ga_27
  16  15  17  2    ga_25
  16  15  14  2    ga_25
  17  15  14  2    ga_27
  17  15  16  2    ga_27
; Connection U1 -> U2
  17  19  20  2    ga_15
  21  20  19  2    ga_15
; U2
  21  23  25  2    ga_27
  21  23  24  2    ga_27
  22  21  23  2    ga_25
  22  21  20  2    ga_25
  23  21  20  2    ga_27
  23  21  22  2    ga_27
; Connection U2 -> U3
  23  25  26  2    ga_15
  27  26  25  2    ga_15
; U3
  27  29  31  2    ga_27
  27  29  30  2    ga_27
  28  27  29  2    ga_25
  28  27  26  2    ga_25
```

```
  29  27  26  2    ga_27
  29  27  28  2    ga_27
; Connection U3 -> U4
  29  31  32  2    ga_15
  33  32  31  2    ga_15
; U4
  33  35  37  2    ga_27
  33  35  36  2    ga_27
  34  33  35  2    ga_25
  34  33  32  2    ga_25
  35  33  32  2    ga_27
  35  33  34  2    ga_27
; Connection U4 -> U5
  35  37  38  2    ga_15
  39  38  37  2    ga_15
; U5
  39  41  43  2    ga_27
  39  41  42  2    ga_27
  40  39  41  2    ga_25
  40  39  38  2    ga_25
  41  39  38  2    ga_27
  41  39  40  2    ga_27
; Connection U5 -> U6
  41  43  44  2    ga_15
  45  44  43  2    ga_15
; U6
  45  47  49  2    ga_27
  45  47  48  2    ga_27
  46  45  47  2    ga_25
  46  45  44  2    ga_25
  47  45  44  2    ga_27
  47  45  46  2    ga_27
; Connection U6 -> U7
  47  49  50  2    ga_15
  51  50  49  2    ga_15
; U7
  51  53  55  2    ga_27
  51  53  54  2    ga_27
  52  51  53  2    ga_25
  52  51  50  2    ga_25
  53  51  50  2    ga_27
  53  51  52  2    ga_27
; Connection U7 -> U8
  53  55  56  2    ga_15
  57  56  55  2    ga_15
; U8
  57  59  61  2    ga_27
  57  59  60  2    ga_27
  58  57  59  2    ga_25
  58  57  56  2    ga_25
  59  57  56  2    ga_27
  59  57  58  2    ga_27
; Connection U8 -> U9
  59  61  62  2    ga_15
  63  62  61  2    ga_15
; U9
  63  65  67  2    ga_27
  63  65  66  2    ga_27
  64  63  65  2    ga_25
  64  63  62  2    ga_25
  65  63  62  2    ga_27
```

```
  65  63  64  2    ga_27

[ dihedrals ]
;  i   j   k   l funct
; Head group
   1   2   3   6  2   gi_1
   1   6   7  12  2   gi_1
   3   1   6   7  2   gi_1
   3   1   4  10  2   gi_1
   6   7  12  10  2   gi_1
   7  12  10   3  2   gi_1
   7   6   8  12  2   gi_1
  10   3  11  12  2   gi_1
  10   3   1   6  2   gi_1
  12  10   3   1  2   gi_1
  12   7  10  13  2   gi_1
; Connection head group-> U1
   6   1  14   7  2   gi_1
; U1
  15  16  17  14  2   gi_1
  17  15  19  18  2   gi_1
; U2
  21  22  23  20  2   gi_1
  23  21  25  24  2   gi_1
; U3
  27  28  29  26  2   gi_1
  29  27  31  30  2   gi_1
; U4
  33  34  35  32  2   gi_1
  35  33  37  36  2   gi_1
; U5
  39  40  41  38  2   gi_1
  41  39  43  42  2   gi_1
; U6
  45  46  47  44  2   gi_1
  47  45  49  48  2   gi_1
; U7
  51  52  53  50  2   gi_1
  53  51  55  54  2   gi_1
; U8
  57  58  59  56  2   gi_1
  59  57  61  60  2   gi_1
; U9
  63  64  65  62  2   gi_1
  65  63  67  66  2   gi_1

[ dihedrals ]
;  ai   aj   ak   al  funct    ph0      cp     mult
; Head group
   1   3   4   5  1    gd_11 ; -C-OA-H
  12   7   8   9  1    gd_11 ; -C-OA-H
; Connection head group-> U1
   1   6  14  15  1   gd_40  ; -CH2-CR1-
  17  15  14   6  1   gd_40  ; -C-CH2-
;U1
  14  15  17  19  1   gd_15  ; -C-C-, double bond
  15  17  19  20  1   gd_34  ; -CH2-CH2-
; Connection U1-> U2
  17  19  20  21  1   gd_40  ; -CH2-CR1-
  23  21  20  19  1   gd_40  ; -C-CH2-
;U2
```

```
  20  21  23  25  1   gd_15  ; -C-C-, double bond
  21  23  25  26  1   gd_34  ; -CH2-CH2-
; Connection U2-> U3
  23  25  26  27  1   gd_40  ; -CH2-CR1-
  29  27  26  25  1   gd_40  ; -C-CH2-
;U3
  26  27  29  31  1   gd_15  ; -C-C-, double bond
  27  29  31  32  1   gd_34  ; -CH2-CH2-
; Connection U3-> U10
  29  31  32  33  1   gd_40  ; -CH2-CR1-
  35  33  32  31  1   gd_40  ; -C-CH2-
;U4
  32  33  35  37  1   gd_15  ; -C-C-, double bond
  33  35  37  38  1   gd_34  ; -CH2-CH2-
; Connection U4-> U5
  35  37  38  39  1   gd_40  ; -CH2-CR1-
  41  39  38  37  1   gd_40  ; -C-CH2-
;U5
  38  39  41  43  1   gd_15  ; -C-C-, double bond
  39  41  43  44  1   gd_34  ; -CH2-CH2-
; Connection U5-> U6
  41  43  44  45  1   gd_40  ; -CH2-CR1-
  47  45  44  43  1   gd_40  ; -C-CH2-
;U6
  44  45  47  49  1   gd_15  ; -C-C-, double bond
  45  47  49  50  1   gd_34  ; -CH2-CH2-
; Connection U6-> U7
  47  49  50  51  1   gd_40  ; -CH2-CR1-
  53  51  50  49  1   gd_40  ; -C-CH2-
;U7
  50  51  53  55  1   gd_15  ; -C-C-, double bond
  51  53  55  56  1   gd_34  ; -CH2-CH2-
; Connection U7-> U8
  53  55  56  57  1   gd_40  ; -CH2-CR1-
  59  57  56  55  1   gd_40  ; -C-CH2-
;U8
  56  57  59  61  1   gd_15  ; -C-C-, double bond
  57  59  61  62  1   gd_34  ; -CH2-CH2-
; Connection U8-> U9
  59  61  62  63  1   gd_40  ; -CH2-CR1-
  65  63  62  61  1   gd_40  ; -C-CH2-
;U9
  62  63  65  67  1   gd_15  ; -C-C-, double bond

[ mapping ]
 1    10 11 12 13
 2     1  2  3  4  5
 3     6  7  8  9
 4    14 15 16 17 18 19
 5    20 21 22 23 24 25
 6    26 27 28 29 30 31
 7    32 33 34 35 36 37
 8    38 39 40 41 42 43
 9    44 45 46 47 48 49
10    50 51 52 53 54 55
11    56 57 58 59 60 61
12    62 63 64 65 66 67
```

Chlorophyll A

```
[ moleculetype ]
; Name nrexcl
CLA      3

[ atoms ]
; nr  type resnr resid atom cgnr charge  mass
   1   CH3  1    CLA   C19   1    0.000  15.0350
   2   CH1  1    CLA   C18   2    0.000  13.0190
   3   CH3  1    CLA   C20   3    0.000  15.0350
   4   CH2  1    CLA   C17   4    0.000  14.0270
   5   CH2  1    CLA   C16   5    0.000  14.0270
   6   CH2  1    CLA   C15   6    0.000  14.0270
   7   CH1  1    CLA   C13   7    0.000  13.0190
   8   CH3  1    CLA   C14   8    0.000  15.0350
   9   CH2  1    CLA   C12   9    0.000  14.0270
  10   CH2  1    CLA   C11  10    0.000  14.0270
  11   CH2  1    CLA   C10  11    0.000  14.0270
  12   CH1  1    CLA    C8  12    0.000  13.0190
  13   CH3  1    CLA    C9  13    0.000  15.0350
  14   CH2  1    CLA    C7  14    0.000  14.0270
  15   CH2  1    CLA    C6  15    0.000  14.0270
  16   CH2  1    CLA    C5  16    0.000  14.0270
  17     C  1    CLA    C3  17    0.000  12.0110
  18   CH3  1    CLA    C4  18    0.000  15.0350
  19   CH2  1    CLA    C2  19    0.000  13.0190
  20   CH2  1    CLA    C1  20    0.160  14.0270
  21    OA  1    CLA   O2A  21   -0.240  15.9994
  22     C  1    CLA   CGA  22    0.460  12.0110
  23     O  1    CLA   O1A  23   -0.380  15.9994
  24   CH2  1    CLA   CBA  24    0.000  14.0270
  25   CH2  1    CLA   CAA  25    0.000  14.0270
  26   CR1  1    CLA   C2A  26    0.000  13.0190
  27   CR1  1    CLA   C3A  27    0.000  12.0110
  28   CH3  1    CLA   CMA  28    0.000  15.0350
  29     C  1    CLA   C4A  29    0.100  12.0110
  30     C  1    CLA   CHB  30   -0.100  12.0110
  31     H  1    CLA    HB  31    0.100   1.0080
  32    NR  1    CLA    NA  32   -0.450  14.0067
  33     C  1    CLA   C1A  33    0.100  12.0110
  34     C  1    CLA   CHA  34    0.000  12.0110
  35   CH1  1    CLA   CBD  35    0.000  13.0190
  36     C  1    CLA   CGD  36    0.460  12.0110
  37     O  1    CLA   O1D  37   -0.380  15.9994
  38    OA  1    CLA   O2D  38   -0.240  15.9994
  39   CH3  1    CLA   CED  39    0.160  15.0350
  40     C  1    CLA   CAD  40    0.330  12.0110
  41     O  1    CLA   OBD  41   -0.330  15.9994
  42     C  1    CLA   C3D  42    0.000  12.0110
  43     C  1    CLA   C2D  43    0.000  12.0110
  44   CH3  1    CLA   CMD  44    0.000  15.0350
  45     C  1    CLA   C4D  45    0.100  12.0110
  46    NR  1    CLA    ND  46   -0.450  14.0067
  47     C  1    CLA   C1D  47    0.100  12.0110
  48     C  1    CLA   CHD  48   -0.100  12.0110
  49     H  1    CLA    HD  49    0.100   1.0080
  50     C  1    CLA   C4C  50    0.100  12.0110
  51    NR  1    CLA    NC  51   -0.450  14.0067
  52     C  1    CLA   C3C  52    0.100  12.0110
  53   CH2  1    CLA   CAC  53    0.000  14.0270
```

```
  54   CH3  1    CLA   CBC  54    0.000  15.0350
  55     C  1    CLA   C2C  55    0.000  12.0110
  56   CH3  1    CLA   CMC  56    0.000  15.0350
  57     C  1    CLA   C1C  57    0.100  12.0110
  58     C  1    CLA   CHC  58   -0.100  12.0110
  59     H  1    CLA    HC  59    0.100   1.0080
  60     C  1    CLA   C4B  60    0.100  12.0110
  61    NR  1    CLA    NB  61   -0.450  14.0067
  62     C  1    CLA   C1B  62    0.000  12.0110
  63     C  1    CLA   C2B  63    0.000  12.0110
  64   CH3  1    CLA   CMB  64    0.000  15.0350
  65     C  1    CLA   C3B  65    0.000  12.0110
  66     C  1    CLA   CAB  66    0.000  12.0110
  67   CH3  1    CLA   CBB  67    0.000  15.0350
  68  MG2+  1    CLA    MG  68    1.000  24.3050

[ bonds ]
; i     j   func
   2    1  2    gb_27
   2    3  2    gb_27
   2    4  2    gb_27
   4    5  2    gb_27
   5    6  2    gb_27
   7    6  2    gb_27
   7    8  2    gb_27
   7    9  2    gb_27
   9   10  2    gb_27
  10   11  2    gb_27
  12   11  2    gb_27
  12   13  2    gb_27
  12   14  2    gb_27
  14   15  2    gb_27
  15   16  2    gb_27
  17   16  2    gb_27
  17   18  2    gb_27
  17   19  2    gb_27
  20   19  2    gb_27
  20   21  2    gb_18
  22   21  2    gb_13
  22   23  2    gb_5
  22   24  2    gb_27
  24   25  2    gb_27
  26   25  2    gb_27
  26   27  2    gb_27
  26   33  2    gb_15
  27   28  2    gb_27
  27   29  2    gb_15
  29   30  2    gb_17
  29   32  2    gb_14
  62   30  2    gb_17
  30   31  2    gb_3
  33   32  2    gb_14
  33   34  2    gb_17
  35   34  2    gb_15
  34   45  2    gb_10
  35   36  2    gb_15
  35   40  2    gb_15
  36   37  2    gb_5
  36   38  2    gb_13
  38   39  2    gb_18
  40   41  2    gb_5
```

```
  40   42  2    gb_10
  42   43  2    gb_10
  42   45  2    gb_10
  43   44  2    gb_27
  43   47  2    gb_10
  45   46  2    gb_9
  47   46  2    gb_9
  47   48  2    gb_10
  48   49  2    gb_3
  50   48  2    gb_10
  50   51  2    gb_9
  50   52  2    gb_10
  57   51  2    gb_9
  52   53  2    gb_27
  52   55  2    gb_10
  53   54  2    gb_27
  55   56  2    gb_27
  55   57  2    gb_10
  57   58  2    gb_10
  58   59  2    gb_3
  60   58  2    gb_10
  60   61  2    gb_9
  60   65  2    gb_10
  62   61  2    gb_9
  62   63  2    gb_10
  63   64  2    gb_27
  63   65  2    gb_10
  65   66  2    gb_27
  66   67  2    gb_27
  68   32  2    gb_35
  68   46  2    gb_35
  68   51  2    gb_35
  68   61  2    gb_35

[ pairs ]
;  i    j  func
   1    5  1
   2    6  1
   3    5  1
   4    7  1
   5    8  1
   5    9  1
   6   10  1
   7   11  1
   8   10  1
   9   12  1
  10   13  1
  10   14  1
  11   15  1
  12   16  1
  13   15  1
  14   17  1
  15   18  1
  15   19  1
  16   20  1
  17   21  1
  18   20  1
  19   22  1
  20   23  1
  20   24  1
  21   25  1
```

```
22   26  1
23   25  1
24   27  1
24   33  1
25   28  1
25   29  1
25   32  1
25   34  1
26   30  1
26   35  1
26   45  1
27   34  1
27   62  1
28   30  1
28   32  1
28   33  1
29   34  1
29   61  1
29   63  1
30   33  1
30   60  1
30   64  1
30   65  1
31   27  1
31   32  1
31   61  1
31   63  1
32   35  1
32   45  1
32   62  1
33   36  1
33   40  1
33   42  1
33   46  1
34   37  1
34   38  1
34   41  1
34   43  1
34   47  1
35   39  1
35   43  1
35   46  1
36   41  1
36   42  1
36   45  1
37   39  1
37   40  1
38   40  1
40   44  1
40   46  1
40   47  1
41   43  1
41   45  1
42   48  1
43   50  1
44   45  1
44   46  1
44   48  1
45   48  1
46   50  1
47   51  1
```

```
  47   52  1
  48   53  1
  48   55  1
  48   57  1
  49   43  1
  49   46  1
  49   51  1
  49   52  1
  50   54  1
  50   56  1
  50   58  1
  51   53  1
  51   56  1
  51   60  1
  52   58  1
  53   56  1
  53   57  1
  54   55  1
  55   60  1
  56   58  1
  57   61  1
  57   65  1
  58   62  1
  58   63  1
  58   66  1
  59   51  1
  59   55  1
  59   65  1
  59   61  1
  60   64  1
  60   67  1
  61   64  1
  61   66  1
  62   66  1
  63   67  1
  64   66  1
  68   26  1
  68   34  1
  68   27  1
  68   30  1
  68   42  1
  68   43  1
  68   48  1
  68   52  1
  68   55  1
  68   58  1
  68   65  1
  68   63  1

[ angles ]
; i   j   k  func
   1   2   3  2   ga_13
   1   2   4  2   ga_13
   3   2   4  2   ga_13
   2   4   5  2   ga_13
   4   5   6  2   ga_13
   5   6   7  2   ga_13
   6   7   8  2   ga_13
   6   7   9  2   ga_13
   8   7   9  2   ga_13
   7   9  10  2   ga_13
```

```
 9  10  11  2   ga_13
10  11  12  2   ga_13
11  12  13  2   ga_13
11  12  14  2   ga_13
13  12  14  2   ga_13
12  14  15  2   ga_13
14  15  16  2   ga_13
15  16  17  2   ga_13
16  17  18  2   ga_19
16  17  19  2   ga_19
18  17  19  2   ga_19
17  19  20  2   ga_19
19  20  21  2   ga_13
20  21  22  2   ga_10
21  22  23  2   ga_33
21  22  24  2   ga_19
23  22  24  2   ga_30
22  24  25  2   ga_13
24  25  26  2   ga_13
25  26  27  2   ga_13
25  26  33  2   ga_13
27  26  33  2   ga_7
26  27  28  2   ga_13
26  27  29  2   ga_7
28  27  29  2   ga_13
27  29  30  2   ga_39
27  29  32  2   ga_7
30  29  32  2   ga_27
29  30  62  2   ga_27
29  32  33  2   ga_7
26  33  32  2   ga_7
26  33  34  2   ga_39
31  30  29  2   ga_19
31  30  62  2   ga_19
32  33  34  2   ga_27
33  34  35  2   ga_39
33  34  45  2   ga_27
35  34  45  2   ga_7
34  35  36  2   ga_13
34  35  40  2   ga_7
36  35  40  2   ga_13
35  36  37  2   ga_30
35  36  38  2   ga_19
37  36  38  2   ga_33
36  38  39  2   ga_10
35  40  41  2   ga_30
35  40  42  2   ga_7
41  40  42  2   ga_30
40  42  43  2   ga_39
40  42  45  2   ga_7
43  42  45  2   ga_7
42  43  44  2   ga_27
42  43  47  2   ga_7
44  43  47  2   ga_27
34  45  42  2   ga_7
34  45  46  2   ga_19
42  45  46  2   ga_19
45  46  47  2   ga_7
43  47  46  2   ga_19
43  47  48  2   ga_39
46  47  48  2   ga_27
```

```
  47  48  50  2   ga_27
  48  50  51  2   ga_27
  48  50  52  2   ga_39
  49  48  47  2   ga_19
  49  48  50  2   ga_19
  51  50  52  2   ga_19
  50  51  57  2   ga_7
  50  52  53  2   ga_27
  50  52  55  2   ga_7
  53  52  55  2   ga_27
  52  53  54  2   ga_13
  52  55  56  2   ga_27
  52  55  57  2   ga_7
  56  55  57  2   ga_27
  51  57  55  2   ga_19
  51  57  58  2   ga_27
  55  57  58  2   ga_39
  57  58  60  2   ga_27
  58  60  61  2   ga_27
  58  60  65  2   ga_39
  59  58  57  2   ga_19
  59  58  60  2   ga_19
  61  60  65  2   ga_19
  60  61  62  2   ga_7
  30  62  61  2   ga_27
  30  62  63  2   ga_39
  61  62  63  2   ga_19
  62  63  64  2   ga_27
  62  63  65  2   ga_7
  64  63  65  2   ga_27
  60  65  63  2   ga_7
  60  65  66  2   ga_27
  63  65  66  2   ga_27
  65  66  67  2   ga_13
  68  32  29  2   ga_34
  68  32  33  2   ga_34
  68  46  45  2   ga_34
  68  46  47  2   ga_34
  68  51  50  2   ga_34
  68  51  57  2   ga_34
  68  61  60  2   ga_34
  68  61  62  2   ga_34
  32  68  46  2   ga_2
  46  68  51  2   ga_2
  51  68  61  2   ga_2
  61  68  32  2   ga_2

[ dihedrals ]
;  i   j   k   l  func
    2   1   4   3   2   gi_2
    7   6   9   8   2   gi_2
   12  11  14  13   2   gi_2
   17  16  18  19   2   gi_1
   22  21  23  24   2   gi_1
   26  25  33  27   2   gi_2
   27  26  28  29   2   gi_2
   29  27  30  32   2   gi_1
   30  29  62  31   2   gi_1
   33  26  32  34   2   gi_1
   34  45  35  33   2   gi_1
   35  34  40  36   2   gi_2
```

```
36  35  37  38   2  gi_1
40  35  41  42   2  gi_1
42  45  43  40   2  gi_1
43  47  44  42   2  gi_1
45  46  42  34   2  gi_1
47  48  46  43   2  gi_1
48  50  47  49   2  gi_1
50  48  51  52   2  gi_1
52  55  53  50   2  gi_1
55  57  56  52   2  gi_1
57  51  58  55   2  gi_1
58  60  57  59   2  gi_1
60  58  61  65   2  gi_1
62  30  61  63   2  gi_1
63  65  64  62   2  gi_1
65  66  63  60   2  gi_1
42  43  47  46   2  gi_1
43  47  46  45   2  gi_1
47  46  45  42   2  gi_1
46  45  42  43   2  gi_1
45  42  43  47   2  gi_1
50  51  57  55   2  gi_1
51  57  55  52   2  gi_1
57  55  52  50   2  gi_1
55  52  50  51   2  gi_1
52  50  51  57   2  gi_1
60  61  62  63   2  gi_1
61  62  63  65   2  gi_1
62  63  65  60   2  gi_1
63  65  60  61   2  gi_1
65  60  61  62   2  gi_1
68  29  33  32   2  gi_3
68  47  45  46   2  gi_3
68  50  57  51   2  gi_3
68  60  62  61   2  gi_3
 5   4   2   1   1  gd_34
 6   5   4   2   1  gd_34
 7   6   5   4   1  gd_34
 9   7   6   5   1  gd_34
10   9   7   6   1  gd_34
11  10   9   7   1  gd_34
12  11  10   9   1  gd_34
14  12  11  10   1  gd_34
15  14  12  11   1  gd_34
16  15  14  12   1  gd_34
17  16  15  14   1  gd_34
15  16  17  19   1  gd_40
20  19  17  16   1  gd_10
21  20  19  17   1  gd_40
19  20  21  22   1  gd_24
24  22  21  20   1  gd_29
25  24  22  21   1  gd_40
26  25  24  22   1  gd_34
33  26  25  24   1  gd_34
29  27  26  25   1  gd_34
25  26  33  34   1  gd_40
26  27  29  32   1  gd_40
27  29  30  62   1  gd_15
27  29  32  33   1  gd_14
63  62  30  29   1  gd_15
34  33  32  29   1  gd_14
```

```
   45  34  33  26   1   gd_10
   40  35  34  33   1   gd_40
   46  45  34  33   1   gd_10
   34  35  36  38   1   gd_40
   34  35  40  42   1   gd_40
   35  36  38  39   1   gd_29
   45  42  40  35   1   gd_10
   43  47  48  50   1   gd_15
   52  50  48  47   1   gd_15
   54  53  52  50   1   gd_40
   51  57  58  60   1   gd_15
   65  60  58  57   1   gd_15


[ mapping ]
; Atoms 29 33 45 47 50 57 60 62 are double mapped. Halve their atomistic mass!!
1      60 61 62
2      60 58 59 57
3      57 51 50
4      50 48 49 47
5      47 46 45
6      45 34 33
7      33 32 29
8      29 30 31 62
9      63 64
10     65 66 67
11     55 56
12     52 53 54
13     42 43 44
14     35 40 41
15     36 37 38 39
16     26 25
17     26 28
18     24 23 22 21 20
19     19 18 17 16
20     15 14 13 12 11
21     10  9  8  7  6
22      5  4  3  2  1
23     68
```

β-carotene

```
[ moleculetype ]
; Name nrexcl
BCR      3

[ atoms ]
; nr  type resnr resid atom cgnr charge mass
   1   CH3  1    BCR   C31   1   0.000  15.0350
   2   CH0  1    BCR    C1   1   0.000  12.0110
   3   CH3  1    BCR   C32   2   0.000  15.0350
   4  CH2r  1    BCR    C2   3   0.000  14.0270
   5  CH2r  1    BCR    C3   3   0.000  14.0270
   6  CH2r  1    BCR    C4   3   0.000  14.0270
   7     C  1    BCR    C5   3   0.000  12.0110
   8   CH3  1    BCR   C33   3   0.000  15.0350
   9     C  1    BCR    C6   3   0.000  12.0110
  10   CR1  1    BCR    C7   3   0.000  13.0190
  11   CR1  1    BCR    C8   4   0.000  13.0190
```

```
  12     C  1    BCR    C9   4   0.000  12.0110
  13   CH3  1    BCR   C34   4   0.000  15.0350
  14   CR1  1    BCR   C10   4   0.000  13.0190
  15   CR1  1    BCR   C11   5   0.000  13.0190
  16   CR1  1    BCR   C12   5   0.000  13.0190
  17     C  1    BCR   C13   6   0.000  12.0110
  18   CH3  1    BCR   C35   6   0.000  15.0350
  19   CR1  1    BCR   C14   6   0.000  13.0190
  20   CR1  1    BCR   C15   6   0.000  13.0190
  21   CR1  1    BCR   C16   6   0.000  13.0190
  22   CR1  1    BCR   C17   6   0.000  13.0190
  23     C  1    BCR   C18   7   0.000  12.0110
  24   CH3  1    BCR   C36   7   0.000  15.0350
  25   CR1  1    BCR   C19   7   0.000  13.0190
  26   CR1  1    BCR   C20   7   0.000  13.0190
  27   CR1  1    BCR   C21   8   0.000  13.0190
  28     C  1    BCR   C22   8   0.000  12.0110
  29   CH3  1    BCR   C37   8   0.000  15.0350
  30   CR1  1    BCR   C23   8   0.000  13.0190
  31   CR1  1    BCR   C24   8   0.000  13.0190
  32     C  1    BCR   C25   8   0.000  12.0110
  33   CH0  1    BCR   C30   8   0.000  12.0110
  34   CH3  1    BCR   C39   9   0.000  15.0350
  35   CH3  1    BCR   C40   9   0.000  15.0350
  36  CH2r  1    BCR   C29   9   0.000  14.0270
  37  CH2r  1    BCR   C28   9   0.000  14.0270
  38  CH2r  1    BCR   C27   9   0.000  14.0270
  39     C  1    BCR   C26   9   0.000  12.0110
  40   CH3  1    BCR   C38  10   0.000  15.0350

[ bonds ]
; i   j  funct
   2   1   2    gb_27
   2   3   2    gb_27
   2   4   2    gb_27
   4   5   2    gb_27
   5   6   2    gb_27
   6   7   2    gb_27
   7   8   2    gb_27
   9   2   2    gb_27
   7   9   2    gb_10
   9  10   2    gb_23
  10  11   2    gb_13
  11  12   2    gb_23
  12  13   2    gb_27
  12  14   2    gb_13
  14  15   2    gb_23
  15  16   2    gb_13
  16  17   2    gb_23
  17  18   2    gb_27
  17  19   2    gb_13
  19  20   2    gb_23
  20  21   2    gb_13
  21  22   2    gb_23
  22  23   2    gb_13
  23  24   2    gb_27
  23  25   2    gb_23
  25  26   2    gb_13
  26  27   2    gb_23
  27  28   2    gb_13
  28  29   2    gb_27
```

```
  28  30   2    gb_23
  30  31   2    gb_13
  31  32   2    gb_23
  32  39   2    gb_10
  39  40   2    gb_27
  38  39   2    gb_27
  37  38   2    gb_27
  36  37   2    gb_27
  33  36   2    gb_27
  33  34   2    gb_27
  33  35   2    gb_27
  32  33   2    gb_27

[ pairs ]
; i   j   funct
   1   5   1    ;   C31   C3
   1   7   1    ;   C31   C5
   1  10   1    ;   C31   C7
   2   6   1    ;    C1   C4
   2   8   1    ;    C1  C33
   2  11   1    ;    C1   C8
   3   5   1    ;   C32   C3
   3   7   1    ;   C32   C5
   3  10   1    ;   C32   C7
   4   7   1    ;    C2   C5
   4  10   1    ;    C2   C7
   5   8   1    ;    C3  C33
   5   9   1    ;    C3   C6
   6  10   1    ;    C4   C7
   7  11   1    ;    C5   C8
   8  10   1    ;   C33   C7
   9  12   1    ;    C6   C9
  10  13   1    ;    C7  C34
  10  14   1    ;    C7  C10
  11  15   1    ;    C8  C11
  12  16   1    ;    C9  C12
  13  15   1    ;   C34  C11
  14  17   1    ;   C10  C13
  15  18   1    ;   C11  C35
  15  19   1    ;   C11  C14
  16  20   1    ;   C12  C15
  17  21   1    ;   C13  C16
  18  20   1    ;   C35  C15
  19  22   1    ;   C14  C17
  20  23   1    ;   C15  C18
  21  24   1    ;   C16  C36
  21  25   1    ;   C16  C19
  22  26   1    ;   C17  C20
  23  27   1    ;   C18  C21
  24  26   1    ;   C36  C20
  25  28   1    ;   C19  C22
  26  29   1    ;   C20  C37
  26  30   1    ;   C20  C23
  27  31   1    ;   C21  C24
  28  32   1    ;   C22  C25
  29  31   1    ;   C37  C24
  30  33   1    ;   C23  C30
  30  39   1    ;   C23  C26
  31  34   1    ;   C24  C39
  31  35   1    ;   C24  C40
  31  36   1    ;   C24  C29
```

```
  31  38   1    ;   C24  C27
  31  40   1    ;   C24  C38
  32  37   1    ;   C25  C28
  33  38   1    ;   C30  C27
  33  40   1    ;   C30  C38
  34  37   1    ;   C39  C28
  34  39   1    ;   C39  C26
  35  37   1    ;   C40  C28
  35  39   1    ;   C40  C26
  36  39   1    ;   C29  C26
  37  40   1    ;   C28  C38

[ angles ]
; i   j   k   funct angle   force.c.
   1   2   3   2    109.5   520.0 ;   C31   C1  C32
   1   2   4   2    109.5   520.0 ;   C31   C1   C2
   1   2   9   2    109.5   520.0 ;   C31   C1   C6
   3   2   4   2    109.5   520.0 ;   C32   C1   C2
   3   2   9   2    109.5   520.0 ;   C32   C1   C6
   4   2   9   2    109.5   520.0 ;    C2   C1   C6
   2   4   5   2    109.5   520.0 ;    C1   C2   C3
   4   5   6   2    109.5   520.0 ;    C2   C3   C4
   5   6   7   2    109.5   520.0 ;    C3   C4   C5
   6   7   8   2    120.0   560.0 ;    C4   C5  C33
   6   7   9   2    120.0   560.0 ;    C4   C5   C6
   8   7   9   2    120.0   560.0 ;   C33   C5   C6
   2   9   7   2    120.0   560.0 ;    C1   C6   C5
   2   9  10   2    120.0   560.0 ;    C1   C6   C7
   7   9  10   2    120.0   560.0 ;    C5   C6   C7
   9  10  11   2    120.0   560.0 ;    C6   C7   C8
  10  11  12   2    120.0   560.0 ;    C7   C8   C9
  11  12  13   2    120.0   560.0 ;    C8   C9  C34
  11  12  14   2    120.0   560.0 ;    C8   C9  C10
  13  12  14   2    120.0   560.0 ;   C34   C9  C10
  12  14  15   2    120.0   560.0 ;    C9  C10  C11
  14  15  16   2    120.0   560.0 ;   C10  C11  C12
  15  16  17   2    120.0   560.0 ;   C11  C12  C13
  16  17  18   2    120.0   560.0 ;   C12  C13  C35
  16  17  19   2    120.0   560.0 ;   C12  C13  C14
  18  17  19   2    120.0   560.0 ;   C35  C13  C14
  17  19  20   2    120.0   560.0 ;   C13  C14  C15
  19  20  21   2    120.0   560.0 ;   C14  C15  C16
  20  21  22   2    120.0   560.0 ;   C15  C16  C17
  21  22  23   2    120.0   560.0 ;   C16  C17  C18
  22  23  24   2    120.0   560.0 ;   C17  C18  C36
  22  23  25   2    120.0   560.0 ;   C17  C18  C19
  24  23  25   2    120.0   560.0 ;   C36  C18  C19
  23  25  26   2    120.0   560.0 ;   C18  C19  C20
  25  26  27   2    120.0   560.0 ;   C19  C20  C21
  26  27  28   2    120.0   560.0 ;   C20  C21  C22
  27  28  29   2    120.0   560.0 ;   C21  C22  C37
  27  28  30   2    120.0   560.0 ;   C21  C22  C23
  29  28  30   2    120.0   560.0 ;   C37  C22  C23
  28  30  31   2    120.0   560.0 ;   C22  C23  C24
  30  31  32   2    120.0   560.0 ;   C23  C24  C25
  31  32  33   2    120.0   560.0 ;   C24  C25  C30
  31  32  39   2    120.0   560.0 ;   C24  C25  C26
  33  32  39   2    120.0   560.0 ;   C30  C25  C26
  32  33  34   2    109.5   520.0 ;   C25  C30  C39
  32  33  35   2    109.5   520.0 ;   C25  C30  C40
  32  33  36   2    109.5   520.0 ;   C25  C30  C29
```

```
  34  33  35   2    109.5   520.0 ;   C39  C30  C40
  34  33  36   2    109.5   520.0 ;   C39  C30  C29
  35  33  36   2    109.5   520.0 ;   C40  C30  C29
  33  36  37   2    109.5   520.0 ;   C30  C29  C28
  36  37  38   2    109.5   520.0 ;   C29  C28  C27
  37  38  39   2    109.5   520.0 ;   C28  C27  C26
  32  39  38   2    120.0   560.0 ;   C25  C26  C27
  32  39  40   2    120.0   560.0 ;   C25  C26  C38
  38  39  40   2    120.0   560.0 ;   C27  C26  C38

[ dihedrals ]
; i   j   k   l  func angle force.c.
   7   6   8   9   2   0.0   167.4 ; imp    C5   C4  C33   C6
   9   2  10   7   2   0.0   167.4 ; imp    C6   C1   C7   C5
  12  14  13  11   2   0.0   167.4 ; imp    C9  C10  C34   C8
  17  19  18  16   2   0.0   167.4 ; imp   C13  C14  C35  C12
  23  25  24  22   2   0.0   167.4 ; imp   C18  C19  C36  C17
  28  30  29  27   2   0.0   167.4 ; imp   C22  C23  C37  C21
  32  31  33  39   2   0.0   167.4 ; imp   C25  C24  C30  C26
  39  40  38  32   2   0.0   167.4 ; imp   C26  C38  C27  C25

[ dihedrals ]
; i  j  k  l  func
   9  2  4  5  1   gd_34
   4  2  9  7  1   gd_34
   2  4  5  6  1   gd_34
   4  5  6  7  1   gd_34
   5  6  7  9  1   gd_34
   6  7  9  2  1   gd_14
   7  9 10 11  1   gd_34
   9 10 11 12  1   gd_14
  10 11 12 14  1   gd_12
  11 12 14 15  1   gd_14
  12 14 15 16  1   gd_12
  14 15 16 17  1   gd_14
  15 16 17 19  1   gd_12
  16 17 19 20  1   gd_14
  17 19 20 21  1   gd_12
  19 20 21 22  1   gd_14
  20 21 22 23  1   gd_12
  21 22 23 25  1   gd_14
  22 23 25 26  1   gd_12
  23 25 26 27  1   gd_14
  25 26 27 28  1   gd_12
  26 27 28 30  1   gd_14
  27 28 30 31  1   gd_12
  28 30 31 32  1   gd_14
  30 31 32 39  1   gd_34
  31 32 39 38  1   gd_14
  32 39 38 37  1   gd_34
  39 38 37 36  1   gd_34
  38 37 36 33  1   gd_34
  37 36 33 32  1   gd_34
  36 33 32 39  1   gd_34

[ mapping ]
1 1 2 3
2 4 5 6
3 7 8 9 10
4 11 12 13 14 15
5 16 17 18 19 20
```

```
6 21 22 23 24 25
7 26 27 28 29 30
8 31 32 39 40
9 36 37 38
10 33 34 35
```

6.6 Martini topologies

Behind all the bonds, angles, proper and improper dihedrals the mean and standard deviations for all distributions obtained from the coarse grain and the mapped atomistic simulation are given. The values are in nanometers (bonds) or degrees (angles, proper and improper dihedrals).

Plastoquinone-9

```
[moleculetype]
; molname       nrexcl
 PLQ          1

[atoms]
;id type resnr residu atom  cgnr charge
  1  C3   1     PLQ9    PLQ1  1   0
  2  SNa  1     PLQ9    PLQ2  2   0
  3  SNa  1     PLQ9    PLQ3  3   0
  4  C3   1     PLQ9    PLQ4  4   0
  5  C3   1     PLQ9    PLQ5  5   0
  6  C3   1     PLQ9    PLQ6  6   0
  7  C3   1     PLQ9    PLQ7  7   0
  8  C3   1     PLQ9    PLQ8  8   0
  9  C3   1     PLQ9    PLQ9  9   0
 10  C3   1     PLQ9   PLQ10 10   0
 11  C3   1     PLQ9   PLQ11 11   0
 12  C3   1     PLQ9   PLQ12 12   0

[bonds]
;  i    j   funct length force.c  CG distr.    UA distr.
   3    4   1     0.40   3000.0 ; 0.391(0.030) 0.402(0.023)
   4    5   1     0.48   6000.0 ; 0.476(0.022) 0.483(0.015)
   5    6   1     0.48   6000.0 ; 0.477(0.021) 0.483(0.012)
   6    7   1     0.48   6000.0 ; 0.477(0.021) 0.484(0.012)
   7    8   1     0.48   6000.0 ; 0.478(0.022) 0.483(0.013)
   8    9   1     0.48   6000.0 ; 0.477(0.021) 0.483(0.014)
   9   10   1     0.48   6000.0 ; 0.477(0.021) 0.483(0.012)
  10   11   1     0.48   6000.0 ; 0.477(0.021) 0.483(0.012)
  11   12   1     0.48   6000.0 ; 0.478(0.021) 0.486(0.012)

[constraints]
;  i    j   funct length  CG distr.    UA distr.
   1    2   1     0.290 ; 0.290(0.004) 0.297(0.003)
   2    3   1     0.340 ; 0.340(0.004) 0.336(0.003)
   1    3   1     0.290 ; 0.290(0.004) 0.294(0.003)

[angles]
;  i    j    k   funct angle force.c CG distr.    UA distr.
   1    3    4   2     145.0  60.0 ;  136.9(10.6) 140.2( 7.8)
   2    3    4   2      90.0  80.0 ;   91.8(10.0)  88.9( 8.4)
   3    4    5   2     130.0  30.0 ;  125.6(18.7) 128.7(16.1)
   4    5    6   2     130.0  30.0 ;  126.7(18.3) 134.0(12.6)
   5    6    7   2     130.0  30.0 ;  127.6(17.7) 134.8(12.1)
```

```
   6    7    8   2     130.0  30.0 ;  127.8(18.4) 134.7(11.9)
   7    8    9   2     130.0  30.0 ;  127.4(18.2) 133.9(12.2)
   8    9   10   2     130.0  30.0 ;  126.7(17.8) 134.2(12.2)
   9   10   11   2     130.0  30.0 ;  127.2(17.6) 134.7(11.8)
  10   11   12   2     130.0  30.0 ;  128.2(18.0) 134.7(11.9)

[dihedrals]
;  i    j    k    l   funct angle force.c multipl. CG distr.    UA distr.
   2    3    4    5   1     0     1.0     2      ; -1.6(116.1) -1.8( 81.2)
   2    3    4    5   1     0     2.0     1      ; -1.6(116.1) -1.8( 81.2)
; The next dihedrals were not defined, because defining
; it doesn't improve the behavior or impairs the
; numerical stability.
;  3    4    5    6                              ;  1.0(101.0)  9.2(123.2)
;  4    5    6    7                              ;  2.3(102.5)  7.0(128.9)
;  5    6    7    8                              ; -1.3(108.0)  2.5(125.6)
;  6    7    8    9                              ; -6.4(102.9)  7.5(124.5)
;  7    8    9   10                              ; -0.4(101.8) 11.1(128.2)
;  8    9   10   11                              ; -3.7( 99.8)  4.3(131.1)
;  9   10   11   12                              ; -1.6(102.7)  0.3(126.6)

[exclusions]
;i  j
 1  4
 2  4
```

Plastoquinol-9

```
[moleculetype]
; molname       nrexcl
 PQol          1

[atoms]
;id  type resnr residu atom   cgnr charge
  1  C3   1     PQol    PLQ1   1   0
  2  SP2  1     PQol    PLQ2   2   0
  3  SP2  1     PQol    PLQ3   3   0
  4  C3   1     PQol    PLQ4   4   0
  5  C3   1     PQol    PLQ5   5   0
  6  C3   1     PQol    PLQ6   6   0
  7  C3   1     PQol    PLQ7   7   0
  8  C3   1     PQol    PLQ8   8   0
  9  C3   1     PQol    PLQ9   9   0
 10  C3   1     PQol   PLQ10  10   0
 11  C3   1     PQol   PLQ11  11   0
 12  C3   1     PQol   PLQ12  12   0

[bonds]
;   i    j funct length force.c   CG distr.    UA distr.
    3    4   1   0.40   3000.0  ; 0.391(0.030) 0.402(0.023)
    4    5   1   0.48   6000.0  ; 0.476(0.022) 0.483(0.015)
    5    6   1   0.48   6000.0  ; 0.477(0.021) 0.483(0.012)
    6    7   1   0.48   6000.0  ; 0.477(0.021) 0.484(0.012)
    7    8   1   0.48   6000.0  ; 0.478(0.022) 0.483(0.013)
    8    9   1   0.48   6000.0  ; 0.477(0.021) 0.483(0.014)
    9   10   1   0.48   6000.0  ; 0.477(0.021) 0.483(0.012)
   10   11   1   0.48   6000.0  ; 0.477(0.021) 0.483(0.012)
   11   12   1   0.48   6000.0  ; 0.478(0.021) 0.486(0.012)

[constraints]
```

```
;   i    j funct length  CG distr.    UA distr.
    1    2   1   0.290 ; 0.290(0.004) 0.297(0.003)
    2    3   1   0.340 ; 0.340(0.004) 0.336(0.003)
    1    3   1   0.290 ; 0.290(0.004) 0.294(0.003)

[angles]
;   i    j    k funct angle  force.c CG distr.    UA distr.
    1    3    4   2   145.0  60.0  ;  136.9(10.6) 140.2( 7.8)
    2    3    4   2    90.0  80.0  ;   91.8(10.0)  88.9( 8.4)
    3    4    5   2   130.0  30.0  ;  125.6(18.7) 128.7(16.1)
    4    5    6   2   130.0  30.0  ;  126.7(18.3) 134.0(12.6)
    5    6    7   2   130.0  30.0  ;  127.6(17.7) 134.8(12.1)
    6    7    8   2   130.0  30.0  ;  127.8(18.4) 134.7(11.9)
    7    8    9   2   130.0  30.0  ;  127.4(18.2) 133.9(12.2)
    8    9   10   2   130.0  30.0  ;  126.7(17.8) 134.2(12.2)
    9   10   11   2   130.0  30.0  ;  127.2(17.6) 134.7(11.8)
   10   11   12   2   130.0  30.0  ;  128.2(18.0) 134.7(11.9)

[dihedrals]
;   i    j    k    l  funct angle force.c multipl. CG distr.    UA distr.
    2    3    4    5  1     0     1.0     2     ; -1.6(116.1) -1.8( 81.2)
    2    3    4    5  1     0     2.0     1     ; -1.6(116.1) -1.8( 81.2)
; The next dihedrals were not defined, because defining
; it doesn't improve the behavior or impairs the
; numerical stability.
;  3    4    5    6                            ;  1.0(101.0)  9.2(123.2)
;  4    5    6    7                            ;  2.3(102.5)  7.0(128.9)
;  5    6    7    8                            ; -1.3(108.0)  2.5(125.6)
;  6    7    8    9                            ; -6.4(102.9)  7.5(124.5)
;  7    8    9   10                            ; -0.4(101.8) 11.1(128.2)
;  8    9   10   11                            ; -3.7( 99.8)  4.3(131.1)
;  9   10   11   12                            ; -1.6(102.7)  0.3(126.6)

[exclusions]
; i    j
 1 4
 2 4
```

Heme b

```
[ moleculetype ]
; molname         nrexcl
HEME              1

[ atoms ]
;id type resnr residu atom cgnr charge
 1  SP1  1     HEM    NA    1   -0.10
 2  SC3  1     HEM    CHA   2    0.00
 3  SP1  1     HEM    NB    3   -0.10
 4  SC3  1     HEM    CHB   4    0.00
 5  SP1  1     HEM    NC    5   -0.10
 6  SC3  1     HEM    CHC   6    0.00
 7  SP1  1     HEM    ND    7   -0.10
 8  SC3  1     HEM    CHD   8    0.00
 9  SC3  1     HEM    C2A   9    0.00
10  SC3  1     HEM    C3A  10    0.00
11  SC3  1     HEM    C2B  11    0.00
12  SC3  1     HEM    C3B  12    0.00
13  SC3  1     HEM    C1C  13    0.00
```

```
14  SC3  1     HEM    C2C  14    0.00
15  Qa   1     HEM    C3C  15   -1.00
16  SC3  1     HEM    C1D  16    0.00
17  SC3  1     HEM    C2D  17    0.00
18  Qa   1     HEM    C3D  18   -1.00
19  SQ0  1     HEM    FE   19    0.40

[ bonds ]
; i    j    func length force.c.  CG distr.    UA distr.
  14   15   1    0.30   10000.0 ; 0.301(0.013) 0.287(0.017)
  17   18   1    0.30   10000.0 ; 0.301(0.013) 0.283(0.015)
   1    9   1    0.27   20000.0 ; 0.270(0.007) 0.274(0.003)
   1   10   1    0.33   20000.0 ; 0.325(0.009) 0.326(0.006)
   9   10   1    0.28   20000.0 ; 0.277(0.008) 0.278(0.010)
   3   11   1    0.27   20000.0 ; 0.270(0.008) 0.274(0.003)
   3   12   1    0.33   20000.0 ; 0.325(0.008) 0.326(0.006)
  11   12   1    0.28   20000.0 ; 0.276(0.008) 0.277(0.010)
   5   13   1    0.27   20000.0 ; 0.268(0.007) 0.274(0.003)
   5   14   1    0.27   20000.0 ; 0.270(0.007) 0.272(0.003)
  13   14   1    0.24   20000.0 ; 0.239(0.006) 0.238(0.005)
   7   16   1    0.27   20000.0 ; 0.269(0.008) 0.274(0.003)
   7   17   1    0.27   20000.0 ; 0.270(0.007) 0.272(0.003)
  16   17   1    0.24   20000.0 ; 0.238(0.006) 0.238(0.005)

[ constraints ]
;  i    j   func length CG distr.     UA distr.
   1    2   1    0.219 ; 0.219(0.004) 0.220(0.002)
   2    3   1    0.219 ; 0.219(0.004) 0.220(0.002)
   3    4   1    0.219 ; 0.219(0.004) 0.220(0.002)
   4    5   1    0.219 ; 0.219(0.004) 0.220(0.002)
   5    6   1    0.219 ; 0.219(0.004) 0.220(0.002)
   6    7   1    0.219 ; 0.219(0.004) 0.220(0.002)
   7    8   1    0.219 ; 0.219(0.004) 0.220(0.002)
   8    1   1    0.219 ; 0.219(0.004) 0.220(0.002)
   1    5   1    0.472 ; 0.474(0.004) 0.476(0.003)
   3    7   1    0.472 ; 0.474(0.004) 0.476(0.003)
   2    4   1    0.436 ; 0.436(0.004) 0.437(0.003)
   4    6   1    0.436 ; 0.436(0.004) 0.437(0.003)
   6    8   1    0.436 ; 0.436(0.004) 0.437(0.003)
   8    2   1    0.436 ; 0.436(0.004) 0.437(0.003)
   1   19   1    0.239 ; 0.237(0.004) 0.239(0.001)
   3   19   1    0.239 ; 0.237(0.004) 0.239(0.001)
   5   19   1    0.239 ; 0.237(0.004) 0.239(0.001)
   7   19   1    0.239 ; 0.237(0.004) 0.239(0.001)
; The next bondss were not defined, because defining
; it doesn't improve the behavior or impairs the
; numerical stability.
;  2   10               0.311(0.015) 0.319(0.011)
;  2   11               0.286(0.012) 0.286(0.005)
;  4   12               0.312(0.015) 0.319(0.011)
;  4   13               0.283(0.009) 0.286(0.005)
;  6   14               0.277(0.009) 0.287(0.004)
;  6   17               0.282(0.013) 0.287(0.004)
;  8   16               0.282(0.009) 0.286(0.005)
;  8    9               0.287(0.012) 0.286(0.005)

[ angles ]
;i  j k  funct angle force.c. CG distr.  UA distr.
   2    1    10 2      68    250 ; 66.4(3.1)  68.3( 2.8)
   2    1     9 2     120   1750 ;120.1(2.2) 121.0( 1.5)
   4    3    12 2      68    250 ; 66.5(3.2)  68.3( 2.8)
```

```
   4    3    11 2     120   1750 ;120.1(2.2) 121.0( 1.5)
   1   10     9 2      54   1500 ; 52.7(1.6)  53.3( 1.1)
   3   12    11 2      54   1500 ; 52.7(1.6)  53.3( 1.1)
   5   13    14 2      64   1500 ; 52.7(1.6)  53.3( 1.1)
   7   16    17 2      64   1500 ; 52.7(1.6)  53.3( 1.1)
  16    7    17 2      52   1000 ; 52.4(1.6)  51.8( 1.3)
  13    5    14 2      52   1000 ; 52.4(1.6)  51.8( 1.3)
   6    5    14 2      68    750 ;123.8(2.3) 120.6( 1.5)
   8    7    17 2     120    750 ;123.8(2.3) 120.6( 1.5)
   8    7    16 2      70   1500 ; 69.8(1.9)  69.8( 1.4)
   4    5    13 2      70   1500 ; 69.8(1.9)  69.8( 1.4)
   1    2     3 2      99   1000 ; 99.9(2.3) 100.0( 1.2)
   3    4     5 2      99   1000 ; 99.3(2.2) 100.2( 1.2)
   5    6     7 2      99   1000 ; 99.6(2.2) 100.1( 1.2)
   7    8     1 2      99   1000 ; 99.5(2.2) 100.2( 1.2)
  13   14    15 2     113    100 ; 82.8(5.3)  92.8(11.4)
  16   17    18 2     113    100 ; 82.8(5.3)  92.8(11.4)
   5   14    15 2     147    300 ;123.8(9.0) 125.3(13.6)
   7   17    18 2     147    300 ;123.8(9.0) 125.3(13.6)
; The next angles were not defined, because defining
; it doesn't improve the behavior or impairs the
; numerical stability.
;  1    9    10                    72.8(2.7)  72.6(2.4)
;  9    1    10                    54.5(2.3)  54.2(2.5)
;  7   16    17                    64.1(1.6)  63.7(1.1)
;  7   17    16                    63.6(2.0)  64.6(1.2)
;  2    3     4                   167.2(2.9) 167.8(2.0)
;  4    5     6                   167.3(2.7) 167.5(2.2)
;  6    7     8                   167.1(2.8) 167.5(2.2)
;  7    8     1                    99.4(2.2) 100.2(1.2)
;  8    1     2                   167.1(2.8) 167.8(2.0)

[ dihedrals ]
;i  j  k  l  funct angle force.c.  CG distr.  UA distr.
  19    1    3    5  2     0.0   2000.0 ;   0.0(1.4)   1.2(6.6)
  19    3    5    7  2     0.0   2000.0 ;  -0.0(1.4)  -1.2(6.6)
  19    5    7    1  2     0.0   2000.0 ;   0.0(1.4)   1.2(6.6)
  19    7    1    3  2     0.0   2000.0 ;  -0.0(1.4)  -1.2(6.6)
   1    3    7   10  2     0.0   1000.0 ;   0.0(2.4)  -0.1(6.7)
   1    3    7    9  2     0.0   1000.0 ;   0.0(2.4)  -0.3(4.7)
   3    5    1   11  2     0.0   1000.0 ;   0.0(2.4)   0.5(4.5)
   3    5    1   12  2     0.0   1000.0 ;  -0.1(2.4)  -0.0(6.7)
   5    7    3   13  2     0.0   1000.0 ;   0.1(2.7)  -1.9(4.3)
   5    7    3   14  2     0.0   1000.0 ;   0.1(2.4)   1.3(4.4)
   7    1    5   16  2     0.0   1000.0 ;   0.0(2.8)   1.9(4.2)
   7    1    5   17  2     0.0   1000.0 ;  -0.0(2.4)  -1.4(4.3)

; The next proper dihedrals were not defined, because defining
; it doesn't improve the behavior or impairs the
; numerical stability.
; 15   14    5    6               2.5(137.3) 110.5(17.0)
; 18   17    7    6                    2.4(137.9) 113.4(14.9)
;  8    7   17   18                    1.7( 75.9) -68.4(15.7)
;  4    5   14   15                    1.5( 76.4) -71.6(17.3)


[ exclusions ]
1 2 3 4 5 6 7 8
2 3 4 5 6 7 8 9 10 11 12
3 4 5 6 7 8
4 5 6 7 8 11 12 13 14
```

```
5 6 7 8
6 7 8 13 14 16 17
7 8
8 9 10 16 17
9 11 12 16 17
10 11 12 16 17
11 13 14
12 13 14
13 16 17
14 16 17
19 2 4 6 8

; To be inserted in the atomistic Gromos topology
[ mapping ]
1   3 20 23      ;20 and 23 are shared: for proper mapping halve the mass!
2   23 27 29     ;23 and 29 are shared: for proper mapping halve the mass!
3   4 29 32      ;29 and 32 are shared: for proper mapping halve the mass!
4   32 36 38     ;32 and 28 are shared: for proper mapping halve the mass!
5   5 38 41      ;38 and 41 are shared: for proper mapping halve the mass!
6   6 8 41       ; 8 and 41 are shared: for proper mapping halve the mass!
7   2 8 11       ; 8 and 11 are shared: for proper mapping halve the mass!
8   11 18 20     ;11 and 20 are shared: for proper mapping halve the mass!
9   21 24
10  22 25 26
11  30 33
12  31 34 35
13  39 42
14  40 43
15  44 45 46 47
16  10 12
17  9 13
18  14 15 16 17
19  1            ;the iron
```

Chlorophyll A

```
[ moleculetype ]
; molname         nrexcl
CLA               1

[ atoms ]
;id  type resnr residu atom cgnr charge
  1  SP3  1     CLA    NA    1   -0.25
  2  SC3  1     CLA    CHA   2    0.00
  3  SP3  1     CLA    NB    3   -0.25
  4  SC3  1     CLA    CHB   4    0.00
  5  SP3  1     CLA    NC    5   -0.25
  6  SC3  1     CLA    CHC   6    0.00
  7  SP3  1     CLA    ND    7   -0.25
  8  SC3  1     CLA    CHD   8    0.00
  9  SC3  1     CLA    C1A   9    0.00
 10  SC3  1     CLA    C2A  10    0.00
 11  SC3  1     CLA    C1B  11    0.00
 12  SC3  1     CLA    C2B  12    0.00
 13  SC4  1     CLA    C1C  13    0.00
 14  SNa  1     CLA    C2C  14    0.00
 15  Na   1     CLA    T1   15    0.00
 16  SC3  1     CLA    C1D  16    0.00
```

```
 17  SC3  1     CLA    C2D  17    0.00
 18  Na   1     CLA    T2   18    0.00
 19  C3   1     CLA    T3   19    0.00
 20  C1   1     CLA    T4   20    0.00
 21  C1   1     CLA    T5   21    0.00
 22  C1   1     CLA    T6   22    0.00
 23  SQ0  1     CLA    FE   23    1.00

[ bonds ]
;  i    j  func length force.c.  CG distr.     UA distr.
   1    9   1    0.27   20000.0 ; 0.267(0.007)  0.264(0.003)
   9   10   1    0.28   20000.0 ; 0.307(0.008)  0.311(0.007)
   1   10   1    0.31   20000.0 ; 0.277(0.008)  0.271(0.010)
   3   11   1    0.27   20000.0 ; 0.267(0.007)  0.264(0.003)
   3   12   1    0.31   20000.0 ; 0.308(0.009)  0.312(0.006)
  11   12   1    0.28   20000.0 ; 0.276(0.008)  0.270(0.009)
   5   13   1    0.23   20000.0 ; 0.231(0.008)  0.225(0.002)
   5   14   1    0.34   20000.0 ; 0.334(0.006)  0.335(0.004)
   6   14   1    0.26   20000.0 ; 0.282(0.007)  0.277(0.006)
  13   14   1    0.28   20000.0 ; 0.265(0.008)  0.262(0.003)
  14   15   1    0.27   10000.0 ; 0.245(0.008)  0.230(0.006)
   7   17   1    0.23   20000.0 ; 0.229(0.009)  0.237(0.006)
   7   16   1    0.23   20000.0 ; 0.256(0.008)  0.258(0.005)
  16   17   1    0.26   20000.0 ; 0.278(0.014)  0.276(0.017)
  17   18   1    0.33    5000.0 ; 0.343(0.019)  0.326(0.022)
  18   19   1    0.42    8000.0 ; 0.411(0.017)  0.421(0.024)
  19   20   1    0.45    8000.0 ; 0.446(0.017)  0.446(0.017)
  20   21   1    0.49    4000.0 ; 0.486(0.025)  0.480(0.023)
  21   22   1    0.49    4000.0 ; 0.487(0.025)  0.481(0.023)

[ constraints ]
;  i    j  func length CG distr.     UA distr.
   1    2   1   0.219 ; 0.219(0.004) 0.216(0.002)
   2    3   1   0.219 ; 0.220(0.004) 0.216(0.002)
   3    4   1   0.219 ; 0.219(0.004) 0.216(0.002)
   4    5   1   0.219 ; 0.219(0.004) 0.213(0.002)
   5    6   1   0.219 ; 0.219(0.004) 0.212(0.003)
   6    7   1   0.219 ; 0.220(0.004) 0.219(0.002)
   7    8   1   0.219 ; 0.220(0.004) 0.227(0.002)
   8    1   1   0.219 ; 0.219(0.004) 0.220(0.002)
   1    5   1   0.472 ; 0.472(0.004) 0.472(0.003)
   3    7   1   0.472 ; 0.471(0.004) 0.470(0.003)
   2    4   1   0.436 ; 0.436(0.004) 0.430(0.003)
   4    6   1   0.436 ; 0.436(0.004) 0.422(0.004)
   6    8   1   0.436 ; 0.436(0.004) 0.443(0.003)
   8    2   1   0.436 ; 0.436(0.004) 0.432(0.003)
   1   23   1   0.239 ; 0.239(0.004) 0.237(0.001)
   3   23   1   0.239 ; 0.239(0.004) 0.237(0.001)
   5   23   1   0.239 ; 0.239(0.004) 0.238(0.001)
   7   23   1   0.239 ; 0.239(0.004) 0.236(0.001)

[ angles ]
;  i    j    k  funct angle force.c. CG distr.  UA distr.
   2    1    9   2     120   2000 ;  120.4( 2.0) 121.1( 1.5)
   2    1   10   2      68    250 ;   64.1( 3.2)  67.7( 2.9)
   1   10    9   2      54   1500 ;   54.0( 1.6)  53.5( 1.4)
   4    3   11   2     120   2000 ;  120.4( 2.1) 120.9( 1.4)
   4    3   12   2      68    250 ;   64.4( 3.2)  67.6( 2.7)
   3   12   11   2      54   1500 ;   53.9( 1.5)  53.3( 1.3)
   1    2    3   2      99   1000 ;   99.0( 2.1) 101.2( 1.2)
   3    4    5   2      99   1000 ;   98.9( 2.1)  99.8( 1.2)
```

```
   5    6    7   2      99   1000 ;   99.0( 2.2) 100.6( 1.3)
   7    8    1   2      99   1000 ;   99.0( 2.1) 100.2( 1.2)
   4    5   13   2      81   1000 ;   81.3( 2.3)  81.4( 1.3)
   5   13   14   2      80   1500 ;   80.7( 1.5)  83.2( 1.5)
  16   17   18   2     118    500 ;  120.7( 3.4) 118.6(14.9)
   6   14   15   2      77    500 ;   80.2( 3.3)  77.2( 4.5)
   5   14   15   2     106    300 ;  108.1( 4.1) 106.4( 7.8)
  17   18   19   2     130    100 ;  116.3( 6.9) 131.3(19.4)
  18   19   20   2     130     25 ;  132.2(20.0) 128.0(22.8)
  19   20   21   2     130     25 ;  129.1(16.8) 131.1(18.6)
  20   21   22   2     130     25 ;  127.4(17.8) 129.9(19.0)
;The next angles were not defined, because defining
;it doesn't improve the behavior or impairs the
;numerical stability.
; 13    5   14                       56.4(1.8)  55.0( 1.4)
;  7   17   16                       54.5(2.6)  57.9( 1.8)
; 17    7   16                       65.3(3.1)  67.0( 2.5)
;  6    5   14                       52.3(2.6)  51.4( 1.0)
;  8    7   17                      109.1(5.7) 114.2( 3.4)
;  8    7   16                       65.4(7.7)  77.2( 2.6)
;  7   17   18                      109.4(6.2) 151.5(11.2)

[ dihedrals ]
;  i    j    k    l  funct angle force.c. CG distr.    UA distr.
   1    3    7   10   2      0.0  1000.0 ; -0.0(2.3)     0.2(6.9)
   1    3    7    9   2      0.0  1000.0 ;  0.1(2.2)    -0.4(3.3)
   3    5    1   11   2      0.0  1000.0 ; -0.1(2.3)    -0.7(3.5)
   3    5    1   12   2      0.0  1000.0 ; -0.0(2.3)     2.3(7.0)
   7    8   16   17   2     45.0  1000.0 ; 44.0(1.9)    42.7(4.2)
;The next dihedrals were not defined, because defining
;it doesn't improve the behavior or impairs the
;numerical stability.
;  23    1    3    5                      12.3(2.0)    -1.6(7.2)
;  23    3    5    7                      12.7(1.9)    -0.5(7.8)
;  23    5    7    1                      12.3(1.9)    -1.6(7.0)
;  23    7    1    3                      12.8(1.9)    -0.6(8.0)
;   5    7    3   13                      -7.4(9.8)     0.9(4.9)
;   6    7    3   14                      -8.1(8.3)     2.3(5.8)
;   7    1    5   16                      -0.3(4.4)    11.9(2.9)
;   7    1    5   17                     -23.7(3.6)   -15.2(2.8)
;  17   18   19   20                    45.6( 85.2) -5.3( 94.5)
;  18   19   20   21                   -21.3(100.9) 14.9(105.7)
;  19   20   21   22                   -19.4(107.3) -3.4( 90.0)

[ exclusions ]
1 2 3 4 5 6 7 8
2 3 4 5 6 7 8 9 10 11 12
3 4 5 6 7 8
4 5 6 7 8 11 12 13 14
5 6 7 8 14
6 7 8 13 14 15 16 17
7 8
8 9 10 16 17
9 11 12 16 17
10 11 12 16 17
11 13 14
12 13 14
13 15 16 17
14 16 17
15 17 18
16 18
```

```
23 2 4 6 8
```

β-carotene

```
[ moleculetype ]
; Name nrexcl
BCR      1

[atoms]
; id type resnr  residu  atom  cgnr  charge
  1  SC3  1      BCR     R1    1     0
  2  SC3  1      BCR     R2    2     0
  3  C4   1      BCR     R3    3     0
  4  C4   1      BCR     T1    4     0
  5  C4   1      BCR     T2    5     0
  6  C4   1      BCR     T3    6     0
  7  C4   1      BCR     T4    7     0
  8  C4   1      BCR     R4    8     0
  9  SC3  1      BCR     R5    9     0
 10  SC3  1      BCR     R6    10    0

[bonds]
;  i    j  funct   length  force.c.  CG distr.     UA distr.
   3    4   1       0.425   10000   ; 0.423(0.017) 0.424(0.009)
   4    5   1       0.501   1250    ; 0.484(0.048) 0.502(0.006)
   5    6   1       0.527   1250    ; 0.516(0.048) 0.529(0.006)
   6    7   1       0.501   1250    ; 0.486(0.046) 0.492(0.007)
   7    8   1       0.425   10000   ; 0.422(0.016) 0.425(0.009)

[constraints]
;  i    j  funct  length   CG distr.       UA distr.
   1    2   1      0.272  ; 0.272(0.004) 0.272(0.004)
   1    3   1      0.292  ; 0.292(0.004) 0.296(0.005)
   2    3   1      0.278  ; 0.278(0.004) 0.279(0.004)
   8    9   1      0.278  ; 0.278(0.004) 0.277(0.004)
   8   10   1      0.292  ; 0.292(0.004) 0.289(0.006)
   9   10   1      0.272  ; 0.272(0.004) 0.272(0.004)

[angles]
;  i   j   k   funct   angle force.c.   CG distr.     UA distr.
   1   3   4   2       97    100     ;   97.5( 9.7)    94.8(5.2)
   2   3   4   2       139   100     ;  141.1(11.4)   139.8(7.6)
   3   4   5   2       148   800     ;  148.0( 6.5)   153.4(6.9)
   4   5   6   2       155   700     ;  155.2( 7.7)   158.9(5.6)
   5   6   7   2       155   700     ;  155.5( 7.9)   151.7(5.4)
   6   7   8   2       148   800     ;  148.4( 6.3)   156.9(6.8)
   7   8   9   2       139   100     ;  141.1(10.9)   141.1(5.6)
   7   8  10   2       97    100     ;   97.6( 9.4)    98.5(4.6)

[dihedrals]
; normal proper dihedrals make the molecule instable due to the
; angles being close to 180deg. They have been switched of while
;waiting for a proper version of gromacs with the special dihedrals
;The distributions have been determined with the dihedrals turned off.
;   i    j    k    l  funct force.c. c0       c1       c2 c3    CG distr.   UA distr.
;   3    4    5    6   11    3000.0 1        -2       1  0 ;   0.7(103.2)  -30.9( 40.4)
;   5    6    7    8   11    3000.0 1        -2       1  0 ;   2.2(104.3) -109.6( 95.3)
;   4    5    6    7   11   13000.0 1         2       1  0 ;   2.2(105.1)   -8.8(157.3)
;   1    3    4    5   11     170.0 0.48255   1.3893  1  0 ;   0.9(101.9)  125.8( 31.0)
```

```
;   2    3    4    5   11     380.0 0.01093   0.2090  1  0 ;  -1.8(102.5)   93.4( 34.2)
;  10    8    7    6   11     170.0 0.48255   1.3893  1  0 ;  -1.9(102.7) -124.3( 71.7)
;   9    8    7    6   11     380.0 0.01093   0.2090  1  0 ;  -4.1(102.6)  -96.4( 27.1)

[improper]
;The next improper dihedrals were not defined, because defining
;it doesn't improve the behavior or impairs the
;numerical stability.
;  i    j    k    l   CG distr.    UA distr.
;   3    4    1    2 ;  0.2( 27.7) 21.3(15.4)
;   8    7    9   10 ;  0.2( 27.7) 21.3(15.4)

[exclusions]
 1  4
 2  4
 9  7
 10 7
```

6.7 Bibliography

[1] J.M Andersson. Photoregulation of the composition, function, and structure of thylakoid membranes. *Ann. Rev. Plant Phys.*, 37:93–136, 1986.

[2] N Nelson and C.F Yocum. Structure and function of photosystems i and ii. *Plant Biol.*, 57:521–565, 2006.

[3] A Guskov, J Kern, A Gabdulkhakov, M Broser, A Zouni, and W Saenger. Cyanobacterial photosystem ii at 2.9-å resolution and the role of quinones, lipids, channels and chloride. *Nat. Struct. Mol. Biol.*, 16:334–342, 2009.

[4] Y Umena, K Kawakami, J.R Shen, and N Kamiya. Crystal structure of oxygen-evolving photosystem ii at a resolution of 1.9 å. *Nature*, 473:55–60, 2011.

[5] S Vassiliev and D Bruce. Toward understanding molecular mechanisms of light harvesting and charge separation in photosystem ii. *Photosynth. Res.*, 97:75–89, 2008.

[6] S Vassiliev, P Comte, A Mahboob, and D Bruce. Tracking the flow of water through photosystem ii using molecular dynamics and streamline tracing. *Biochemistry*, 49:1873–1881, 2010.

[7] S Vassiliev and D Bruce. A protein dynamics study of photosystem ii: The effects of protein conformation on reaction center function. *Biophys. J.*, 90:3062–3073, 2006.

[8] P Palencar, T Prudnikova, F Vacha, and M Kuty. The effects of light-induced reduction of the photosystem ii reaction center. *J. Mol. Model*, 15:923933, 2009.

[9] S.J Marrink, H.J Risselada, S Yefimov, D.P Tieleman, and A.H de Vries. The martini force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 111:7812–7824, 2007.

[10] L Monticelli, S.K Kandasamy, X Periole, R.G Larson, D.P Tieleman, and S.J Marrink. The martini coarse-grained force field: extension to proteins. *J. Chem. Theory Comput.*, 4:819–834, 2008.

[11] Z Sovoviá A.H de Vries S.J Marrink C.A Lopez, F van Eerden. Lipids of the thylakoid membrane. *in preparation*, 2012.

[12] K Karki and D Roccatano. Molecular dynamics simulation study of chlorophyll a in different organic solvents. *J. Chem. Theory Comput.*, 7:1131–1140, 2011.

[13] L Eriksson and A Laaksonen. Molecular dynamics simulations of plastoquinone in solution. *Mol. Phys.*, 99:247–253, 2001.

[14] F Autenrieth, E Tajkhorshid, J Baudry, and Z Luthey-Schulten. Classical force field parameters for the heme prosthetic group of cytochrome c. *J. Comput. Chem.*, 25:1613–1622, 2004.

[15] P Palencar, F Vacha, and M Kuty. Force field development on pigments of photosystem 2 reaction centre. *Photosynthetica*, 43:417–420, 2005.

[16] M Ceccarelli, P Procacci, and M Marchi. An ab initio force field for the cofactors of bacterial photosynthesis. *J. Comput. Chem.*, 24:129–142, 2003.

[17] L Zhang, D.A Silva, Y Yan, and X Huang. Force field development for cofactors in the photosystem ii. *J. Comp. Chem.*, 33:1969–1980, 2012.

[18] W.F Van Gunsteren, S Billeter, A.A Eising, P.H Hünenberger, P Krüger, A.E Mark, W.R.P Scott, and I.G Tironi. *Biomolecular Simulation: The GROMOS96 manual and user guide*. BIOMOS, 1996.

[19] A.J Rzepiela, M Louhivuori, C Peter, and S.J Marrink. Hybrid simulations: combining atomistic and coarse-grained force fields using virtual sites. *Phys. Chem. Chem. Phys.*, 13:10437–10448, 2011.

[20] B Hess, C Kutzner, D van der Spoel, and E Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4:435–447, 2008.

[21] I.G Tironi, R Sperb, P.E Smith, and W.F Van Gunsteren. A generalized reaction field method for molecular dynamics simulations. *J. Chem. Phys.*, 102:5451–5459, 1995.

[22] H.J.C Berendsen, J.P.M Postma, W.F Van Gunsteren, A DiNola, and J.R Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684, 1984.

[23] B Hess, H Bekker, H.J.C Berendsen, and J.G.E.M Fraaije. Lincs: a linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18:1463–1472, 1997.

[24] B Hess. P-lincs: A parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.*, 4:116–122, 2008.

[25] A.J Rzepiela, L.V Schäfer, N Goga, H.J Risselada, A.H de Vries, and S.J Marrink. Reconstruction of atomistic details from coarsegrained structures. *J. Comput. Chem.*, 31:1333–1343, 2010.

[26] A.W Schuttelkopf and D.M.F Aalten. Prodrg: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Cryst.*, 60:1355–1363, 2004.

[27] A.K Malde, L Zuo, M Breeze, M Stroet, D Poger, P.C Nair, C Oostenbrink, and A.E Mark. An automated force field topology builder (atb) and repository: version 1.0. *J. Chem. Theory Comput.*, 7:4026–4037, 2011.

[28] C Oostenbrink, A Villa, A.E Mark, and W.F Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos forcefield parameter sets 53a5 and 53a6. *J. Comput. Chem.*, 25:1656–1676, 2004.

[29] M Schmidt, K Baldridge, J.A Boatz, S.T Elbert, M.S Gordon, J.H Jensen, S Koseki, N Matsunaga, K.A Nguyen, S Su, T.L Windus, M Dupuis, and J.A Montgomery. General atomic and molecular electronic structure system. *J. Comput. Chem.*, 14:1347–1363, 1993.

[30] M Swart, P.T van Duijnen, and J.G Snijders. A charge analysis derived from an atomic multipole expansion. *J. Comput. Chem.*, 22:79–88, 2001.

[31] F Martyn, I Bush, H.J.J van Dam, P Sherwood, J.M.H Thomas, J.H Van Lenthe, R.W.A Havenith, and J Kendrick. The gamess-uk electronic structure package: algorithms, developments and applications. *Mol. Phys.*, 103:719–747, 2005.

[32] S.J Marrink, M Fuhrmans, H.J Risselada, and X Periole. *Coarse graining of condensed phase and biomolecular systems*, chapter The MARTINI force field. CRC Press, 2009.

[33] S.O Yesylevskyy, L.V Schäfer, D Sengupta, and S.J Marrink. Polarizable water model for the coarse-grained martini force field. *PLoS Comput. Biol*, 6:e1000810, 2010.

[34] N.M Garrido, A.J Queimada, M Jorge, E.A Macedo, and I.G Economou. 1-octanol/water partition coefficients of n-alkanes from molecular simulations of absolute solvation free energies. *J. Chem. Theory Comput.*, 5:2436–2446, 2009.

[35] W.F.D Bennet C Arnarez T.A Wassenaar L.V Schaäfer X. Periole D.P Tieleman S.J Marrink D.H de Jong, G Singh. Improved parameters for the martini coarse-grained protein force field. *submitted*, 2012.

[36] X Periole, M Cavalli, and S.J Marrink. Combining an elastic network with a coarse-grained molecular force field: Structure, dynamics, and intermolecular recognition. *J. Chem. Phys.*, 5:2531–2543, 2009.

[37] I Sakurai, J.R Shen, J Leng, S Ohashi, M Kobayashi, and H Wada. Lipids in oxygen-evolving photosystem ii complexes of cyanobacteria and higher plants. *J. Biochem.*, 140:201, 2006.

[38] O Berger, O Edholm, and F Jähnig. Molecular dynamics simulations of a fluid bilayer of dipalmitoylphosphatidylcholine at full hydration, constant pressure, and constant temperature. *Biophys. J.*, 72:2002–2013, 1997.

[39] X Daura, K Gademann, B Jaun, D Seebach, W.F van Gunsteren, and A.E Mark. Peptide folding: when simulation meets experiment. *Angew. Chem. Int. Edit.*, 38:236–240, 1999.

[40] F Himo, G.T Babcock, and L.A Eriksson. Conformational analysis of quinone anion radicals in photosystem ii and photosynthetic bacteria. *J. Phys. Chem.*, 103:3745–3749, 1999.

[41] J de Vogel, D Jonker-Termont, M.B Katan, and R van der Meer. Natural chlorophyll but not chlorophyllin prevents heme-induced cytotoxic and hyperproliferative effects in rat colon. *J. Agric. Food Chem.*, 135:1995–2000, 2005.

[42] D Cooper, D Webb, and J.C Peters. Evaluation of the potential for olestra to affect the availability of dietary phytochemicals. *J. Nutr.*, 127:1699S1709S, 1997.

[43] K Liao and M Yin. Individual and combined antioxidant effects of seven phenolic agents in human erythrocyte membrane ghosts and phosphatidylcholine liposome systems: importance of the partitioning coefficient. *J. Agric. Food Chem*, 48:2266–2270, 2000.

[44] P.R Rich and R Harper. Partition coefficients of quinones and hydroquinones and their relation to biochemical reactivity. *FEBS Lett.*, 269:139–144, 1990.

[45] R Wang, Y Gao, and L Lai. Calculating partition coefficient by atom-additive method. *Perspectives in Drug Discovery and Design*, 19:47–66, 2000.

[46] R Wang, Y Fu, and L Lai. A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Comp. Sci.*, 37:615–621, 1997.

[47] I.V Tetko and P Bruneau. Application of alogps to predict 1-octanol/water distribution coefficients, logp, and logd, of astrazeneca in-house database. *J. Pharm. Sci.*, 93:3103–3110, 2004.

[48] I.V Tetko and V.Y Tanchuk. Application of associative neural networks for prediction of lipophilicity in alogps 2.1 program. *J. Chem. Inf. Comput. Sci.*, 42:1136–1145, 2002.

[49] M.I Bulacu. *Molecular dynamics studies of entangled polymer chains*. PhD thesis, University of Groningen, 2008.

[50] R Weast. *CRC handbook of chemistry and physics*. CRC Press, 1986.

[51] S.J Marrink, A.H de Vries, and A.E Mark. Coarse grained model for semiquantitative lipid simulations. *J. Phys. Chem. B*, 108:750–760, 2004.

[52] B.R Green and D.G Durnford. The chlorophyll-carotenoid proteins of oxygenic photosynthesis. *Annu. Rev. Plant Biol.*, 47:685–714, 1996.

[53] M Jemioła-Rzemińska, M Pasenkiewicz-Gierula, and K Strzałka. The behaviour of -carotene in the phosphatidylcholine bilayer as revealed by a molecular simulation study. *Chem. Phys. Lipids*, 135:27–37, 2005.

7

Improved Parameters For The Martini Coarse-Grained Protein Force Field

This chapter is based upon the manuscript:
Improved Parameters For The Martini Coarse-Grained Protein Force Field by Djurre H. de Jong, Gurpreet Singh, W.F. Drew Bennett, Clement Arnarez, Tsjerk A. Wassenaar, Lars V. Schäfer, Xavier Periole, D. Peter Tieleman, and Siewert J. Marrink, J. Chem. Theory Comput., **2012**, **DOI:**10.1021/ct300646g

Abstract

The Martini coarse-grained force field has been successfully used for simulating a wide range of (bio)molecular systems. Recent progress in our ability to test the model against fully atomistic force fields, however, has revealed some shortcomings. Most notable, phenylalanine and proline were too hydrophobic, and dimers formed by polar residues in apolar solvents did not bind strongly enough. In this chapter we reparametrize these residues either through reassignment of particle types or by introducing embedded charges. The new parameters are tested with respect to partitioning across a lipid bilayer, membrane binding of Wimley-White peptides, and dimerization free energy in solvents of different polarity. In addition, we improve some of the bonded terms in the Martini protein force field that lead to a more realistic length of α-helices and to improved numerical stability for poly-alanine and glycine repeats. The new parameter set is denoted Martini version 2.2.

7.1 Introduction

The use of coarse-grained (CG) models in a variety of simulation techniques has proven to be a valuable tool to probe the spatial and temporal evolution of systems on the mesoscale, beyond what is feasible with traditional all-atom (AA) models. A large diversity of coarse-graining approaches is available; they range from qualitative, often solvent-free models, via more realistic explicit solvation models, to models including chemical specificity (for recent overview see Refs. [1, 2, 3, 4, 5]). Models within this latter category are typically parameterized based on comparison to atomistic simulations, using iterative Boltzmann[6, 7], force matching[8, 9], conditional reversible work[10], or minimization of relative entropy approaches[11]. Our own model, coined the Martini force field[12, 13, 14], has also been developed in close connection with atomistic models, especially considering the bonded interactions. However, the philosophy of our coarse-graining approach is different. Instead of focusing on an accurate

reproduction of structural details at a particular state point for a specific system, we aim for a broader range of applications without the need to reparameterize the model each time. We do so by extensive calibration of the non-bonded interactions of the chemical building blocks against experimental data, in particular thermodynamic data such as oil/water partitioning coefficients, since processes such as lipid self-assembly, peptide-membrane binding, and protein-protein recognition depend critically on the degree to which the constituents partition between polar and non-polar environments. The use of a consistent strategy for the development of compatible CG and atomic-level force fields is of additional importance for its intended use in multi-scale applications[15, 16, 17, 18, 19]. The overall aim of our coarse-graining approach is to provide a simple model that is computationally fast and easy to use, yet flexible enough to be applicable to a wide range of (bio)molecular systems.

The name Martini of the force field was coined in 2007 with the release of version 2.0[13] for lipids. The subsequent extension to peptides and proteins[14] was released as version 2.1. The Martini model is based on an approximate four-to-one mapping, i.e., on average four heavy atoms plus associated hydrogens are represented by a single interaction center, with an exception for ring-like molecules. The mapping of ring-like fragments or small molecules (e.g., benzene, cholesterol, and several of the amino acid side chains) is not possible with the general four-to-one mapping approach. Such molecules are therefore mapped with a higher resolution of up to two-to-one. The Martini model considers four main types of interaction sites: polar (P), non-polar (N), apolar (C), and charged (Q). Within a main type, subtypes are distinguished either by a letter denoting the hydrogen-bonding capabilities (d = donor, a = acceptor, da = both, 0 = none) or by a number indicating the degree of polarity (from 1 = low polarity to 5 = high polarity). Small (ring-type) particles are denoted with a prefix S. To improve the treatment of electrostatic interactions, polarizable water models were also recently introduced for use with Martini[20, 21].

Recently, progress in computational power has allowed for more extensive testing of the Martini force field with respect to all-atom models. Singh and Tieleman[22] compared relative binding free energies of the Wimley-White (WW) pentapeptides to an experimentally derived free energy scale. Overall the Martini model predicts the relative binding of these peptides to a lipid membrane in close agreement with the experimental data[23], with notable exceptions for the charged residues, as well as phenylalanine and proline. The first category can be improved by resorting to the polarizable water model, but both phenylalanine and proline are apparently too hydrophobic. In chapter 3 we studied the dimerization free energy of amino acid side chains in solvents of different polarity. Here, also the overall performance of Martini is quite good in comparison to all-atom force fields such as Gromos and OPLS. Again, exceptions are the aromatic side chains, which are too hydrophobic. In addition, charged and polar interactions in a low dielectric medium are too weak compared to the atomistic models. Furthermore, ongoing refinement of partitioning free energy profiles across a water/bilayer interface by MacCallum et al.[24] reveal a significantly underestimated interfacial binding of the polar side chains asparagine and glutamine in Martini.

In the current chapter we aim to fix some of these shortcomings, largely pertaining to the protein force field. These include *i)* new topologies for proline, phenylalanine, and tryptophan side chains to improve partitioning free energies, *ii)* introduction of an off-centre charge model for a more realistic description of contact pairs of oppositely charged residues, *iii)* parameterization of polarized beads for polar side chains to improve dimerization in apolar environments and interfacial binding, and *iv)* some adjustment of bonded terms to improve the length of standard α-helices and increase numerical stability for poly-alanine and glycine repeats. The new version of the force field will be denoted Martini 2.2 (or 2.2P in combination with the polarizable water model).

The rest of this chapter is organized as follows. In section 7.2 the methods are outlined, providing details about the simulation set-ups used for refining the parameters. In

section 7.3 the results are presented, subdivided into sections dealing with apolar residues (Phe, Trp, Pro), then charged (Glu, Asp, Arg, Lys, His) and polar ones (Thr, Ser, Asn, Gln, His), followed by a section on the bonded terms, and a section on the new script used for implementation. A short conclusive paragraph (section 7.4) ends this chapter..

7.2 Methods

General

The molecular dynamics (MD) simulations described in this chapter were performed using the GROMACS software package[25], version 4.x. The scheme developed for the Martini model[12, 13, 14] was used: non-bonded interactions are cut-off at a distance of 1.2 nm with smooth switching of the interactions and forces from 0.0 to 1.2 nm for the Coulomb potential and 0.9 to 1.2 nm for the Lennard-Jones (LJ) potential. The pair-list update frequency was set to once per 10 steps. A time step of 20-30 fs was used in most cases, which is adequate for preserving energy and temperature in Martini simulations[26]. Note, the simulation times reported in the chapter are plain simulation times. Constant temperature and pressure was maintained using weak coupling to a bath[27]. Standard input files as well as the newly developed parameters can be downloaded from `http://cgmartini.nl`.

Partitioning free energy

Potentials of mean force (PMFs) for amino acid side chain analogues (SCAs) across a DOPC (dioleoyl-phosphatidylcholine) bilayer were calculated using umbrella sampling as described in MacCallum et al.[24]. Small bilayer patches were simulated with 72 DOPCs, 1200 water beads, and one side chain. Polarizable water was used in cases of polar/charged side chains. For each PMF, we ran 81 independent simulations with a harmonic restraint on the distance between the side chain and the center of the DOPC bilayer in the z dimension. The distance was varied from -4 nm to 4 nm with an 0.1 nm spacing and a 1000 kJ mol^{-1} nm^{-2} force constant. Each simulation was run for 100 ns. The weighted histogram analysis method (WHAM)[28] was used to calculate the free energy profiles.

Partitioning free energies of SCAs between water and oil (decane) were calculated using thermodynamic integration. Separate boxes with a single SCA solvated by 334 CG water beads or 122 decane molecules were set up. In both solvents 2 ns simulations were run at each lambda point, using 11 equally spaced points. The electrostatic and van der Waals interactions were switched off separately when charged molecules were present in the system. The free energies and the corresponding errors were calculated using the Bennett Acceptance Ratio as implemented in the g_bar analysis tool of GROMACS. The partitioning free energy, ΔG_{part}, was obtained by subtracting the free energy in oil from the free energy in water.

Dimerization free energy

The dimerization free energy, ΔG_{dim}, of pairs of amino acid SCAs was computed as described in chapter 3. We prepared systems consisting of two amino acid SCAs solvated in a cubic unit-cell with an edge length 3.0 nm, filled with either water or decane. For each pair considered, we determined the PMF as a function of the side chains center of mass (COM) distance. Simulations were run with the COM distances between SCAs constrained in the range 0.3-1.5 nm with a 0.025 nm interval. At each distance the system was simulated for 2 ns from which the first 50 ps were discarded as equilibration period. Over the remaining simulation time the mean constraining force was calculated using the constraint pulling code

implemented in GROMACS and integrated as described by Hess[29]. From these PMFs, the dimerization free energy, ΔG_{dim}, was obtained using

$$\Delta G_{\mathrm{dim}} = -k_B T \ln \left(\frac{4\pi R_{\max}^3 \int_0^{r_c} r^2 g(r) dr}{3v^{theta} \int_{r_c}^{R_{\max}} r^2 g(r) dr} \right) \tag{7.1}$$

where k_b is the Boltzmann constant, T is the simulation temperature, r is the SCAs COM distance, r_c is the dimer-monomer cut-off (defined by the distance at which the PMF reaches its first maximum), $Ri_{\max}$ is the maximum distance considered, v^{θ} is the standard volume (1.66 nm^3, equivalent to 1 mol L^{-1}), and $g(r)$ is the radial distribution function which is calculated from the PMF using:

$$g(r) = e^{-\frac{PMF(r)}{k_B T}} \tag{7.2}$$

Statistical errors are calculated using a Monte Carlo procedure: 10000 PMF profiles are generated using mean and standard deviation of the constraining force for every distance, r. For all profiles ΔG_{dim} is calculated using equation 7.1. Over all values of ΔG_{dim} the mean and standard error are calculated.

Binding of WW peptides

Wimley-White (WW) peptides are pentapeptides with the sequence Ac-WLXLL, where X denotes a variable residue. The binding of these peptides to a POPC (palmitoyl-oleoyl-PC)/water interface was studied experimentally[23]. The relative binding free energies $\Delta\Delta G_{WW}$ of these peptides[23] with respect to X=Ala provides an energy scale that is useful as yardstick to gauge the relative surface affinity of different residues. Here we calculated $\Delta\Delta G^{WW}$ to test our new parameters using a protocol recently established by Singh and Tieleman[22]. In short, for amino acids Phe, Trp, and Pro the $\Delta\Delta G_{WW}$ are obtained using a combination of free energy perturbation (FEP) and multiple Bennett acceptance ratio (MBAR) methods, following the thermodynamic cycle as described in detail elsewhere[22]. For polar residues, the PMF profiles for translocating the entire peptide from the interface to the bulk water were computed using the distance between the COM of the POPC bilayer and the COM of the peptide as the reaction coordinate. The COM of the peptide was held at its relative position by applying a harmonic potential with a force constant of 1000 kJ mol^{-1} nm^2. After an initial equilibration, data were collected over 600-1200 ns time periods depending on the convergence of the PMFs. In some cases, multiple simulations were performed to collect better statistics. The data was divided into windows of 100 ns to compute the free energies and standard errors, using the weighted histogram analysis method (WHAM) as implemented in GROMACS. The free energy of adsorption of residue X at the interface from the bulk was calculated as

$$\Delta G^{WW} = -k_B T \ln \left(\int_{Z_f}^{Z^s} e^{-\frac{\Delta G(z)}{k_B T}} dz \right) \tag{7.3}$$

where z_s=1.0 nm and z_f=4.0 nm define the POPC/water interface and bulk water, respectively. The integrations were carried out numerically using Simpson's rule. The relative binding free energy of residue X, $\Delta\Delta G_{\mathrm{WW}}(\mathrm{X})$, is obtained by subtracting $\Delta G_{\mathrm{WW}}(\mathrm{Ala})$ from $\Delta G_{\mathrm{WW}}(\mathrm{X})$.

7.3 Results

Change of amino acid particle type for Phe, Trp, and Pro

Based on dimerization data of amino acid side chain analogues136 and interfacial binding of Wimley-White (WW) peptides[23], it has become clear that especially Phe and Pro, and to a lesser extent Trp, are too hydrophobic in the Martini 2.1 force field.

For Phe, the behavior can be greatly improved by changing the particle type from SC4 to the slightly more polar SC5. The dimerization free energy in aqueous solution ($\Delta G_{\mathrm{dim,\ water}}$) increases from -4.5 kJ mol^{-1} to -3.0 kJ mol^{-1} which is in better agreement with the value of -1.6 kJ mol^{-1} predicted using an atomistic model (table 7.3). The dimerization free energy in less polar solvents ($\Delta G_{\mathrm{dim,\ oil}}$) is hardly affected. The relative binding free energy of the WW peptide ($\Delta\Delta G_{\mathrm{WW}}$) also improves and gets within k_bT of the experimental value (see Table 7.3). The partitioning of the Phe side chain analogue across the water/bilayer interface was also computed. The new topology shows an improved overall binding profile. For consistency, we also investigated the use of the SC5 particle type to model benzene (note, in Martini benzene and Phe are represented the same way). The original model[13] underestimates the partitioning of benzene in water; the partition free energy of benzene between butane and water is 22 kJ mol^{-1} compared to the experimental value[30] of 12.4 kJ mol^{-1} (using cyclohexane). With the new model, the value drops to 10 kJ mol^{-1}, slightly too hydrophilic. Note that the properties of pure benzene solvent are not affected by the change of particle type. The SC5 self-interaction is identical to the SC4 self-interaction (see the full interaction matrix[13]).

Although Trp performs quite well in the Martini 2.1 parameterization, the properties can actually be somewhat improved using a similar particle reassignment as for Phe. In changing the original SC4-SC4-SC4-SP1 to SC5-SC5-SC4-SNd, ΔG^{dim} (water) changes from -4.7 kJ mol^{-1} to -4.0 kJ mol^{-1}, compared to the atomistic value (table 7.3) of -3.3 kJ mol^{-1}. The added benefit of the new particle assignment is the presence of an explicit hydrogen bond donor group (SNd), which better reflects the underlying chemical nature of the Trp side chain. Considering the WW peptides, the new particle assignment for Trp does not change the already good agreement with the experimental binding free energy (Table 7.3). The partitioning profile along the bilayer normal shows significant improvement, in line with the results obtained for Phe (Table 7.3). For the other aromatic residues, His and Tyr, the agreement between atomistic data and the Martini force field is already quite good and no further improvement was attempted.

In the case of Pro, a number of different particle assignments were tried, also involving the backbone bead. We focussed on improving the WW peptide binding free energy. It is clear that the original assignment Na-AC2 is too hydrophobic, and much better results are obtained increasing the polarity of the backbone and/or the polarity of the side chain (Table 7.3). Based on the arguments that *i)* the backbone polarity should be less than that of a regular side chain due to the reduced h-bonding propensity, and *ii)* the side chain analogue is actually propane that should be kept rather apolar, we settled on the P4-C3 combination. Note that the P4 particle type for the Pro backbone only applies to the case where the residue is part of an unstructured chain. As part of α-helix or β-strand less polar particle types are used (see ref. [14]) which remain valid in Martini 2.2.

In summary, we reparameterized the residues Phe, Trp, and Pro, improving their self-association behavior, binding of the respective pentapeptides, and partitioning across the membrane. Future tests should reveal whether or not these parameters lead to a generic improvement.

Table 7.1: Overview of parameters and thermodynamic properties of amino acid side chain analogues: Binding free energy difference $\Delta\Delta G^{\mathrm{WW}}$ of Wimley-White peptides to a POPC/water interface, partitioning free energy ΔG^{part} between water and oil, and dimerization free energy ΔG^{dim} in either water or oil. All values in kJ mol^{-1}. Standard errors for $\Delta\Delta G^{\mathrm{WW}}$ are indicated, standard errors are about 1 kJ mol^{-1} for ΔG^{part} and about 0.1 kJ mol^{-1} for ΔG^{dim}. [b] Italic font denotes results for Martini 2.1, bold font is used for the final set of parameters of Martini 2.2. In cases of polar and charged residues, data are obtained using the polarizable water model and Martini 2.2P. Reference (Ref.) data from: [c] Experimental values[23] with partitioning free energy of Ala set to zero; [d] Experimental data[31, 32]; [e] Atomistic MD data[33]; [f] In case of Pro the first CG particle type is for the backbone.

SC		Type (charge)	$\Delta\Delta G_{WW}$	$\Delta G_{part,d}$	ΔG_{dim} water[e]	ΔG_{dim} oil[e]
Phe	Ref.		5.4±0.3	12	-1.6	-2.9
	CG.	*SC4-SC4-SC4*	12.2±0.1	21	-4.5	-1.3
		SC5-SC5-SC5	7.7±0.1	10	-3.0	-1.7
Trp	Ref.		8.5±0.4	9	-3.3	-3.3
	CG.	*SC4-SP1-SC4-SC4*	9.2±0.1	10	-4.7	-3.0
		SC4-SNd-SC5-SC5	9.4±0.1	8	-4.0	-2.7
Pro	Ref.		-1.2±0.6	-	-	-
	CG.	*Na-AC2*	7.6±0.1	20	-	-
		P4-AC2	4.1±0.1	20	-	-
		P4-C3	1.9±0.1	12	-	-
		P4-C5	0.7±0.1	9	-	-
Thr	Ref.		0.1±0.4	-11	0.2	-5.8
	CG.	*P1*	-1.9±0.1	-12	0.0	-2.3
		N0 (0.36)	-0.3±0.3	-12	-0.5	-4.0
		Nda (0.31)	2.3±0.3	-13	-0.5	-4.2
Ser	Ref.		0.2±0.4	-14	1.6	-5.9
	CG.	*P1*	-1.9±0.1	-12	0.0	-2.3
		N0 (0.40)	-0.5±0.3	-14	-0.2	-5.2
Asn	Ref.		-1.0±0.4	-28	-0.1	-17.3
	CG.	*P5*	-2.7±0.1	-31	0.3	-4.2
		Nda (0.51)	1.9±0.7	-28	-0.2	-20.6
		Nda (0.46)	2.0±0.4	-23	-0.4	-13.9
		N0 (0.54)	-1.3±0.3	-27	-0.2	-18.1
Gln	Ref.		-1.7±0.4	-25	-1.2	-17.2
	CG.	*P4*	-2.0±0.1	-23	-0.1	-3.4
		Nda (0.42)	2.4±0.2	-20	-0.2	-7.2
		N0 (0.51)	-1.1±0.5	-24	-0.6	-14.6
His+	Ref.		-	-	1.0	-
	CG.	C3-Qd	-	-66	0.4	-
		C3-Qd (off-centre)	-	-90	0.5	-
Lys	Ref.		-4.2±0.7	-	1.0	-
	CG.	*C3-Qd*	-3.6±0.1	-55	0.7	-
		C3-Qd (off-centre)	-3.0±0.3	-73	0.2	-
Glu	Ref.		-7.7±0.5	-	1.7	-
	CG.	*Qa*	-5.2±0.1	-71	4.7	-
		Qa (off-centre)	-7.2±0.3	-88	4.0	-
Arg	Ref.		-3.4±0.7	-	1.5	-
	CG.	*N0-Qd*	-3.0±0.1	-70	-0.2	-
		N0-Qd (off-centre)	-2.5±0.3	-89	0.2	-
Asp	Ref.		-4.4±0.4	-	3.9	-
	CG.	*Qa*	-5.3±0.1	-71	4.7	-
		Qa (off-centre)	-6.4±0.4	-88	4.0	-

Improving charged residues by putting the charge off-centre

The absence of partial charges in the standard Martini water model warrants the use of a global screening constant of ϵ_r =15. Together with the smooth shifting of the electrostatic

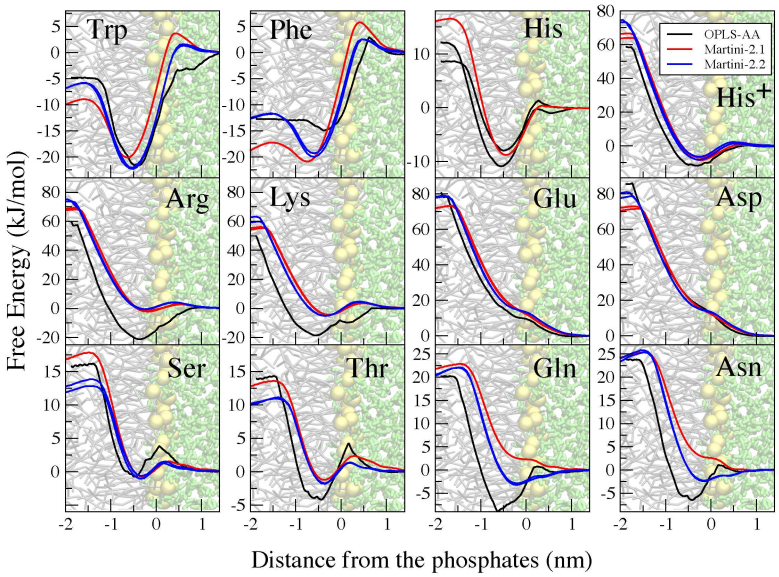

Figure 7.1: Potentials of mean force (PMF) for side chain analogues across a DOPC bilayer interface. Each PMF was set to 0 kJ mol^{-1} in bulk water, and to 0 nm at the peak in the phosphate density for each bilayer. In the background, one leaflet of the DOPC bilayer is hown, with the lipids depicted as grey lines, water as green lines, and the phosphate beads shown by yellow spheres. Black lines denote results obtained with the OPLS all-atom force field, and red/blue lines with Martini 2.1/2.2. In case of polar and charged side chains, Martini 2.1P/2.2P is used. PMFs obtained with the Martini 2.2 models were obtained for both leaflets, and are shown as independent profiles to give an estimate of the accuracy. PMFs for neutral His modeled with the OPLS force field are obtained for both the δ and ϵ protonated form (δ protonated has the lowest energy at the bilayer midplane).

interaction toward the cut-off (1.2 nm), this treatment results in a distance dependent dielectric screening. However, charged particles feel the same screening independent of their environment. Consequently, charge-charge interactions in an apolar medium are severely underestimated. This effect becomes evident from Figure 7.2, which shows the PMFs for the association of the Lys-Lys and Lys-Glu pairs in an apolar solvent, both for Martini and two atomistic models. In Martini 2.1 a relatively shallow contact pair is observed for oppositely charged residues, whereas the atomistic force fields show very strong binding. In contrast, Martini 2.1 predicts a stable contact pair for like charged residues, at odds with the global repulsion seen with the more detailed models.

Improvement of this unphysical situation is obtained using the polarizable Martini water model (compare PMFs with Martini 2.1 to 2.1P in Figure 7.2). The Lys-Lys contact pair disappears, and the Lys-Glu binding strength increases. However, Coulombic interactions fall off as 1/r and thus the bigger size of the CG beads (defined by their van der Waals radius, ≈0.26 nm) limits the approach of the charged beads to ≈0.5 nm, whereas in atomistic models

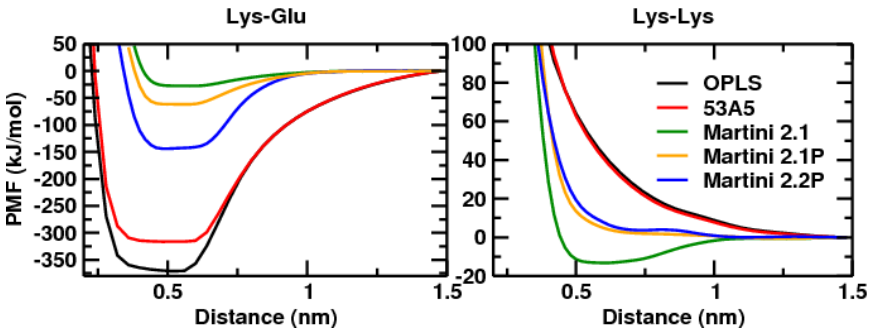

Figure 7.2: Potential of mean force (PMF) for LysGlu (left) and LysLys (right) in decane. PMFs are plotted against the centre of mass distance between the side chain analogue pairs. Black/red lines are obtained with the atomistic force fields OPLS/GROMOS, green/orange/blue lines with Martini 2.1/2.1P/2.2P. Errors do not exceed the thickness of the lines.

the charges can come much closer. Therefore, the interaction strength between the charged atomistic side chains is still much stronger.

To remedy this issue, we designed an alternative model for charged side chains in which the electrostatic and van der Waals interactions are carried by two different particles. The two particles are connected by a constrained bond of length 0.11 nm, as illustrated in Figure 7.3, and have a mass of 36 amu. Using this off-centre set-up, the charges may come closer and the interaction increases by approximately a factor of 3 in the case of the Lys-Glu pair (compare Martini 2.1P to 2.2P in Figure 7.2). The Lys- Lys pair, for which close contact of charges is not favourable, is not further improved.

The new parameters do not change the already good behavior observed with Martini 2.1P in regard to the partitioning of the charged particles along the bilayer interface. The same is true for the binding of the WW peptides, for which the relative binding free energy is reproduced to within 1 k_BT from the experimental values for the charged residues in both Martini 2.1P and 2.2P. Dimerization free energies of charged residues in water remain also largely unaffected and in overall good agreement with the atomistic data.

For histidine the charged form had not been parameterized for Martini 2.1 [14]. Following the general pattern used for the other charged amino acids in Martini 2.1 and using the topology of neutral His, we defined a side chain topology for His+ consisting of a three bead ring: SC4-SP1-SQd. The bonded interactions are unchanged compared to the neutral form, the SC4 bead is bound to the protein backbone and the SQd bead carries a +1 e charge. In addition to this topology in line with the Martini 2.1 force field, we also defined a topology for His+ where the charge was placed off-centre. Both His+ topologies have a strongly negative oil/water partitioning free energy (-66 and -90 kJ mol^{-1}, respectively) and show very similar membrane partitioning behavior (Fig. 1) and dimerization free energies in water and decane consistent with the atomistic data (cf. Table 1).

We conclude that the charge off-centre set-up describes the behavior of charged residues at the distance of contact more realistically. In particular there is a drastic improvement of the SCA dimerization free energy in solvents of low polarity. The off-centre model will be set as default in Martini 2.2P. Whether or not this approach could also improve description

of other charged beads in Martini (e.g., lipid head groups) remains to be tested.

Improving polar residues using polarized particles

Just like interactions between charged beads, interactions between polar particles (P type) in an oil-like medium are grossly underestimated in the Martini force field. Alternatively, we introduce embedded dipoles to polar residues, in line with the polarizable Martini water model. Figure 7.3 shows the typical set-up of a polarized residue. It consists of one virtual site and two real sites. The virtual site is the centre of the LJ interactions and is defined as the geometrical centre of the two real sites. The real sites carry equal partial charges of opposite sign and interact via a Coulomb potential, except between each other. They do not have LJ interactions. Both have a mass of 36 amu and are bound to each other by a 0.28 nm constraint. This bond defines a fixed dipole moment that can thus only contribute to orientational polarization, in contrast to the case of polarizable Martini water where the charges can move independently. The magnitude of the embedded charges is used as a fitting parameter, together with the particle type of the virtual site which determines the strength of the LJ interactions. The resulting dipole moment therefore has no direct physical meaning. Due to the addition of electrostatic interactions, the contribution from the LJ interactions is reduced compared to the standard Martini model.

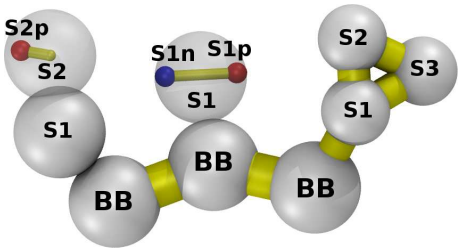

Figure 7.3: Schematic drawing of a Lys-Asn-Phe peptide demonstrating the new topologies for charged and polar amino acids in Martini 2.2P. As an example of charged residues, the Lys side chain consists of two beads S1 and S2, carrying the van der Waals interaction. An additional bead is introduced, S2p, which carries the full positive charge of Lys but has no van der Waals interaction. It is bound to S2 using a 0.11 nm constraint. For polar residues like Asn, the side chain consists of one main particle S1 carrying the van der Waals interaction. It is a virtual site positioned in the middle of two additional sites S1n and S1p that carry a negative and a positive partial charge and no van der Waals interaction. Their mutual distance is constrained at 0.28 nm. The topology of aromatic residues like Phe consists of three beads S1, S2, S3 and remains identical to the topology in Martini 2.1. For each side chain, the S1 particle is bonded to the peptide backbone bead (BB).

We fitted the parameters in first instance by reproducing the experimental oil/water partitioning free energies. The partitioning free energies are very sensitive to the magnitude of the charges. For Thr and Ser, charges in the range 0.3-0.4 e match the experimental data. For Asn and Gln slightly bigger charges 0.4-0.55 e are required, reflecting the higher polarity of these side chains.

We computed the dimerization free energy of polar SCA pairs using the polarized particle type (Table 7.3) and compared to values obtained with atomistic force fields. It is evident that the self-association of each of the polar residues in an apolar solvent (ΔG^{dim}, oil) is improved. In particular Asn switches from weak (around 4 kJ mol^{-1}) to strong (14-20 kJ mol^{-1}) type of association which is consistent with the atomistic data. The dimerization of Thr and Ser also improves by about 2-3 kJ mol^{-1}. Self-association in aqueous environment (ΔG^{dim},water) is largely unaffected and remains in good agreement with the atomistic data for all polar side

chains. Similar ΔG^{dim} can be obtained with Nda and N0 particle types by increasing the dipole charges in N0 compared to Nda.

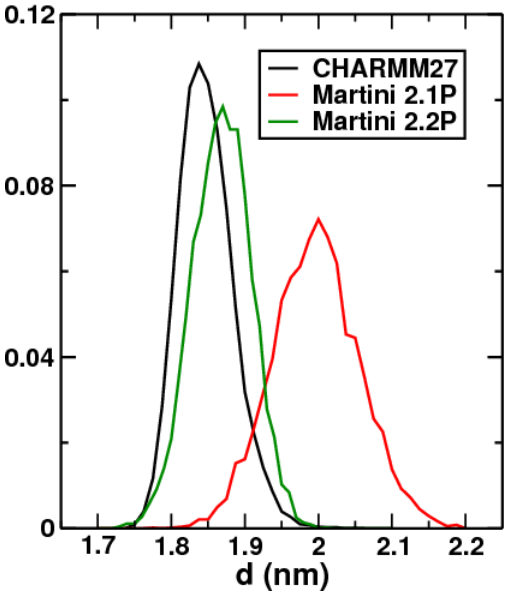

Figure 7.4: Probability distribution of the length of a (Leu)17 α-helix obtained from all-atom simulation (CHARMM27, black curve) and from Martini 2.1/2.2 CG simulations (red/green curves). The length is estimated from the distance (d) between the backbone of Leu3 and Leu15.

We further tested the polarized side chains by considering their partitioning behavior across a lipid membrane. The profiles for Thr and Ser (Figure 7.2) match the atomistic data very well, as does the non-polarized version of Martini. In fact, Martini 2.1 is slightly better in that respect. For Asn and Gln, however, the profiles significantly improve. Notably the free energy minimum at the water/lipid interface is now reproduced, although still too shallow. Asn and Gln require the use of an Nda particle type to observe this minimum. In the case of Ser/Thr, the N0 particle type is appropriate.

Finally, the binding free energy of the WW peptides, $\Delta\Delta G^{\mathrm{WW}}$, was calculated using the polarized side chains. Overall, Martini 2.1 predicts too weak binding of these pentapeptides as compared to the experimental data, although the differences are less than $k_B T$[22]. Perfect agreement can be obtained in the case of Ser and Thr using a polarized particle of N0 type. For Asn and Gln, the N0 particle also gives the best results. As discussed above, the Nda particle that is needed to reproduce the interfacial minimum of the PMFs across the membrane, actually results in too strong binding of the respective pentapeptides by about 1.5 $k_B T$.

The art of coarse-graining is in the compromise: one can not always get everything right at the same time. Here in the case of Asn and Gln side chains we decide to give priority to the partitioning profiles along the membrane normal, which show that Gln and Asn bind strongly at the lipid/water interface. This feature is only reproduced using an Nda particle type. One may argue that the relative weak binding of the Gln and Asn pentapeptides, suggested by the experimental data from Wimley and White, is caused by the restricted orientational freedom of the central residues. Whereas a free Asn or Gln SCA can align its hydrophobic moment along the membrane normal, as part of a pentapeptide this is not possible, providing a possible explanation for the reduced binding strength.

We also tested the use of polarized particles for the neutral form of histidine. However this change did not improve the partitioning behavior over the membrane (data not shown), while the oil-water partitioning free energy deviated significantly more from the experimental value as compared to Martini 2.1. For the topology with the best cross bilayer profile (SC4-SP1-SNda with ± 0.20 e partial charges) the partitioning free energy was -15 kJ mol^{-1} versus

-20 kJ mol^{-1} experimentally obtained. The neutral form of His was not tested with respect to WW peptides and dimerization free energies. For these reasons, we decided to leave the neutral His topology unchanged with respect to Martini 2.1.

In summary, in this section we introduced the concept of polarized particles to remedy some of the shortcomings of polar residues in Martini. Due to the addition of charged interaction sites, the new model will be slightly more expensive than the original model. In practice the polar residues (Thr, Ser, Asn, Gln) only constitute a very small fraction of a protein, which is itself usually only a small part of the simulation system. The additional costs will thus be negligible. However, the polarized particles require the use of the polarizable version of Martini water, which slows down the computation by up to a factor of three (for systems largely composed of water). Hence, in practice, the use of polarized particles should depend on the presumed importance of polar residues in the system of interest. Whether or not the concept of polarized particles can be used to replace polar particles in general, i.e. also as part of other bio-molecules, is currently under investigation. Preliminary endeavours aimed at reproducing membrane poration energies and transition states consistent with all-atom results, a known shortcoming of the Martini lipid model, did not show much improvement yet.

Improving the backbone of α-helices and the stability of poly-Ala and poly-Gly sequences

In the Martini 2.1 protein force field[14], the bonded interactions (bond lengths, bond angles, dihedral angles) were parameterized to match as closely as possible the corresponding distributions extracted from the protein data bank (PDB) for a large number of protein structures (ca. 2000). To obtain a good agreement of the relaxed CG structures with those from the PDB, the bonded parameters for the polypeptide backbone need to depend on the secondary structure. This dependency was implemented in the model through the equilibrium values and force constants of the angles and dihedrals between consecutive backbone beads, which differ for helical, extended, or coiled structures (Table 3 in Monticelli *et al.*[14]). For the bond between two neighboring backbone beads, the force constant was also made dependent on the secondary structure, whereas the equilibrium bond length was set to 0.35 nm in Martini 2.1, irrespective of the secondary structure. However, from the distance distribution extracted from the PDB (Figure 3a in Monticelli *et al.*[14]), it is clear that this approximation may hold quite well for extended and coiled structures, but the distribution between two backbone beads in helices peaks at a shorter distance of about 0.31 nm. In the new Martini 2.2 force field described in this work, the distance between two neighboring backbone beads in helical structures is set to 0.31 nm and treated by a constraint instead of a flexible harmonic bond. This set-up more accurately describes the length of helical structures (see below), and is also more consistent with the secondary structure dependence of the other bonded parameters (angles, dihedrals).

Figure 7.4 shows the length distributions obtained from 100 ns MD simulations of a $(Leu)_{17}$ α-helix (T = 300 K, gas phase). This poly-Leu repeat was used as a representative example of a stable α-helix; however, the described effect is general and not restricted to certain amino acids. In the CG simulations, the helicity was imposed on the entire structure via the dihedral angles, which is the standard procedure used in the Martini force field. In the all-atom simulation, the CHARMM27 force field[34] with the CMAP[35] correction was used. The α-helical structure was stable throughout the 100 ns, with some transient fraying of the terminal 1-2 residues. Thus, to compare all-atom and CG simulations, we analyzed the distance between the backbone of residues 3 and 15. Prior to analysis, the all-atom trajectory was converted to the CG representation, although the difference between the backbone-backbone and Cα-Cα length distributions was negligible. Figure 4 shows that

the Martini 2.1 force field yields too long α-helices (average distance 1.99 nm, red curve) as compared to the all-atom simulation (1.84 nm, black curve), due to the too large distance between neighboring backbone beads. In addition, the CG structure appears to be slightly too flexible, with a standard deviation of the distance distribution of 0.06 nm as compared to 0.04 nm in the all-atom simulation. Introducing a constraint bond length of 0.31 nm in Martini 2.2 (green curve) brings both the average (1.86 nm) as well as the width of the distribution (standard deviation 0.04 nm) into good agreement with the all-atom results.

Another issue pertains to the behavior of poly-Ala and Gly repeats. In the original parameterization of Martini 2.1, the value of the force constant (K_{BB}) used in the harmonic potential of the backbone-backbone bond is a function of the secondary structure of the residues. The relative flexibility of the loop regions (where coil, bend and turn classification are often found) was modelled by a small K_{BB}. This increased flexibility of the bonded terms goes with the increased non-bonded interactions of the same secondary structure types. For instance the coil and bend backbone particle type are P5 that has a very strong self-interaction. This combination of bonded and nonbonded parameters led in some specific cases to the local collapse of the protein backbone most often causing numerical instability. The collapse of the backbone results from the 1-3 non-bonded (LJ) interactions taking over the 1-2 and 2-3 (weakened) bonded terms. In most cases the presence of a side chain bead prevents backbone beads to come close. However, in cases in which two Gly and Ala residues are consecutive, the bond between them would collapse to a value of about 0.1 nm. The solution adopted in Martini 2.2 is the systematic increase of K_{BB} in the coil/turn/bend secondary structure type from 200/400/500 to 1250 kJ mol^{-1} nm^{-2}. This prevents the collapse of the bonds of consecutive Gly and Ala. To compensate for the associated decrease of flexibility the bending angle applied to three consecutive backbone beads in coil/turn/bend is reduced from 25 to 20 kJ mol^{-1}. The new set-up was tested on pentapeptides with sequence AlaAlaAlaAlaAla, AlaAlaValAlaAla and AlaAlaAlaAlaVal and proved to be stable over microsecond time scales using conventional simulation set-ups. It is important to note that this parameterization of the flexible part of proteins applies to folded proteins and should not be taken as general parameterization for flexible protein regions such as large unfolded domains or intrinsically disordered proteins. In such cases we expect that a more elaborated parameterization would be required.

Testing of new parameters on a soluble peptide

To test the behavior of the new class of polarized and charged particles in a soluble peptide, we ran simulations of a small 16 residue helical peptide based on the N-terminal helix of the GCN4 leucine zipper (PDB entry code: 2ZTA). Residues 8 (Lys) and 12 (Leu) were mutated to polar Asn, bringing the total to five polar residues (Gln4, Asn8, Asn12, Ser14, Asn16) and six charged residues (Lys3, Glu6, Asp7, Glu9, Glu10 and Lys14), modelled as polarized and off-centre charge particles, respectively. We also used the shorter backbone bond in this simulation. The helical conformation was restrained by using the standard Martini bonded potentials for α-helices. Figure 7.5 shows the histograms of the COM distances between pairs of polar (left panel) and charged (right panel) side chains. The results are compared to a fully atomistic simulation using the CHARMM27 force field[34] with the CMAP[35] and to Martini 2.1P. For the polar residues there are two main differences between the Martini 2.1P and Martini 2.2P. First the peaks observed at short distance (side chains pointing towards each other) and large distance (side chains pointing away from each other) shift to a smaller average distance in Martini 2.2P, thus better matching the position of the peaks obtained from the atomistic simulations. Second, the distributions of the pairs (Asn12-Ser14 and Ser14-Asn16) become narrower. However, the peak found at d$\approx$0.65 nm (side chains pointing in the same direction) in the atomistic Asn8-Asn12 pair is still not observed in the CG simulations.

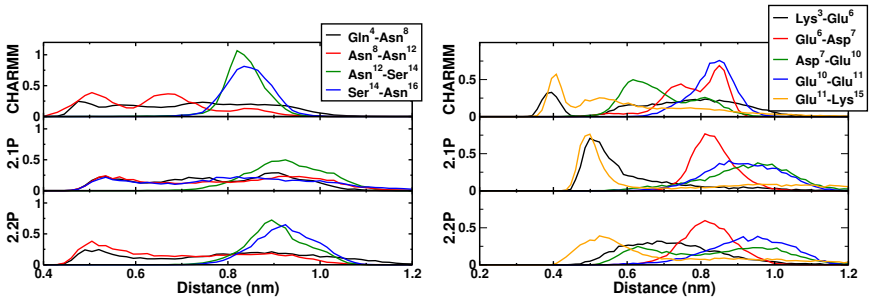

Figure 7.5: Distance distribution of the side chains of pairs of polar (left) and pairs of charged residues (right) that are part of the small helical GCN4 peptide. The distributions for the old Martini 2.1P, the new Martini 2.2P, and the all-atom CHARMM27 force field are compared.

In case of the charged side chains the situation is less clear. For the oppositely charged Lys3-Glu6 and Glu11-Lys15 pairs the peak observed at short distance becomes broader or completely disappears when using Martini 2.2P. For the oppositely (Glu6-Asp7) and equally charged (Glu10-Glu11) direct neighbors the same orientations are sampled with the new parameters. For the oppositely charged Asp7-Glu10 pair, a new orientation at close distance is sampled, thus better reproducing the atomistic distribution.

Although the overall behavior of the side chains for this particular peptide appear to have improved with our new model, it is also clear that our CG model can not capture some of the fine details of the distributions as seen when using an atomistic model. Additional testing is required to assess whether or not the Martini model can be further improved in this respect.

Implementation

To facilitate the use of the Martini force field, an auxiliary program was developed called *martinize.py*. This program offers a one-step solution for coarse-graining atomistic structures, yielding coarse grain structures and corresponding topologies. The program contains easily editable tables for mapping atoms to CG beads and for assigning atom types and bonded parameters based on residue and secondary structure type. The secondary structure of a protein sequence can be specified explicitly or can be inferred from the structure by a call to the DSSP program . Other features include the possibility to specify disulfide bridges, adding arbitrary links between beads, and support for writing structures and topology files using the Elnedyn approach. The *martinize.py* script can be downloaded from `http://cgmartini.nl`.

The user can also specify the version of the Martini force field that is used: Martini 2.1 or 2.2 with standard water, or 2.1P and 2.2P in combination with polarizable water. The default of Martini 2.2P will be the use of polarized particles to model polar side chains and the charge off-centre model for charged side chains. Both set-ups can be optionally reverted to regular particle types.

7.4 Conclusions

Bassed on recent data in which the Martini protein force field was compared to all-atom force fields and experimental data, we reparameterized a number of side chains. For Phe we now use the SC5 particle type that makes this aromatic residue slightly more polar. Similarly, Trp behavior is improved with the new SC5-SC5-SC4-SNd assignment. For Pro, the side chain is made slightly more polar (C3) and at the same time the polarity of the backbone bead is increased (P4). Furthermore, we introduced a class of polarized particles to model the polar but neutral Asn, Gln, Ser, and Thr residues. This greatly improves their dimerization free energy in low dielectric solvents. We also presented a new model for charged residues with an off-centre charge. This leads to an improvement of the potentials of mean force between two oppositely charged residues in apolar medium. Both the polarized particle types and charge off-set models should be used in combination with the polarizable Martini water model. Finally we changed some of the bonded parameters to provide a better description of the length of an α-helix and to improve on numerical stability of Gly and Ala repeats. All systems simulated in this study are stable with time steps up to at least 20 fs.

The changes described in the current chapter, combined with the parameters from versions 2.1 and 2.1P that have not changed, define version 2.2 and 2.2P of the Martini protein force field. The new protein force field is still fully compatible with the Martini 2.0 lipid and carbohydrate force fields. A generic script has been developed in which these changes are implemented, allowing for a straightforward set-up of CG simulations based on an atomistic input structure. Further testing is required to verify whether or not the changes proposed here are generic improvements. With the help of many Martini users around the globe, we are constantly trying to further optimize the model. Planned changes include a change in form of the non-bonded interaction potential to improve the surface tension of polar solvents and reduce the general over structuring seen with the current force field, as well as the introduction of polarized particles in the peptide backbone that will allow secondary structure transitions to occur.

7.5 Bibliography

[1] G.A Voth. *Coarse graining of condensed phase and biomolecular systems*. CRC Press, 2009.

[2] J Padding and W Briels. Systematic coarse-graining of the dynamics of entangled polymer melts: the road from chemistry to rheology. *J. Phys-Condens. Mat.*, 23:233101, 2011.

[3] M Deserno. Mesoscopic membrane physics: Concepts, simulations, and selected applications. *Macromol. Rapid Comm.*, 30:752–771, 2009.

[4] C Clementi. Coarse-grained models of protein folding: toy models or predictive tools? *Cur. Opin. Struc. Biol.*, 18:10–15, 2008.

[5] M Klein and W Shinoda. Large-scale molecular dynamics simulations of self-assembling systems. *Science*, 321:798, 2008.

[6] A Lyubartsev and A Laaksonen. Calculation of effective interaction potentials from radial distribution functions: A reverse monte carlo approach. *Phys. Rev. E*, 52:3730–3737, 1995.

[7] D Reith, M Pütz, and F MüllerPlathe. Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem.*, 24:1624–1636, 2003.

[8] S Izvekov, A Violi, and G.A Voth. Systematic coarse-graining of nanoparticle interactions in molecular dynamics simulation. *J. Phys. Chem. B*, 109:17019–17024, 2005.

[9] W Noid, J Chu, G Ayton, V Krishna, S Izvekov, G.A Voth, A Das, and H.C Andersen. The multiscale coarse-graining method. i. a rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.*, 128:244114, 2008.

[10] E Brini and N van der Vegt. Chemically transferable coarse-grained potentials from conditional reversible work calculations. *J. Chem. Phys.*, 137:154113, 2012.

[11] A Chaimovich and M Shell. Coarse-graining errors and numerical optimization using a relative entropy framework. *J. Chem. Phys.*, 134:094112, 2011.

[12] S.J Marrink, A.H de Vries, and A.E Mark. Coarse grained model for semiquantitative lipid simulations. *J. Phys. Chem. B*, 108:750–760, 2004.

[13] S.Jan Marrink, H.J Risselada, S Yefimov, D.P Tieleman, and A.H de Vries. The martini force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 111:7812–7824, 2007.

[14] L Monticelli, S.K Kandasamy, X Periole, R.G Larson, D.P Tieleman, and S.J Marrink. The martini coarse-grained force field: extension to proteins. *J. Chem. Theory Comput.*, 4:819–834, 2008.

[15] J Chu, G Ayton, S Izvekov, and G Voth. Emerging methods for multiscale simulation of biomolecular systems. *Mol. Phys.*, 105:167–175, 2007.

[16] S Nielsen, R Bulo, P Moore, and B Ensing. Recent progress in adaptive multiscale molecular dynamics simulations of soft matter. *Phys. Chem. Chem. Phys.*, 12:1240112414, 2010.

[17] A. Rzepiela, M Louhivuori, C Peter, and S.J Marrink. Hybrid simulations: combining atomistic and coarse-grained force fields using virtual sites. *Phys. Chem. Chem. Phys.*, 13:1043710448, 2011.

[18] C Peter and K Kremer. Multiscale simulation of soft matter systems–from the atomistic to the coarse-grained level and back. *Soft Matter*, 5:4357–4366, 2009.

[19] S Kamerlin and A Warshel. Multiscale modeling of biological functions. *Phys. Chem. Chem. Phys.*, 13:10401–10411, 2011.

[20] S.O Yesylevskyy, L.V Schäfer, D Sengupta, and S.J Marrink. Polarizable water model for the coarse-grained martini force field. *PLoS Comput. Biol.*, 6:e1000810, 2010.

[21] Z Wu and Q Cui. A new coarse-grained model for water: The importance of electrostatic interactions. *J. Phys. Chem. B*, 114:10524–10529, 2010.

[22] G Singh and D.P Tieleman. Using the wimley-white hydrophobicity scale as a direct quantitative test of force fields: the martini coarse-grained model. *J. Chem. Theory Comput.*, 7:23162324, 2011.

[23] W Wimley and S White. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.*, 3:842–848, 1996.

[24] J.L MacCallum, W.F Bennett, and D.P Tieleman. Distribution of amino acids in a lipid bilayer from computer simulations. *Biophys. J.*, 94:3393–3404, 2008.

[25] B Hess, C Kutzner, D van der Spoel, and E Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4:435–447, 2008.

[26] S.J Marrink, X Periole, D.P Tieleman, and A.H de Vries. Comment on "on using a too large integration time step in molecular dynamics simulations of coarse-grained molecular models" by m. winger, d. trzesniak, r. baron and w. f. van gunsteren, phys. chem. chem. phys., 2009, 11, 1934. *Phys. Chem. Chem. Phys.*, 12:2254, 2010.

[27] H.J.C Berendsen, J.P.M Postma, W.F Van Gunsteren, A DiNola, and J.R Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684, 1984.

[28] S kumar, D Bouzida, R Swendsen, P Kollman, and J.M Rosenberg. The weighted histogram analysis method for freeenergy calculations on biomolecules. i. the method. *J. Comput. Chem.*, 13:1011–1021, 1992.

[29] B Hess, C Holm, and N.F.A van der Vegt. Osmotic coefficients of atomistic nacl (aq) force fields. *J. Chem. Phys.*, 124:164509, 2006.

[30] J.S Hub, B de Groot, and D van der Spoel. g_wham a free weighted histogram analysis implementation including robust error and autocorrelation estimates. *J. Chem. Theory Comput.*, 6:37133720, 2010.

[31] A radzicka and R Wolfenden. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry*, 27:1664–1670, 1988.

[32] R Wolfenden, L Andersson, and P Cullis. Affinities of amino acid side chains for solvent water. *Biochemistry*, 20:849–855, 1981.

[33] D.H de Jong, X Periole, and S.J Marrink. Dimerization of amino acid side chains: lessons from the comparison of different forcefields. *J. Chem. Theory Comput.*, 8:1003–1014, 2012.

[34] A.D MacKerell, N Banavali, and N Foloppe. Development and current status of the charmm force field for nucleic acids. *Biopolymers*, 56:257–265, 2001.

[35] A.D MacKerell, M Feig, and C.L Brooks. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.*, 25:1400–1415, 2004.

8

On the Use of Mie Non-Bonded Coarse Grain Potentials

Abstract

Coarse grain molecular force fields used in MD simulations provide the advantage over classic force fields to allow simulations of bigger systems for longer times. The popular Martini coarse grain force field, although extensively used, does not perform optimal in several aspects: It has a too high freezing temperature, a too low air-water surface tension and too much structure in the the water radial distribution function. The latter of those problems also occur for other solvent. In order to improve on those aspects we have investigated the use of shifted Mie-potentials. We show that using a Mie 4-2 potential gives a surface tension of air-water interfaces in better agreement with experiments when compared to the Lennard-Jones potential currently used in the Martini model. Furthermore, it gives a much lower melting temperature of water in comparison to the current model. The correct density of water and linear alkanes are maintained. Using the Mie 4-2 potential thermodynamic properties are calculated for different coarse grain beads as a first step towards an revised version of the Martini model.

8.1 Introduction

For contemporary molecular dynamics computer simulations to reach the time and length scales of interest for most (bio)molecular systems, coarse grain (CG) force fields have become commonplace[1]. Several methods are available to derive a CG potential, notable examples being *inverse Boltzmann*, [2] *force matching* [3, 4] or systematic parametrization against thermodynamic data. The former two methods yield potentials that are strongly system dependent. In contrast the latter method provides so called *transferable force fields*, which are parametrized such that different molecules may be combined and simulated using the same force field. A transferable force field allows one to (relatively) quickly parameterize new types of molecules and combine those with existing molecules.

A popular example of a transferable CG force fields is the Martini model, developed by Marrink and coworkers. For this force field parameters are available for, among others, water[5, 6], lipids[6], proteins[7, 8], sugars[9], fullerenes[10] and polymers[11]. The Martini force field applies on average a mapping of four heavy atoms to one coarse grain bead. The non-bonded interactions between the beads have been parametrized against thermodynamic data. In particular the heat of vaporization, hydration and partitioning between polar and apolar phases have been considered. Bonded interactions are partially parametrized against structural information, such as lipid bilayer properties, and partially based upon simulations

using atomistic force fields. A more detailed describtion of the implementation of the Martini forcefield can be found in chapter 1.

A major determining factors for the behavior of a force field is the choice of the interaction potential. In the Martini force field mentioned, a Lennard-Jones potential[12] is used:

$$V(r_{ij}) = 4\epsilon \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right), \tag{8.1}$$

where $V(r_{ij})$ is the potential energy between particles i and j at the distance r_{ij}, ϵ the minimum value of the potential and σ_{ij} the distance at which the potential is zero. The first term inside the brackets (the *"12-term"*) represents the repulsive interactions, whereas the second term (the *"6-term"*) represents the attractive interactions. The LJ potential is a special form of the more general Mie potential[13]:

$$V(r_{ij}) = \left(\frac{n}{n-m} \right) \left(\frac{n}{m} \right)^{m/(n-m)} \epsilon \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{n} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{m} \right), \tag{8.2}$$

where the exponents, n and m, are 12 and 6 in the case of the Lennard-Jones potential. Note that both n and m should be positive integers and that n should always be larger than m. The Lennard-Jones potential is often used in atomistic force fields[14], because it represents the interactions between single atoms quite well: the *6-term* is derived from the attractive London-forces[15], whereas the *12-term* is a reasonable representation of the repulsive forces and is computationally cheap to calculate, by taking the square of the *6-term*.

In the Martini force field the interactions are only calculated up to 1.2 nm for computational efficiency. In order to assure a continuous potential (and subsequent force) the potential is gradually shifted to zero starting from 0.9 nm and up to 1.2 nm[16, 17]. The Martini model shows some unwanted characteristics related to the choice of the potential. First, due to the relative steep shape of the Lennard-Jones potential the radial distribution function (RDF) of solvents, like water, shows too much long range structure[18](see also Figure 8.2 below). Second, due to the surplus of long range structure, the melting temperature of Martini water is too high[5]: $T_{melt} = 290 \pm 5$K ($T_{exp} = 273$K). Third, the hydration free energy of all beads is too low in comparison to the molecular building blocks used in the parametrization (see Table 3, Marrink *et al.*[6]). And fourth, due to the lack of long range interaction (in contrast to long range *structure*), the water-air surface tension is too low[6].

As several studies[19, 20, 21] have pointed out, this can be solved by using a *"softer"* potential: the attractive well should have less steep flanks and the attraction at longer distance should be stronger. Different approaches are conceivable to obtain such a potential. Van Hoof *et al.* used the CUMULUS coarse graining method[20], where they obtained tabulated (non-analytical) potentials with a 4:1 mapping for water and different solutes. The method uses a trajectory of an atomistic simulation, in which neighboring small molecules such as water and ions are grouped together. On the resulting mapped trajectory an iterative Boltzmann inversion[2] procedure is applied to obtain the CG potential. Disadvantages of this method are the non-analytical shape of the potential, which makes it less easy to implement and the specificity of the potential: even though, as the authors note, the potentials are similar when derived for different systems, there is no obvious way to derive a potential that is representative for a range of molecules.

Chiu *et al.* obtained a softer transferable CG potential in a similar way as done for the Martini potential[19]. They parameterized water with a 4:1 mapping against thermodynamic data, but using a Morse potential[22] instead of a LJ-potential. In addition the authors obtain parameters for alkanes using a mixed 4:1 and 3:1 mapping. The Morse potential has a less steep profile and is more attractive at longer distances, which gives rise to better solvent properties. A disadvantage of the work by Chiu *et al.* is the lack of a shifting function,

which makes the potential non-zero at 1.2 nm. A much larger cut-off (1.6 nm) has to be used, making calculations considerably slower.

He *et al.* have explored a wide range of Mie-potentials for water with different mapping ratios (1 to 4 water molecules per CG bead) and different cut-off distances[21]. They parameterized the potentials against density, compressibilty and surface tension of water and concluded that reasonable values for those three properties may be obtained with different mappings, Mie-exponents and cut-offs. In addition they find that a mapping of more than 3:1 does not give reasonable results with any of the parameters tested. However in this study, non-shifted potentials were used, which causes the potentials to be non-zero at the cut-off. Because of this the authors did not explore the use of Mie-potentials with low exponents, as they have a longer ranged interaction and consequently a too large energy gap at the cut-off.

Here we investigate the possibility to use Mie-potentials with different exponents n and m with a 4:1 mapping in combination with a shift function. The objective is to derive a potential that could replace the current Martini potential without the need to replace the interaction matrix. In other words, it should be possible to simulate previously parameterized molecules with the new non-bonded potential using the current mapping and bonded interactions. To this end different bead-types will be parametrized, corresponding to the standard bead types of Martini (see Table 1 from Marrink *et al.*[6]). The choice of Mie-exponents and ϵ and σ will be determined based upon comparison to experimental data. In the first phase liquid density, compressibility and liquid-air surface tension for water and octane will be used to obtain parameters for the beads representative of these liquids. Next, the densities for alkanes of differing length will be matched to obtain, on average, correct behavior for all alkanes. And finally the heat of vaporization, hydration and partitioning between different solvents will be matched for all the different bead types as was done for the Martini model. The rest of this chapter is organized as follows: First, we describe the methods used (Section 8.2), next the results will be presented and discussed following the three phases described above (Section 8.3). The chapter will end by a short concluding section (Section 8.4).

8.2 Methods

Simulation setup

All simulations were performed using the Gromacs simulation package[23] (version 4.5.x). The tabulated non-bonded potentials used were calculated between 0 and 1.2 nm using equation 8.2 with a shift function as implemented in Gromacs [17] applied between 0.9 and 1.2 nm. For molecules consisting of more than one bead Martini bonded parameters (described in [5]) were used. Coarse grain masses were set to match the molecular masses, unless mentioned otherwise. A time step of 20 fs was used. The neighborlist is updated every 10 steps, with a neighborlist cut-off of 1.4 nm, 0.2 nm longer than the non-bonded cut-off to ensure energy conservation. Temperature and pressure were kept constant by coupling to an external bath[24, 25], using coupling constants $\tau_t = 1.0$ ps^{-1} and $\tau_p = 3.0$ ps^{-1} for temperature and pressure, respectively. All simulations were done at a temperature of 298.15 K.

Liquid properties

As representative of polar and apolar phases, water and octane were chosen. Water consists of a single interaction site (a P4 bead), corresponding to four water molecules. Octane (consisting of two beads each representing four carbon atoms) was preferred over butane (one bead; four carbon atoms), since the boiling point of butane is below our simulation

temperature, 298.15K[26]. The choice of octane was preferred over lowering the simulation temperature, in order to parameterize everything at the same temperature. Both octane sites are represented by a C1 bead, connected by a harmonic bond 0.47 nm and 1250 kJ nm^{-2} force constant.

For the determination of the optimal σ and ϵ for Mie-potentials with different exponents, an approach similar to He *et al.*[21] was used. The protocol consisted of four steps and was used for both water and octane. First, a solvent box containing 1000 CG water beads (equivalent to 4000 water molecules) or 318 octane molecules at the experimental density was created, roughly corresponding to a boxsize of 5x5x5 nm in both cases. The system was relaxed at constant particle number, volume and temperature (NVT ensemble) for 500 ps. Second, the relaxed system was simulated for 4 ns at constant particle number, pressure and temperature (NPT ensemble). From this run the last 1500 ps were used to calculate the average pressure, temperature, volume and density of the system. Third, the Z-dimension of the relaxed system was increased threefold to create and air-liquid interface. The resulting system was simulated for 4 ns in an NVT-ensemble. The last 1500 ps were used to calculate the surface tension, γ, using[21]:

$$\gamma = \frac{1}{2} Z \langle P_z - \frac{(P_x + P_y)}{2} \rangle, \tag{8.3}$$

where Z is the length of the box in the Z dimension and P is the pressure with the subscripts indicating the dimension along which the pressure was measured. Fourth, a system with a volume equal to the equilibrium volume determined in the second step plus and minus 1% (dubbed S_{small} and S_{large} respectively) was simulated for 4 ns in an NVT-ensemble. From these simulations the compressibilty, β, was calculated using[27]:

$$\beta = \frac{V_{small} - V_{large}}{V_{average} \cdot (\langle P_{large} - P_{small} \rangle)} \tag{8.4}$$

where V and P are volume and pressure and the subscripts indicate the small or large system, or the average determined in the second step.

For every combination of the Mie-exponents (n and m) these four steps were repeated for a range of values for σ (0.45-0.55 nm) and ϵ (4.5-15.5 kJ mol^{-1}) in steps of 0.005 nm and 0.1 kJ mol^{-1}, respectively. The calculated values for the density (second step), surface tension (step three) and compressibility (step four) were plotted against σ and ϵ, together with the experimental values. From this initial scan an approximate value for both σ and ϵ was determined, around which a second series of simulations was made with a finer σ/ϵ-grid. From the fine grid optimal parameter values were chosen by hand, instead of by fitting the surfaces as done in [21]. In order to keep the model simple all beads were to have the same van der Waals radius, and thus σ was required to be equal for water and octane. After a value for σ and ϵ for water had been determined, the octane ϵ was adjusted to optimize the octane properties.

Radial distribution function

Radial distribution functions (RDFs) were calculated using the Gromacs tool g_rdf. A reference RDF was obtained from an atomistic simulation, for which an approach similar to Fuhrmans *et al.*[18] was used. Four SPC-water molecules were simulated *bundled* together by harmonic bonds between all oxygens in the cluster, while applying a slightly increased C_{12} parameter to maintain the correct water properties. The centers of mass of the water bundles are mapped to coarse grain sites. An RDF is calculated for the thus obtained CG trajectory. For octane, atomistic simulations using the Gromos 43A3 force field[27] are mapped to CG trajectories. RDFs are calculated from these mapped trajectories as well.

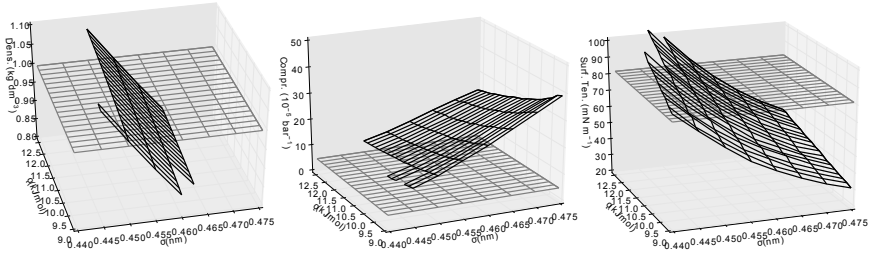

Figure 8.1: Example plots showing the density (left), compressibility (middle) and surface tension of water as a function of sigma and epsilon for the Mie potential with exponents 4-2. Black wire frame are values obtained from simulations. Grey wire frame is the experimental value.

Linear alkane densities

The densities of linear alkanes were calculated for CG molecules with 1 (butane), 2 (hexane and octane), 3 (decane and dodecane), and 4 (tetradecane and hexadecane) beads. For the molecules with the same number of beads the bead masses were set to match the total molecular mass and all beads in one molecule had the same mass. For these liquids a system of approximately 5x5x5 nm was simulated for 9 ns, of which the first 5 ns were discarded as relaxation period.

Heats of vaporization, hydration and partitioning

Based on the final choice of m and n, free energies of vaporization, hydration and partitioning were determined for interaction strengths covering the entire range of polar to non-polar CG beads using the free energy perturbation (FEP) method[28]. In FEP the van der Waals interactions between a single solute bead in a box of 1000 solvent molecules, controlled by a parameter λ, are slowly turned off. λ was switched from one (full interactions) to zero (no interactions) in 21 windows. Each window was simulated for 500 ps. A soft-core (sc) potential modification term was used to smooth the transition along λ. The smoothest transition curve was found using sc-α=0.025, sc-σ=0.455 nm, and the sc-exponent=1. For each λ-window energies were calculated for the current window and neighboring windows to either side, using the foreign-lambda code implemented in Gromacs[23]. The total free energies were calculated using the "Bennet Acceptance Ratio"[29], implemented in the Gromacs tool g_bar.

8.3 Results and Discussion

Our first target is to find optimal Mie potentials that could reproduce the experimental values for the density, surface tension and compressibility of water (996 kg m^{-3}, 71.2 mN m^{-1} and 4.6×10^{-5} bar^{-1}, respectively) and octane (698 kg m^{-3}, 21.8 mN m^{-1} and 12.9×10^{-5} bar^{-1}, respectively) at 298.15 K. Additionally the melting temperature of the model should be sufficiently low in order to prevent the freezing of systems simulated at temperatures close to the experimental melting temperature. To this end, we systematically varied σ and ϵ at different Mie exponents n and m and measured the system properties.

Table 8.1: List of optimal σ and ϵ values for water obtained for different m and n Mie-exponents and corresponding water properties. Experimental values are obtained from [26]. Values between brackets are relative deviations from experimental values.

Pot. (m-n)	σ(nm)/ ϵkJ mol^{-1}	Dens. (kg dm^{-3})	Surf.Ten. (mN m^{-1})	Compr. (10^{-5}bar^{-1})	Melt. T. (K)
Exp.	-/-	0.997	72.0	4.6	273.15
5-4	0.4725/7.10	1.009(1.2%)	70.9(-1.5%)	7.8(-68.6%)	266
5-3	0.4825/5.40	1.005(0.8%)	72.1(-0.2%)	10.0(117%)	210
5-2	0.4725/6.35	1.016(2.0%)	71.5(-0.7%)	11.5(150%)	180
4-3	0.4675/9.10	0.997(0.0%)	72.4(0.5%)	10.1(120%)	213
4-2	0.4550/12.3	0.993(-0.4%)	72.5(0.7%)	11.9(159%)	193
3-2	0.4425/15.5	0.992(-0.5%)	72.7(1.0%)	14.4(212%)	163

Table 8.2: List of optimal σ and ϵ values for octane obtained for different m and n Mie-exponents and corresponding octane properties. Experimental values are obtained from [26]. Values between brackets are relative deviations from experimental values.

Pot. (m-n)	σ(nm)/ ϵ(kJ mol^{-1})	Dens. (kg dm^{-3})	Surf.Ten. (mN m^{-1})	Compr. (10^{-5}bar^{-1})
Exp.	-/-	0.699	21.8	12.9
5-4	0.4725/2.4	0.700 (0.3%)	14.9 (-32%)	40.7 (216%)
5-3	0.4825/3.6	0.681 (-2.6%)	21.6 (0.7%)	11.5 (11%)
4-3	0.4675/3.4	0.732 (4.8%)	21.5 (1.5%)	34.5 (170%)
4-2	0.4550/4.4	0.723 (3.5%)	20.8 (4.7%)	40.6 (216%)
3-2	0.4425/4.9	0.699 (0.0%)	17.1 (-17.1%)	54.6 (325%)

Parametrization of water and octane

Figure 8.1 shows the three dimensional plots of density, compressibility and surface tension for the Mie 4-2 potential. Based on these plots optimal values for σ and ϵ were determined. As already pointed out by He *et al.*[21], it was not possible to obtain correct values for all three properties for a 4:1 mapping. Therefore, we here focused on density and surface tension, since those properties are of key importance for many bio-molecular applications. Tables 8.1 and 8.2 list the best values obtained for σ and ϵ for a few potentials with low n and m parameters. Since we decided to use the same σ for different bead types, octane has the same σ as water. For the Mie 4-2 potential (our final choice, see below), properties for two values of ϵ are reported. Although octane has a better performance in terms of density with $\epsilon = 4.4$ kJ mol^{-1}, the average density over all linear alkanes tested here is better when using $\epsilon = 5.35$ kJ mol^{-1} (see below).

Softer potential reduces over structuring

Figure 8.2 shows the RDFs of water (top panel) and octane (bottom panel) for a selected number of potentials. For both liquids, the RDFs show less structure for potentials with lower m & n-exponents. In water the improvement over the Martini 12-6 potential for the first solvation shell (peak around 0.5 nm) is considerable for all Mie-potentials with lower exponents, however the difference to the bundled atomistic water remains large. Also, all RDFs show only a minor decrease of structure at larger distance. The improvements between the different soft potentials are relatively small. For octane, the RDFs for the soft potentials are much closer to the mapped atomistic RDF, and even show too little structure at larger distance. The softest potential shown (Mie 3-2) also considerably underestimates the first

solvation shell.

The lack of long range structure and relatively low first solvation peak in the RDF of water demonstrates the special nature of water. Clusters of four water molecules can interpenetrate each other, while CG water beads, representative of such a cluster, can not interpenetrate. Other liquids with a density similar to water will not show this interpenetration and subsequently show a more structured RDF. For example for ethanediol, which in Martini is represented by the same P4 particle as water, the first solvation peak in the RDF of a mapped atomistic simulation is high when compared to bundled water (see Figure 8.2). This difference in behavior of water makes it difficult to model water with the same potential form as other molecules.

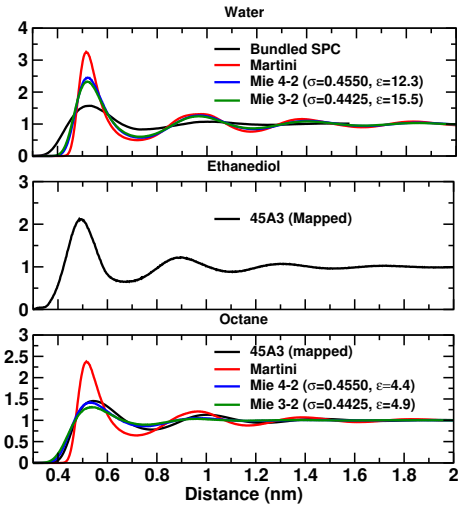

Figure 8.2: Water (top), ethanediol (middle) and octane (bottom) radial distribution functions from simulations using atomistic (black line), Martini 2.0 (red line) and Mie potentials with exponents 4-2 (blue) or 3-2 (green). Water and ethanediol are represented by the same (P4) bead in Martini.

Comparing Mie 4-2 to other potentials

To obtain a reasonably unstructured water RDF but avoid underestimating structure in octane, in the remainder of this work we continue using the Mie 4-2 potential. Figure 8.3 shows this potential as obtained for water and compares it to the water potential used by the Martini force field[5], the potential obtained by Chiu *et al.*[19] and the tabulated potential obtained by Van Hoof *et al.*[20]. The potential by He *et al.* is not included in the comparison since it uses a 3:1 mapping. Of the four potentials shown, the current Martini LJ12-6 potentials has its minimum at the smallest distance ($\approx$ 0.53 nm), whereas the tabulated potential has it at the largest distance ($\approx$ 0.73 nm). The Mie 4-2 and the LJ12-6 potential are comparable in depth, even though their epsilon is greatly different; due to the shift applied to make the potential zero at the cut-off, the Mie potentials with low exponents get shifted up strongly for any distance. At small distance (the *"repulsive regime"*) the Mie 4-2, LJ12-6 and the Morse potential behave comparably, while the tabulated potential is by far the least steep. At larger distance the LJ12-6 potential is the least attractive, and the other potentials behave similarly. The Morse potential is not shifted, and thus requires a longer cut-off distance (1.6 nm).

Properties of linear alkanes with Mie 4-2

Using the obtained interaction potential, we calculated the densities for a series of linear alkanes shown in Figure 8.4. Only the densities of alkanes with an even number of carbons

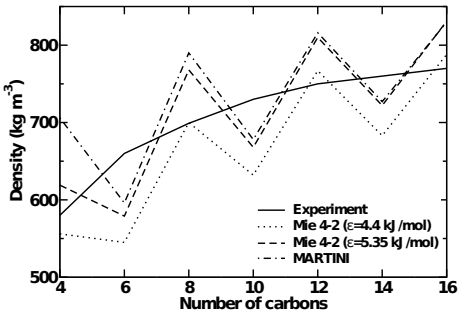

Figure 8.4: Alkanes densities. The solid line represents experimental values obtained from [26]. The dotted line is obtained with the Mie 4-2 potential with $\epsilon = 4.4$ kJ mol^{-1}. The dashed line is obtained with the Mie 4-2 potential with $\epsilon = 5.35$ kJ mol^{-1}. The dash-dotted line is obtained with the original Martini LJ12-6 potential. The zigzag pattern is caused by two consecutive alkanes being represented by the same number of beads in Martini with different bead masses.

have been measured. Since in Martini C1-beads correspond to 4 carbon atoms, two subsequent alkanes in the measured series are represented by the same number of beads. The masses of the beads are set to match the molecule mass and thus differentiate the molecules. Together this causes the zigzag behavior seen in Figure 8.4. With octane as our original target fluid, the average density of the alkanes is consequently too high. In order to correct for this the density is increased by raising ϵ for the interaction between two C1-beads from 4.4 kJ mol^{-1} to 5.35 kJ mol^{-1} (Table 8.2 shows the corresponding density, compressibility and surface tension for octane.) The new alkane densities (Figure 8.4, dashed line), on average, match the experimental densities and are very close to the densities obtained with the Martini model.

Interpolating interaction strengths

In the Martini model different levels of interaction strength are defined, numbered O to IX. The levels differ only in the Lennard-Jones ϵ parameter (except level IX which has a higher value for σ). Using those 10 levels the interactions between all beads are defined. For the Mie 4-2 potential, the same 10 levels need to be defined. The interactions strengths obtained above by fitting to water and alkane properties, correspond to C1-C1 and P4-P4 interactions, levels I and IV in the Martini model. In order to investigate the influence of changing different interaction levels, three sets of ϵ values were obtained for the remaining levels by roughly extra- and interpolating from levels I and IV. The three sets are denominated $\mathbb{A}, \mathbb{B}$ and $\mathbb{C}$, the obtained ϵ values are given in Table 8.3. The shape of the potentials belonging to these sets can be found in Figure 8.5. σ was 0.455 nm for all levels and all sets except level IX for which σ was scaled to 0.5 nm for all sets. For all three sets ϵ for level I was kept the same in order to preserve the correct water properties. Level IV was varied slightly, thus varying

Table 8.3: Three sets of ϵ values corresponding to 10 interactions levels. Values are in kJ mol^{-1}. σ is 0.455 nm for all levels and all sets except level IX where σ 0.5 nm for all three sets.

	O	I	II	III	IV	V	VI	VII	VIII	IX
$\mathbb{A}$	13.3	12.3	7.8	6.1	5.35	4.7	4.25	4.0	3.9	3.9
$\mathbb{B}$	13.3	12.3	9.3	7.5	5.9	5.4	4.9	4.5	3.9	3.9
$\mathbb{C}$	13.3	12.3	9.5	8.7	6.4	6.0	6.0	4.0	2.0	2.0

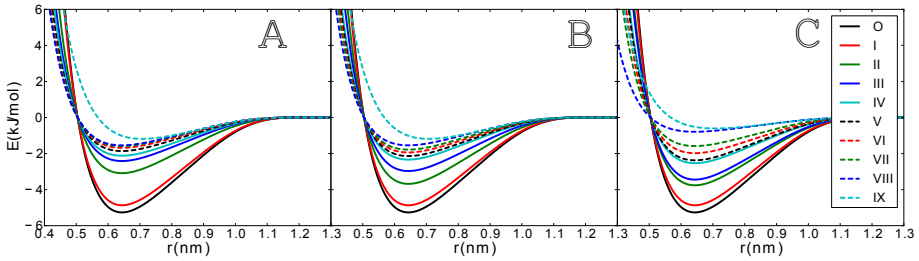

Figure 8.5: Plots of the three sets of Mie potentials. The corresponding σ and ϵ parameters can be found in Table 8.3.

the properties of the linear alkanes. Set $\mathbb{A}$ uses the ϵ value obtained by matching the alkane densities.

Polar liquids with Mie 4-2

Except for the linear alkanes several other liquids have been parametrized in Martini. Of those (di-ethyl)ether, chloroform and octanol are of special interest; these solvents form the basis for the parametrization of cross interactions in Martini. Ether and chloroform are represented by a single N0 and C4 bead, octanol is represented by one C1 and one P1 bead connected via a 0.47 nm bond with 1250 kJ mol^{-1} force constant. Using the values for ϵ in set $\mathbb{A}$ (see Table 8.3) liquid properties for those solvents are measured. Note that both the N0-N0, C1-C1, and C4-C4 interactions are level IV, and only the P1-C1 and P1-P1 interaction levels have been obtained by interpolation. Table 8.4 shows the obtained densities, compressibilities and surface tension ($\epsilon = 5.35$ kJ mol^{-1} is used for the interaction between two C1 beads in octanol).

There is no general trend for the change of density for the three liquids; the final densities are reasonably close the experimental values. The compressibilities are increased for all three solvents, thus worsening the comparison to experiment. However, as mentioned above, the compressibility is not the main target in this work. The surface tension of the three liquids increases, leading to a better match to experimental values as compared to the current Martini force field.

Table 8.4: Properties of polar liquids. Values for density, isothermal compressibility and air-liquid surface tension are reported for the new Mie 4-2 potential (Mie.), the Martini LJ 12-6 potential (Mar.) and experimental (Exp.). Experimental values have been obtained from [26]. **a** experiment at 20°**b** experiment at 0°

Molecule	**Density (gr dm^{-3})**			**Compressibility (10^{-5}bar^{-1})**			**Surface Tension (mN m^{-1})**		
	Mie.	Mar.	Exp.	Mie.	Mar.	Exp.	Mie.	Mar.	Exp.
Ether	805	899	713	63	32	19.70	15	12	16.4
Chloroform	1293	1445	1492	57	32	9.74	15	11	27.14[a]
Octanol	906	889	824	27	17	6.82[b]	32	19	27.53[a]

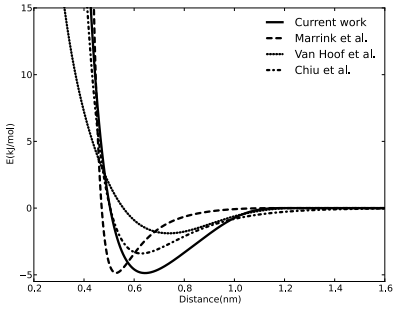

Figure 8.3: Comparison of the non-bonded potential shapes used for water. Solid line: Shifted Mie 4-2 potential obtained in this work. Dashed line: Shifted LJ12-6 potential used by the original Martini force field[5]. Dash-dotted line: Morse potential obtained by Chiu *et al.*[19]. Dotted line: Tabulated potential obtained by Van Hoof *et al.*[20].

Diffusion rates

As mentioned in chapter 1 of this thesis, the diffusion rate of molecules in coarse grain models is not well understood and hard to quantitatively predict. To investigate the effect of a softer potential on the diffusion, we calculated self-diffusion rates for different liquids using the interactions levels of set $\mathbb{A}$ defined in Table 8.3. The obtained rates are shown in Table 8.5. The diffusion rates obtained with the Mie 4-2 potential are in general higher than those obtained with Martini or the experimental values. The speed-up of Mie 4-2 compared to Martini ranges from a factor 2.5 (octanol) to 3.7 (hexadecane), a rather small range. Compared to the experimental values the speed-up factor ranges from 3.4 (octane) to 43.8 (octanol). This large difference is similar to that found for Martini (see chapter 1).

Table 8.5: Comparison of self diffusion rates of liquids (10^{-5} cm^2 s^{-1}) obtained for the 4-2 Mie potential (Mie), the current Martini model (Mar.) and Experiment (Exp.). Values are obtained at 298 K, unless mentioned otherwise. **a.** Value has been multiplied by 4 to account for the mapping of four water molecules to one bead. **b.** Values from [6]. Values were extrapolated from temperature-dependent data. **c.** Values from [30]. **d.** Values from [31].

Molecule	Mie	Mar.	Exp.
water	16[a]	5.6[a]	2.3[b]
octane	6.7	2.3	2[b]
hexadecane	2.6	0.7	-
chloroform	12.1	4.3	2.1[c]
ether	15.6	5.5	4.5[c]
octanol	5.7	2.3	0.13[d]

Thermodynamic properties

Tables 8.6 and 8.7 present the free energies of vaporization (ΔG^{vap}), hydration (ΔG^{hyd}), and partitioning (ΔG^{part}) between water (W) and apolar phases (H, hexadecane; C, chloroform; E, ether; O, octanol) obtained for the three sets of the interaction levels, $\mathbb{A}, \mathbb{B}$ and $\mathbb{C}$ defined in Table 8.3.

All three sets of parameters show considerably better agreement with the experimental free energies of vaporization and hydration than the original Martini model. The vaporization free energies are more negative for all bead types. Sets $\mathbb{B}$ and $\mathbb{C}$ perform slightly better for the Nda, Nd and Na particles. Set $\mathbb{B}$ is on average for those particles a better match (-18 kJ mol^{-1} against -22 kJ mol^{-1} for set $\mathbb{C}$). Set $\mathbb{C}$ gives slightly better values for the more apolar particles, although the difference between $\mathbb{B}$ and $\mathbb{C}$ is small (-15 versus -14 kJ mol^{-1}). Set $\mathbb{C}$ gives considerably lower values for the hydration free energies for Nda, Nd and Na particles, in better agreement with experimental values. Also on the apolar particles, set $\mathbb{C}$ has better agreement (lower energies) with experiment. Whereas the new potentials perform better for the vaporization and hydration free energies, they on average perform worse for the different

			ΔG^{vap}					ΔG^{hyd}					ΔG^{part}_{HW}				
type	building block	examples	Exp.	Mar.	$\mathbb{A}$	$\mathbb{B}$	$\mathbb{C}$	Exp.	Mar.	$\mathbb{A}$	$\mathbb{B}$	$\mathbb{C}$	Exp.	Mar.	$\mathbb{A}$	$\mathbb{B}$	C
Q_{da}	H_3N^+-C_2-OH	ethanolamine (protonated)							-25	-39	-39	-39		<-30	-59	-62	-65
Q_d	H_3N^+-C_3	1-propylamine (protonated)							-25	-39	-39	-39		<-30	-59	-62	-65
	Na^+-C_3	sodium (hydrated)							-25	-39	-39	-39		<-30	-59	-62	-65
Q_a	PO_4^-	phosphate							-25	-39	-39	-39		<-30	-59	-62	-65
	Cl^-HO	chloride (hydrated)							-25	-39	-39	-39		<-30	-59	-62	-65
Q_0	C_3N^+	choline							-25	-39	-39	-39		<-30	-59	-62	-65
P_5	H_2N-C_2=O	acetamide	sol	sol	-37	-37	-37	-40	-25	-39	-39	-39	-27	-28	-38	-40	-49
P_4	HOH (x4)	water	-27	-18	-33	-33	-33	-27	-18	-33	-33	-33	-25	-23	-32	-33	-43
	H_2N-C_2-OH	ethanediol	-35	-18	-33	-33	-33	-33	-18	-33	-33	-33	-21	-23	-32	-33	-43
P_3	H_2N-C_2=O	acetic acid	-31	-18	-33	-33	-33	-29	-18	-33	-33	-33	-19	-21	-31	-30	-35
	C-NH-C=O	methylformamide	-35	-18	-33	-33	-33		-18	-33	-33	-33		-21	-31	-30	-35
P_2	C_2-OH	ethanol	-22	-16	-19	-25	-25	-21	-14	-7	-15	-16	-13	-17	-5	-12	-18
P_1	C_3-OH	1-propanol	-23	-16	-19	-25	-25	-21	-14	-7	-15	-16	-9	-11	-4	-10	-8
		2-propanol	-22	-16	-19	-25	-25	-20	-14	-7	-15	-16	-10	-11	-4	-10	-8
N_{da}	C_4-OH	1-butanol	-25	-16	-14	-18	-25	-20	-9	2	-5	-13	-5	-7	5	0	-5
N_d	H_2N-C_3	1-propylamine	-17	-13	-14	-18	-22	-18	-9	2	-5	-13	(-6)	-7	5	0	-5
N_a	C_3=O	2-propanone	-17	-13	-14	-18	-22	-16	-9	2	-5	-13	-6	-7	5	0	-5
	C-NO_2	nitromethane	-23	-13	-14	-18	-22	-17	-9	2	-5	-13	-6	-7	5	0	-5
	C_3=N	proprionitrile	-22	-13	-14	-18	-22	-17	-9	2	-5	-13	-5	-7	5	0	-5
	C-O-C=O	methylformate	-16	-13	-14	-18	-22	-12	-9	2	-5	-13	(-6)	-7	5	0	-5
	C_2HC=O	propanal		-13	-14	-18	-22	-15	-9	2	-5	-13	-4	-7	5	0	-5
N_0	C-O-C_2	methoxyethane	-13	-10	-12	-14	-15	(-8)	-2	5	2	0	(1)	-2	8	8	8
C_5	C_3-SH	1-propanethiol	-17	-10	-12	-14	-15		1	7	4	2		5	13	13	12
	C-S-C_2	methyl ethyl sulfide	-17	-10	-12	-14	-15	-6	1	7	4	2	(7)	5	13	13	12
C_4	C_2=C_2	2-butyne	-15	-10	-12	-14	-15	-1	5	9	6	3		9	14	15	14
	C=C-C=C	1,3-butadiene		-10	-12	-14	-15	2	5	9	6	3	11	9	14	15	14
	C-X_4	chloroform	-18	-10	-12	-14	-15	-4	5	9	6	3	(7)	9	14	15	14
C_3	C_2=C_2	2-butene		-10	-12	-14	-15		5	9	6	3		13	19	18	16
C_3	C_3-X	1-chloropropane	-16	-10	-12	-14	-15	-1	5	9	6	3	12	13	19	18	16
		1-bromopropane	-16	-10	-12	-14	-15	-2	5	9	6	3		13	19	18	16
C_2	C_3	propane	gas	-10	-12	-14	-15	8	10	10	8	9		16	20	20	22
C_1	C_4	butane	-11[a]	-10	-12	-14	-15	9	14	10	10	13	18	18	20	22	26
		isopropane	gas	-10	-12	-14	-15	10	14	10	10	13		18	20	22	26

Table 8.6: Free energies of vaporization ΔG^{vap}, hydration ΔG^{hyd} and partitioning (ΔG^{part}) between water (W) and apolar phases (H, Hexadecane) are compared to experimental values. All experimental (Exp) values are used as compiled in [6]. Values for the original Martini force field (Mar.) are taken from [6]. $\mathbb{A}$, $\mathbb{B}$ and $\mathbb{C}$ indicate the values calculated here with the three sets of interactions levels. The experimental vaporization free energy was calculated from the vapor pressure p_{vap} using $\Delta G^{vap} = k_B T \ln(p_{vap}/k_B T C_M)$, where C_M denotes the molar concentration of the liquid . The experimental free energies of hydration were compiled from various sources based on log P values. The temperature for the experimental data is in the range 298-300 K, except where indicated. Simulation data were obtained at 298.15K. Experimental properties between parentheses are estimates obtained from comparison to similar compounds. The statistical accuracy of the free energies obtained from simulations is ± 1 kJ mol^{-1}. **a** The temperature for the experimental data is 273 K.

type	building block	examples	ΔG^{part}_{CW} Exp.	Mar.	$\mathbb{A}$	$\mathbb{B}$	$\mathbb{C}$	ΔG^{part}_{EW} Exp.	Mar.	$\mathbb{A}$	$\mathbb{B}$	$\mathbb{C}$	ΔG^{part}_{OW} Exp.	Mar.	$\mathbb{A}$	$\mathbb{B}$	$\mathbb{C}$
Q_{da}	H_3N^+-C_2-OH	ethanolamine (protonated)		-18	-33	-31	-29		-13	-27	-26	-24		-18	-13	-17	-18
Q_d	H_3N^+-C_3	1-propylamine (protonated)		-18	-33	-31	-29		-13	-27	-26	-24		-18	-13	-17	-18
	Na^+-C_3	sodium (hydrated)		-18	-33	-31	-29		-13	-27	-26	-24		-18	-13	-17	-18
Q_a	PO_4^-	phosphate		-18	-33	-31	-29		-13	-27	-26	-24		-18	-13	-17	-18
	Cl^-HO	chloride (hydrated)		-18	-33	-31	-29		-13	-27	-26	-24		-18	-13	-17	-18
Q_0	C_3N^+	choline		-18	-33	-31	-29		-13	-27	-26	-24		-18	-33	-33	-32
P_5	H_2N-C_2=O	acetamide	(-20)	-18	-33	-31	-29	-15	-13	-27	-26	-24	-8	-10	-1	-4	-4
P_4	HOH (x4)	water		-14	-27	-24	-22	-10	-7	-21	-19	-18	-8	-9	-15	-13	-15
	H_2N-C_2-OH	ethanediol		-14	-27	-24	-22		-7	-21	-19	-18	-8	-9	-15	-13	-15
P_3	H_2N-C_2=O	acetic acid	-9	-10	-25	-22	-20	-2	-6	-21	-19	-18	-1	-7	-15	-12	-14
	C-NH-C=O	methylformamide		-10	-25	-22	-20		-6	-21	-19	-18	-5	-7	-15	-12	-14
P_2	C_2-OH	ethanol	-5	-2	5	-2	-1	-3	1	9	8	13	-2	-2	11	-12	3
P_1	C_3-OH	1-propanol	-2	-2	5	-2	-1	0	1	9	8	13	1	-1	9	6	6
		2-propanol	-2	-2	5	-2	-1	-1	1	9	8	13	0	-1	9	6	6
N_{da}	C_4-OH	1-butanol	2	0	14	9	0	4	2	18	18	3	4	3	18	16	10
N_d	H_2N-C_3	1-propylamine	(1)	0	14	9	0	(-3)	2	18	18	3	(3)	3	18	16	10
N_a	C_3=O	2-propanone	1	0	14	9	0	-1	2	18	18	3	-1	3	18	16	10
	C-NO_2	nitromethane		0	14	9	0		2	18	18	3	-2	3	18	16	10
	C_3=N	proprionitrile		0	14	9	0		2	18	18	3	1	3	18	16	10
	C-O-C=O	methylformate	(4)	0	14	9	0	(-1)	2	18	18	3	(0)	3	18	16	10
	C_2HC=O	propanal		0	14	9	0	2	2	18	18	3	3	3	18	16	10
N_0	C-O-C_2	methoxyethane		6	17	16	15	(3)	6	21	25	15	(3)	5	15	16	18
C_5	C_3-SH	1-propanethiol		10	19	18	17		10	24	27	17		6	16	14	13
	C-S-C_2	methyl ethyl sulfide		10	19	18	17		10	24	27	17	(9)	6	16	14	13
C_4	C_2=C_2	2-butyne		13	21	20	19		13	25	29	19	9	9	18	16	14
	C=C-C=C	1,3-butadiene		13	21	20	19		13	25	29	19	11	9	18	16	14
	C-X_4	chloroform	14	13	21	20	19		13	25	29	19	11	9	18	16	14
C_3	C_2=C_2	2-butene		13	21	20	19		13	25	29	19	13	14	19	17	16
C_3	C_3-X	1-chloropropane		13	21	20	19		13	25	29	19	12	14	19	17	16
		1-bromopropane		13	21	20	19		13	25	29	19	12	14	19	17	16
C_2	C_3	propane		15	22	22	22		14	18	19	22	14	16	18	18	21
C_1	C_4	butane		18	22	24	26		14	16	18	24	16	17	17	16	24
		isopropane		18	22	24	26		14	16	18	24	16	17	17	16	24

Table 8.7: Free energies of partitioning (ΔG^{part}) between water (W) and apolar phases (C, chloroform; E, ether; O, octanol) are compared to experimental values. All experimental (Exp.) values are used as compiled in [6]. Values for the original Martini force field (Mar.) are taken from [6]. $\mathbb{A}$, $\mathbb{B}$, and $\mathbb{C}$ indicate the values calculated here with the three sets of interactions levels. The experimental free energies of partitioning were compiled form various sources based on log P values. The temperature for the experimental data is in the range 298-300 K, except where indicated. Simulation data were obtained at 298.15K. Experimental properties between parentheses are estimates obtained from comparison to similar compounds. The statistical accuracy of the free energies obtained from simulations is ± 1 kJ mol^{-1}.

partitioning free energies. Especially the polar (P) particles are too polar and apolar (C) particles are too apolar. For the intermediate (N) particles set $\mathbb{C}$ gives results comparable to the original Martini force field. The situation is most complicated for octanol, since it consists of both a polar and a apolar bead. This causes the large difference between the partitioning free energies of Q0 (-33 kJ mol^{-1}) and P5 (-4 kJ mol^{-1}). In octanol also the partitioning of N-type particles of set $\mathbb{C}$ is too high. For the more apolar (C-type) particles, the sets $\mathbb{A}$ and $\mathbb{B}$ perform better than set $\mathbb{C}$. The too negative free energy values for the polar beads are a direct result of the lower free energies of hydration of those beads. To correct for this, without changing the hydration free energy, the interaction energy of apolar bead with the apolar solvents should be increased compared to the current set $\mathbb{C}$.

8.4 Conclusions

The LJ 12-6 potential used for the popular Martini force field does not perform optimal in some aspects, among which the surface tension and melting temperature of water. In this work we explored the possibility to replace the current potential by a softer Mie potential. Potentials have been derived based on the liquid properties and RDFs of water and octane. Although none of the potentials perfectly reproduce the density, surface tension, compressibility and RDF of both water and octane, the overall properties do improve when compared to the current Martini potential. The first solvation shell in the RDF of water is much less pronounced with all of the tested Mie potentials, albeit still higher than the mapped atomistic RDF. The first solvation shell of atomistic octane is matched well by the softer potentials and is even too low for the Mie potential with exponents 3 and 2. Based on this observation the 4-2 potential was selected. Using this potential the optimal σ and ϵ for the water (P4 bead) self interaction are 0.455 nm and 12.3 kJ mol^{-1}, giving a density and surface tension within one percent of the experimental value. This same sigma is used for all bead types. Based on the comparison to the density of linear alkanes, ϵ for the C1 bead self interaction is adjusted to 5.35 kJ mol^{-1}.

Using the obtained sigma and epsilon values and three sets of estimates for the epsilon values for the remaining beads, hydration, solvation and partitioning free energies were calculated for all Martini bead types. Overall the hydration and vaporization free energies improve in comparison to the Martini force field, whereas the partitioning free energies worsen. Although none of the sets gave completely satisfying free energies, the values are reasonably close to experimental values and show variation between the sets. In other words, the two interaction levels obtained from the liquid properties of water and octane seem to leave enough room to tune the remaining interaction levels and bring the free energies in accordance with the experimental values.

Although the current results obtained with the Mie-potentials are promising there are many remaining issues. First, the difference between the RDF of water obtained with the new potential and mapped SPC water, is still considerable. Since a (global) softer potential worsens the properties of more apolar solvents, such as octane, it might be interesting to use a separate, possibly non-analytical, potential for the water self-interactions, while maintaining the current potential for other interactions. Second, although the much too low melting temperature of water will have no influence on (normal) simulations at room temperature, other phase transition temperatures, such as the fluid to gel transition of lipid bilayers, might be affected as well. Since this transition temperature is already quite low in Martini [32], this would be undesirable. Third, the proposed changes in non-bonded parameters might have an indirect influence on the bonded interactions. These last two points might require the re-parametrization of the Martini bonded interactions, a considerable amount of work.

Currently further efforts are being made to investigate the possibility of combining a tabulated water potential with the obtained Mie potentials, to find the optimal set of ϵ parameters and investigate the influence of the new potential on previously parameterized molecules, such as lipids and proteins.

8.5 Bibliography

[1] G.A Voth. *Coarse graining of condensed phase and biomolecular systems*. CRC Press, 2009.

[2] A Soper. Empirical potential monte carlo simulation of fluid structure. *Chem. Phys.*, 202:295–306, 1996.

[3] S Izvekov, M Parrinello, C.J Bernham, and G.A Voth. Effective force fields for condensed phase systems from ab initio molecular dynamics simulation: A new method for force-matching. *J. Chem. Phys.*, 120:1–19, 2004.

[4] W.G Noid, J.W Chu, G.S Ayton, V Krishna, S Izvekov, G.A Voth, A Das, and H.C Andersen. The multiscale coarse-graining method. i. a rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.*, 128:244114, 2008.

[5] S.J Marrink, A.H de Vries, and A.E Mark. Coarse grained model for semiquantitative lipid simulations. *J. Phys. Chem. B*, 108:750–760, 2004.

[6] S.J Marrink, H.J Risselada, S Yefimov, D.P Tieleman, and A.H de Vries. The martini force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 111:7812–7824, 2007.

[7] L Monticelli, S.K Kandasamy, X Periole, R.G Larson, D.P Tieleman, and S.J Marrink. The martini coarse-grained force field: extension to proteins. *J. Chem. Theory Comput.*, 4:819–834, 2008.

[8] W.F.D Bennet C Arnarez T.A Wassenaar L.V Schaäfer X. Periole D.P Tieleman S.J Marrink D.H de Jong, G Singh. Improved parameters for the martini coarse-grained protein force field. *submitted*, 2012.

[9] C.A López, A.J Rzepiela, A.H de Vries, L Dijkhuizen, P.H Hünenberger, and S.J Marrink. Martini coarse-grained force field: Extension to carbohydrates. *J. Chem. Theory Comput.*, 5:3195–3210, 2009.

[10] H Lee and R.G Larson. Molecular dynamics study of the structure and interparticle interactions of polyethylene glycol-conjugated pamam dendrimers. *J. Phys. Chem. B*, 113:13202–13207, 2009.

[11] G Rossi, I Giannakopoulos, L Monticelli, N.K.J Rostedt, S.R Puisto, C Lowe, A.C Taylor, I Vattulainen, and T Ala-Nissila. *Macromolecules, title = A MARTINI Coarse-Grained Model of a Thermoset Polyester Coating,*, 44:6198–6208, 2011.

[12] J.E Jones. On the determination of molecular fields. ii. from the equation of state of a gas. *P. Roy. Soc. A-Math. Phys.*, 106:463–477, 1924.

[13] G Mie. Zur kinetischen theorie der einatomigen körper. *Ann. Phys-Berlin*, 316:657–697, 1903.

[14] O Guvench and A.D MacKerell. Comparison of protein force fields for molecular dynamics simulations. *Method. Mol. Biol.*, 443:63–88, 2008.

[15] F London. The general theory of molecular forces. *Trans. Faraday Soc.*, 33:8–26, 1937.

[16] E Lindahl, B Hess, and D van der Spoel. Gromacs 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.*, 7:306–317, 2001.

[17] R Baron, D Trzesniak, A.H de Vries, A Elsener, S.J Marrink, and W.F van Gunsteren. Comparison of thermodynamic properties of coarsegrained and atomiclevel simulation models. *ChemPhysChem*, 8:452–461, 2007.

[18] M Fuhrmans, B.P Sanders, S.J Marrink, and A.H de Vries. Effects of bundling on the properties of the spc water model. *Theor. Chem. Acc.*, 125:335–344, 2010.

[19] S Chiu, H Scott, and E Jakobsson. A coarse-grained model based on morse potential for water and n-alkanes. *J. Chem. Theory Comput.*, 6:851–863, 2010.

[20] B van Hoof, A.J Markvoort, R.A van Santen, and P Hilbers. The cumulus coarse graining method: Transferable potentials for water and solutes. *J. Phys. Chem. B*, 115:10001i–10012, 2011.

[21] X He, W Shinoda, R DeVane, and M.L Klein. Exploring the utility of coarse-grained water models for computational studies of interfacial systems. *Mol. Phys.*, 108:2007–2020, 2010.

[22] P.M Morse. Diatomic molecules according to the wave mechanics. ii. vibrational levels. *Phys. Rev.*, 34:57–64, 1929.

[23] B Hess, C Kutzner, D van der Spoel, and E Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4:435–447, 2008.

[24] G Bussi, D Donadio, and M Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126:14101, 2007.

[25] M Parrinello and A Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52:7182–7190, 1981.

[26] R Weast. *CRC handbook of chemistry and physics*. CRC Press, 1986.

[27] L.D Schuler, X Daura, and W.F van Gunsteren. An improved gromos96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.*, 22:1205–1218, 2001.

[28] R.W Zwanzig. Hightemperature equation of state by a perturbation method. i. nonpolar gases. *J. Chem. Phys.*, 22:1420–1428, 1954.

[29] C.H Bennett. Efficient estimation of free energy differences from monte carlo data. *J. Comput. Phys.*, 22:245–268, 1976.

[30] S.A Sanni, C.J.D Fell, and H.P Hutchison. Diffusion coefficients and densities for binary organic liquid mixtures. *J. Chem. Eng. Data*, 16:424–427, 1971.

[31] L Guendouz, S Leclerc, A Retournard, A Hedjiedj, and D Canet. Single-sided radiofrequency field gradient with two unsymmetrical loops: Applications to nuclear magnetic resonance. *Rev. Sci. Instrum.*, 79:123704, 2008.

[32] S.J Marrink, H.J Risselada, and A.E Mark. Simulation of gel phase formation and melting in lipid bilayers using a coarse grained model. *Chem. Phys. Lipids*, 135:223–244, 2005.

Outlook

The three main focus areas of this thesis (testing of the Martini forcefield, improvements to the forcefield and application to biological problems.) offer plenty opportunities for further investigation.

Focus areas 1 & 2: Testing and improving

The study of interactions between single amino acids in the Martini forcefield (chapter 3), should be taken to a next level by looking at the interactions between whole proteins and peptides. Two approaches are conceivable here: first, determining the preferred binding interfaces of protein pairs for which a crystal structure of the complex is known [1] can elucidate to which extent Martini correctly models the interface between proteins. This work has recently been started in cooperation with Felipe Franco. Very interesting in this respect is to study the effect of the new protein parameters proposed in chapter 7 of this thesis.

Second, it should be tested if in CG models like the Martini forcefield the binding between (soluble) proteins in general is too strong, as has previously been hypothesized [2]. Comparing the binding free energy of a small peptide obtained using both atomistic forcefields and the Martini forcefield could give this insight. Since obtaining a converged free energy of binding between peptides from atomistic simulations is still a formidable task, it might be sufficient to restrict the orientations of the proteins and focus on a single interface.

Another informative test of the Martini protein forcefield concerns the simulation of unstructured peptides. Measuring the average radius of gyration of such a molecule and comparing these values to available experimental data allows for testing and improving the coil parameters. In addition to the rough global information of the radius of gyration, more specific data is available in the form of secondary structure propensities obtained from NMR experiments [3]. This test might not only give a better representation of coil regions in proteins but can even open the way to studying intrinsically disordered proteins [4].

Besides the developments mentioned above major improvements are possible by continuing to look for better CG potentials, as described in chapter 8. One promising option that deserves further investigation is the combination of (clustered water) structure based potentials [5] for the interactions between water and thermodynamics based potentials like the Martini forcefield for all other interactions. Optimally these two types of potentials can be combined and can additionally be combined with the existing bonded parameters that are readily available for the Martini forcefield. This would give the best of two types of CG forcefields, without the need of reparameterizing a complete forcefield.

The development and testing of this next generation Martini forcefield would greatly benefit from a comprehensive test set, comparing a spectrum of properties for systems con-

taining some of the many different molecules that have been parametrized in Martini. At a first glance, the set should contain: *i)* Vaporization, hydration and partitioning free energy calculations for all Martini beads. These would have to reproduce the free energies that were reported in the original Martini paper [6] and form the basis of the forcefield. *ii)* A set of pure liquids (e.g.: water, octanol, hexadecane, chloroform and ether), for which properties such as density, compressibility and self-diffusion rates should be calculated. For water and alkanes also the surface tensions (vapor/liquid and water/alkane surface tension is of importance. *iii)* Single component bilayers containing simple lipids (like DPPC, POPE) measuring properties like bilayer thickness, area per lipid, tail order parameters, and diffusion rates. *iv)* Ternary, phase-separating mixtures (e.g. DPPC:DLiPC:Cholesterol), for which system properties like the line tension should be calculated. *v)* A set of small systems to measure properties such as the binding between amino acids, the partitioning of amino acids over a bilayer and the Wimbley-White partitioning coefficient. *vi)* Protein fragments to study secondary structure characteristics such as helix length, pitch, and coil flexibility. *vii)* A set of proteins, simulated with and without elastic network, to study protein stability over time, measuring properties such as the RMSD, RMSF-profiles and principle components of the dynamics. *viii)* A set of miscellaneous (small) molecules, mainly testing the numerical stability of these molecules and possibly some molecule dependent properties, such as partitioning free energies, location in a membrane or protein binding.

Focus area 3: Applications

The applications of the Martini forcefield to the study of membrane and protein organization has proven successful, as described in the chapters 4 and 5. The investigation of the influence of gangliosides on the domain partitioning of peptides and proteins (chapter 5) is currently hampered by *i)* the relatively small size of the simulated systems and the resulting small liquid ordered and liquid disordered domains. And *ii)* the use of DLiPC lipids in stead of the DOPC lipids found *in vivo* and used in most experiments. The first problem is within reach of the available computational power today. With respect to the second issue, Cesar Lopez is currently developing new Martini parameters for DOPC, that will hopefully fix the lack of domain separation between fully saturated lipids and DOPC.

The addition of more components, like gangliosides, are a first step towards *in vivo* complexity. The real challenge lies in the simulation of a realistic plasma membrane, containing dozens of lipids types and possibly even more different proteins. Composing and setting up such a system will be greatly facilitated by automated tools like the *martinize.py* script described in chapter 7 and the *insane.py* script developed by Tsjerk Wassenaar. Some of the problems that will have to be solved before setting up such a system are the lack of Martini parameters for the glycosylphosphatidylinositol (GPI) anchor, a common way to anchor proteins to membranes, and the effect of the cytoskeleton on plasma membranes that is absent in our current membrane simulations.

The parameters for thylakoid cofactors, presented in chapter 6, in combination with existing Martini parameters for proteins and parameters for thylakoid lipids recently developed by Cesar Lopez [7], open the way for simulations of the photo system II in the thylakoid membrane. For this system many interesting questions are within reach of the Martini model. Examples are *i)* how plastoquinone leaves and enters the active site of the protein complex and how it behaves inside the membrane, *ii)* identification of specific lipid binding sites on the protein (similar to the work Clement Arnarez recently did on the respiratory chain protein complex [8]) and *iii)* how dynamic the different protein complexes in the thylakoid membrane are with respect to each other.

Bibliography

[1] H Hwang, T Vreven, J Janin, and Z Weng. Protein–protein docking benchmark version 4.0. *Proteins*, 78:3111–3114, 2010.

[2] K.M Ravikumar, W Huang, and S Yang. Coarse-grained simulations of protein-protein association: An energy landscape perspective. *Biophys. J.*, 103:837–845, 2012.

[3] K Tamiola and F.A.A Mulder. Using nmr chemical shifts to calculate the propensity for structural order and disorder in proteins. *Biochem. Soc. T.*, 40:1014–1020, 2012.

[4] A Ghavami, E van der Giessen, and P.R Onck. Coarse-grained potentials for local interactions in unfolded proteins. *J Chem Theory Comput*, 9:432–440, 2012.

[5] B van Hoof, A.J Markvoort, R.A van Santen, and P.A.J Hilbers. The cumulus coarse graining method: Transferable potentials for water and solutes. *J. Phys. Chem. B*, 115:10001–10012, 2011.

[6] S.J Marrink, H.J Risselada, S Yefimov, D.P Tieleman, and A.H de Vries. The martini force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 111:7812–7824, 2007.

[7] C.A Lopez, Z Sovova, F.J van Eerden, A.H de Vries, and S.J Marrink. Martini force field parameters for glycolipids. *J Chem Theory Comput*, 9, 2013.

[8] C Arnarez, X Periole, and S.J Marrink. Accepted. *Scientific Reports*, 2013.

Summary

The organizational principles of life are among the great, unsolved mysteries of today's science: How do small, simple molecules such as water, lipids and proteins organize to form enormously complex organisms, like bacteria, plants or even humans? In this thesis we studied this complexity by means of computer simulations, a well known but slightly unorthodox technique to study biological problems. In our simulations we see the self-assembly of lipids into membranes, the formation of domains within these membranes and, for the first time, the recruitment of proteins into one of those domains. This domain, known as the "liquid ordered" or "raft" domain, has been hypothesized to play an important role in many processes in the cell such as signaling.

The length and time scales of these organizational processes (tens of nanometers and microseconds) are outside the reach of most commonly used computer models. Therefore we used a computer model, or "forcefield", that has been recently developed in Groningen, dubbed the "Martini" forcefield. This forcefield combines the required chemical specificity with computational efficiency by grouping together multiple atoms in super-atoms. This so called 'coarse graining' procedure reduces the computational costs of the model by a factor 1000 as compared to atomistic models.

The novelty of this approach requires that the forcefield is constantly being tested and improved. Therefore, in addition to the applications, this thesis also describes the testing of the forcefield and the implementation of corrections and new developments. The thesis is separated in four parts.

In the introductory chapter 1 an historical overview of molecular simulation techniques is given and the inner workings of the Martini forcefield are introduced.

The second part, consisting of chapters 2 and 3, describes the critical testing of the Martini protein forcefield. The binding constants of amino acid sidechains are compared to those obtained with atomistic forcefields and to binding constants obtained from crystal structures.

The third part describes the application to the study of the organizational principles of the cell membrane. In chapter 4 we identify the driving forces of partitioning of small peptides in simple but domain separated bilayers. In chapter 5 we study more complex bilayer mixtures and find that lipid anchors attached to proteins influence the partitioning behavior of these proteins. When a small percentage of special lipids, so called gangliosides, are added, the partitioning is even inverted: peptides that initially partitioned to the liquid disordered bilayer phase, are now being dragged into the liquid ordered phase.

The final three chapters describe new developments of the Martini forcefield. Chapter 7 details the new parameters for the current protein forcefield. Special topologies are developed to improve the behavior of polar and charged amino acids and a script to easily setup a Martini protein system is described. In chapter 6 parameters are developed for the cofactors

of Photosystem II. These molecules play an important role in the conversion of sunlight into chemical energy, taking place in the thylakoid membrane of plants ans cyano-bacteria. The new parameters allow for the study of this system with the Martini model. Finally, chapter 8 describes the search for a better coarse grain potential. This potential will allow future coarse grain forcefields to even more accurately describe biological molecules and enable the study of an even wider range of systems with the Martini forcefield.

Samenvatting

De manier waarop leven zich organiseert, behoort tot de grote, onopgeloste mysteries van de hedendaagse wetenschap: Hoe kunnen kleine, simpele moleculen zoals water, lipiden en eiwitten samenkomen om enorm complexe organismen, zoals bacteriën, planten of zelfs mensen te vormen? In dit proefschrift bestuderen we deze complexiteit door middel van computer simulaties; een beproefde maar enigzins onorthodoxe methode om biologische vraagstukken te bestuderen. In onze simulaties zien we de *self-assembly* van lipiden tot membranen, het vormen van domeinen binnen deze membranen en, voor de eerste keer, het verzamelen van eiwitten in een van deze domeinen. Van dit domein, ook wel het 'vloeibaar geordende' of *'raft'* domein genoemd, wordt gedacht dat het een belangrijke rol speelt bij een verscheidenheid aan processen binnen de cel, zoals het doorgeven van signalen.

De afstanden en tijdspanne waarop deze organisatieprocessen zich afspelen (tientallen nanometers en microseconden) vallen buiten het bereik van gebruikelijke computer modellen (de zogenaamde 'krachtenvelden'). Daarom gebruiken we het recentelijk in Groningen ontwikkelde 'Martini'-krachtenveld. Dit krachtenveld combineert de benodigde chemische specificiteit met computationele efficientie door het samenvoegen van meerdere atomen in superatomen. Deze *'coarse graining'* methode reduceert de computationele kosten met een factor 1000 in vergelijking met atomistische modellen.

Het gebruik van deze *state-of-the-art* methode zorgt er ook voor dat het krachtenveld aan voortdurende testen en verbeteringen onderhevig is. Daarom beschrijft dit proefschrift naast de eerder beshreven toepassingen ook het testen van het krachtenveld en de implementatie van correcties en nieuwe ontwikkelingen. Het proefschrift bestaat dan ook uit vier gedeelten.

In het introductie hoofdstuk 1 wordt een historisch overzicht gegeven van moleculaire simulatie technieken en wordt de werking van het Martini krachtenveld uitgelegd.

In het tweede gedeelte, de hoofdstukken 2 en 3, wordt het kritisch testen van het Martini eiwitkrachtenveld beschreven. De bindingsconstanten van aminozuur zijketens worden vergeleken met atomistische krachtenvelden en met bindingsconstanten verkregen uit kristal structuren.

Het derde gedeelte beschrijft de toepassing van het Martini krachtenveld bij het bestuderen van de manier waarop de celmembraan zich organiseert. In hoofdstuk 4 ontrafelen we de drijvende krachten achter de segregatie van kleine peptiden in simpele, maar in domeinen gescheiden bilagen. In hoofdstuk 5 bestuderen we complexere bilaag mengsels en zien we dat het bevestigen van lipide ankers aan eiwitten segregatie gedrag van die eiwitten beïnvloed. Wanneer een klein percentage gangliosiden, een speciaal soort lipiden, wordt toegevoegd, inverteert het segregatie gedrag zelfs; peptiden die zich eerst in het 'vloeibaar ongeordende' domein bevonden, worden nu het 'vloeibaar geordende' domein ingetrokken.

De laatste drie hoofdstukken beschrijven nieuwe ontwikkelingen met betrekking tot het Martini krachtenveld. Hoofdstuk 7 beschrijft nieuwe parameters voor het huidige eiwitkracht-

enveld. Speciale topologieën worden ontwikkeld voor polaire en geladen aminozuren en er wordt een script beschreven om gemakkelijk een Martini eiwit systeem op te zetten. In hoofdstuk 6 worden parameters ontwikkeld voor de cofactoren van het fotosysteem II. Deze moleculen spelen een belangrijke rol bij het omzetten van zonlicht in chemische energie zoals die plaatsvindt in de thylakoïde membraan van planten en blauwalgen. De nieuw ontwikkelde parameters maken het mogelijk om deze systemen met het Martini krachtenveld te bestuderen. Tot slot beschrijft hoofdstuk 8 de zoektocht naar een betere *coarse grain* potentiaal. Dankzij een dergelijk potentiaal kunnen toekomstige *coarse grain* krachtenvelden biologische moleculen nog beter beschrijven en kan een nog groter scala aan systemen worden bestudeerd met het Martini krachtenveld.

List of Publications

1. Norbert Kucerka, Drew Marquardt, Thad A. Harroun, Mu-Ping Nieh, Stephen R. Wassall, Djurre H. de Jong, Lars V. Schäfer, Siewert J. Marrink, and John Katsaras. *Cholesterol in Bilayers with PUFA Chains: Doping with DMPC or POPC Results in Sterol Reorientation and Membrane-Domain Formation*, Biochemistry, **2010**, *49*, 7485-7493
2. Lars V. Schäfer, Djurre H. de Jong, Andrea Holt, Andrzej J. Rzepiela, Alex H. de Vries, Bert Poolman, J. Antoinette Killian, and Siewert J. Marrink. *Lipid packing drives the segregation of transmembrane helices into disordered lipid domains in model membranes*, Proceedings of the National Academy of Science, **2011**, *108*(4), 1343-1348
3. Djurre H. de Jong, Lars V. Schäfer, Alex H. De Vries, Siewert J. Marrink, Herman J.C. Berendsen, Helmut Grubmüller. *Determining Equilibrium Constants for Dimerization Reactions from Molecular Dynamics Simulations*, Journal of Computational Chemistry, **2011**, *32* (9), 1919-1928
4. Djurre H. de Jong, Xavier Periole, and Siewert J. Marrink. *Dimerization of Amino Acid Side Chains: Lessons from the Comparison of Different Force Fields*, Journal of Chemical Theory and Computation, **2012**, *8*, 1003–1014
5. Mingyang Hu, Djurre H. de Jong, Siewert J. Marrink, and Markus Deserno. *Gaussian curvature elasticity determined from global shape transformations and local stress distributions: A comparative study using the MARTINI model*, Faraday Discussions, **2012**, *161*, 333–345, **DOI:**10.1039/C2FD20087B
6. Djurre H. de Jong, Cesar A. Lopez, and Siewert J. Marrink. *Molecular view on protein sorting into liquid- ordered membrane domains mediated by gangliosides and lipid anchors*, Faraday Discussions, **2013**, *161*, 347–363, **DOI:**10.1039/C2FD20086D
7. Djurre H. de Jong, Gurpreet Singh, W.F. Drew Bennett, Clement Arnarez, Tsjerk A. Wassenaar, Lars V. Schäfer, Xavier Periole, D. Peter Tieleman, and Siewert J. Marrink, *Improved Parameters For The Martini Coarse-Grained Protein Force Field*, Journal of Chemical Theory and Computation, **2013**, *9*, 687-697
8. Djurre H. de Jong, Tom van der Berg, Clement Arnarez, Xavier Periole, and Siewert J. Marrink, *Parameterization of thylakoid cofactors at atomistic and coarse grain resolution*, *In preparation*
9. Helgi I. Ingolfsson, Cesar A. Lopez, Jaakko J. Uusitalo, Djurre H. de Jong, Srinivasa M. Gopal, Xavier Periole, and Siewert J. Marrink *The Power of Coarse-Graining in Biomolecular Simulations*, WIREs Computational Molecular Science, **2013**, *In Press*