Diss. ETH No. 13686

Computer Simulation of Biomolecules: Investigation of Molecular Environment and Simulation Parameters

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZURICH

for the degree of
Doctor of Natural Sciences

presented by
URS STOCKER
Dipl. Chem. ETH

born June 11, 1971
citizen of Neftenbach and Schönenberg, Switzerland

accepted on the recommendation of
Prof. Dr. Wilfred F. van Gunsteren, examiner
Prof. Dr. Herman J. C. Berendsen, Prof. Dr. Philippe H. Hünenberger,
co-examiners

2000

Seite Leer /
Blank leaf

Für meine Mutter

und meinen Vater

Seite Leer /
Blank leaf

Acknowledgments

I wish to especially thank Wilfred van Gunsteren for giving me the opportunity to work in the *Gruppe für informatikgestützte Chemie (IGC)*. With his deep knowledge and unsurpassable optimism, Wilfred not only introduced me into the world of computer science and statistical dynamics, also the non-scientific discussions at the several informal occasions broadened my view of the world.

I am very grateful to Philippe H. Hünenberger. He was always willing to help with problems and answer every question one could imagine. We had a great time inside and outside the laboratory.

I thank Prof. Herman J. C. Berendsen, one of the fathers of molecular simulation, for his willingness to serve as a co-referent.

I am grateful to Katrin Spiegel, Tomas Hansson, and Daniel Juchli for their contributions to chapters 4, 5, and 6.

Martin Borer, I would like to thank for the administration and the smooth running of the C4 cluster, on which many of the calculations presented in this study were done.

My thanks go also to all the people I met in IGC and I got friends with, especially Heiko, Regula, Hubi, the members of the *independent C62istan*, Alan, Christian, Christine, Dirk, Ilario, Lukas, Roland, Salomon, Tomas, and Xavier for the interesting time spent in the group, but also in the several pubs, chatting, drinking, and singing.

Special thanks go to my friends inside and outside the ETH who accompanied me through my life: Adi, Armin, Ecki, Fränzi, Jeami, Küb, Mägi, Max, Niggi, Räff, Ruedi, Scherro, Schwank, Silver, Stineli, Terr, Tim, Töme, and Zufi. Without them, life would not have been the same.

Seite Leer /
Blank leaf

Contents

Seite Leer /
Blank leaf

Kurzfassung

Moleküldynamiksimulationen haben in den letzten Jahrzehnten einen immer grösseren Stellenwert in Biochemie und Biophysik erreicht. Allerdings werden, wenn klassisch simuliert wird, mehrere Annahmen und Näherungen getroffen. Das Ziel dieser Arbeit ist es, die Auswirkungen einger dieser Näherungen genauer zu untersuchen, wozu im Speziellen die Umgebung des simulierten Moleküls und der Einfluss der verschiedenen Simulationsparameter untersucht wurden.

Kapitel 1 enthält ein kurze Einführung in Moleküldynamiksimulationen. Die zugrundeliegende Theorie und das Kraftfeld, welches die Wechselwirkungen zwischen den Atomen beschreibt werden erklärt. Im letzten Abschnitt werden einige häufige Einsatzmöglichkeiten beschrieben.

In Kapitel 2 wird die Konvergenz als Funktion der Zeit verschiedener Moleküleigenschaften untersucht. Als Testsystem wurde der Kristall des Proteins Ubiquitin für 2ns simuliert. Die Gesamtenergie, die Abweichungen der simulierten Struktur von der experimentell bestimmten Kristallstruktur, die Fluktuationen der Atome um ihre mittlere Position und die Fluktuationen der Diederwinkel wurden untersucht. Die meisten dieser Eigenschaften konvergieren in der kurzen simulierten Zeit, nur die strukturelle Abweichung von der Kristallstruktur hat nach 2ns noch keinen Endwert erreicht.

In Kapitel 3 werden die Auswirkungen von Kraftfeldparameteränderungen auf ein simuliertes Protein untersucht. Dazu wurde das Protein *hen egg white lysozyme* (HEWL) mit dem GROMOS96 Kraftfeld für 2ns simuliert. Die Konvergenz verschiedener Eigenschaften wurde beurteilt und die konvergierten Werte dann mit experimentell bestimmten Werten verglichen. Es wurden Daten von Röntgenstrukturanalyse- und NMR-Experimenten verwendet: NOE Atom-Atom Distanzgrenzen, $^3J_{HN\alpha}$-Kopplungskonstanten und Ordnungsparameter von NH Bindungsvektoren in Haupt- und Seitenketten. Die Werte, die mit dem GROMOS96 Kraftfeld errechnet wurden, wurden zusätzlich mit Werten aus einer früheren Simulation verglichen. Die Simulation mit dem GROMOS96 Kraftfeld stimmt besser mit den experimentellen NMR Daten und auch mit der Kristallstruktur überein als die frühere Simulation, wo eine verbesserte Version des GROMOS87 Kraftfeldes eingesetzt wurde.

Proteinkristalle haben, verglichen mit Kristallen kleinerer Moleküle, einen hohen Wassergehalt. Darum wird allgemein angenommen, dass sich Proteinmoleküle in Kristall und Lösung sehr ähnlich verhalten. In Kapitel 4 werden die Auswirkung der zwei verschiedenen Umgebungen auf das Protein *hen egg white lysozyme* (HEWL) untersucht. Zwei Simulationen von HEWL, beide von 2ns Länge, eine davon im Kristall, die andere in wässriger Lösung, werden miteinander verglichen. Die Resultate aus beiden Simulationen werden experimentell bestimmten Werten gegenübergestellt. Es wurden Daten von Röntgenstrukturanalyse- und NMR-Experimenten verwendet: kristallographische Temperaturfaktoren, NOE Atom-Atom Distanzgrenzen, $^3J_{HN\alpha}$-Kopplungskonstanten und Ordnungsparamter von NH Bindungsvektoren in Haupt- und Seitenketten. Die Resultate der beiden Simulationen sind sehr ähnlich. Die Kristallsimulation reproduziert allerdings Kristall- und NMR-Daten etwas besser.

In Kapitel 5 werden 5 MD-Simulationen dazu verwendet, die Auswirkung der molekularen Umgebung auf das Verhalten des Proteins Ubiquitin zu bestimmen. Da Ubiquitin nur leicht geladen ist, sollte die Gegenwart von Gegenionen keinen grosse Effekt haben. Zwei der Simulationen wurden mit Gegenionen durchgeführt, drei ohne. Wenn man in Simulationen Ladungen

verwendet, sollte man darauf achten, dass die Auswirkungen, die dadurch entstehen, dass man die Wechselwirkungen nur bis zu einem bestimmten Abschneideparameter berechnet, minimiert werden. In vier der vorliegenden Simulationen wurde dafür der Reaktionsfeldansatz benutzt, in einer der Kristallsimulationen wurden die Wechselwirkungen einfach abgeschnitten. Da die Ladungsdichte in Kristallen viel höher ist als in Lösungen, wo auch mehr Wasser vorhanden ist, welches die Ladungen abschirmt, sollten Kristallsimulationen empfindlicher auf Gegenionen und auch auf die Gegenwart eines Reaktionsfeldes sein.

In biomolekularen Simulationen werden die verschiedenen erscheinenden Frequenzen von Kraftfeldtermen bestimmt. Eine Übersicht dieser Frequenzen wird in Kapitel 6 gegeben. Dadurch, dass man die höchsten Frequenzen ausschaltet, kann man in Gleichgewichts- und Verfeinerungssimulationen einen längeren Zeitschritt erreichen, muss allerdings eine Einbusse in der physikalischen Bedeutung in Kauf nehmen. Die verschiedenen Kraftfeldterme, die man in molekularen Simulationen einsetzt, können geschwächt oder die korrespondierenden Freiheitsgrade festgehalten werden. Dadurch kann ein grösstmöglicher Zeitschritt von 5fs für eine Gleichgewichtssimulation erreicht werden, in der statische Gleichgewichtseigenschaften, im Gegensatz zu dynamischen Eigenschaften, nicht beeinflusst werden. Der längste Zeitschritt, der mit einem sehr weichen Kraftfeld und erhöhten Wasserstoffmassen erreicht werden kann, ist 14fs. Dieser Parametersatz sollte allerdings nur in Verfeinerungssimulationen eingesetzt werden, wo die unphysikalischen Auswirkungen des Kraftfelds von untergeordneter Bedeutung sind.

Der Gebrauch anisotroper Druckkopplung erlaubt eine Deformation der simulierten Zelle. Die Auswirkungen anisotroper Druckkopplung auf ein kleines System mit vielen Ladungen wird in Kapitel 7 untersucht. Endliche und unendliche periodische Randbedingungen werden eingesetzt. 27 gleich geladene Partikel (mit verschiedenen Ladungen) wurden in Wasser und Argon simuliert, um die abschirmenden Effekte der unterschiedlichen dielektrischen Durchlässigkeiten zu untersuchen. Die Startzellen wurden in den meisten Simulationen deformiert. In den verdünntesten Systemen, 27 Partikel mit einer Ladung von $1e$ und den zwei Simulationen bei höherer Temperatur, 400K, blieb die Startzelle erhalten. Bei den Simulationen in Argon mit endlichen periodischen Randbedingungen wurde eine sehr langsame Konvergenz beobachtet, die Simulationen in Wasser fanden die optimale Form der Simulationszelle schneller. Die Simulation mit unendlichen periodischen Randbedingungen, wo auch eine andere optimale Form gefunden wurde, konvergierten noch schneller.

Summary

Molecular dynamics simulations have become more and more important during the last few decades. However, within the framework of classical dynamics, several approximations are made. The studies presented in this thesis aim at a deeper understanding of the effects of some of the various approximations made. Special emphasis was put on molecular environment and simulation parameters.

Chapter 1 gives a brief introduction into molecular dynamics. The basic theory is described together with the functional form of the interaction function used. The last paragraph describes some common applications of molecular dynamics.

In chapter 2, the degree of convergence of different molecular properties is examined. As a test system, the protein ubiquitin was simulated for 2ns in crystalline environment. The properties examined were energies, deviation from the experimentally derived crystal structure, atom-positional fluctuations and dihedral angle fluctuations. Most properties turn out to be converged on the short time-scale, 2ns, simulated. The atom-positional deviation from the X-ray structure, on the other hand, did not converge within 2ns.

In chapter 3, the effect of parameter changes in a molecular force field is examined. First, the degree of convergence of different molecular properties is investigated using a simulation trajectory of a 2ns simulation. The protein hen egg white lysozyme (HEWL) was simulated using the GROMOS96 force field. Converged properties were compared to values of quantities derived from X-ray and NMR experiments: NOE atom-atom distance bounds, $^3J_{HN\alpha}$-coupling constants and backbone and side-chain order parameters. In addition, quantities calculated from the GROMOS96 trajectory were compared to values obtained from an earlier simulation. The GROMOS96 simulation shows better agreement with the NMR data and also with the X-ray crystal structure of HEWL than a previous simulation, which was based on an earlier version of the GROMOS force field.

As protein crystals generally possess a high water content, it is assumed that the behaviour of a protein in solution and in crystal environment is very similar. In chapter 4, the effects of the different molecular environments on the protein hen egg white lysozyme (HEWL) is examined. Two 2ns simulations of HEWL, one in crystalline form and the other in solution, are compared to one another and to experimental data derived from both, X-ray and NMR experiments, such as crystallographic B-factors, NOE atom-atom distance bounds, $^3J_{HN\alpha}$-coupling constants, and ^{1}H-^{15}N bond vector order parameters. Both MD simulations give very similar results. The crystal simulation reproduces X-ray and NMR data slightly better than the solution simulation.

In chapter 5, five MD simulations of the protein ubiquitin are analysed to determine the effects of the molecular environment on the behaviour of a protein. Since the protein is only slightly charged, the presence of counterions is not expected to have a big effect. Two of the simulations were carried out with counterions, the other three without. Using charged species, one should pay attention to effects induced by the use of a cut-off radius when evaluating non-bonded interactions. In four of the simulations, the reaction field approach was used, one of the crystals was simulated without reaction field, truncating interactions at the cut-off radius. Since in a crystal, the charge density is much higher than in solution and there is less solvent present to screen these charges, simulations in crystal are expected to be more sensitive to the presence of counterions and to inclusion of reaction-field forces than simulations in aqueous solution.

In biomolecular simulation, the different force field terms determine the frequencies found in a particular simulation. The spread of frequencies is reviewed in chapter 6. By eliminating the high frequencies, longer time steps can be achieved in equilibrium and refinement (non-equilibrium) simulations at the expense of decreasing physical correctness of the simulated system. The different force field terms used in molecular simulations can be softened or the corresponding degrees of freedom can be rigidified leading to a maximum time step of 5fs for a simulation in which static equilibrium properties are not affected by the changes in the force field parameters introduced, but dynamical quantities are. Using a very smoothed, soft interaction function and increased hydrogen masses, a time step of 14fs can be reached. However, it should only be used in refinement simulations, in which the unphysical properties of the force field and dynamics are of minor importance.

Use of anisotropic pressure coupling allows for a deformation of the simulated system. The effect of anisotropic pressure coupling on small systems containing highly charged particles is analysed in chapter 7. Finite and infinite periodic boundary conditions were used. 27 identically charged particles were simulated in water and argon using various charges per particle to investigate the dampening effect of the different dielectric permittivities. A deformation of the initial box is observed in almost all the simulations. Only the most dilute systems, 27 particles with charges of $+1e$ in water, and the two simulations at 400K did not deform. The simulations in argon using finite periodic boundary conditions, converge very slowly. The simulations in water need shorter time to find the optimal box shape. The simulations using infinite periodic boundary conditions again converge quicker and lead to a different box shape than the simulations using finite periodic boundary conditions.

Publications

This thesis led to the following publications:

Chapter 2:

Urs Stocker and Wilfred F. van Gunsteren,
"Molecular Dynamics Simulation of Protein Crystals: Convergence of Molecular Properties of Ubiquitin"
Intl. Tables for Crystallography, Volume F (2000), in press.

Chapter 3:

Urs Stocker and Wilfred F. van Gunsteren,
"Molecular Dynamics Simulation of Hen Egg White Lysozyme: a Test of the GROMOS96 Force Field against NMR data"
Proteins, **40**, (2000), 145-153

Chapter 4:

Urs Stocker, Katrin Spiegel, and Wilfred F. van Gunsteren,
"On the similarity of properties in solution or crystalline state: a molecular dynamics study of Hen Lysozyme"
Journal of Biomolecular NMR (2000), in press.

Chapter 5:

Urs Stocker, Tomas Hansson, and Wilfred F. van Gunsteren,
"Effect of boundary conditions and treatment of electrostatics: crystal and solution molecular dynamics of ubiquitin"
for submission to *European Biophysical Journal*

Chapter 6:

Urs Stocker, Daniel Juchli, and Wilfred F. van Gunsteren,
"Increasing the time step and efficiency of molecular dynamics simulations: optimal solutions for equilibrium simulations or structure refinement of large biomolecules"
submitted to *Journal of Biomolecular NMR*

Chapter 7:

Urs Stocker, Philippe H. Hünenberger and Wilfred F. van Gunsteren,
"Anisotropic pressure coupling applied to small, highly charged systems"
for submission to *Molecular Simulation*

Chapter 1

Introduction

1.1 Computational chemistry

> *"Every attempt to employ mathematical methods in the study of chemical questions must be considered profoundly irrational and contrary to the spirit of chemistry. If mathematical analysis should ever hold a prominent place in chemistry - an aberration which is happily almost impossible - it would occasion a rapid and widespread degeneration of that science."*
> A. Comte, 1830, in "Philosophie Positive"

Although, like Comte feared already 170 years ago, mathematics and with it computation have become more and more important in almost all the fields of chemistry, one cannot say that this development did induce the decline of chemistry. Actually, the opposite is true. Theory and computation have been and are still a perfect complement to experiments and both techniques benefit from each other pushing chemistry to an ever increasing importance in everybody's life.

Computational chemistry knows different levels of theory or abstraction which all have their particular advantages and limitations. Quantum-chemical calculation, describing the highest (most fundamental) level of theory includes explicit treatment of (valence) electrons. Quantum-chemical calculations can be used to describe spectral properties of atoms and molecules, or chemical reactions. As chemistry is the science of matter, its properties, and transformation of matter, quantum chemistry theoretically describes "real" chemistry. Unfortunately its predictions are computationally very demanding. With increasing number of atoms, and thus electrons considered, the computational effort required increases much faster than the number of particles considered, in a range of N^3 to N^7 (where N is the number of particles), depending on the level of theory used. So, only properties of very small systems can be calculated accurately and even with the fast increase of computational power, large systems will always be out of reach of high-level quantum-chemical calculations. So, in order to be able to treat large systems, one has to make further approximations. In molecular dynamics (MD) [1] one does not treat electrons separately, but together with the corresponding nuclei. This leads to point charges moving in the field of the rest of the simulated system. This neglect of electronic degrees of freedom renders it impossible to follow chemical reactions, but allows for much larger systems to be examined. Due to the slower motions of nuclei compared to electrons, the time-scales accessible to these classical simulations is orders of magnitude longer. Within the time scales accessible to

molecular dynamics, some system properties reach a dynamical equilibrium and can be related to macroscopic properties using the laws of statistical mechanics.

1.2 Theory of molecular dynamics

1.2.1 Equations of motion

In classical molecular dynamics, atoms are basically treated as point charges having a mass. These atoms move in the field originating from all the other atoms in the system according to Newton's equations of motion.

$$\vec{f}_i = m_i \frac{d\vec{v}}{dt} \tag{1.1}$$

with

$$\vec{f}_i = -\frac{\partial}{\partial \vec{r}_i} V(\vec{r}), \tag{1.2}$$

where $\vec{f}_i$ is the force acting on particle i with mass m_i, velocity $\vec{v}_i$ and position $\vec{r}_i$. $V(\vec{r})$ is the potential energy of the system which depends on the coordinates $\vec{r}$ of all particles of the system and on the force field used.

1.2.2 The force field

The empirical force field used in molecular dynamics can be split into two parts. In the following equations, the dependence of the potential energy on the coordinates $\vec{r}$ and the force field parameters s is included.

$$V^{phys}(\vec{r};s) = V^{b}(\vec{r};s) + V^{n}(\vec{r};s) \tag{1.3}$$

where V^{phys} is the physical interaction function used and the bonded, V^{b}, and the non-bonded, V^{n}, parts of it can be further subdivided.

$$V^{b}(\vec{r};s) = V^{bonds}(\vec{r};s) + V^{angles}(\vec{r};s) + V^{imp}(\vec{r};s) + V^{tors}(\vec{r};s). \tag{1.4}$$

In the GROMOS96 [2] force field used during this study, the different bonded terms have the following functional forms.

$$V^{bond}(\vec{r};s) = \sum_{n=1}^{N_b} \frac{1}{4} K_{b_n} [b_n^2 - b_{0_n}^2]^2 \tag{1.5}$$

with K_{b_n} the force constant of bond n, b_n and b_{0_n} the actual and ideal bond length, respectively, of all N_b bonds of the simulated system.

$$V^{angles}(\vec{r};s) = \sum_{n=1}^{N_\theta} \frac{1}{2} K_{\theta_n} [cos\theta_n - cos\theta_{0_n}]^2 \tag{1.6}$$

where K_{θ_n} is the force constant of angle n, θ_n and θ_{0_n} are the actual and ideal bond angle, respectively, of all N_θ angles of the system.

$$V^{imp}(\vec{r};s) = \sum_{n=1}^{N_\xi} \frac{1}{2} K_{\xi_n} [\xi_n - \xi_{0_n}]^2 \tag{1.7}$$

with K_{ξ_n} the force constant of improper dihedral n, ξ_n and ξ_{0_n} the actual and ideal value for improper dihedral n, respectively, of all N_ξ improper dihedrals of the simulated system. Improper dihedrals are used in molecular dynamics to restrain planar groups to planarity, since the other forces do not fulfill this task sufficiently. In addition, improper dihedrals are used to impose the correct chirality and tetrahedral geometry of centers consisting of a united CH atom with three ligands in the simulation.

$$V^{tors}(\vec{r};s) = \sum_{n=1}^{N_\varphi} K_{\varphi_n}[1 + cos(\delta_n)cos(m_n\varphi_n)] \tag{1.8}$$

in which δ_n is a phase shift (0 or 180°) and m_n the multiplicity of the selected torsion (or proper dihedral) angle n with force constant K_{φ_n} and value φ_n.

The non-bonded interaction consists of different terms as well. The electrostatic part is treated as follows.

$$V^C(\vec{r};s) = \sum_{i=1}^{N}\sum_{j<i}^{N} \frac{1}{4\pi\varepsilon_0\varepsilon_1}\frac{q_i q_j}{r_{ij}} \tag{1.9}$$

The summation runs over all pairs of atoms, q_i and q_j are the charges of atoms i and j respectively, and r_{ij} is the distance between the two atoms. The Lennard-Jones interaction part is treated as follows.

$$V^{LJ}(\vec{r};s) = \sum_{i=1}^{N}\sum_{j<i}^{N}\left[\frac{C_{12}(i,j)}{r_{ij}^6} - C_6(i,j)\right]\frac{1}{r_{ij}^6} \tag{1.10}$$

The summation runs again over all pairs of atoms, C_{12} and C_6 are the repulsive and attractive parts of the Lennard-Jones potential energy term which depend on the selected atom pair and on the type (purely van der Waals or partly electrostatic) of the selected interaction.

The parameters used in the force field, force constants, ideal values for the various quantities, charges, and van der Waals parameters are either derived from theoretical (quantum-chemically computed) and experimental values, or fitted such that macroscopic quantities of a simulated system agree with experimentally measured values for these quantities, such as the heat of vaporization, mixing enthalpy, or density. All the summations in the presented terms in principle run over all of the selected quantities. However, for theoretical or practical reasons, parts of the summations are generally excluded.

Bonds are usually constrained to their ideal values using the SHAKE [3] algorithm. This has two major advantages: (i) at room temperature, bonds are in their quantum-mechanical ground state, so keeping the bond lengths fixed is probably a better approximation to the real world than treating them as classical harmonic oscillators [4]. Besides, bond vibrations would be the fastest motions in a simulated system. Removing these high frequencies leads to an increase of the simulation time step Δt one is able to use by a factor of four, allowing for much longer simulation times at a comparable computational effort.

Atoms connected to each other through a bond (1-2 interactions) or through two bonds (1-3 interactions), and 1-4 interactions between atoms in or attached to aromatic rings are excluded from the non-bonded interactions since at these short distances, the concept of Lennard-Jones interactions and point charges for atoms is no longer valid. 1-4 van der Waals interactions outside aromatic rings are reduced for the same reasons.

Non-bonded interactions which involve a sum over all pairs of atoms are generally truncated at a given cut-off radius R_c. In order to compute non-bonded interactions reasonably accurate, a

large value for this radius ($>$ 1.4nm) has to be chosen. As most of the computational cost of a classical simulation arises from the non-bonded interactions, in GROMOS96 a twin-range cut-off approach is used. Only the interactions between atoms within a short range cut-off radius R_{cp} are calculated every step. The interactions between atoms lying between the short-range cut-off and the longer-range system cut-off are only evaluated every, usually, 5 steps (10fs) thus taking them approximately into account without extensively increasing the computational work to be done. As the potential energy at the long-range cut-off is non-zero, truncating these interactions brings noise into the system. This is especially true for the very long-ranged, r^{-1}, electrostatic interactions. In order to reduce this noise, a reaction field correction is usually applied, especially in systems containing highly charged atoms. The reaction field behaves like an electrostatic continuum outside the cut-off radius, with an adjustable relative dielectric permittivity.

When simulating small systems, like in molecular dynamics, boundary effects are not negligible and have to be taken into account. This can be done in three ways:

(i) Using vacuum boundary conditions is the cheapest possibility since only the molecule of interest is simulated. However, one usually wants to simulate biomolecules in their natural environment, in solution. The solvent will have an influence on the behaviour of the molecule of interest, so vacuum boundary conditions are the crudest approximations and are, with the computer power nowadays available, no longer used for equilibrium simulations.

(ii) If the simulated biomolecule is extremely large and the system (solute and solvent) gets too big to be simulated at a reasonable computational cost, only part of the system can be simulated using full MD. Outside this core region, in the so-called extended wall region, atoms are restrained to their initial positions to minimize the effect of the vacuum boundary conditions. Outside the extended wall shell, however, there is still vacuum. The part of interest, where full MD is performed, should be as far away as possible from the vacuum, at least as far as the cut-off radius R_c

(iii) The most common treatment of boundaries is applying periodic boundary conditions. A space filling box (e.g. a cube or a truncated octahedron) represents the whole space. An atom leaving the box at one side enters it at the opposite side. This way of treating boundaries introduces artificial periodicity into the system, as usually molecules in solution or liquids are to be simulated. To minimize the effect of this, the box should be large enough that an atom can never interact directly with its own periodic copy, thus the box should in every direction be at least twice as large as the cut-off radius R_c. When simulating real crystals, so when one wants to include interactions with periodic images, infinite periodic boundary conditions, like Ewald summation [1] or the particle-particle particle-mesh technique [5] should be used.

1.2.3 Integration of the equations of motion

In order to integrate Newton's equations of motion, the time is discretized into simulation time steps Δt. The forces $\vec{f}_i$ acting on the atoms are the negative derivatives of the potential energy function $V^{phys}(\vec{r};s)$. The most widely used algorithm to integrate the equations of motion in molecular dynamics is the leap-frog algorithm, although other algorithms are also in use. Using

the leap-frog scheme, one gets

$$\vec{v}_i(t+\frac{\Delta t}{2}) = \vec{v}_i(t-\frac{\Delta t}{2}) + \frac{1}{m_i}\vec{f}_i\Delta t \tag{1.11}$$

for the velocities $\vec{v}_i$ and

$$\vec{r}_i(t+\Delta t) = \vec{r}_i(t) + \vec{v}_i(t+\frac{\Delta t}{2})\Delta t \tag{1.12}$$

for the coordinates. Although, in principle, energy is conserved in such a simulation, the use of the several approximations such as constraints or a cut-off radius introduces noise, so the system is usually coupled to a temperature bath using a weak coupling scheme [6].

$$\frac{dT(t)}{dt} = \frac{1}{\tau_T}(T_0 - T(t)) \tag{1.13}$$

with $T(t)$ and T_0 the actual and the reference temperature, respectively, and τ_T the temperature relaxation time. In order to adjust the temperature of the system, the velocities of all atoms are scaled by a factor $\lambda(t)$

$$\lambda(t) = \left[1 + \frac{2c_v^{df}}{k_B}\frac{\Delta t}{\tau_T}\left[\frac{T_0}{T(t)} - 1\right]\right]^{\frac{1}{2}} \tag{1.14}$$

where c_v^{df} is the heat capacity per degree of freedom and k_B is Boltzmanns constant. Since the heat capacity of every degree of freedom is usually not known, the quantity $\frac{2c_v^{df}}{k_B}\frac{1}{\tau_T}$ is used as an adjustable parameter to obtain the desired strength of coupling to the temperature bath.

Often one also wants to maintain the pressure using a similar weak coupling approach. For adapting the pressure, the positions of the particles are scaled by the factor

$$\mu(t) = \left[1 - \kappa_T\frac{\Delta t}{\tau_p}[P_0 - P(t)]\right]^{\frac{1}{3}} \tag{1.15}$$

with κ_T the isothermal compressibility of the system, P_0 and $P(t)$ the reference and actual pressure and τ_P the pressure relaxation time. Since in a mixed system such as a protein in solution, the compressibility is usually not known, the quantity $\frac{\kappa_T}{\tau_P}$ is used as an adjustable parameter to obtain the desired coupling strength to the pressure bath. Since the definition of the pressure depends on the kinetic energy which again depends on particle velocities, pressure coupling should be weaker than temperature coupling.

1.3 Applications of molecular dynamics

Provided that the force field used is accurate enough and that the simulation time accessible is long enough for system or molecular properties to converge, thus that an equilibrium simulation can be reached, properties calculated from a molecular dynamics trajectory can be compared to experimentally determined values. In addition, one can then analyse details that are not experimentally observable. One of the great advantages of simulations is that one can access shorter time scales and also smaller space resolution than in experiment. A structure determined by X-ray crystallography or NMR spectroscopy for example represents an average over seconds to minutes in time and over very many molecules. With molecular dynamics, on the other hand,

a single or a few protein molecules can be examined for time scales ranging from a few femtoseconds up to, currently, about 10-100ns. Molecular dynamics can also be used to interpret experimental data, as was done for example in Schiffer et al. [7] where a molecular dynamics simulation showed that seemingly different results coming from X-ray and NMR experiments were in fact explainable with the same molecular dynamics trajectory.

In addition to equilibrium simulations, MD is nowadays standardly used in structure refinement of proteins where an extra potential energy term is added to the physical force field [8]

$$V^{tot}(\vec{r};s) = V^{phys}(\vec{r};s) + kV^{special}(\vec{r};s). \tag{1.16}$$

This special term represents the contribution of experimentally determined properties (e.g. NOE distances or J-coupling constants from NMR experiments or intensities obtained from X-ray diffraction experiments). Ideally, if force fields were "perfect", the weight given to the $V^{special}(\vec{r};s)$ term could approach zero.

Overall, even if progress is still made in the field of molecular dynamics simulation, one can say that MD simulation nowadays belongs to the standard repertoire of investigation methods in biophysics and biological chemistry. Experiment and theory have, especially in this field, strongly benefited from each other and are expected to do so in the future.

Chapter 2

Molecular Dynamics Simulation of Protein Crystals: Convergence of Molecular Properties of Ubiquitin

2.1 Abstract

A unit cell of ubiquitin was simulated for 2ns using molecular dynamics to investigate the degree of convergence of different molecular properties in crystals. Energies, deviation from the experimentally derived crystal structure, atomic positional fluctuations and dihedral angle fluctuations were analysed. Most of the examined properties are converged after 2ns. The atomic positional deviation from the X-ray structure has not converged within 2ns.

2.2 Introduction

Molecules in crystals are often believed to have very rigid structure due to their ordered packing. Therefore, investigation of molecular motion of such systems is often considered to be of little interest. In contrary to small molecule crystals, the solvent concentration in protein crystals is rather high, usually about half of the crystal consists of water. Thus, one could compare protein crystals with very concentrated solutions and would expect non-negligible atomic motion. The atomic mobility in proteins can be investigated by experiment (X-ray diffraction, NMR) or by molecular simulation.

Although todays experimental techniques are very advanced, they are only able to examine time- and ensemble averaged structures and properties. With simulations, in contrary, one can go beyond that and look at the motions of a single molecule in the pico- and nanosecond time regime. Such simulations however, have only become possible since experiments were able to produce high resolution data and provide good starting structures of biologically relevant systems. Depending on what kind of properties one is interested in, there exist different methods of simulation. Equilibrium properties can be obtained using Monte Carlo (MC) and molecular dynamics (MD) simulation techniques, but with the latter motions can be observed too. Current interest of the simulation community mainly focuses on proteins in their natural environment - in solution. Thus, the force fields used are parametrized to mimic the behaviour and function of proteins in solution and not many crystal simulations have yet been performed. Consequently, a

crystal environment provides an excellent occasion to test the force field on a task it should be able to carry out, but where it is not biased by the parametrization.

Apart from the pure observation of properties, MD simulation is used in structure refinement. In refinement, be it X-ray crystallographic or NMR, a special term is added to the standard physical force field to reflect the presence of experimental data.

$$V(\vec{r}) = V^{phys}(\vec{r}) + V^{special}(\vec{r}) \tag{2.1}$$

In NMR, a variety of properties can be measured and each of these can be used in the definition of an additional term that restrains the generated structures to reproduce given experimental values. Refinement procedures exist that use NOE [8] [9], J-value [10] and chemical shift [11] restraints. In crystallography, X-ray intensities are used to generate the restraining energy contribution [12] [13]. Combined NMR/X-ray refinement uses both, solution and crystal data [7].

As in experiment averages over time and molecules are measured, instantaneous restraints can lead to artificial rigidity in the molecular system [14]. This can be circumvented by restraining only time or ensemble averages to the value of the measured quantity instead of instantaneous values. Time averaging was applied to NOEs [14] and J-values [10] in NMR structure determination and to X-ray intensities in crystallography [15] [16] [17]. Ensemble averaging was applied in NMR refinement [18] [19]. For a more detailed discussion on restrained MD simulations we refer to the literature [20] [21].

The very first - unrestrained - MD simulations of a protein in a crystal were carried out in the early eighties [22] [23]. The protein concerned was bovine pancreatic trypsin inhibitor (BPTI), a small (58 residue) protein for which high resolution X-ray diffraction data was available. The initial level of simulation was to neglect solvent, using vacuum boundary conditions. This has been improved gradually by the inclusion of Lennard-Jones particles at the density of water as solvent [23], to let the protein feel random forces and friction from the outside as well as a slightly attractive external field. The next step was to use a simple (simple point charge, SPC) water model [24]. Further improvement was achieved by incorporating counter ions into the modeled systems to obtain overall charge neutrality [25].

Despite these early attempts, only few unrestrained crystal simulations have been reported in the literature up to now and the ones known concern one to four protein molecules, simulating one unit cell [26] [27]. The time range covered is about 100ps.

In this work, the current state of MD simulation of protein crystals is illustrated. A full unit cell of ubiquitin containing four ubiquitin and 692 water molecules was simulated for a period of two nanoseconds. Since the simulation is an order of magnitude longer than crystal simulations in the literature, it offers the possibility to analyze convergence of different properties as function of time as well as of the number of protein molecules. Secondly, converged properties can be compared to experimental values to test the GROMOS96 force field [28]. Finally, the obtained motions can be analyzed to obtain a picture of the molecular behaviour of ubiquitin in a crystalline environment.

2.3 Methods

Ubiquitin consists of 76 amino acids with 602 non-hydrogen atoms. Hydrogen atoms attached to carbon atoms are incorporated into these and the remaining 159 hydrogen atoms which are able to form hydrogen bonds are treated explicitly. Ubiquitin crystallises in the orthorhombic

space group $P2_12_12_1$ with a=5.084nm, b=4.277nm and c=2.895nm. There is 1 molecule in the asymmetric unit. The protein was crystallised at pH 5.6, so GLU and ASP amino acids were chosen to be deprotonated, LYS, ARG and HIS residues protonated, leading to a charge of +1 electron charge per chain. Because this is a small value compared to the size of the system, no counter ions were added. Four chains of ubiquitin building up a full unit cell of the crystal were simulated together with 692 water molecules modelled using the SPC water model [29]. 232 water molecules were placed at crystallographically observed water sites, the remaining 460 were added to obtain the experimental density leading to a system size of 3044 protein atoms and 5120 atoms in total.

As a starting point, the crystal structure of ubiquitin solved at 1.8Å resolution [30] was used (entry 1UBQ of the Brookhaven Protein Database [31]). The non-crystallographic water molecules were added using a minimum distance of 0.220605nm between non-hydrogen protein atoms or crystallographic water oxygens and the closest oxygen atom of the added solvent which was taken from a periodic box with equilibrated water [28]. Initial velocities were assigned from a Maxwell-Boltzmann distribution at 300K. Protein and solvent were coupled separately to temperature baths of 300K with a coupling time of 0.1ps [6]. No pressure coupling was applied, but a short run (results not shown) with pressure coupling showed no significant change in the box volume. Bonds were kept rigid using the SHAKE method [3] with a relative geometric tolerance of 10^{-4}. Long range forces were treated using twin range cutoff radii of 0.8nm and 1.4nm [32]. The pairlist for non-bonded interactions was updated every 10fs. No reaction field correction was applied. All simulations were carried out using the GROMOS96 package and force field [28].

As a first step, the system was minimised for 20 cycles using the steepest descent method, restraining the protein atoms to their initial positions with a force constant of 25000kJ/(mol*nm^2). This minimised structure was then pre-equilibrated in several short MD runs of 500 steps of 0.002ps each, gradually lowering the restraining force constant from 25000kJ/(mol*nm^2) to zero. Then the time origin was reset to zero and the whole unit cell was simulated for 2ns. The time step was 0.002ps and every 500th configuration was stored for evaluation.

The first 400ps of the run were treated as equilibration time using the remaining 1.6ns for analysis.

2.4 Results

In Figure 2.1, the non-bonded contributions to the total potential energy are shown. Non-bonded interactions are formed by Lennard-Jones and electrostatic interactions. Solvent and solute internal respectively and solute-solvent interaction energies are shown separately. All of these look converged after about 100ps. Solvent-solvent energy stays at its initial value during the whole simulation as the water molecules had time to relax in the pre-equilibration period, when the proteins were kept fixed by the restraining potential. Protein internal interactions are weakened in the first some hundred picoseconds, whereas protein-solvent energy is lowered. The proteins adapt to the already relaxed solvent. This effect disappears after about 200ps from where on the system can be viewed being equilibrated as far as energies are concerned. The distribution of kinetic versus potential energy and the total (bonded and non-bonded) energy of the system are relaxed even earlier (results not shown).

Figure 2.2 shows, however, that not all properties converge as fast as energies do. Root mean square atom positional deviations (RMSD) from the X-ray structure are presented for the four individual chains for both, Cα and all atoms. Several relaxation periods can here be distinguished.

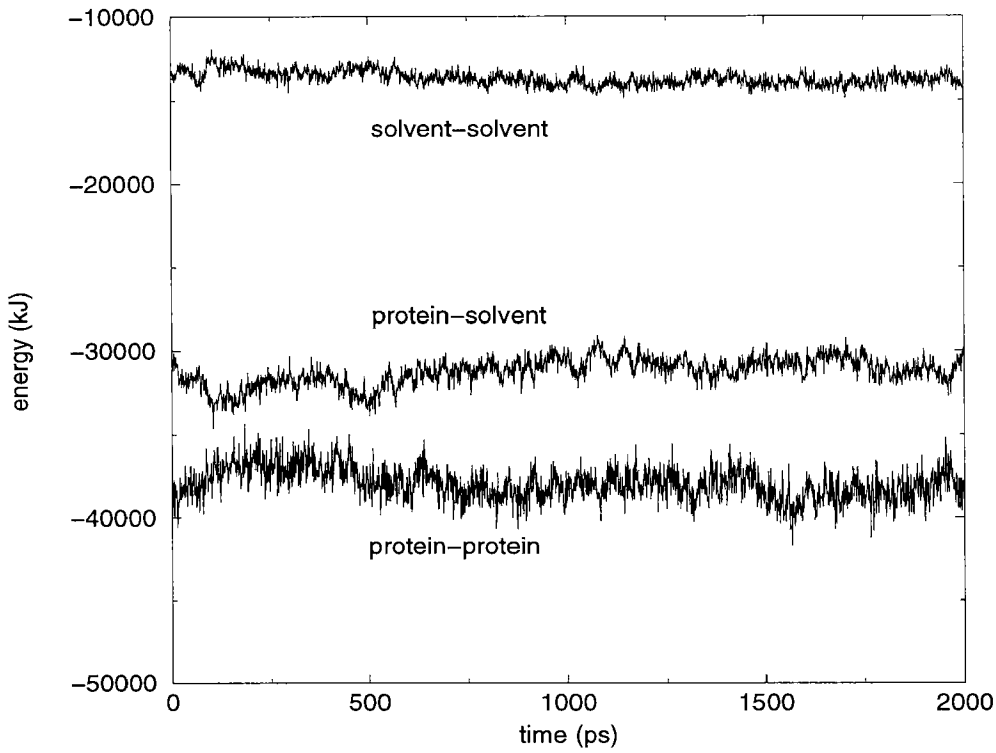

Figure 2.1: *Total nonbonded energies (in kJ/mol) of the system as a function of time*

After the initial increase, occurring during the first 50ps of the simulation, a plateau value is reached which is more or less stable until 300ps. The values reached are 0.12nm for Cα atoms and 0.20nm if all atoms are taken into account. These numbers are in agreement with results obtained in crystal simulations of other proteins of a few hundred picoseconds length reported in the literature [24] [25] [26] [27] [33]. After 300ps, however, the values do increase again slowly. For the Cα atoms, we can recognise a second plateau from 300 to 850ps. During the same period of time, RMSD for all atoms still increase monotonically. After 850ps a final plateau value is reached. For the second nanosecond of the simulation (1000 - 2000ps), RMSD are 0.21nm for Cα and 0.29nm for all atoms respectively. Concerning RMSD, chain 1 is an exception. There is a strong increase after 1200ps, due to a movement of a particular part of the chain which will be addressed later. To be sure that the RMSD values are really converged or to find out if there are other - higher or even lower - plateaus after the simulated 2ns, even longer runs would have to be performed.

Although the RMSD in this long simulation become larger than usually observed in the course of short simulations, the hydrogen bonding pattern and thus the secondary structure is well reproduced (Table 2.1-2.4). Most of the hydrogen bonds reported [30] show high occupancies during the whole simulation, especially the ones within secondary structure elements. Only 6 out of the 44 hydrogen bonds present in the X-ray structure disappear or are converted into other ones in the simulation. Hydrogen bonds in the α-helix from residues 23-34 show very large occupancies, from 75% at the N-terminal part to well over 90% inside the helix. Only its

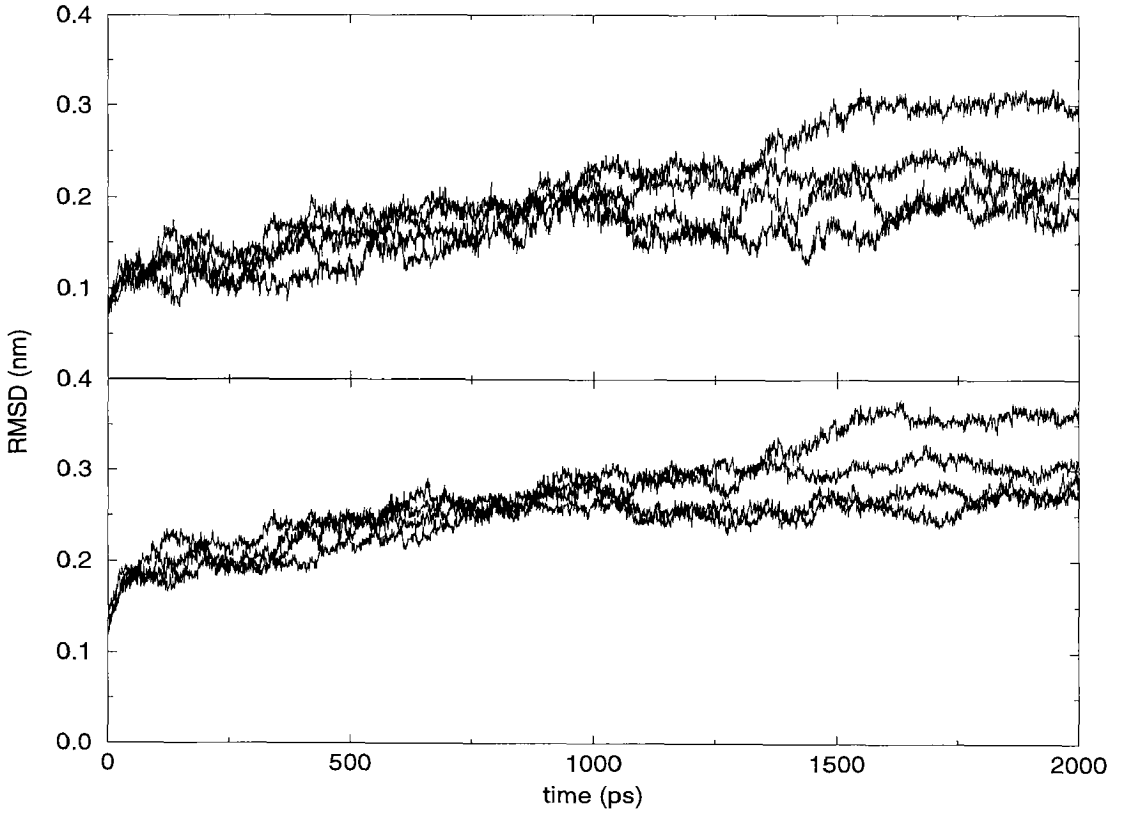

Figure 2.2: *Root mean square atomic positional deviations (RMSD) in nm from the X-ray structure of the four different protein molecules in the unit cell as function of time. Rotational and translational fitting was applied over the Cα atoms of residues 1 to 72. The top graph shows the deviations for the Cα atoms, the bottom graph the ones for all atoms.*

C-terminal end shows a little instability, the α-helix being deformed towards a 3_{10}-helix. The β-sheet pattern is - apart from chain 1 in the region of residues 49-64 - as stable as the α-helix. Occupancies are from 55% up to 95%. The six hydrogen bonds not reproduced (five 3_{10} and one α-helical) can be rationalised as follows. The bridges 10-7 and 65-62 are part of the most mobile regions of the protein. These regions are residues 7-10, 51-54 and 62-65 [30]. The hydrogen bond at the C-terminal end of the α-helix (35-31) disappears completely and the preceding one is changed, indicating that the end of the α-helix is deformed towards a 3_{10}-helix. The donor of the bond 40-37 is replaced by residue 41 and the 4 residue 3_{10}-helix that was stabilised by hydrogen bonds 58-55 and 59-56 is replaced by an α-helical hydrogen bond 59-55. The high occupancy of this particular bond and the complete absence of the two experimental ones indicates an early rearrangement in this part of the structure (before the analysis period) which is stable during the rest of the run.

Backbone-sidechain and sidechain-backbone hydrogen bonds are less well reproduced than the backbone internal ones. There are some which are present during 80-90% of the time but there are others present less than 50%. Two out of seven hydrogen bonds in which a backbone atom is the donor do not show up in the simulation. Both of them involve the OG1 atom of THR7 as acceptor. The donor atoms of these two hydrogen bonds are the backbone nitrogen atoms of

Intramolecular Hydrogen Bonds		X-ray	MD, Molecule			
			1	2	3	4
Backbone	Backbone					
3Ile H	15Leu O	100	94	94	95	98
4Phe H	65Ser O	100	85	69	87	77
5Val H	13Ile O	100	80	90	87	93
6Lysh H	67Leu O	100	85	82	88	94
7Thr H	11Lysh O	100	65	49	54	62
8Leu H	69Leu O	0	5	52	19	55
10Gly H	7Thr O	100	0	0	0	0
13Ile H	5Val O	100	86	76	70	87
15Leu H	3Ile O	100	87	92	72	82
17Val H	1Met O	100	68	39	79	51
21Asp H	18Glu O	100	68	84	84	90
23Ile H	54Arg O	100	0	74	89	92
24Glu H	52Asp O	100	58	69	63	84
26Val H	22Thr O	100	92	69	78	61
27Lysh H	23Ile O	100	94	97	98	99
28Ala H	24Glu O	100	71	71	84	89
29Lysh H	25Asn O	100	91	79	94	88
30Ile H	26Val O	100	92	76	94	91
31Gln H	27Lysh O	100	85	53	66	93
32Asp H	28Ala O	100	82	27	87	77
33Lysh H	29Lysh O	100	23	13	81	51
33Lysh H	30Ile O	0	59	23	7	19
34Glu H	30Ile O	100	95	54	64	86
35Gly H	31Gln O	100	0	0	0	0
36Ile H	34Glu O	0	62	50	28	35
40Gln H	37Pro O	100	0	0	0	0
41Gln H	37Pro O	0	68	56	72	20
41Gln H	38Pro O	100	14	25	14	50
42Arg H	70Val O	100	82	82	83	88
44Ile H	68Hish O	100	84	96	93	95
45Phe H	48Lysh O	100	20	74	77	91
48Lysh H	45Phe O	100	24	62	59	44
50Leu H	43Leu O	100	29	88	92	85
54Arg H	51Glu O	100	20	60	19	69
56Leu H	21Asp O	100	0	90	81	81
57Ser H	19Pro O	100	3	78	86	83
58Asp H	55Thr O	100	0	0	0	0
59Tyr H	55Thr O	100	58	86	92	85
59Tyr H	56Leu O	100	0	0	0	0
60Asn H	57Ser O	100	38	34	60	58
61Ile H	56Leu O	100	67	7	63	56
64Glu H	2Gln O	100	0	42	6	95
65Ser H	62Gln O	100	0	0	0	0
67Leu H	4Phe O	100	69	74	87	70
68Hish H	44Ile O	100	62	68	83	89
69Leu H	6Lysh O	100	79	72	92	90
70Val H	42Arg O	100	91	89	90	91
72Arg H	40Gln O	100	79	59	85	78

Table 2.1: *Occurrence of intramolecular hydrogen bonds (in%) during the final 1.6ns. The criteria for a hydrogen bond to be present are: angle donor - hydrogen - acceptor $\geq$ 135°, distance hydrogen - acceptor $\leq$ 0.25nm. Hydrogen bonds are shown if they are either present in the X-Ray structure or if at least one of the four protein molecules examined shows the hydrogen bond of interest for at least 50% of the time.*

residues THR9 and LYS11 which both have high experimental B-factors, the one of THR9 is 18.32Å^2, the one of LYS11 is lower, 13.56Å^2. These values should be compared to the mean experimental B-factors of 10.73Å^2 for the backbone atoms and 13.41Å^2 for all protein atoms respectively. If a sidechain atom is the donor, three out of five hydrogen bonds are not found in the simulation. All of these involve the sidechain nitrogen atom of lysine residues as donor, the experimental B-factors of which are even higher than before, ranging from 23.92Å^2 for the NZ atom of LYS48 up to 30.06Å^2 for the NZ atom of LYS33. From the four sidechain-sidechain hydrogen bonds, no one is observed as in the crystal, the 54-58 hydrogen bond is replaced by a 55-58 one. The three hydrogen bonds we do not see in the simulation all have very mobile atoms as acceptors, experimental B-Factors are bigger than 25Å^2, and donors with B-factors of 15.47Å^2 for the NZ atom of LYS27 and over 20Å^2 for the NZ atoms of LYS11 and the NH atom of ARG54.

The one intermolecular hydrogen bond (Table 2.5) in the starting structure which is as well not seen in the simulation has as donor the sidechain nitrogen atom of LYS6 with an experimental

Intramolecular Hydrogen Bonds		X-ray	MD, Molecule			
			1	2	3	4
Backbone	Sidechain					
2Gln H	64Glu OE2	0	63	7	84	0
9Thr H	7Thr OG1	100	0	0	0	0
11Lysh H	7Thr OG1	100	0	0	0	0
18Glu H	21Asp OD2	100	80	3	0	0
20Ser H	18Glu OE2	0	0	0	55	0
25Asn H	22Thr OG1	100	31	13	61	38
51Glu H	59TYR OH	100	46	87	56	76
55Thr H	58Asp OD1	100	29	62	22	75
58Asp H	55Thr OG1	100	53	76	72	86
64Glu H	64Glu OE2	0	55	6	16	0

Table 2.2: *Occurrence of intramolecular hydrogen bonds (in%) during the final 1.6ns. The criteria for a hydrogen bond to be present are: angle donor - hydrogen - acceptor $\geq$ 135°, distance hydrogen - acceptor $\leq$ 0.25nm. Hydrogen bonds are shown if they are either present in the X-Ray structure or if at least one of the four protein molecules examined shows the hydrogen bond of interest for at least 50% of the time.*

Intramolecular Hydrogen Bonds		X-ray	MD, Molecule			
			1	2	3	4
Sidechain	Backbone					
29Lysh HZ2	16Glu O	100	0	0	0	0
33Lysh HZ2	14Thr O	100	0	0	0	0
41Gln HE21	27Lysh O	100	81	91	47	71
41Gln HE22	36Ile O	100	90	89	60	83
48Lysh HZ3	46Ala O	100	0	0	0	0

Table 2.3: *Occurrence of intramolecular hydrogen bonds (in%) during the final 1.6ns. The criteria for a hydrogen bond to be present are: angle donor - hydrogen - acceptor $\geq$ 135°, distance hydrogen - acceptor $\leq$ 0.25nm. Hydrogen bonds are shown if they are either present in the X-Ray structure or if at least one of the four protein molecules examined shows the hydrogen bond of interest for at least 50% of the time.*

B-factor of 20.55Å^2 and as acceptor the sidechain oxygen atom of GLU51 with B-factors of 32.13Å^2 and 33.44Å^2 respectively. Most of the hydrogen bonds not reproduced in the simulation contain at least one mobile atom. These do not stay at their equilibrium positions all the time, although they can still stabilise the structure on average.

In Figure 2.3, the deviation of the Cα atoms of the different chains from the X-ray structure and from the mean MD structure are examined together with the deviation of the mean MD structure from the X-ray structure. Overall, the single proteins stay close to the experimental structure, but parts of them deviate quite substantially. The region involving residues 7-11 which is experimentally (B-factors) known to be mobile, shows in three out of the four chains a RMSD for Cα atoms that lies at 0.3nm or, for chain 2 even 0.5nm. Chain 3, in contrast is close to the

Intramolecular Hydrogen Bonds		X-ray	MD, Molecule			
			1	2	3	4
Sidechain	Sidechain					
11Lysh HZ2	34Glu OE2	100	0	0	0	0
20Ser HG	18Glu OE2	0	0	0	60	0
27Lysh HZ2	52Asp OD2	100	0	0	0	0
49Gln HE21	16Glu OE1	100	0	0	0	0
54Arg HH12	58Asp OD1	100	0	0	0	0
55Thr HG1	58Asp OD1	0	44	83	29	86

Table 2.4: *Occurrence of intramolecular hydrogen bonds (in%) during the final 1.6ns. The criteria for a hydrogen bond to be present are: angle donor - hydrogen - acceptor $\geq$ 135°, distance hydrogen - acceptor $\leq$ 0.25nm. Hydrogen bonds are shown if they are either present in the X-Ray structure or if at least one of the four protein molecules examined shows the hydrogen bond of interest for at least 50% of the time.*

Intermolecular Hydrogen Bonds		X-ray	MD, Molecule			
			1-4	2-3	3-2	4-1
6 Lys HZ3	51 Glu OE1	100	0	0	0	0
12 Thr HG1	18 Glu OE1	0	56	57	75	34
49 Gln H	8 Leu O	0	10	34	0	67
68 Hish HE2	32 Asp OD2	0	0	0	53 (3-1)	13 (4-2)
71 Leu H	58 Asp O	0	65	0	0	0

Table 2.5: *Occurrence of intermolecular hydrogen bonds (in%) during the final 1.6ns. The criteria for a hydrogen bond to be present are: angle donor - hydrogen - acceptor $\geq$ 135°, distance hydrogen - acceptor $\leq$ 0.25nm. Hydrogen bonds are shown if they are either present in the X-ray structure or if at least one of the four protein molecules examined shows the hydrogen bond of interest for at least 50% of the time.*

X-ray structure and the mean MD structure is closer to the crystal than the single chains are, suggesting that the simulation does not systematically deviate from experiment but that different regions of conformational space are visited by the distinct chains. These regions are quite far off but the deviations average out if the whole system is considered. The same argument holds for the very mobile region between residues 47 and 64 where chain 1 deviates dramatically from both the mean MD and the X-ray structure. The other chains differ in that part of the protein less from the crystal than from the mean MD structure, keeping the simulation average close to the X-ray structure. For all single chains and for the averaged simulation structure, the deviation is largest in the C-terminal region. These parts are ill-defined in experiment as well with occupancies of 0.45 for residues 73 and 74 or 0.25 for the terminal two glycines respectively. Other parts of the protein, especially the stable secondary structure elements stay close to the X-ray structure. The whole α-helix, even its C-terminal part which was deformed to a 3_{10}-helix, deviates by maximally 0.08 nm from the crystal although, as seen before, the single chains may be less close to the experimental structure. The β-sheet regions stay close to the X-ray structure as well. As with the helix, residues 1-7, 40-45 and 64-72 stay within 0.1nm RMSD from the X-ray structure.

The β-strands formed by residues 10-17 and 48-50 are not as similar to experiment since they lie close to mobile regions and are thus influenced by neighbouring mobile residues. For the strand formed by residues 10-17 already from residue 12 onwards, the same structural similarity is reached as for all other secondary structure elements and residues 48-50 are again strongly influenced by the moving part of chain 1.

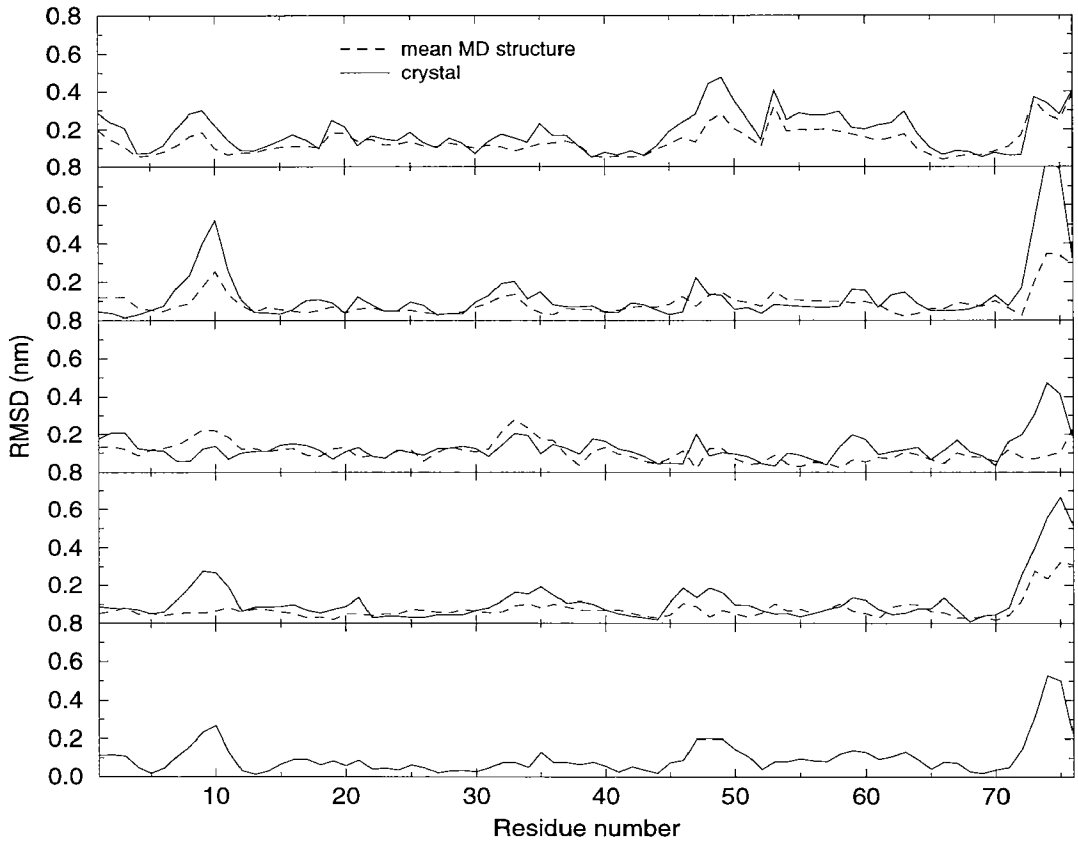

Figure 2.3: *Root mean square Cα atom positional deviation (RMSD) in nm as a function of the residue number using the final 1.6ns of the simulation. RMSD of the different chains versus the mean molecular dynamics structure and the X-ray structure are shown. The first four graphs show the RMSD of the four different chains, the bottom graph shows the RMSD of the mean MD structure versus the X-ray structure*

In Figure 2.4, the impact of different fitting protocols on atomic mean-square positional fluctuations is examined. B-factors are related to root mean square positional fluctuations (RMSF) according to

$$B_i = \frac{8\pi^2}{3}\langle(\vec{r}_i - \langle\vec{r}_i\rangle)^2\rangle \tag{2.2}$$

with the angles indicating time or time and ensemble averages. Molecule 4 was selected because it is more stable than the other chains. RMSF for the Cα atoms were calculated using the simulation trajectory directly, after applying translational fitting using the Cα atoms of the well-defined residues 1-72 and as well by using both, rotational and translational fitting. For the bottom graph, full fitting was applied using all Cα atoms. Removal of the overall translational

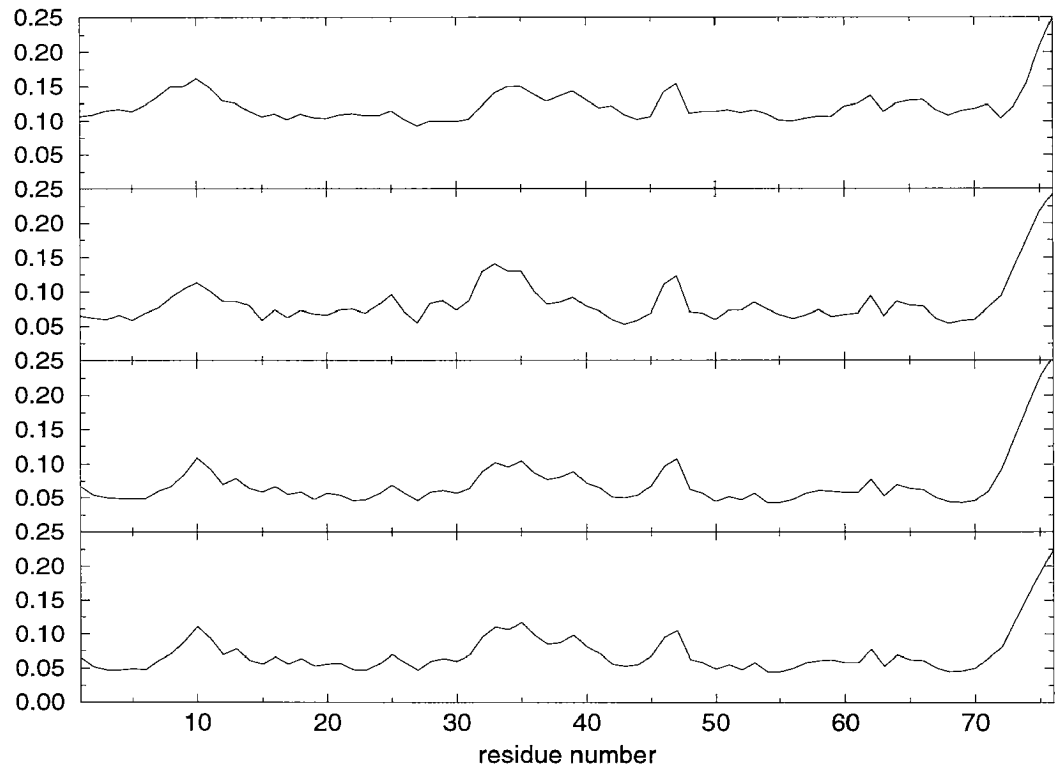

Figure 2.4: *Root mean square Cα atom positional fluctuations (RMSF) in nm for chain 4 are shown as a function of residue number. RMSF are averaged over the final 1.6ns. In the top graph, no fitting was applied, in the second graph, translational fitting was applied using the Cα atoms of residues 1 to 72 and in the third one, the rotational component of motion was removed additionally. In the bottom graph, translational and rotational fitting was applied over all Cα atoms (1-76).*

component of motion reduces the positional fluctuations on average by 0.04nm. Only the RMSF at the end and right after the large α-helix formed by residues 23-34 are nearly not affected by the introduction of translational fitting. In contrast, it is exactly this region where the fluctuations are substantially lowered by additionally introducing a rotational fit. The regions before residue 27 and after residue 42 are only slightly affected by the removal of overall rotation. These findings suggest that the protein as whole translates by about 0.04nm and the α-helix region is held back in its initial position thus rotating compared to the rest of the protein. Introduction of the four C-terminal residues into the fitting procedure does only affect their own RMSF and the ones of the rotating part of the molecule, indicating that these four residues move together with the rest of the molecule. The atom positional fluctuations obtained by applying a full (rotational and translational) fitting are determined by internal motion only. The largest RMSF for residues 1-72 are 0.12nm. RMSF of the 2 C-terminal glycines are 0.26nm if the last four residues are excluded from the fitting and 0.22nm otherwise.

If the same properties are examined but averaged over all the chains, similar trends can be observed (Figure 2.5). If no fitting is applied, the RMSF of 0.24nm on average indicates that the

different molecules show relative translation and rotation. After translational fitting is applied, the mean RMSF drops to 0.18nm. Thus the molecules translate within the unit cell. If the rotational component of overall motion is removed, the whole helix region is much less mobile than before and the mean RMSF drops to 0.14nm. The same holds for the region 47-64 dominated by the rotation of part of chain 1. Fluctuations are generally much larger than before where only 1 chain was observed, again indicating that the distinct chains behave uncorrelated. The peaks of the RMSF averaged over all chains are around 0.22nm compared to 0.27nm when overall rotation is still present. Thus, in addition to internal rotations, there occurs a relative rotation of the different molecules. If the fit is not only applied over the well-defined Cα-atoms of residues 1-72, the RMSF value gets slightly higher - apart of the C-terminal region - but this influence is small, the mean RMSF staying at 0.14nm. The relative heights of the peaks however differ which shows that it is crucial to have a standard fitting protocol defined which must not be changed during the course of analysis.

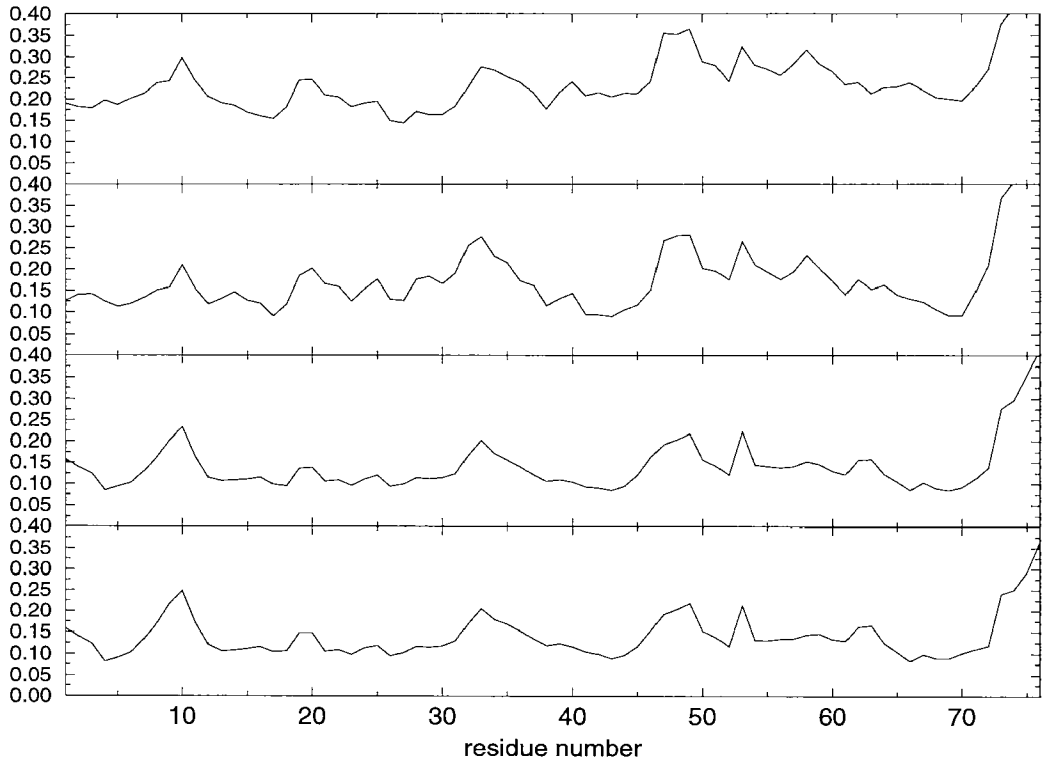

Figure 2.5: *Root mean square Cα atom positional fluctuations (RMSF) in nm are shown using the same fitting protocols as in figure 2.4 but averaged over all the four protein molecules.*

In Figure 2.6, we concentrate again on molecule 4. Comparing different averaging periods using different starting points, it can be seen that - in general - the later the averaging period is chosen, the less movement is observed. During the period 800ps-1200ps, only the small region between residues 9 and 12 shows more mobility than between 400ps and 800ps. The rest of the molecule shows - in the stable regions - the same mobility as in the earlier time period. In the parts that are most mobile between 400ps and 800ps, the motions decrease significantly after the

latter time point, indicating that equilibrium is reached. Focussing on the longer averaging periods, 400ps-1200ps vs. 1200ps-2000ps, we see that over the whole chain mobility is comparable indicating clear equilibrium as far as internal motions are concerned. The fluctuations during the 400ps-1200ps period are of the same size as the ones of the shorter subperiods 400ps-800ps and 800ps-1200ps. They are thus determined by movements on a time scale shorter than 400ps.

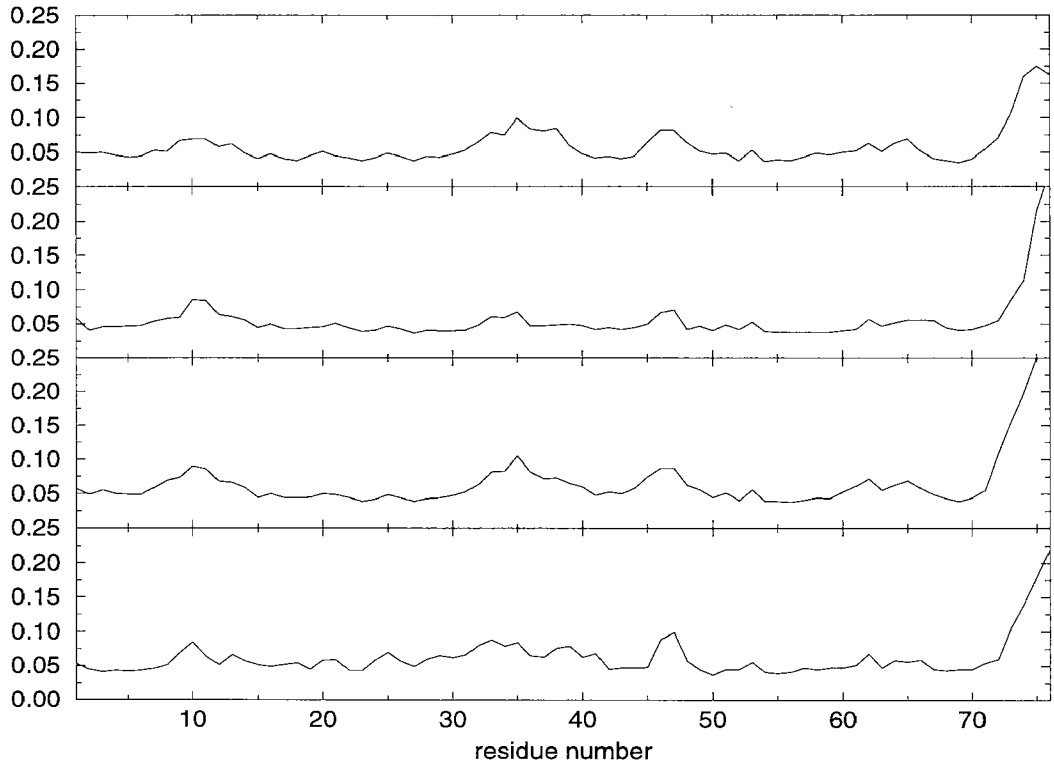

Figure 2.6: *Root mean square Cα atom positional fluctuations (RMSF) in nm are shown for chain 4 with full transrotational fitting over the Cα atoms of residues 1 to 72. Different averaging periods are compared. The top graph shows RMSF using the period from 400ps - 800ps, the second graph from 800ps - 1200ps. For the third graph, both of the previous periods are averaged (400ps - 1200ps) and the for the bottom one, the period from 1200ps - 2000ps was used.*

In Figure 2.7, the same properties as in the previous ones are observed but now the whole unit cell is taken into account. Comparing RMSF between 400ps and 800ps with the ones of the following 400ps period, not much difference is seen. The later period however shows - contrary to when only a single chain was examined - somewhat more mobility than the earlier one. This difference shows that - although the configurations of the single chains converge quite rapidly - different parts of phase space are visited by the different chains. The fact that the fluctuations in the 400ps-1200ps period are between the ones of the two shorter analysis periods or in some places even slightly lower than both of them is another indication that the single chain movements are converged on these short timescales. In the last 800ps of the simulation, the RMSF are substantially higher than in the 800ps window before that. All of the peaks can be found in one of the single chains. If only one of the four molecules differs strongly from the

other three, this one determines the magnitude of the fluctuations of the average. The peak at residue 10 comes from chain 2, the ones around 20 and the whole region 47 - 64 are determined by chain 1. The peak at residue 33 origins in chain 3 which at this point differs substantially from the mean MD structure (figure 2.3).

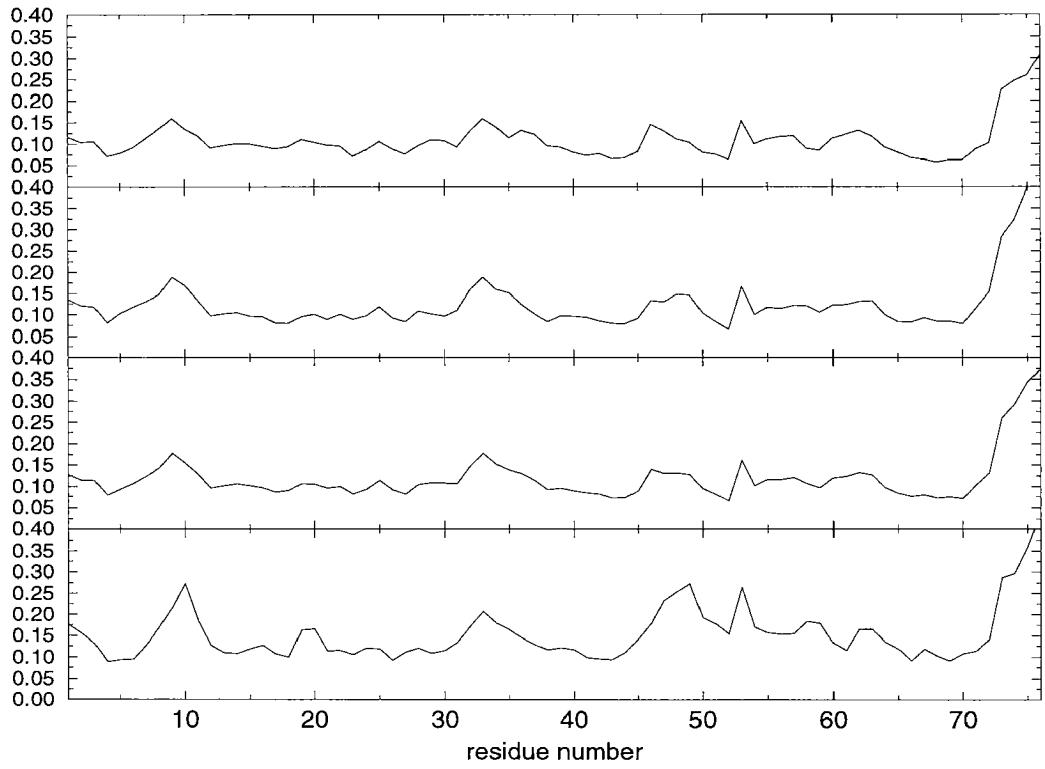

Figure 2.7: *Root mean square Cα atom positional fluctuations (RMSF) in nm are shown using the same averaging periods as in figure 2.6 but averaged over all the four protein molecules.*

Figure 2.8 describes the atomic root mean square positional fluctuations for the Cα atoms of the four protein molecules during the 1.6ns analysis period and corresponding values defined by equation 2.2 obtained from the experimental B-factors belonging to the crystal structure. Rotational and translational fitting was applied using the Cα atoms of residues 1 to 72 and the fluctuations were averaged over the final 1.6ns. The stable secondary structure elements show as little mobility in the simulation as inferred from experiment. The more mobile parts of the proteins show enhanced mobility in the simulation as well but the magnitude of the fluctuations is overestimated. The movements of the single chains can be rationalized as follows. In protein one, the whole region from GLY47 onwards rotates around the remarkably stable axis formed by residues 41-46. This part lies, as all the flexible parts, on the outside of the protein. Residues 19 and 20 which are stable in all but this single chain are in contact with this moving region. This rotation which tends to compact the protein occurs during the 200ps period between 1350 and 1550ps after the start of the simulation in which the RMSD increase significantly (Figure 2.2). Chain two is more stable than chain 1 but some residues are still very mobile. The end of the unwinding helix shows large fluctuations. In the course of this deformation, the sidechain

nitrogen of LYS11 moves from close to the OE atom of GLU34 towards the backbone oxygen of LYS33 which causes GLY10 to adapt its position. A similar but smaller motion occurs in protein molecule four. Both lysines, LYS33 and LYS63 are fully exposed to the solvent and have no intramolecular contacts. Also in protein three, the flexible residues are not part of secondary structure elements and are located on the outside of the protein. The backbone oxygen of GLN62 that moves in all the four chains, has in addition the closest contact to another heavy atom: the OG1 atom of SER65 is only 2.51Å away and the van der Waals repulsion of these atoms causes them to move further away from one another. The mobile residues in chain four are again in contact with the solvent, GLY35, GLY47, GLN62 and the end of the helix pulling GLY10 out of its equilibrium position. The terminal residues of all the protein molecules are very mobile as experimentally observed in the crystal.

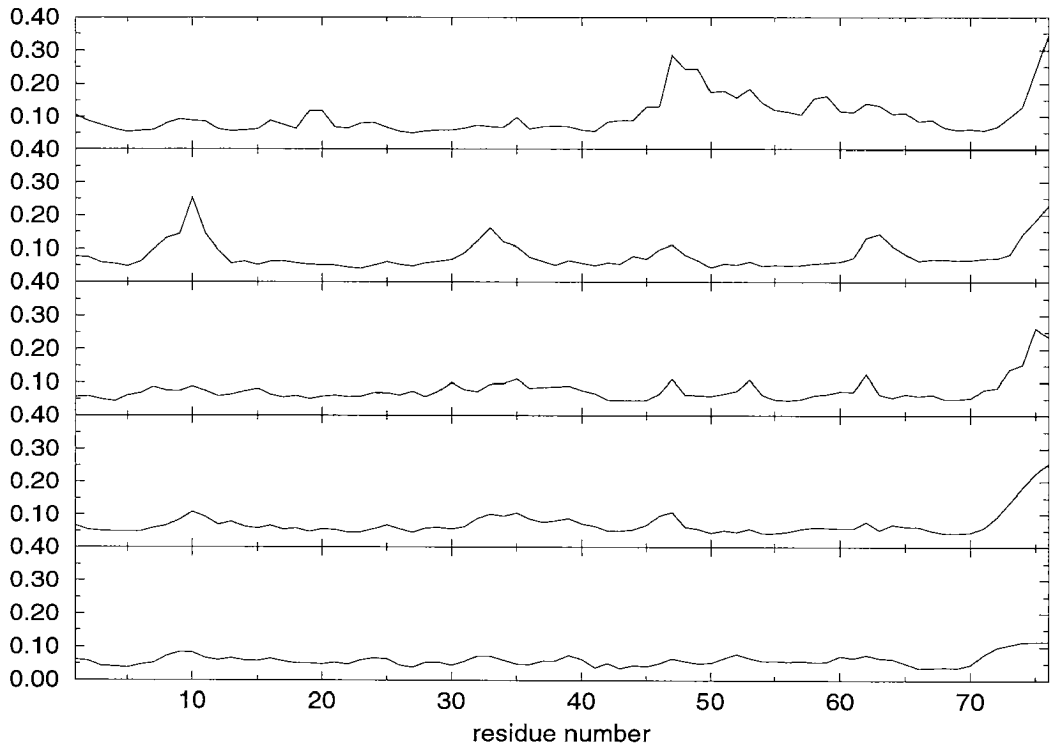

Figure 2.8: *Root mean square Cα atom positional fluctuations (RMSF) in nm for the different chains as function of residue number. Full trans- and rotational fitting was applied over the Cα atoms of residues 1 to 72 using the final 1.6ns of the simulation. The four top graphs show the four different protein molecules, the bottom graph shows corresponding values defined by equation 2.2 obtained from experimental B-factors.*

Backbone dihedral angle fluctuations and transitions are examined in tables 2.6 and 2.7 using different relaxation periods. After the first 400ps of analysis, the φ/ψ dihedral angle fluctuations differ only slightly but if longer averaging times are chosen, the different protein molecules show larger differences from one another. These fluctuations also increase for longer analysis times indicating that they are not yet converged after 2ns. In the period from 800ps to 1200ps, chain 3

Molecule	400-800ps φ/ψ	400-1200ps φ/ψ	400-2000ps φ/ψ
1	18.4/22.9	19.5/23.7	31.0/33.6
2	17.2/17.0	18.6/18.7	23.6/26.8
3	18.5/20.4	25.6/26.3	35.3/37.5
4	19.7/18.8	19.4/20.3	21.6/28.8
all	26.1/26.2	28.0/28.6	35.2/38.1

Table 2.6: *Root mean square fluctuations of backbone ϕ and ψ dihedral angles (in degrees) for the different molecules using different time averaging periods. The bottom lines show the averages over all four protein molecules.*

	120°		
Mol.	400-800ps	400-1200ps	400-2000ps
1	46.5	45.4	47.7
2	40.5	41.5	47.3
3	50.5	57.1	51.3
4	44.8	46.4	46.4
all	45.6	47.6	48.2
	60°		
1	245.5	246.6	289.3
2	271.5	272.1	261.3
3	381.5	381.0	348.3
4	356.8	325.4	325.4
all	313.8	306.3	306.1
	all		
1	292.0	292.0	336.9
2	312.0	313.7	308.6
3	432.0	438.1	399.6
4	401.5	371.8	371.8
all	359.4	353.9	354.2

Table 2.7: *Number of backbone dihedral angle transitions for the four different protein molecules using different time periods. The transitions are shown on a per 100ps basis. dihedral angles with 3-fold and 6-fold potential energy wells are distinguished. The bottom line shows the averages over all protein molecules.*

shows a large increase in mean square dihedral angle fluctuations whereas the Cα atom positional RMSD with respect to the X-ray structure during the same time fluctuate around a plateau value. Thus there is a lot of flexibility without the simulation structure diverging from the experimental one. Protein molecule 3, for example, shows the largest φ/ψ fluctuations of all the 4 molecules and the lowest RMSD of Cα atoms at the end of the simulation (Figure 2.2) indicating that it explores phase space around the equilibrium structure. If in contrary, as in molecule 1 after 1200ps, the Cα atom positional RMSD with respect to the X-ray structure increase significantly, larger dihedral angle fluctuations are observed as well.

Concerning relaxation, similar observations as before can be made analysing dihedral angle transitions (table 2.7). The number of transitions for the different chains differ by about 20%. Within a single chain however, the number of transitions increases proportionally to the observation time. Again, the protein molecules showing the most transitions do not have the largest Cα atom positional RMSD from the X-ray structure. Thus, only certain dihedral angle flips induce large movements that determine the RMSD value.

In figure 2.9 the number of water oxygens with given atomic root mean square positional fluctuations for the oxygen atom are shown. The time evolution and the shape of the curves are like for bulk water, a gaussian distribution with the maximum at larger RMSF and higher standard deviation when using longer averaging times. Using the diffusion constant of bulk SPC water at 300K of $3.9 \cdot 10^{-3}$nm^2/psec [34], the root mean square positional fluctuation for an average water molecule would be 1.25nm for a 400ps period, 1.77nm for a 800ps period and 2.5nm for a 1600ps period. Comparison of these values with the distributions in figure 2.9 indicates that the motion of most of the crystal water molecules is restricted by the crystalline environment.There are water molecules which during long time periods (in the order of some 100ps) stay at defined positions. Several water molecules can also be observed that move very little, then leave a stable site but later turn less mobile again which shows that they stay close to a crystallographically determined site which they revisit after being replaced by a different water molecule shortly.

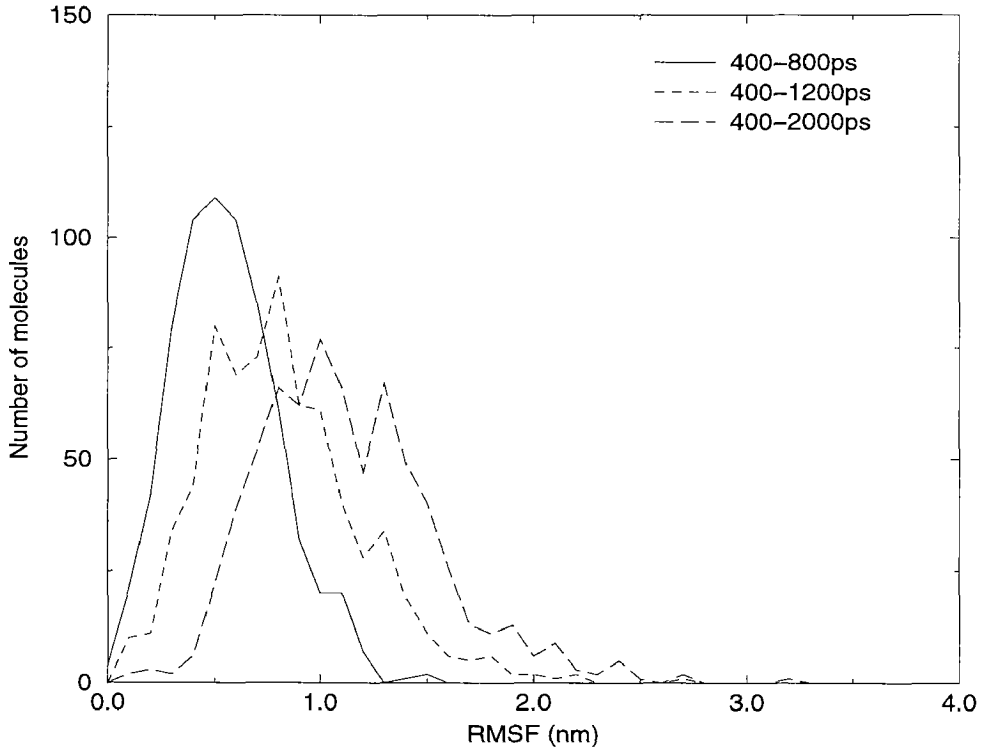

Figure 2.9: *The number of water molecules with a given root mean square oxygen positional fluctuation are shown for different averaging periods.*

2.5 Conclusions

The present molecular dynamics simulation showed very fast convergence in energy, within about 100ps. Other properties, such as dihedral angle fluctuations and backbone atomic positional fluctuations converge on an intermediate timescale of hundreds of picoseconds. Root mean square deviations of the simulated protein molecules from the starting structure need longer time - about 1ns - to reach a plateau value. Some observed motions suggest that longer simulations are necessary to obtain convergence for all molecular properties in computer simulations. These findings show that the convergence of a quickly relaxed property of the system, such as energies, is not a good indicator of the global convergence of a molecular dynamics simulation.

The GROMOS96 force field used in this simulation turned out to largely reproduce the secondary structure and the relative internal mobility of ubiquitin, overestimating the magnitude of the fluctuations in the most mobile regions of the protein. On the other hand, the different protein molecules translate and rotate relative to one another indicating that the force field would not be able to reproduce the experimental melting temperature of this crystal.

2.6 Acknowledgments

The authors wish to thank Dr. Thomas Huber for fruitful discussions. Financial support was obtained from the Schweizerischer Nationalfonds (project 21-41875.94), which is gratefully acknowledged.

Seite Leer /
Blank leaf

Chapter 3

Molecular Dynamics Simulation of Hen Egg White Lysozyme: a Test of the GROMOS96 Force Field against NMR data

3.1 Abstract

Biomolecular force fields for use in molecular dynamics (MD) simulations of proteins, DNA, or membranes are generally parametrized against ab-initio quantum-chemical and experimental data for small molecules. The application of a force field in a simulation of a biomolecular system, such as a protein in solution, may then serve as a test of the quality and transferability of the force field. Here, we compare various properties obtained from two MD simulations of the protein hen egg white lysozyme (HEWL) in aqueous solution using the latest version, GROMOS96, of the GROMOS force field and an earlier version, GROMOS87+, with data derived from NMR experiments: NOE atom-atom distance bounds, $^3J_{HN\alpha}$-coupling constants and backbone and side-chain order parameters.

The convergence of these quantities over a $2ns$ period is considered, and converged values are compared to experimental ones. The GROMOS96 simulation shows better agreement with the NMR data and also with the X-ray crystal structure of HEWL than the GROMOS87+ simulation, which was based on an earlier version of the GROMOS force field.

3.2 Introduction

Molecular dynamics (MD) computer simulations are increasingly used to gain insight into molecular motion at the atomic level. The accuracy of the simulated properties depends primarily on the quality of the atomic interaction function or force field that is used, and on the extent of sampling of the conformational space and convergence of different quantities obtained within the simulation period. Force fields used in biomolecular simulation can be parametrized in different ways. Some force fields are parametrized using experimental data on small molecules in the condensed phase, e. g. to reproduce properties of liquids, liquid mixtures and small molecules in solution. The parameter set obtained is then used to simulate biologically relevant molecules

or molecular systems, such as proteins or DNA in solution, or membranes. To test the quality of a biomolecular force field, atomic level properties obtainable from NMR experiments on biomolecules, such as NOE distance bounds, ^{3}J-coupling constants, or order parameters S^2, can be computed from a molecular dynamics trajectory and compared to experimental values.

With increasing theoretical and experimental knowledge and computer power also biomolecular force fields evolve and should show increased accuracy. In the present work, a $2ns$ simulation of hen egg white lysozyme (HEWL) based on the GROMOS96 force field [28, 35] is compared to a simulation [36] with an enhanced version, GROMOS87+ [36] of the previous GROMOS87 [37] force field. In an earlier study of HEWL [36], a slight modification (GROMOS87+) of the GROMOS87 parameter set [37] was shown to improve the agreement with NMR data. Here, we show that the GROMOS96 parameter set [28, 35] improves the agreement obtained in [36]. The major differences between the latter two versions of the GROMOS force field are the following (Table 3.1).

In the GROMOS force field, the van der Waals interaction is modelled in terms of a Lennard-Jones function, V_{LJ}, with C12 and C6 parameters,

$$V_{LJ}(r_{ij}) = C12(i,j) \cdot r_{ij}^{-12} - C6(i,j) \cdot r_{ij}^{-6}. \tag{3.1}$$

The parameters for the C12 and C6 pairwise interaction between two (united) atoms i and j at a distance $r_{ij} = |\vec{r}_i - \vec{r}_j|$ are derived from the atomic parameters by a multiplication combination rule,

$$C12(i,j) = (C12(i,i))^{1/2}(C12(j,j))^{1/2}, \tag{3.2}$$

and similarly for the C6 parameters, where different values for $(C12(i,i))^{1/2}$ and $(C12(j,j))^{1/2}$ can be chosen when combining different atom types i and j depending on the nature (purely van der Waals or partly electrostatic) of the pairwise interaction. Calculation of the solvation free energy of alkane molecules in simple point charge (SPC) water [29] led to a change in the $C12(CH_x, O_{water})$ parameter since the alkane solvation energy was too favourable with the GROMOS87 force field. In addition, aromatic hydrogen atoms were treated explicitly in GROMOS87+ in order to mimic the quadrupole moment of aromatic rings. These two changes led to an improved simulation of HEWL in aqueous solution [36]. However, protein simulations using the GROMOS87 and GROMOS87+ force fields showed slightly too much motion of the amino acid peptide planes [36, 38]. Therefore, in GROMOS96, the dihedral angle force constant K_φ for the φ, ψ backbone dihedrals was redetermined and slightly (about $k_BT/4$) increased. The dihedral angle force is the negative derivative with respect to coordinates of the dihedral angle potential energy term

$$V^{trig}(\varphi) = K_\varphi[1 + cos(\delta)cos(m\varphi)] \tag{3.3}$$

in which δ is a phase shift (0 or 180°) and m the multiplicity (6 for φ and ψ) of the selected dihedral angle φ. Finally, the van der Waals parameters of CH_x united atoms were redetermined by fitting of the simulated heat of vaporisation and density to experimental values for liquid n-alkanes [39]. The major differences between the GROMOS87 [37], GROMOS87+ [36], and GROMOS96 [28, 35] force fields have been summarized in table 3.1.

A second condition for obtaining accurate prediction of properties from a molecular simulation is sufficient sampling of conformational space, leading to converged values of different atomic and molecular or system properties. Here, the convergence of calculated NOE distance bounds, ^{3}J-coupling constants, and main- and side-chain ^{1}H-^{15}N order parameters is analysed by calculating the selected property using different parts of the simulation trajectory and time averages extending over $800ps$ and $1700ps$. When comparing the properties of HEWL calculated

Force field Force field version code	GROMOS87 37C4	GROMOS87+ 37C4+	GROMOS96 43A1
Explicit treatment of aromatic H-atoms			
charge $q_{HC} = -q_C$	-	0.14	0.10
$(C12(HC,HC))^{1/2}$	-	0.123	0.123
$(C6(HC,HC))^{1/2}$	-	0.0092	0.0092
van der Waals parameters for use in (polar, non-polar) pairs			
$(C12(OA,OA))^{1/2}$	0.8611	0.8611	1.125
$(C12(OW,OW))^{1/2}$	0.8611	1.623	1.544
for use in any pair			
$(C12(CH1,CH1))^{1/2}$	8.470	8.470	3.373
$(C6(CH1,CH1))^{1/2}$	0.1118	0.1118	0.06148
$(C12(CH2,CH2))^{1/2}$	5.944	5.944	5.077
$(C6(CH2,CH2))^{1/2}$	0.09538	0.09538	0.08429
$(C12(CH3,CH3))^{1/2}$	5.114	5.114	5.794
$(C6(CH3,CH3))^{1/2}$	0.09421	0.09421	0.09958
Dihedral angle torsional parameters for dihedrals of type -N-CH_x- and -CH_x-C-			
force constant K_φ	0.4184	0.4184	1
phase shift δ for -N-CH_x-	180	180	180
phase shift δ for -CH_x-C-	0	0	0
multiplicity m	6	6	6

Table 3.1: *Major differences between the three GROMOS force fields for biomolecular simulation, GROMOS87 [37], GROMOS87+ [36], and GROMOS96 [28, 35]. GROMOS non-bonded atom type names: HC, hydrogen bound to carbon; C, bare carbon; OA, hydroxyl, sugar or ester oxygen; OW, water oxygen; CH1, aliphatic or sugar CH-group or united atom; CH2, aliphatic or sugar CH_2-group; CH3, aliphatic or sugar CH_3-group; N, peptide nitrogen or N_ε in ARG; Units: charge in e; $(C12)^{1/2}$ in 10^{-3} $(kJmol^{-1}nm^{12})^{1/2}$; $(C6)^{1/2}$ in $(kJmol^{-1}nm^6)^{1/2}$; K_φ in $kJmol^{-1}$; δ in degree; For further explanation see text and equations 3.1 - 3.3. Minor differences involve the functional forms of the covalent bond-stretching and bond-angle bending interactions and the definition of charge groups between which the non-bonded interaction is calculated.*

with the GROMOS96 force field to the ones calculated in [36] using the GROMOS87+ force field, the trajectory between $300ps$ and $1100ps$ was used as in [36].

3.3 Methods

Lysozyme consists of 129 amino acids with 1001 non-hydrogen atoms. Hydrogen atoms attached to aliphatic carbon atoms are incorporated into these (the united atom approach), and the remaining 321 hydrogen atoms are treated explicitly. The protein was simulated at pH 6. The amino

Simulation of HEWL	previous [36]	present
Force field (version code)	GROMOS87+(37C4+)	GROMOS96(43A1)
Number of atoms or molecules		
solute atoms	1321	1322
water molecules	4463	7113
counter ions (Cl^-)	0	9
total	14710	22670
Truncated octahedron box length (nm)	6.71	7.7392
PDB code starting structure	2LZT	1AKI
Simulation length (ps)	1100	2000
Simulation time step (fs)	2	2
Treatment of long range forces		
pair-list cut-off radius R_{cp} (nm)	0.8	0.8
pair-list update frequency (ps^{-1})	100	100
non-bonded interaction cut-off R_{cl} (nm)	1.4	1.4
Poisson-Boltzmann reaction field (PBRF)	no	yes
beyond radius R_{rf} (nm)	-	1.4
using relative dielectric permittivity ε_{rf}	-	54

Table 3.2: *Major differences between the simulation parameters and set-up of the previous [36] and present MD simulations of HEWL. See also table 3.1 and the text. PDB: protein data bank.*

acids GLU and ASP were taken to be deprotonated, LYS, ARG, and HIS residues were protonated, leading to a charge of +9 electron charges per chain. The crystal structure of lysozyme (entry 1AKI [40] of the Brookhaven Protein Database [31]) determined at 1.5Å resolution [41] was used as a starting structure. Truncated octahedron boundary conditions were used with a box length of $7.7392nm$ between the quadratic surfaces. 7122 SPC water molecules [29] were added from an equilibrated cubic box containing 216 water molecules [28]. The added water molecules were selected such that no water oxygen atom is closer than $0.23nm$ to a non-hydrogen atom of the protein or another water oxygen atom. The system, protein and water, was initially energy minimised for 100 cycles using the steepest descent method. The protein atoms were harmonically restrained [28] to their initial positions with a force constant of $25000kJ/(mol\ nm^2)$. The minimised structure was then pre-equilibrated in a short MD run of 100 steps of $0.002ps$, still restraining protein atom positions. Initial velocities were assigned from a Maxwell-Boltzmann distribution at 300K. Protein and solvent were coupled separately to temperature baths of 300K with a coupling time of $0.1ps$ [6]. No pressure coupling was applied. A follow-up simulation (results not shown) including pressure coupling showed no significant change in the box volume. Bonds were kept rigid using the SHAKE method [3] with a relative geometric tolerance of 10^{-4}. Long-range forces were treated using twin-range cutoff radii $R_{cp} = 0.8nm$ for the charge-group [28] pair-list and $R_{cl} = 1.4nm$ for the longer-range non-bonded [32] interactions. The pair-list for the (short-range) non-bonded interactions and the longer-range forces were updated every $10fs$. Reaction-field forces were included [42] originating from a dielectric continuum beyond a radius of $R_{rf} = 1.4nm$ using a self-consistent relative dielectric permittivity $\varepsilon_{rf} = 54$ for SPC water [43]. Nine counter ions were added by replacing water molecules. The water molecule having the highest electrostatic field at the water oxygen site, after the successive placement of previous ions, was replaced by a chloride anion. After having introduced the ions, the energy

was again minimised using 100 steps of steepest descent and protein atom position constraining. The total size of the system was 1322 protein atoms, 9 chloride anions and 7113 water molecules leading to a total system size of 22670 atoms. The simulation was carried out over $2ns$ with a time step of $0.002ps$. Every 500th step, the configuration was saved. The first $300ps$ of the simulation were treated as equilibration period, different time periods of the remaining $1.7ns$ were used for analysis. All simulations were performed using the GROMOS96 package and force field [28, 44].

The 1158 NOE distances, the 95 $^3J_{HN\alpha}$-coupling constants, and the 124 backbone and 28 side-chain S^2 order parameters were calculated from the trajectory as in [36]. The experimental values were taken from [45–47].

3.4 Results

3.4.1 Convergence of calculated properties

In figure 3.1, the root-mean-square deviation (RMSD) of the simulated structure from the X-Ray structure is shown. After an initial increase, a first plateau is reached, where the RMSD stabilises at $0.14nm$ for Cα atoms and $0.21nm$ for all atoms. The plateau reaches from $100ps$ to $700ps$. After that, the RMSD increases to $0.18nm$ for Cα atoms and $0.26nm$ for all atoms. After $800ps$, until the end of the simulation at $2000ps$, the values fluctuate around these values.

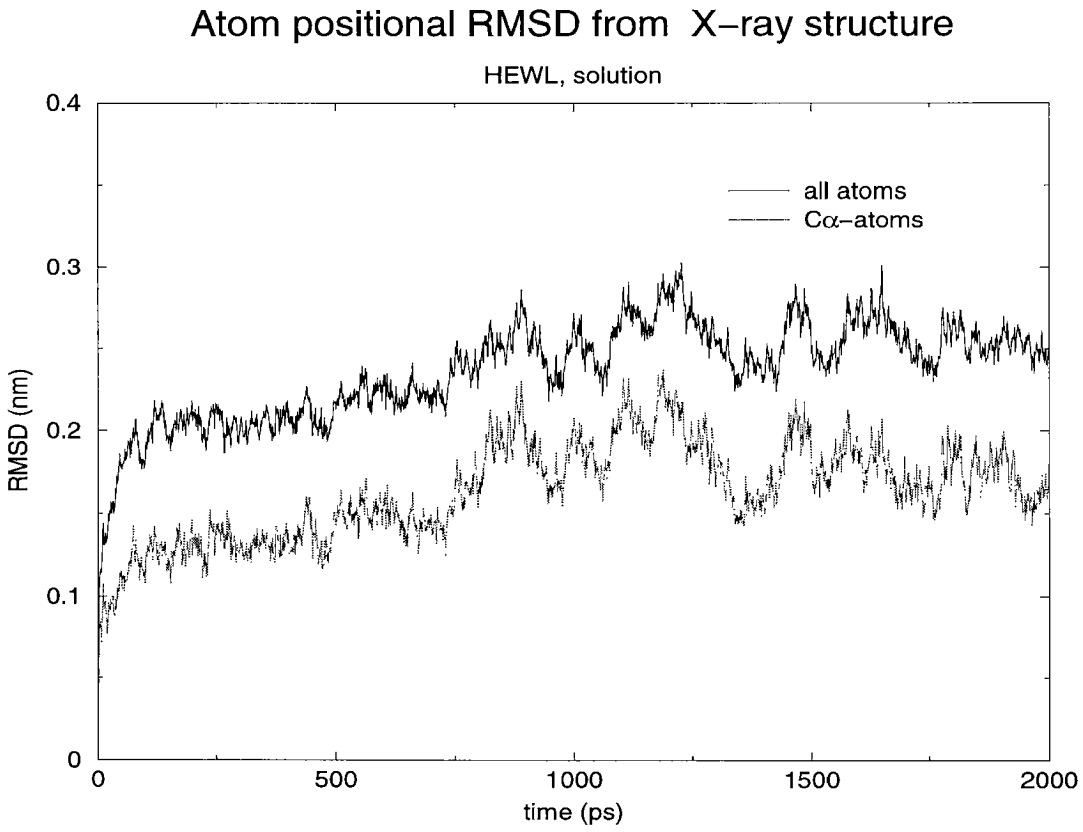

Figure 3.1: *Root-mean-square atom positional deviation (RMSD) in nm from the X-ray (crystal) structure [40] as function of time in ps. Rotational and translational fitting was applied using all 129 Cα atoms. RMSD of Cα atoms (dotted line) and of all atoms (solid line) are shown.*

NOE distance bound violations are summarized in table 3.3. The average over the final $800ps$ of the simulation shows more and larger violations than the one over $300-1100ps$. Also the mean violation increases. As for calculating distance violations, r^{-3} averaging is used, a few configurations with a low r-distance strongly determine the average, violations do not increase significantly comparing the periods $300-1100ps$ and $300-2000ps$. The r^{-3} averaging makes NOE distance violations a quickly converging property of the system.

	number of violations			mean violation
Simulation	$> 0.05nm$	$> 0.1nm$	$> 0.3nm$	R_E - R_0
GROMOS87+ (W2 [36]) ($300-1100ps$)	94	54	7	0.14
GROMOS96 ($300-1100ps$)	77	45	8	0.11
GROMOS96 ($1200-2000ps$)	88	65	14	0.14
GROMOS96 ($300-2000ps$)	75	49	10	0.12

Examples of NOE violations

NMR data		exp	GROMOS87+ [36]		GROMOS96					
			$300-1100ps$		$300-1100ps$		$1200-2000ps$		$300-2000ps$	
NOE pair		R_0	R_E	R_E - R_0	R_E	R_E - R_0	R_E	R_E - R_0	R_E	R_E - R_0
2Hγ2	- 38HN	0.40	0.58	0.18	0.79	0.39	0.76	0.36	0.78	0.38
8Hγ	- 17Hγ	0.47	1.04	0.57	0.97	0.50	0.99	0.52	0.98	0.51
15Hε1	- 88Hγ1	0.54	0.95	0.41	0.73	0.19	0.85	0.31	0.78	0.24
15Hε1	- 88Hδ	0.40	0.84	0.44	0.62	0.22	0.82	0.42	0.70	0.30
23Hζ	- 104HN	0.76	1.13	0.37	0.58	0	0.53	0	0.56	0
28Hε3	- 88Hδ	0.85	0.91	0.06	1.19	0.34	0.64	0	0.65	0
92Hγ2	- 96Hα	0.40	0.64	0.24	0.87	0.47	0.87	0.47	0.87	0.47
95Hα	- 108Hζ2	0.30	0.68	0.38	0.62	0.32	0.61	0.31	0.62	0.32
95Hβ	- 108Hζ2	0.40	0.85	0.45	0.75	0.35	0.75	0.35	0.75	0.35
95Hβ	- 108Hη2	0.35	0.84	0.49	0.73	0.38	0.78	0.43	0.75	0.40
99Hβ	- 108Hζ3	0.45	0.74	0.29	0.85	0.40	0.86	0.41	0.84	0.39

Table 3.3: *NOE distances and bound violations computed from MD trajectories and experimentally determined NOE distance bounds (1158). Results are given for an earlier simulation [36] and three analysis periods of the present simulation. R_0 is the experimentally determined distance bound [45–47]. R_E is the distance from the trajectory using r^{-3} averaging. The symbol R_E-R_0 indicates a (mean) violation, calculated as a mean of the larger value of R_E-R_0 and zero. So, 0 indicates no violation. Distances and NOE bounds are in nm*

Most of the 95 $^3J_{HN\alpha}$ coupling constants do not change very much between the two $800ps$ analysis periods. Exceptions are the coupling constants of residues VAL2 ($10.0Hz$), ASN65 ($9.4Hz$), ASP101 ($7.0Hz$), and ASN103 ($8.2Hz$). Also the coupling constants from the final $1700ps$, the whole analysis period, are, except for the mentioned residues, close to the ones from the shorter periods. Torsional angle librations within a potential energy well have a small effect on mean calculated ^{3}J-coupling constants [48]. So, if no jumps over rotational barriers occur, averaged ^{3}J-coupling constant values are converged on a subnanosecond to nanosecond time-scale.

The same holds for backbone ^{1}H-^{15}N order parameters (S^2). S^2 values calculated for different time periods (Fig. 3.3) are very similar to each other with a handful of exceptions. This behaviour

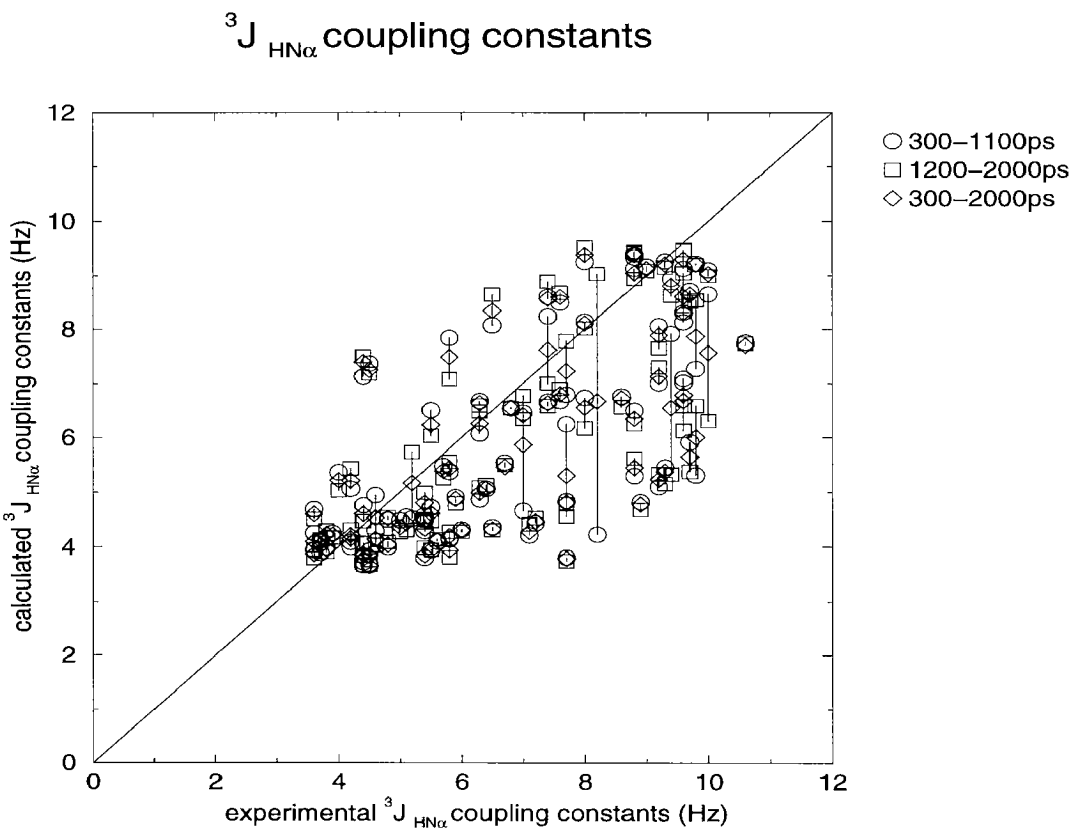

Figure 3.2: *Comparison of 95 experimental [45] and calculated $^3J_{HN\alpha}$ coupling constants (in Hz) averaged over different time windows:* ○ $300-1100ps$; □ $1200-2000ps$; ◇ $300-2000ps$. *Corresponding calculated $^3J_{HN\alpha}$ coupling constants are connected with lines.*

is expected for N-H bond vectors that sample the same region of configuration space in both of the shorter analysis periods. If the order parameter of the longer analysis period is about the average of the ones for the two $800ps$ subperiods, the extent of motion in one of the two subperiods (the one with low S^2 values) encompasses the extent of motion in the other subperiod. For some residues, e.g. SER72, ILE88, or CYS127, the first subperiod shows larger disorder, for other residues, e.g. ALA42, SER50, ILE78, or GLY102, the second subperiod does. If, on the other hand, a torsional angle transition over a potential energy barrier occurred between the two subperiods or around the end of the first subperiod or around the beginning of the second subperiod, the S^2 value averaged over both subperiods can indicate less order than is found within the sub-periods, as is the case for residues GLY67, ASN103, and THR118. Order parameters of ^1H-^{15}N bonds in side chains (Fig. 3.4) calculated using different time windows differ more from one another than in the backbone. This reflects the larger mobility, and thus, that a larger region of configuration space is accessible to these bond vectors. Some of the values (e.g. TRP28(0.90), ASN46(0.62), ASN59(0.78), or TRP61(0.28), ARG125(0.05), experimental S^2 values between parentheses) are converged within $800ps$, indicating sufficient sampling for the particular side chain, but for other residues (e.g. GLN57(0.82), ASN103(0.26), or ASN113(0.47)), sampling periods longer than a few nanoseconds are required.

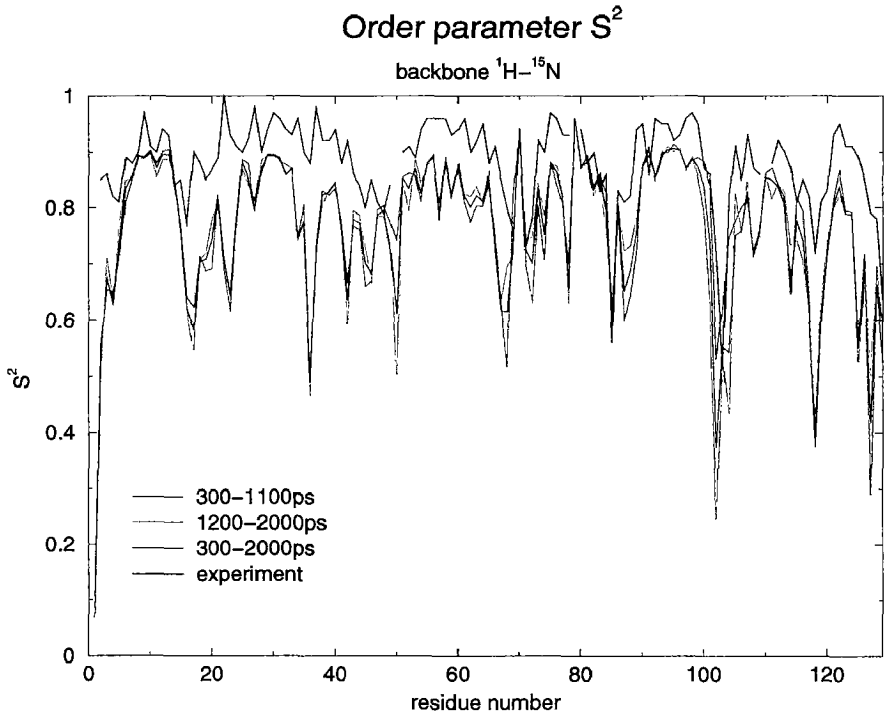

Figure 3.3: *Backbone ^1H-^{15}N order parameters (S^2) as function of residue number. Averages over different analysis time windows are shown together with experimental data [47]. There is no experimental value available for residues LYS1, SER50, PRO70, PRO79, and ALA110. Blue line, exp; green line, $300-1100ps$; red line, $1200-2000ps$; black line, $300-2000ps$*

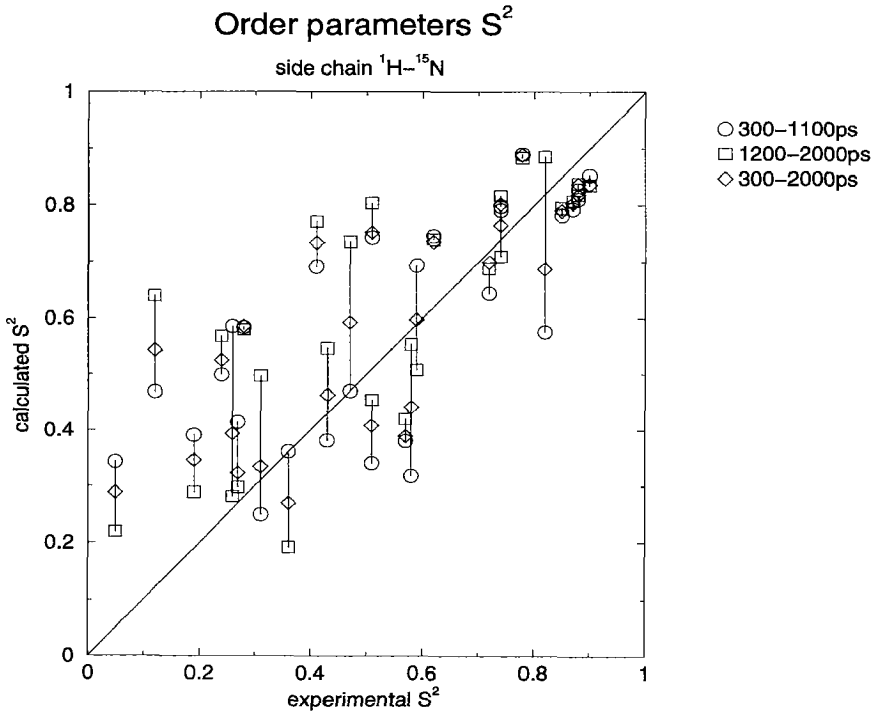

Figure 3.4: *Comparison of 28 experimental [47] ^1H-^{15}N order parameters (S^2) of side chain NH groups with values calculated over different time windows: ○ $300-1100ps$; □ $1200-2000ps$; ◇ $300-2000ps$. Corresponding calculated S^2 are connected with lines. For NH_2 groups the average of the order parameters for the two NH vectors is displayed*

3.4.2 Comparison of the GROMOS87+ and GROMOS96 simulations with experimental data

When comparing the $2000ps$ GROMOS96 simulation with the $1100ps$ GROMOS87+ simulation of [36], a corresponding analysis period, $300-1100ps$, is used, if not stated otherwise. The atom positional RMSD from the X-ray structure (Fig. 3.5) is smaller in the GROMOS96 simulation than in the GROMOS87+ one showing that the re-engineered force field keeps the trajectory closer to the experimentally derived crystal structure. If structures are averaged over $50ps$ windows and the atom positional RMSD from the X-ray structure is calculated for these averaged structures (Fig. 3.5), high-frequency motions are averaged out and the RMSD values decrease slightly compared to when single structures are considered (Fig. 3.1). In this case, the average RMSD values for the $800-2000ps$ period are $0.18nm$ for Cα atoms and $0.24nm$ for all atoms.

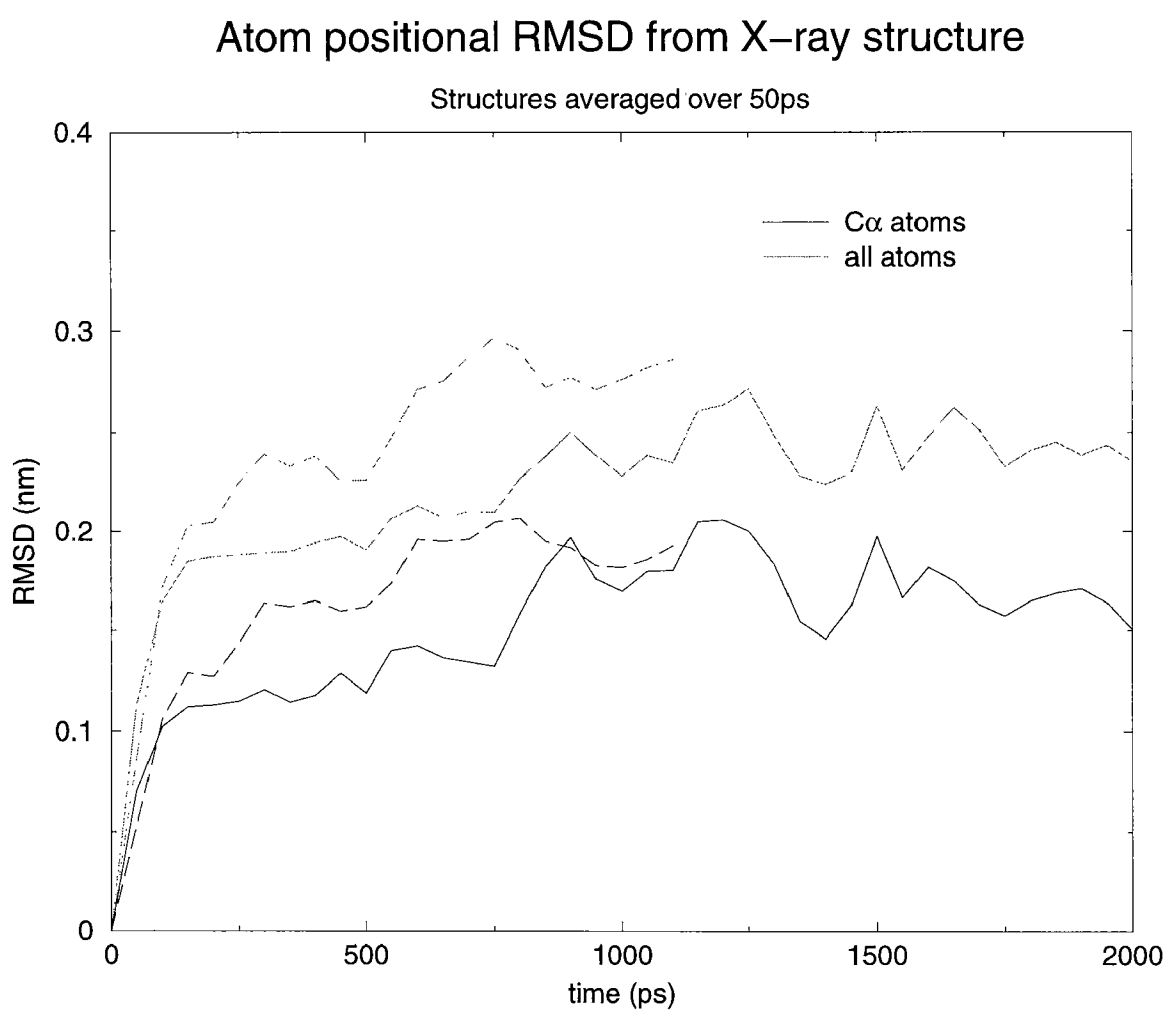

Figure 3.5: *Root-mean-square atom positional deviation (RMSD) in nm from the X-ray (crystal) structure as function of time in ps. Rotational and translational fitting was applied using all 129 Cα atoms. Structures were averaged over 50ps windows. RMSD of Cα atoms (solid line) and of all atoms (dashed line) of these averages are shown for the present (GROMOS96) simulation (thick lines) and the earlier (GROMOS87+(W2)) simulation [36] (thin lines).*

For NOE distance violations, an improvement compared to the previous simulation is observed (Table 3.3). For the GROMOS96 simulation, the average violation is $0.011nm$ compared to $0.014nm$ for the GROMOS87+ one.

$^3J_{HN\alpha}$ coupling constants calculated in the GROMOS96 simulation reproduce the 95 experimental values with a root-mean-square-deviation of $1.72Hz$ compared to $1.83Hz$ in the GROMOS87+ simulation (Fig. 3.6). These values depend on the potential energy force field term for the backbone dihedral angle φ, for which the force constant was slightly increased between the two force fields used.

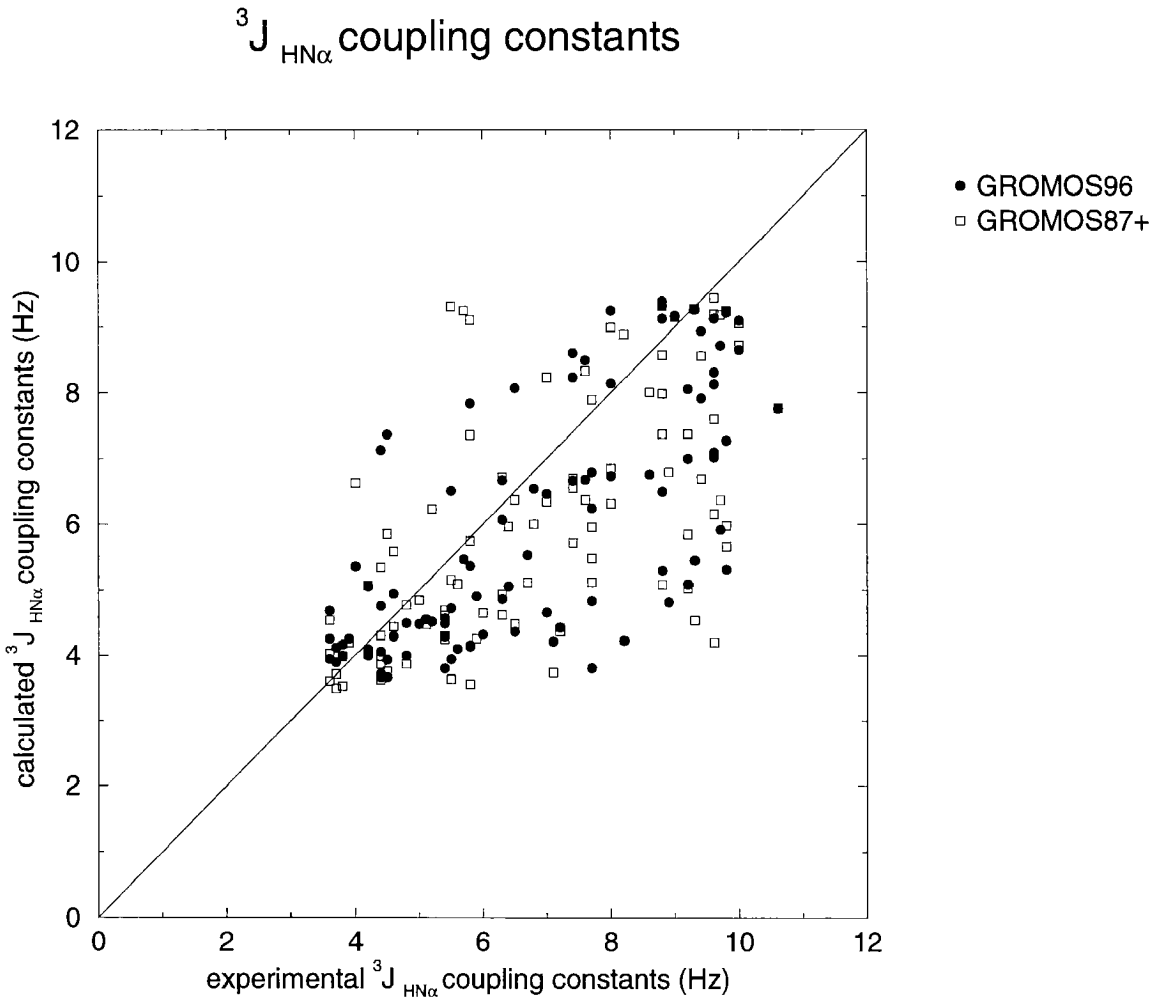

Figure 3.6: *Comparison of 95 experimental [45] and calculated $^3J_{HN\alpha}$ coupling constants for the GROMOS96 (•) and the GROMOS87+ (□) simulations, averaged over* $300 - 1100ps$

Backbone order parameters (Fig. 3.7) show a more significant improvement from the GROMOS87+ simulation to the GROMOS96 one when comparing with S^2 derived from experiment. The values in regions of stable secondary structure elements are better reproduced in the GROMOS96 simulation.

The side-chain order parameters calculated from the GROMOS87+ and GROMOS96 simulations (Fig. 3.8) show similar degree of agreement with experimental data. Side-chain order parameters are more sensitive to the length of the averaging period (Fig. 3.4). So, better agreement with experiment is not only to be expected from force field improvement, but also from longer averaging periods.

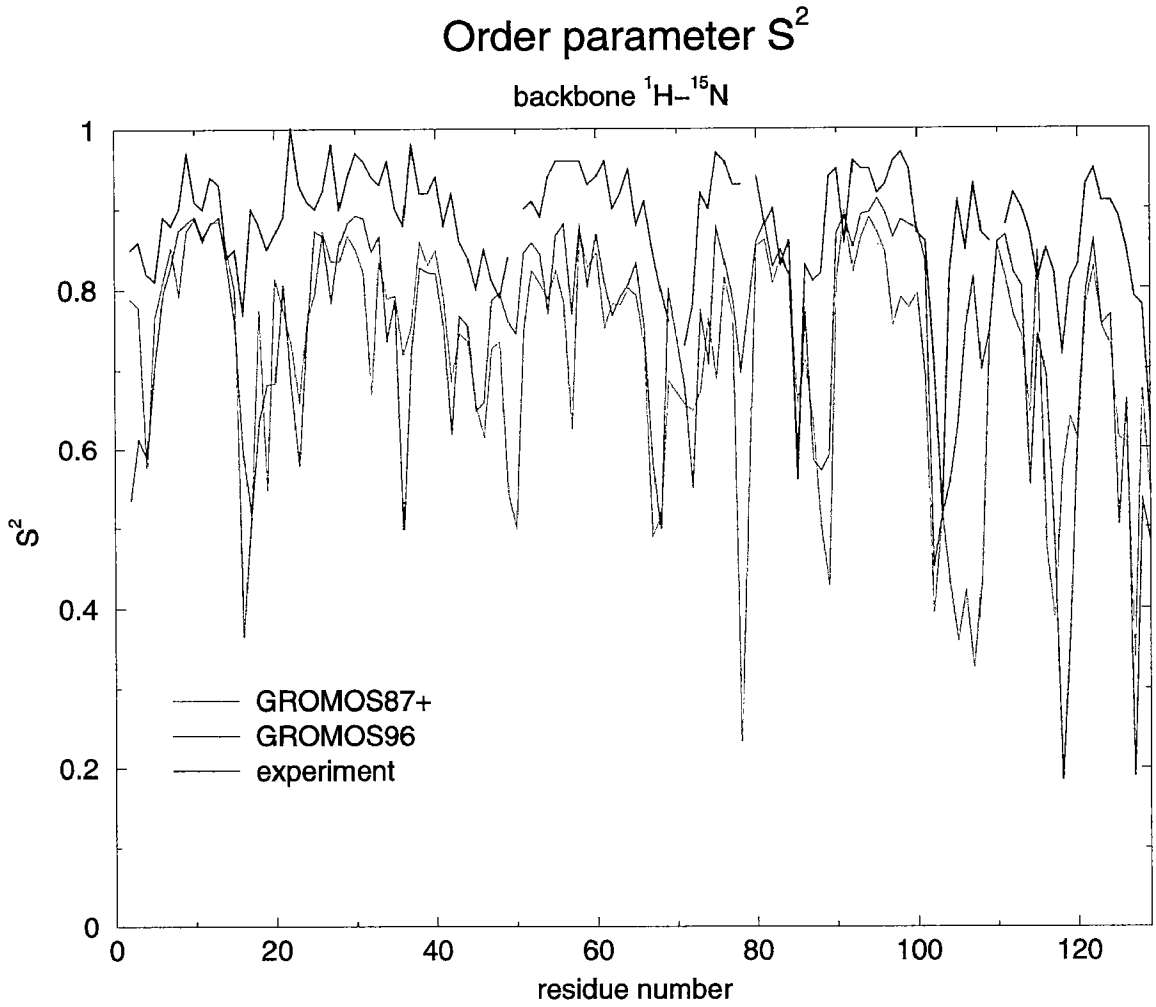

Figure 3.7: *Backbone ^1H-^{15}N order parameters (S^2) as function of residue number. The values determined over the $300-1100ps$ periods of the GROMOS87+ and GROMOS96 simulations are compared with experimental data [47]. There is no experimental value available for residues LYS1, SER50, PRO70, PRO79, and ALA110. Blue line, exp; green line, GROMOS96; red line, GROMOS87+*

3.5 Discussion

Although continuously increasing computer power enables increasingly longer simulations of biomolecular systems, such as proteins or DNA in solution and membranes, the attainable simulation periods are still too short, a few nanoseconds, for all system and atomic properties to converge. Fast relaxing properties, such as system energies, the macromolecular radius of gyration (results not shown) generally converge within a few $100ps$. The atom-positional root-mean-square deviation from the starting (X-ray) structure usually converges after a nanosecond. Properties depending on rare structural transitions may even take longer to converge and are, therefore, still out of simulation range. $^3J_{HN\alpha}$-coupling constants and backbone ^{1}H-^{15}N order parameters generally converge within $1-2ns$, whereas the sampling for order parameters of side-chain bonds needs longer time.

The GROMOS96 force field is able to maintain the experimentally derived X-ray structure better than the previous versions (GROMOS87, GROMOS87+) of the force field. This is due to

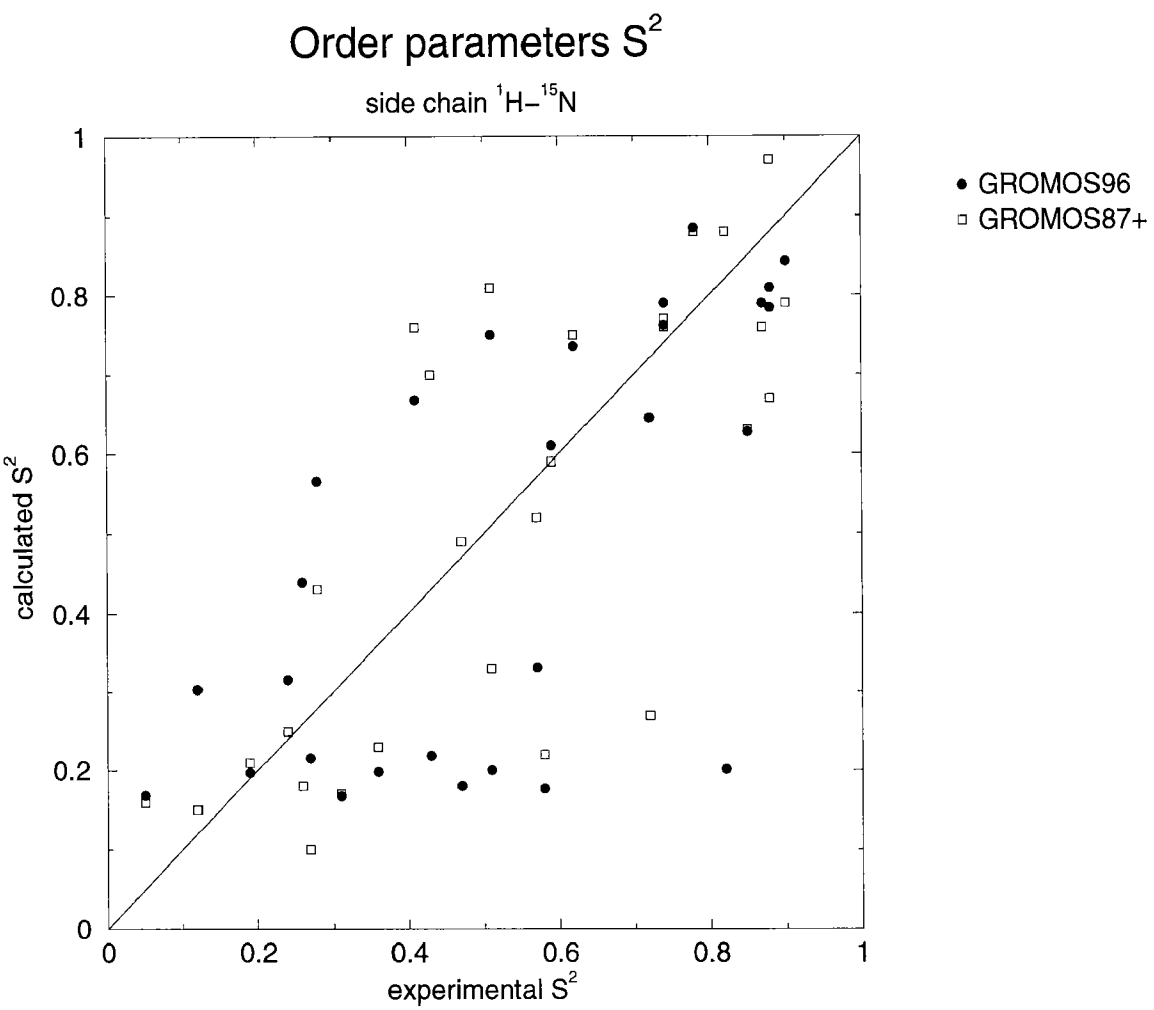

Figure 3.8: *Comparison of 28 experimental [47] ^1H-^{15}N order parameters S^2 of side-chain NH groups with values calculated using the $300-1100ps$ period of the GROMOS96(•) and GROMOS87(□) simulations. For NH_2 groups the average of the order parameters for the two NH vectors is displayed*

the improvement made in the GROMOS force field:(i) The packing of aromatic rings has been improved by the introduction of explicit hydrogens with partial charges generating a quadrupole moment;(ii) The balance between apolar-apolar, apolar-polar, and polar-polar pair interactions has been improved by a refitting of van der Waals parameters to liquid data [39];(iii)The mobility of φ and ψ torsional angles has been slightly reduced by a slight increase of the torsional angle interaction force constants.

The agreement with the 1158 NOE bounds derived from experiment is only marginally improved going from the GROMOS87+ to the GROMOS96 force field. This is not surprising since satisfying NOE bounds is not very sensitive to details of a force field. As expected, the 95 experimental $^3J_{HN\alpha}$-coupling constants and 124 order parameters for the backbone are slightly better reproduced by the GROMOS96 force field than by the GROMOS87+ one. The agreement with the 28 side-chain order parameters derived from NMR experiments is equally good for both force field versions.

The present results for hen egg white lysozyme illustrate the progress that is still made in the field of force field development for biomolecular systems. They also illustrate the different

sensitivities of different quantities to particular changes in force field parameters and to the extent of sampling of conformational space.

3.6 Acknowledgments

The authors wish to thank Katrin Spiegel and Dr. Lorna J. Smith for making data available and useful discussions. Financial support was obtained from the Schweizerischer Nationalfonds, project number 21-41875.94, which is gratefully acknowledged.

Seite Leer /
Blank leaf

Chapter 4

On the Similarity of Properties in Solution or Crystalline State: a Molecular Dynamics Study of Hen Lysozyme

4.1 Abstract

As protein crystals generally possess a high water content, it is assumed that the behaviour of a protein in solution and in crystal environment is very similar. This assumption can be investigated by molecular dynamics (MD) simulation of proteins in the different environments.

Two $2ns$ simulations of hen egg white lysozyme (HEWL) in crystal and solution environment are compared to one another and to experimental data derived from both, X-ray and NMR experiments, such as crystallographic B-factors, NOE atom-atom distance bounds, $^3J_{HN\alpha}$-coupling constants, and ^{1}H-^{15}N bond vector order parameters. Both MD simulations give very similar results. The crystal simulation reproduces X-ray and NMR data slightly better than the solution simulation.

4.2 Introduction

Molecular dynamics (MD) simulations of biomolecular systems, such as proteins, DNA, or membranes, are generally performed in their natural environment, in aqueous solution, or, in the case of membranes, as a bilayer with water on both sides or as micelles. Most protein structures, however, are determined by X-ray crystallography, where the molecule of interest is in a different environment. In contrast to small molecule crystals, in protein crystals, the water content is high (generally around 50%). In addition, in protein crystals, only few water molecules occupy well-defined sites, most of the solvent is disordered and thus comparable to the protein environment in solution. Third, if proteins are crystallized in different space groups, the structures are usually not very different from one another. Fourth, for a number of small proteins the structure has been determined in the crystalline state by X-ray diffraction as well as in aqueous solution by NMR spectroscopy [49–60] and structural differences appear to be minor, e. g. involving loop or side chain conformations. So, it is generally assumed that proteins in crystals and in solution have, apart from polar side-chains that take part in crystal packing contacts, a very similar struc-

ture in the two different environments. The motions in the crystal, especially side-chain motions, are expected to be of lesser amplitude due to crystal packing contacts.

Molecular dynamics simulation provides a good tool to investigate structural and dynamical differences on a sub-nanosecond time-scale. In an early study [61] compared the structural properties of bovine pancreatic trypsin inhibitor (BPTI) on a picosecond time scale by comparing the results of two $25ps$ MD simulations of BPTI, one in crystalline form and the other in aqueous solution. To our knowledge, however, since then no systematic investigation of simulations of a protein in crystal and solution, with a detailed analysis of the differences observed, has been published. In the present work, two $2ns$ MD simulations of hen egg white lysozyme (HEWL) are compared. An orthorhombic unit cell containing 4 protein molecules was simulated, and the results compared to those of a simulation of the protein in solution.

Data computed from the two simulations are compared with data derived from X-ray crystallography [40], such as atomic positions, B-factors, and hydrogen bonding patterns, or with data derived from NMR experiments [45–47], such as NOE atom-atom distance bounds, $^3J_{HN\alpha}$-coupling constants, and ^{1}H-^{15}N bond vector order parameters. In addition, the data from the two simulations are compared to each other.

4.3 Methods

4.3.1 Solution simulation

Lysozyme consists of 129 amino acids with 1001 non-hydrogen atoms. Hydrogen atoms attached to aliphatic carbon atoms are incorporated into these (the united atom approach), and the remaining 321 hydrogen atoms are treated explicitly. The protein was simulated at pH 6. The amino acids GLU and ASP were taken to be deprotonated, LYS, ARG, and HIS residues were protonated, leading to a charge of +9 electron charges per protein molecule. The crystal structure of lysozyme (entry 1AKI [40] of the Brookhaven Protein Database [31]) determined at 1.5Å resolution [41] was used as a starting structure. Truncated octahedron periodic boundary conditions were used with a box length of $7.7392nm$ between the quadratic surfaces. 7122 SPC water molecules [29] were added from an equilibrated cubic box containing 216 water molecules [28]. The added water molecules were selected such that no water oxygen atom is closer than $0.23nm$ to a non-hydrogen atom of the protein or another water oxygen atom. The system, protein and water, was initially energy minimised for 100 cycles using the steepest descent method. The protein atoms were harmonically restrained [28] to their initial positions with a force constant of $25000kJ/(mol\ nm^2)$. The minimised structure was then pre-equilibrated in a short MD run of 100 steps of $0.002ps$, still restraining protein atom positions. Initial velocities were assigned from a Maxwell-Boltzmann distribution at 300K. Protein and solvent were coupled separately to temperature baths of 300K with a coupling time of $0.1ps$ [6]. No pressure coupling was applied. A follow-up simulation (results not shown) including pressure coupling showed no significant change in the box volume. Bonds were kept rigid using the SHAKE method [3] with a relative geometric tolerance of 10^{-4}. Long-range forces were treated using twin-range cut-off radii $R_{cp} = 0.8nm$ for the charge-group [28] pair-list and $R_{cl} = 1.4nm$ for the longer-range non-bonded [32] interactions. The pair-list for the (short-range) non-bonded interactions and the longer-range forces were updated every $10fs$. Reaction-field forces were included [42] originating from a dielectric continuum beyond a radius of $R_{rf} = 1.4nm$ using a self-consistent relative dielectric permittivity $\varepsilon_{rf} = 54$ for SPC water [43]. Nine counter-ions were added by replac-

ing water molecules in the following way. The water molecule having the highest electrostatic field at the water oxygen site, after the successive placement of previous ions, was replaced by a chloride anion. After having introduced the ions, the energy was again minimised using 100 steps of steepest descent and protein atom position constraining. The total size of the system was 1322 protein atoms, 9 chloride anions and 7113 water molecules leading to a total system size of 22670 atoms. The simulation was carried out over $2ns$ with a time step of $0.002ps$. Every 500th step, the configuration was saved. The first $300ps$ of the simulation were treated as equilibration period, the remaining $1.7ns$ were used for analysis. All simulations were performed using the GROMOS96 force field (version 43A1) and software [28, 44]. The simulation parameters have been summarized in table 4.1.

simulation type	crystal	solution
protein molecules	4	1
counterions (Cl^-)	36	9
water molecules	1715	7113
number of atoms	10469	22670
periodic boundary conditions	orthorhombic	truncated octahedron
box lengths (nm)	5.9062	7.7392
	6.8451	7.7392
	3.0517	7.7392
box volume (nm^3)	123.38	231.77
mass density (gcm^{-3})	1.203	1.023
water content (by mass, %)	34.57	89.76
timestep (fs)	2	
relative SHAKE precision	10^{-4}	
temperature coupling at (K)	300	
using coupling time (ps)	0.1	
pressure coupling	no	
pair-list cut-off radius R_{cp} (nm)	0.8	
pair-list update frequency (ps^{-1})	100	
non-bonded interaction cut-off R_{cl} (nm)	1.4	
Poisson-Boltzmann reaction field (PBRF)	yes	
beyond radius R_{rf} (nm)	1.4	
using relative dielectric permittivity ε_{rf}	54	

Table 4.1: *Simulation parameters used in the crystal and the solution simulation of hen egg white lysozyme*

4.3.2 Crystal simulation

For the crystal simulation, basically the same set-up was used as for the solution simulation. The differences are the geometry of the periodic box containing the simulated system and the number of protein molecules, ions, and water molecules. Four protein molecules related by the crystallographic symmetry $P2_12_12_1$ were placed in the orthorhombic unit cell with a=5.9062nm, b=6.8451nm, and c=3.0517nm. 78 crystallographically observed water molecules were added

together with their symmetry-related equivalents. After that, 1439 water molecules were added as described in the previous subsection leading to a water content of 42.5% (v/v). Energy minimisation, placing of the ions, and pre-equilibration simulations were performed as described in the previous subsection. The simulation parameters have been summarized in table 4.1.

4.3.3 Analysis

The 1158 NOE distances, the 95 $^3J_{HN\alpha}$-coupling constants, and the 124 backbone and 28 side-chain S^2 order parameters were calculated from the trajectory as in [36]. The S^2 order parameters were calculated using a $200ps$ averaging window (moving through the whole $1700ps$ analysis period) which approximates the time scale of the N-H bond vector motions which determine the S^2 order parameters as derived from the NMR experiments for HEWL [62]. Similarly, $200ps$ windows were used to compute the r^{-3} averages for the NOE distances. The experimental values were taken from [45–47].

4.4 Results

Root-mean-square atom positional deviations (RMSD) from the starting (X-ray) structure are shown in figure 4.1. In the top graph, values for C_α atoms of the solution simulation and of the four molecules in the crystal simulation are displayed, in the bottom graph values for all atoms. RMSD values for the solution simulation are converged after $800ps$ for both, C_α and all atoms. In the crystal simulation, the RMSD of C_α atoms are converged earlier, after $500ps$, whereas the side-chain RMSD only converge after $1ns$. The crystal simulation stays - in terms of RMSD - closer to the X-ray structure than the solution simulation. In the second nanosecond of the simulations, the RMSD for C_α atoms is $0.13nm$ compared to $0.16nm$, the RMSD for all atoms is $0.22nm$ compared to $0.26nm$ in the crystal and the solution simulation, respectively. The different molecules in the crystal simulation show, after being converged, very similar behaviour. In the solution simulation, larger structural fluctuations are observed, which should be reflected in the calculated B-factors (fig. 4.2). Atomic isotropic B-factors from simulation trajectories were calculated from atom-positional root-mean-square fluctuations for C_α atoms according to

$$B_i = \frac{8\pi}{3}\langle(\vec{r}_i - \langle\vec{r}_i\rangle)^2\rangle. \tag{4.1}$$

When comparing atomic B-factors resulting from structure refinement based on crystallographic X-ray diffraction data with B-factors obtained through (4.1) from the atomic mean square positional fluctuations calculated from a MD trajectory, it should be kept in mind, that these two types of B-factors are not wholly comparable [63].

(1) The configuration space sampled in the experiment is incompletely sampled in the simulation since: (i) crystallographic B-factors derived from X-ray data include static disorder due to averaging over a collection of molecules; (ii) the simulation time (nanoseconds) is much shorter than the data aquisition time (>seconds).

(2) Crystallographic B-factors form an incomplete measure of the motion since: (i) they are a measure of the spread of the electron density as a function of position in the crystal, irrespective of which particular atom is contributing to the electron density (i. e. they result from fitting on the electron density map); (ii) they are, in general, restricted to a given maximum value during structure refinement, whereas the simulated B-factors are true atomic positional fluctuations,

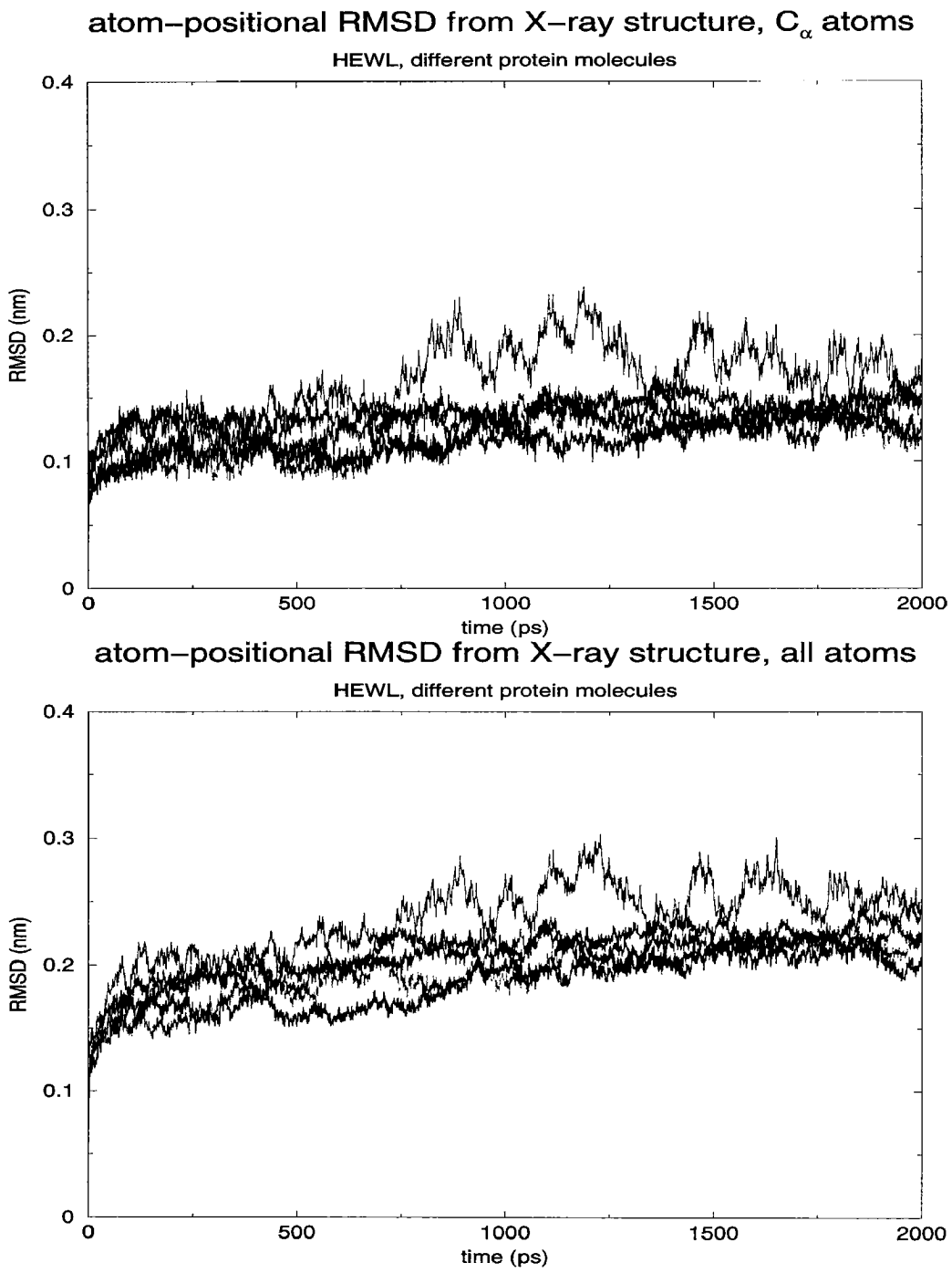

Figure 4.1: *Root-mean-square atom positional deviation (RMSD) in nm from the X-ray (crystal) structure (Carter et al., 1997) as function of time in ps. Rotational and translational fitting was applied using all 129 C_α atoms. RMSD of C_α atoms (top graph) and of all atoms (bottom graph) are shown. RMSD of the solution simulation are shown in indigo, of the four chains in the crystal simulations are shown in black (molecule 1), red (molecule 2), green (molecule 3), and blue (molecule 4).*

which may in principle grow infinitely with atomic mobility; (iii) systematic errors in the data, e. g. due to absorption, extinction and thermal diffuse scattering, may not have been corrected.

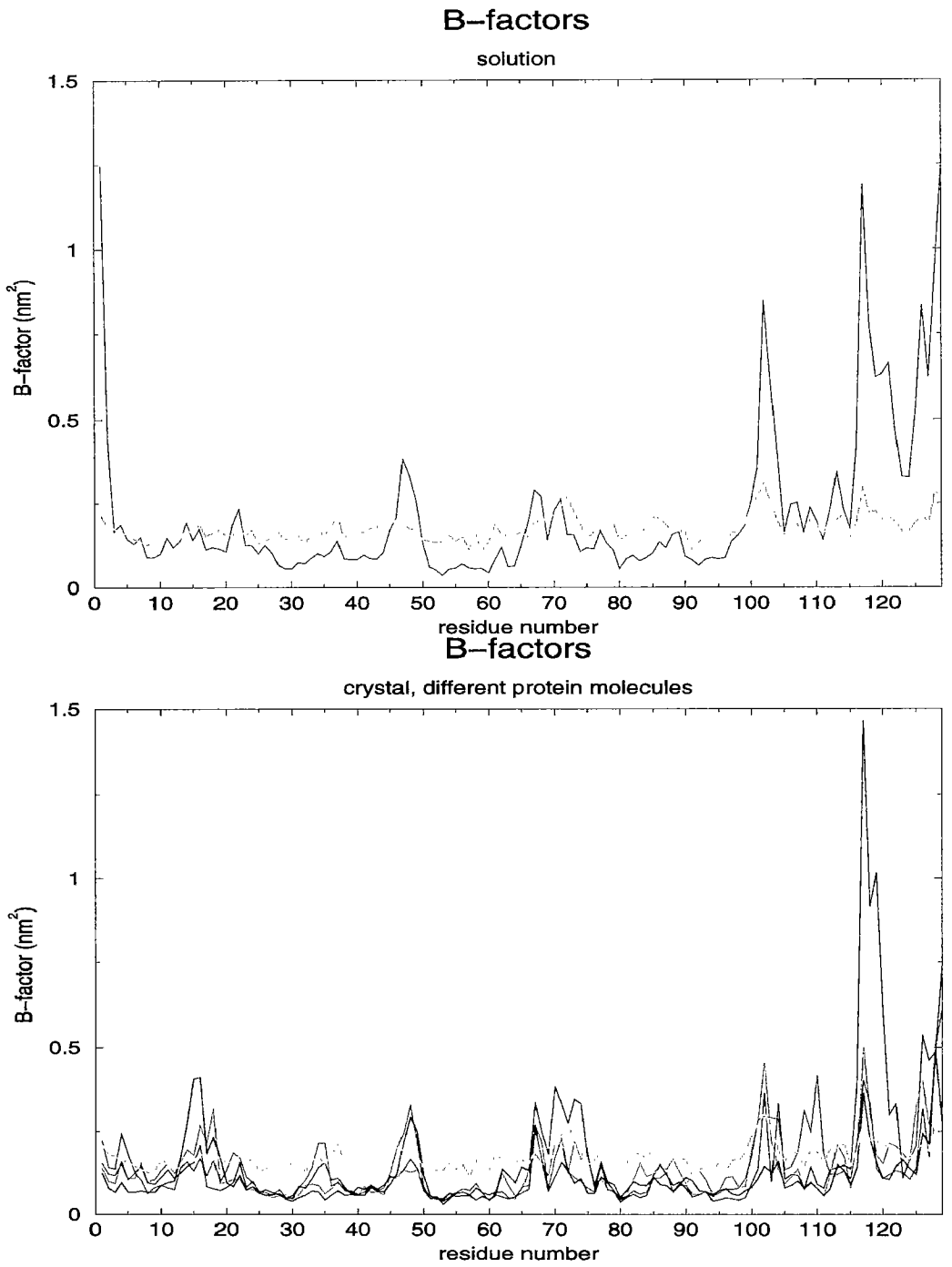

Figure 4.2: *Isotropic B-factors for backbone C_α-atoms in nm^2. In the top graph, B-factors calculated from the solution simulation (indigo) are shown together with experimentally derived values (orange). In the bottom graph, B-factors for the different protein molecules calculated from the crystal simulation are shown in black (molecule 1), red (molecule 2), green (molecule 3), and blue (molecule 4) together with experimentally derived values (orange).*

Although the backbone is more mobile in the solution simulation, the agreement of calculated and experimentally derived B-factors is similar for both simulations (fig. 4.2). Residues having large experimental B-factors also show enhanced mobility in the simulation. Secondary structure regions (α-helices 4GLY-14ARG, 25LEU-36SER, 88ILE-99VAL, 108TRP-115CYS, and 3_{10}-helix 119ASP-124ILE, and β-sheet 42ALA-60SER, see table 4.2) are stable and thus, in both simulations, not much flexibility is observed for these parts of the chain. Exceptions are residue 1LYS in the solution simulation which in the crystal lies close to some crystallographically observed water molecules, and the considerably rearranged loop region on the outside of the protein between residues 115CYS and 119ASP. The overall trends in mobility are thus correctly sampled, magnitudes, however, are overestimated. Flexible regions show more mobility in the simulations compared to experimentally derived B-factors, stable regions show less mobility, which might be due to the fact that single-molecule B-factors were calculated from the simulations, whereas the experimentally derived B-factors contain contributions from many molecules.

Residue		Percentage hydrogen bonding					
Donor	Acceptor	*X−ray*	*cry*1	*cry*2	*cry*3	*cry*4	*sol*
3PHE	38PHE	100.0	91.0	91.4	90.6	86.2	70.5
8LEU	4GLY	100.0	30.9	94.4	97.9	98.2	97.1
9ALA	5ARG	100.0	3.4	93.8	84.4	92.3	90.8
10ALA	6CYS	100.0	91.5	96.5	96.6	97.5	98.7
11ALA	7GLU	100.0	64.4	86.8	89.9	89.2	85.3
12MET	8LEU	100.0	99.1	99.1	99.1	99.2	99.3
13LYS	9ALA	100.0	99.2	97.9	95.8	96.9	96.6
14ARG	10ALA	100.0	94.9	91.2	93.8	87.7	84.3
15HIS	11ALA	0	68.6	40.0	55.5	46.8	73.2
15HIS	12MET	0	18.1	36.9	27.1	36.1	11.6
16GLY	12MET	0	30.9	19.8	14.1	41.6	56.4
16GLY	13LYS	100.0	35.4	39.3	51.1	28.9	32.0
17LEU	12MET	100.0	25.9	25.1	53.3	15.1	33.4
20TYR	17LEU	100.0	47.1	51.2	0	51.4	35.6
20TYR	23TYR	0	25.1	11.6	78.5	17.7	44.6
22GLY	19ASN	100.0	40.2	27.7	0	46.1	28.8
22GLY	41GLN	0	39.5	52.8	42.7	22.7	0
23TYR	20TYR	100.0	64.4	68.5	70.3	62.9	69.0
27ASN	24SER	100.0	26.4	32.2	52.9	53.3	26.8
28TRP	24SER	0	7.4	30.0	15.9	10.3	37.6
29VAL	25LEU	100.0	95.1	89.2	89.2	84.9	94.8
30CYS	26GLY	100.0	98.4	95.2	99.1	98.8	98.3
31ALA	27ASN	100.0	97.2	96.9	98.0	96.7	97.5
32ALA	28TRP	100.0	93.4	97.6	88.6	92.7	96.2
33LYS	29VAL	100.0	79.0	81.9	34.5	88.5	93.0
34PHE	30CYS	100.0	66.6	69.1	28.4	52.1	69.4
35GLU	31ALA	100.0	9.2	93.2	3.9	93.0	87.3
36SER	32ALA	100.0	0	19.1	0	22.2	25.5
37ASN	34PHE	0	40.0	2.4	5.5	27.6	22.2
38PHE	32ALA	100.0	83.4	59.2	75.2	77.3	81.5
39ASN	36SER	0	71.6	22.2	43.5	8.0	9.4

Residue		Percentage hydrogen bonding					
Donor	Acceptor	*X−ray*	*cry*1	*cry*2	*cry*3	*cry*4	*sol*
40THR	1LYS	100.0	94.0	91.3	91.1	87.1	10.1
42ALA	39ASN	0	8.5	3.4	9.3	20.7	27.2
42ALA	40THR	0	46.2	50.9	52.8	31.9	33.7
44ASN	52ASP	100.0	93.5	83.4	89.5	51.6	89.2
46ASN	50SER	100.0	75.0	65.9	0	54.2	69.8
49GLY	46ASN	100.0	67.4	76.9	0	72.5	80.2
52ASP	44ASN	100.0	75.3	92.9	85.9	90.8	74.8
53TYR	58ILE	100.0	97.4	96.8	94.2	96.2	95.8
54GLY	42ALA	100.0	90.7	78.2	82.1	42.2	68.4
57GLN	54GLY	100.0	78.6	0	31.6	10.1	84.0
58ILE	53TYR	100.0	98.7	95.4	92.6	79.4	98.4
60SER	51THR	100.0	87.2	81.9	89.9	86.6	92.8
63TRP	59ASN	100.0	72.1	75.4	52.1	83.5	84.5
64CYS	59ASN	0	65.8	65.8	18.5	29.5	17.5
64CYS	60SER	0	10.3	9.0	11.8	23.6	43.4
65ASN	78ILE	100.0	83.2	92.9	81.2	77.0	80.0
73ARG	61ARG	100.0	82.6	77.3	76.8	79.4	90.3
74ASN	61ARG	0	51.5	52.8	0	43.4	0
75LEU	62TRP	100.0	91.7	97.4	95.9	91.3	91.8
76CYS	63TRP	100.0	75.2	64.4	81.7	90.1	91.5
77ASN	74ASN	100.0	56.4	58.8	53.1	41.3	72.7
78ILE	76CYS	0	29.9	83.8	74.0	73.1	6.2
80CYS	65ASN	100.0	94.8	84.5	93.5	90.9	93.6
82ALA	79PRO	100.0	17.4	46.0	32.8	34.9	6.5
83LEU	79PRO	0	21.4	15.7	9.1	3.8	80.8
83LEU	80CYS	100.0	21.1	40.2	27.6	35.8	2.6
84LEU	81SER	100.0	5.6	11.6	12.0	19.4	5.4
84LEU	80CYS	0	78.8	32.4	56.9	46.4	82.6
85SER	81SER	0	43.1	51.9	38.5	44.8	39.3
85SER	82ALA	0	18.5	21.8	23.9	25.9	30.2
89THR	87ASP	0	0	0	2.8	0	25.6
92VAL	88ILE	100.0	93.1	60.1	95.8	88.2	3.6
93ASN	89THR	100.0	96.9	91.1	98.2	92.8	95.8
94CYS	90ALA	100.0	97.8	96.9	98.7	98.4	98.3
95ALA	91SER	100.0	98.5	95.2	98.8	96.7	98.1
96LYS	92VAL	100.0	94.9	97.1	95.1	97.6	98.1
97LYS	93ASN	0	69.8	77.6	82.2	80.0	80.6
98ILE	94CYS	100.0	93.8	90.4	93.4	93.6	95.2
99VAL	95ALA	100.0	91.1	93.7	94.7	96.5	69.2
100SER	96LYS	0	95.6	91.2	89.8	91.2	93.8
101ASP	98ILE	100.0	24.6	42.1	25.5	29.9	5.2
101ASP	97LYS	0	66.2	41.7	52.7	58.8	68.5
104GLY	98ILE	0	0	0	0	0	34.9
106ASN	103ASN	100.0	69.3	17.5	0	6.6	55.1
107ALA	104GLY	0	17.6	79.4	42.7	40.1	7.9

Residue		Percentage hydrogen bonding					
Donor	Acceptor	*X−ray*	*cry*1	*cry*2	*cry*3	*cry*4	*sol*
108TRP	105MET	100.0	80.5	82.9	67.2	79.5	81.4
112ARG	108TRP	100.0	78.8	84.0	29.9	74.5	71.6
113ASN	109VAL	100.0	0	92.9	13.6	93.6	82.1
114ARG	110ALA	100.0	0	26.6	40.9	72.7	65.5
115CYS	110ALA	0	60.1	89.6	18.6	89.6	87.2
115CYS	111TRP	100.0	26.8	7.1	62.9	5.8	9.2
116LYS	111TRP	100.0	43.7	95.6	85.3	96.5	79.7
118THR	115CYS	100.0	13.4	19.4	13.3	16.2	14.3
122ALA	119ASP	100.0	3.4	0	2.9	7.1	4.7
123TRP	119ASP	0	79.9	26.3	3.2	69.9	39.3
123TRP	120VAL	100.0	7.6	27.0	49.5	4.4	11.1
124ILE	120VAL	0	92.9	40.2	3.6	63.3	67.4
124ILE	121GLN	100.0	0	19.2	50.7	10.6	14.5
125ARG	121GLN	0	0	16.4	73.7	5.6	12.1
125ARG	122ALA	0	4.1	17.7	17.5	36.4	25.1
125ARG	123TRP	0	17.4	21.3	0	8.3	24.5
127CYS	124ILE	100.0	12.5	28.2	17.2	18.1	6.0

Table 4.2: *Occurrence of backbone-backbone hydrogen bonds in the X-ray structure, the crystal simulation, and the solution simulation. Intramolecular hydrogen bonds are listed if they are present in the X-ray crystal structure (X−ray), if they are present for more than 20% of the simulation time in the solution simulation (sol), or if they are present for more than 20% simulation time averaged over the four molecules (cry*1−4*) in the crystal simulation. A hydrogen bond is counted if the distance between the hydrogen atom and the acceptor atom is maximally 0.25nm and the angle between donor atom, hydrogen atom and acceptor atom is larger than 135°.*

Table 4.2 shows the backbone hydrogen bonding pattern in the X-ray structure and the percentage of hydrogen bonding as obtained by averaging over the single molecule trajectories from the simulations. A hydrogen bond is counted if the distance between the hydrogen atom and the acceptor atom is maximally 0.25*nm* and the angle between donor atom, hydrogen atom and acceptor atom is larger than 135°. Hydrogen bonds are very well reproduced in both simulations. Generally, the solution simulation shows stronger hydrogen bonding than the crystal simulation. Of the 63 hydrogen bonds found in the X-ray structure, 59 show occupancies of over 20% in the solution simulation or over 20% average occupancy in the crystal simulation. The hydrogen bonds not reproduced can be rationalized as follows: 84LEU-81SER is replaced by 84LEU-80CYS and 85SER-81SER. The other three hydrogen bonds lost in both simulations all lie in the very mobile region from residue 115CYS onwards. 118THR-115CYS is partly replaced by 118THR-114ARG (not shown in table 4.2), 122ALA-119ASP is replaced by 123TRP-119ASP. Also the two following hydrogen bonds, 123TRP-120VAL and 124ILE-121GLN, are, although still present over 20% in at least one of the simulations, shifted from a 3_{10}-helical pattern towards an α-helical structure: 124ILE-120VAL and 125ARG-121GLN. The hydrogen bond 127CYS-124ILE at the end of this helix shows slightly less than 20% occupancy in the simulations. The

two hydrogen bonds showing the highest occupancies in the simulations without being recognized in the X-ray structure are 97LYS-93ASN and 100SER-96LYS, where the chosen geometric criterion for counting a hydrogen bond is only slightly missed (donor-acceptor distances of $0.258nm$ and $0.262nm$ for the two hydrogen bonds). These residues are part of an α-helix running from residue 88ILE to residue 99VAL.

NOE distance bound violations are summarized in table 4.3. The average violation for the analysis period, $300ps$-$2000ps$, is lower for all chains in the crystal simulation compared to the solution simulation. Also, the number of large violations is always lower in the crystal simulation, reflecting the larger backbone fluctuations in the solution simulation.

protein molecule	NOE bound violations			
	number			mean
	$> 0.05nm$	$> 0.1nm$	$> 0.3nm$	$< R_E - R_0 > (nm)$
*cry*1	65	37	6	0.009
*cry*2	57	35	3	0.008
*cry*3	67	39	7	0.009
*cry*4	56	33	4	0.007
sol	75	48	9	0.010

Table 4.3: *Number of NOE distance bound violations larger than a given value computed from the crystal simulation and the solution simulation. The total number of experimentally determined NOE bounds R_0 is 1158 [45–47]. R_E is the distance from the trajectory using r^{-3} averaging over 200ps windows throughout the whole 300ps-2000ps analysis period. The symbol R_E-R_0 indicates a (mean) violation, calculated as a mean of the larger value of R_E-R_0 and zero. Average violations are in nm. The four molecules in the crystal unit cell are indicated by the symbol cry*1−4*, the one molecule in solution simulation by the symbol sol.*

The 95 $^3J_{HN\alpha}$-coupling constants (fig. 4.3) are slightly better reproduced in the solution simulation (RMSD of calculated versus experimentally derived J-coupling constants are $1.70Hz$ in the solution simulation and $1.72Hz$, $1.77Hz$, $1.88Hz$, and $1.64Hz$ for the four chains in the crystal simulation). The higher mobility in solution allows for a larger part of phase space being sampled and thus, averaging may be better. The X-ray crystal structure reproduces the experimental J-coupling constants well with an RMSD of $0.88Hz$, as is seen in the lower right corner of figure 4.3 where J-coupling constants calculated from the X-ray structure are compared with J-coupling constants obtained from NMR experiments. Calculated J-coupling constants are generally slightly lower than experimentally derived ones but longer simulation times could bring these values closer to each other.

Backbone order parameters, displayed in figure 4.4, are equally well reproduced in both simulations. The secondary structure elements correspond to large calculated S^2 values. Order parameters calculated from the simulation are generally slightly lower compared to those determined using NMR data. In the crystal simulation, different chains generally have very similar order parameters with a handful of exceptions. This happens, if one protein molecule has sampled a particular movement, while an other has not. This effect is even stronger for side-chain order parameters (fig. 4.5). Slow motions observed only once in a simulation lead to different order parameters for the same residue in the four protein molecules. The crystal simulation

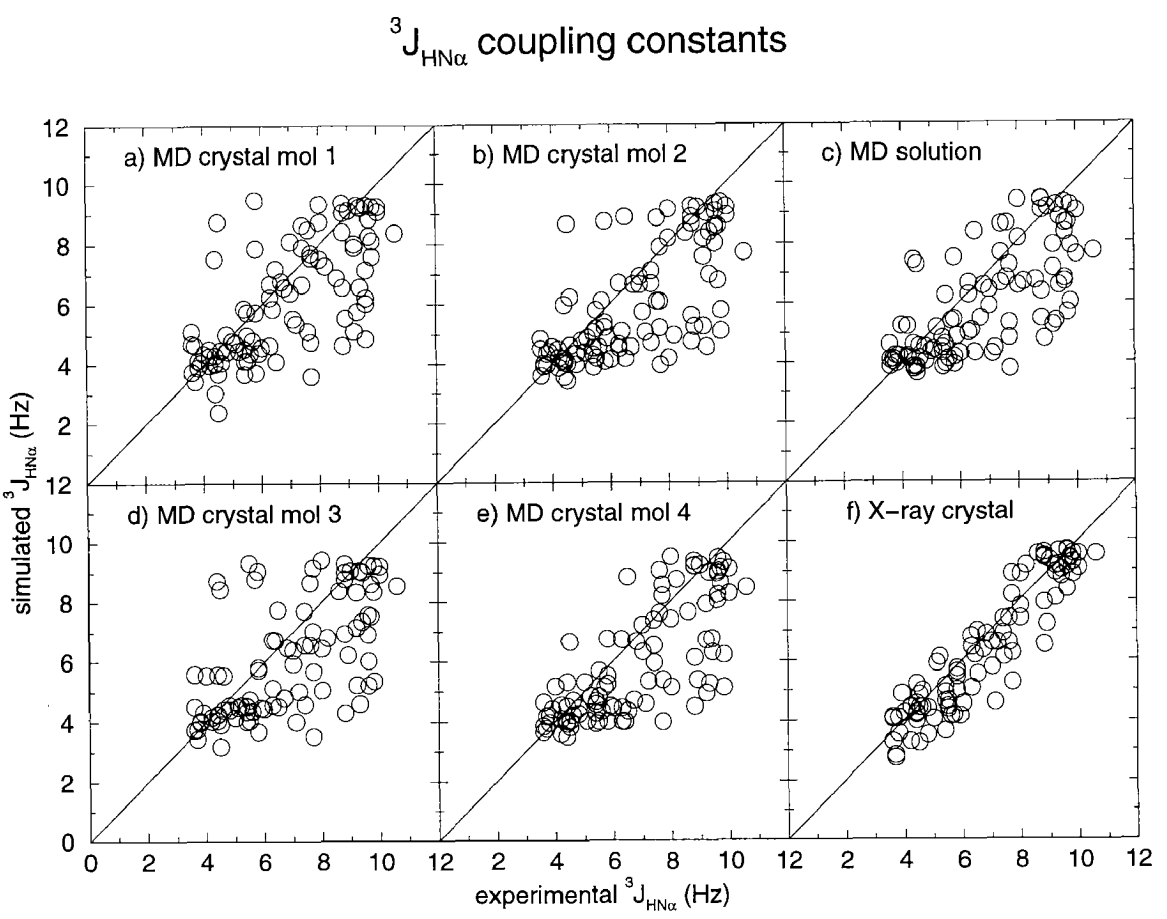

Figure 4.3: *Comparison of 95 experimental (Smith et al., 1991) and calculated $^3J_{HN\alpha}$ coupling constants (in Hz). In graphs a), b), d), and e), $^3J_{HN\alpha}$ coupling constants calculated from the crystal simulation are displayed for the 4 different protein molecules in the unit cell. In graph c), values calculated from the solution simulation are shown. Graph f) shows J-values calculated from the experimentally determined X-ray crystal structure.*

shows significantly larger side-chain order parameters than experimentally derived ones. Its order parameters are also larger than in the solution simulation which is expected since side-chains involved in crystal packing contacts will be more rigid than in solution. Order parameters calculated from the solution simulation correspond very well to experimentally derived ones.

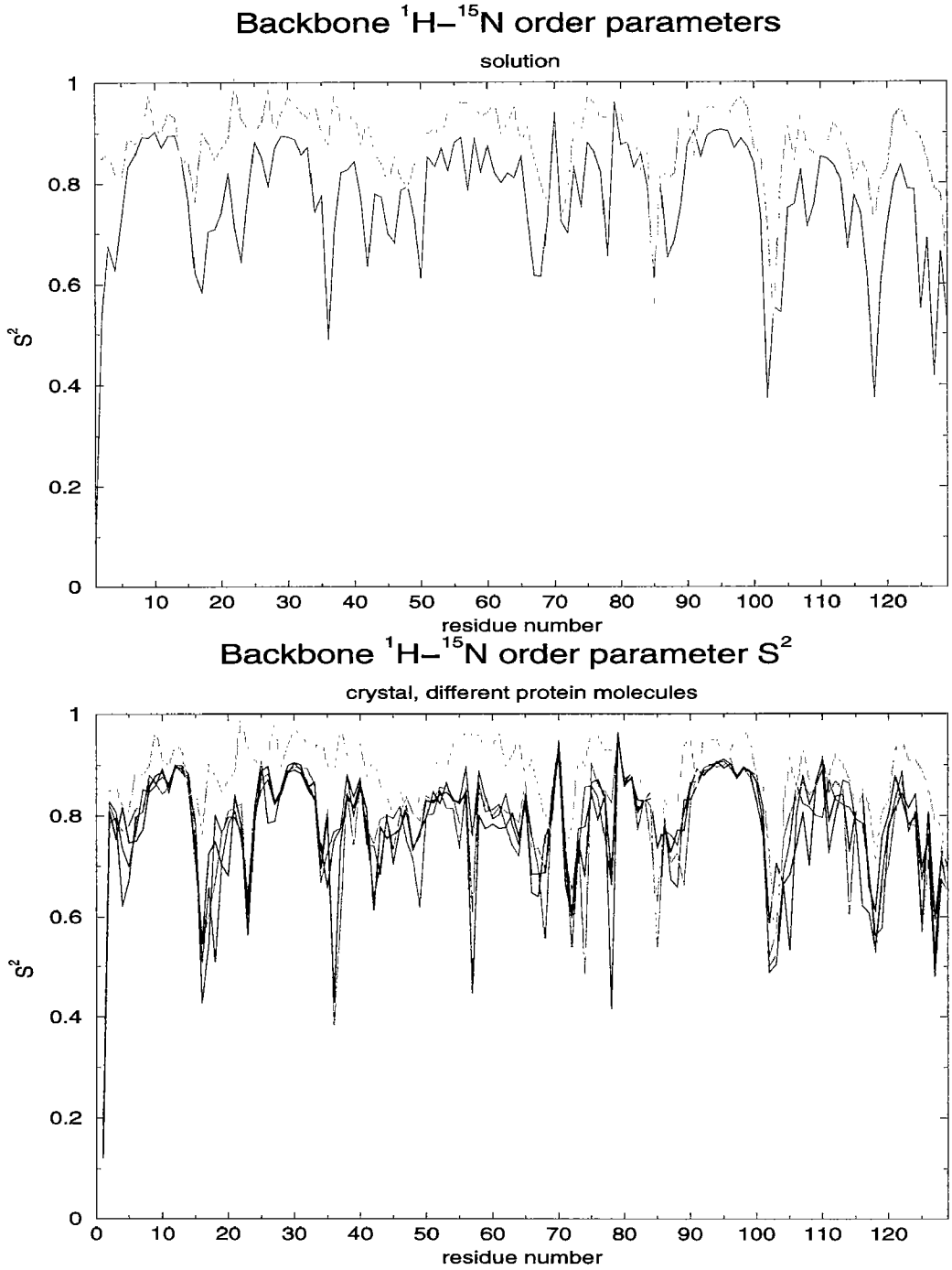

Figure 4.4: *Backbone ^1H-^{15}N order parameters (S^2) as function of residue number. The order parameters were calculated using a 200ps averaging window moving through the whole 1700ps analysis period. In the top graph, order parameters calculated from the solution simulation (indigo) are shown together with experimentally derived values (orange). In the bottom graph, order parameters for the 4 different protein molecules calculated from the crystal simulation are shown in black (molecule 1), red (molecule 2), green (molecule 3), and blue (molecule 4) together with experimentally derived values (orange). There is no experimental value available for residues LYS1, SER50, PRO70, PRO79, and ALA110.*

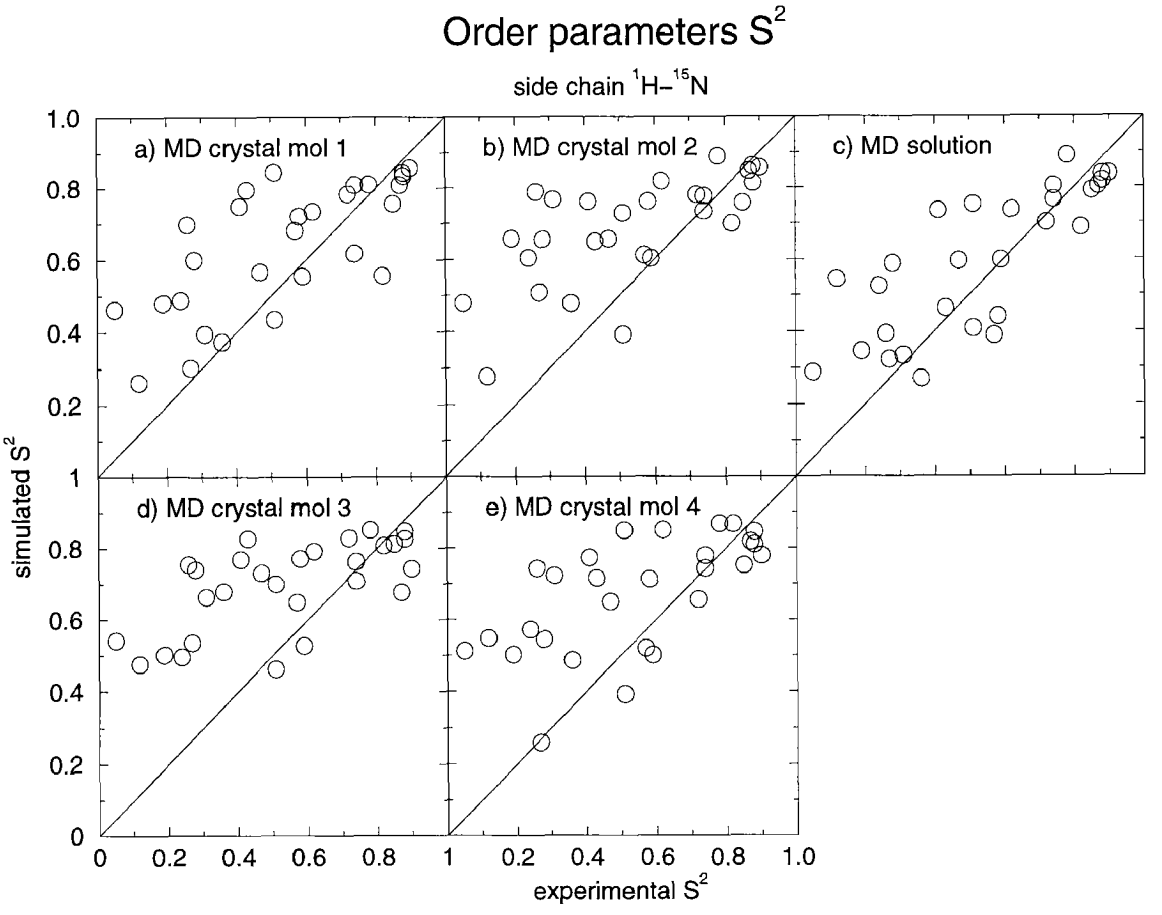

Figure 4.5: *Comparison of 28 experimental [47] ^1H-^{15}N order parameters S^2 of side-chain NH groups with calculated values. The order parameters were calculated using a 200ps averaging window moving through the whole 1700ps analysis period. In graphs a), b), d), and e), $^3J_{HN\alpha}$ coupling constants calculated from the crystal simulation are displayed for the 4 different protein molecules in the unit cell. In graph c), values calculated from the solution simulation are shown. For NH_2 groups the average of the order parameters for the two NH vectors is displayed*

4.5 Conclusions

The two $2ns$ molecular dynamics simulations of hen egg white lysozyme (HEWL), one in solution, the other in crystalline environment give very similar results, slight differences are, however, present. Not unexpectedly, the four protein molecules in the crystal simulation stay closer to the X-ray structure. They also fulfill atom-atom distance bounds derived from NMR solution data significantly better than the solution simulation. B-factors, representing the magnitude of atom positional fluctuations and order parameters S^2 which are determined by orientational mobility of N-H bond vectors, show the same degree of agreement with experimentally derived data for both simulations. The root-mean-square deviation from the X-ray structure shows larger fluctuations for the solution simulation indicating enhanced sampling of configuration space and maybe better converged J-coupling constants than in the crystal simulation. The backbone hydrogen bonding network is very well maintained in both simulations. Occupancies of intramolecular hydrogen bonds are even higher in the solution simulation, probably because in the crystal, the

higher atom density leads to less flexibility: the protein in solution is less restrained than in the crystal. The four different molecules in the crystal show highly similar properties, especially in stable regions of secondary structure. In the more mobile regions of the molecule, a rare event, such as a jump over a rotational barrier of a dihedral angle, can lead to different observations between single molecules. With increasing simulation length, more of these events would be observed, and the same events could also happen in another protein molecule. Longer simulations would thus enhance the similarity between the different molecules in the crystal simulation.

The present results for hen egg white lysozyme confirm that proteins in crystalline environment and in solution show very similar behaviour. Fluctuations are somewhat larger in solution, structural properties are almost identical.

4.6 Acknowledgments

The authors wish to thank Dr. Lorna J. Smith for making data available and useful discussions and Prof. Dr. Alan E. Mark for fruitful discussions. Financial support was obtained from the Schweizerischer Nationalfonds, project number 21-41875.94, which is gratefully acknowledged.

Chapter 5

Effect of Boundary Conditions and Treatment of Electrostatics: Crystal and Solution Molecular Dynamics of Ubiquitin

5.1 Abstract

Five molecular dynamics simulations of ubiquitin under different conditions probe effects of the molecular environment, especially regarding electrostatics. Three simulations of protein crystals are compared to two simulations where a single protein molecule is well solvated by bulk water. Simulations of crystals provide good tests of the methodology, since they should exactly reproduce crystallographic experiments, and since the higher charge density in a crystal more readily exposes any problems with the treatment of electrostatics. The use of a reaction field term for long-range electrostatics is found to be necessary to reproduce the experimental crystal structure. Addition of neutralizing counterions stabilizes the protein structures further in the simulations.

5.2 Introduction

The dynamics of the protein ubiquitin in crystal and in solution is here studied by molecular dynamics simulations. Two simulations are performed in a solvent environment and three in a crystalline environment. The simulations differ especially with regard to the electrostatic properties of the surroundings of the proteins. One of the crystal simulations was performed disregarding all long-range electrostatic interactions. The resulting protein structures are shown to deviate strongly from crystallographic experimental data. In the remaining simulations, we investigate the effects of a crystalline environment versus a solution environment, and of the addition of a neutralizing counterion.

Among the forces considered in molecular dynamics simulations, the electrostatic interactions are the most long-ranged. The Coulomb interaction energy between two charged particles decreases with distance only as r^{-1}, which makes also interactions between charges rather far from one another relatively important. This is problematic in practical simulations, where the usual method for limiting the computational time needed is to restrict the number of non-bonded interactions actually included in the calculations, usually by employing a cut-off radius. Every

particle in the simulation is then only considered to interact with other particles which are closer than a certain distance.

Such a cut-off scheme is not only practical, but virtually mandatory in cases where one avoids boundary effects by simulating periodically replicated copies of a system. The number of interactions in the periodic system would be infinite, were one not to restrict them somehow. The cut-off radius is usually chosen so as to implicitly enforce a "minimum-image" treatment of the periodic boundary condition, meaning that a particle should only interact with the closest of several possible periodic neighbor images, and never interact with an image of itself, to avoid pseudo-crystal simulations of non-crystalline systems.

A common refinement of the cut-off scheme involves grouping particles into electroneutral groups, as far as possible, and employing a cut-off radius at which interactions between all particles in two such "charge groups" are simultaneously turned on or off. This leads effectively to the groups interacting as dipoles, making their interactions decrease more rapidly with distance and thereby allowing smaller cut-off radii with retained accuracy.

The other non-bonded interaction term is the Lennard-Jones term used for modelling dispersive interactions and atomic core repulsions. These terms are usually of the form

$$V = C_{12} r^{-12} - C_6 r^{-6} \tag{5.1}$$

and thus decrease much more rapidly with distance than the electrostatic terms, but since the interactions are on average always negative (attractive) for non-polar molecules in standard pressure equilibrium, contributions neglected outside a medium-size cut-off radius can sum up to non-negligible values [39].

In the GROMOS96 molecular modelling program package [2, 64], which was used throughout this work, a charge group based twin-range cut-off scheme is implemented [32]. An outer cut-off radius, which we usually take to be 1.4nm, is used to determine which non-bonded interactions are calculated every 5 time steps (each of which is 2 fs). Simultaneously to the calculations with the outer cut-off radius, a list is constructed of group interactions which fall inside an inner cut-off radius, here we employed 0.8nm, and the interactions on this "non-bonded list" are recalculated every time step. This of course filters away fast variations in mid-range interactions, but allows a relatively large cut-off radius to be used practically by reducing the mid-range interaction calculation effort by a factor of 5.

The remaining consideration is of course how to treat interactions outside the (outermost) cut-off radius. The simplest and least accurate way is to neglect them totally, which is not uncommonly found in the literature. More advanced approaches include more or less approximate treatment of the electrostatic interactions outside of the cut-off radius. Methods that make use of infinite lattice sums [65] include the standard Ewald summation (see e.g. [1]), the particle-particle-particle-mesh method (p3m) of [5] and the particle-mesh Ewald summation [66].

An alternative approach, which avoids the periodicity artifacts of infinite lattice sums, but at the cost of using a mean-field approximation, is implemented in the GROMOS96 program package. In this scheme, a generalized Poisson-Boltzmann reaction field is used to model the interaction of each atom with everything outside its cut-off radius. These distant surroundings are then approximated by a dielectric continuum characterized by a given dielectric permittivity ε_{rf} and a given inverse Debye screening length κ (or ionic strength) [2, 42, 64]. The electrostatic interaction (Coulomb plus reaction field) between two atoms i and j with charges q_i and q_j at a

distance r_{ij} is in this model given by

$$V^{CRF} = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_1}\left[\frac{1}{r_{ij}} - \frac{\frac{1}{2}C_{rf}(r_{ij})^2}{R_{rf}^3} - \frac{(1-\frac{1}{2}C_{rf})}{R_{rf}}\right] \tag{5.2}$$

with

$$C_{rf} \equiv \frac{(2\varepsilon_1 - 2\varepsilon_{rf})(1+\kappa R_{rf}) - \varepsilon_{rf}(\kappa R_{rf})^2}{(\varepsilon_1 + 2\varepsilon_{rf})(1+\kappa R_{rf}) + \varepsilon_{rf}(\kappa R_{rf})^2} \tag{5.3}$$

The second term represents the Poisson-Boltzmann reaction field contribution due to the presence of the dielectric continuum with permittivity ε_{rf} and inverse Debye screening length κ outside the reaction field cut-off radius R_{rf}. The third term is a constant that makes the electrostatic energy zero for $r_{ij} = R_{rf}$. The permittivity inside the reaction field cut-off sphere is $\varepsilon_1 = 1$.

We have chosen to study the effects of the electrostatic environment on the molecular dynamics of the protein ubiquitin, which is small and thus easy to treat and which has a well determined, stable protein structure. It is also a very interesting molecule of great biological significance.

Ubiquitin is a 76 residue protein, which is probably present in all eukaryotic cells [67, 68]. All 76 residues are perfectly conserved between animal species, and only 3 residues differ between animals and plants. It is in other words the most conserved protein known. Ubiquitin seems to be primarily involved in ATP-dependent intracellular non-lysosomal protein degradation, marking a protein molecule for degradation by amide linkage of the carboxy terminus of ubiquitin to the N-terminal nitrogen of the target protein [69]. Ubiquitination of lysine ε-nitrogens also occurs. Commonly, the ubiquitins are themselves ubiquitinated at lysine 48, forming ubiquitin chains. Constitutive ubiquitination is observed for a subpopulation of histone H2A and for specific ribosomal proteins. For reviews of the fascinating ubiquitin system see [70–73].

The extreme conservation between species leads one to assume that the precise structure is very critical for the function of this protein, and the crystallographically determined structure of ubiquitin is indeed quite advanced for such a small protein, exhibiting three and one-half turns of α-helix, a short piece of 3_{10}-helix and a mixed β-sheet of five strands, all in a compact arrangement with tight hydrogen-bonding. The structure has been determined to a resolution of 1.8 Å [30].

Ubiquitin is itself very resistant to proteinase degradation by for example trypsin, chymotrypsin, subtilisin and carboxypeptidase [68], and is stable under a wide range of pH and temperature conditions [74]. This makes it reasonable to assume that the structure is very stable and fluctuates very little. This has been confirmed by nuclear magnetic resonance spectroscopy; for the entire protein except the four carboxyterminal residues, the ^{15}N relaxation of ubiquitin backbone amide nitrogens exhibits high generalized order parameters, especially for NH groups involved in the main chain hydrogen bonding network [75].

The small size of the protein makes it feasible to simulate several protein chains in a box, allowing the copies of the protein to move relatively independently but still in a crystal environment. The crystallographic rectangular prism unit cell, $a = 5.084$nm, $b = 4.277$nm, $c = 2.895$nm, is a suitable box for the purpose, containing 4 protein chains.

We are obviously not the first group performing molecular dynamics simulations of ubiquitin, although all earlier simulations have modelled solution environments, rather than crystals. Our fore-runners have used quite varying approaches to the surroundings of the protein. Braatz *et al.* [76] used only the 58 water molecules given in the crystallographic model, representing the rest of the surrounding water with an *ad hoc* continuum model; a linear distance-dependent

electrostatic permittivity was used and the total charge of each acidic or basic functional group was set to zero. Alonso & Daggett [77] did simulations to mimic a system at pH 2 including 13 chloride counterions. They used explicit water or an explicit water-methanol mixture in a periodic box of unknown size, but had a cut-off radius of 0.8nm for non-bonded interactions without any additional treatment of longer-ranged forces. Abseher *et al.* [78] used explicit water molecules in a periodic box (5.0nm × 5.0nm × 5.0nm) with Ewald summation of electrostatics as recommended by [79, 80]. Lienin *et al.* [81] had explicit water in a periodic box and an 0.8nm cut-off for non-bonded interactions, but no treatment of longer-ranged forces. Unfortunately, they chose a cubic periodic box with an edge length of 4.665nm, while the diameter of the ubiquitin molecule in its X-ray structure is 4.49nm. Due to the periodicity condition, the protein molecules are then only 0.175nm apart, a much shorter distance than the 0.8nm cut-off radius used. This then leads to aggregation of the periodic images of the protein, which renders the simulated results questionable. Juranic *et al.* [82] used a similar setup, but chose an edge length of 5.60nm, which avoids the aggregation problem. They used a 1.0nm non-bonded interaction cut-off without treating longer-ranged forces.

This work aims to remedy the lack of simulations of ubiquitin crystals, and compare these to solution simulations. The effects of using a reaction field treatment of longer-ranged electrostatic interactions and of adding counterions are investigated.

5.3 Methods

Simulations of ubiquitin and analyses of the results were performed using the GROMOS96 program suite [2, 64]. The force-field parameters used were the GROMOS96 explicit solvent simulation force-field 43A1, in which aliphatic hydrogens are united with the atom to which they are bound, while polar hydrogens and aromatic ring hydrogens are modelled explicitly.

Starting crystallographic coordinates were acquired from the Protein Data Bank [83], structure 1UBQ [30], which is refined to 1.8 Å resolution. The protonation state of the protein was modelled for neutral or slightly acidic pH, i.e., all arginine, lysine and histidine residues in the proteins were positively charged, and all aspartate and glutamate residues negatively charged. The rationale for these choices was that the structure determination was done at pH 5.6 and ambient temperature [30].

For the two solution simulations, a single protein chain was treated by truncated octahedron periodic boundary conditions. The box dimensions of 6.89145nm were chosen large enough to ensure that all protein atoms lie well outside the non-bonded cut-off radius of any protein atom in the periodically replicated copies of the protein. In the three crystal simulations, the four protein chains found in the rectangular prism unit cell of dimensions 5.084nm × 4.277nm × 2.895nm were simulated, with the unit cell as the rectangular periodic box.

A single protein chain, together with the 58 water molecules belonging to that chain in the crystallographic model, were for the solution simulations surrounded by explicit solvent (water) molecules represented with the simple point charge (SPC) water model [29] by immersing the system and its truncated octahedron box in a larger equilibrated water box, and deleting all water molecules with their oxygen atoms within 0.23 nm distance from any non-hydrogen protein atom or crystallographic water oxygen atom. This resulted in a total of 5050 water molecules in the truncated octahedron box.

Similarly, for the crystal simulations, the 232 crystallographical model waters were retained, and the unit cell was flooded with equilibrated SPC water, but here a slightly different value

was used, retaining all water molecules within 0.220605nm of protein or crystallographic water. This value led to the addition of exactly 460 water molecules, as needed for a total of 692 water molecules giving the experimental density of 1.23 g cm^{-3}.

In the two simulations where counterions were added, one chloride ion was needed to neutralize each protein chain, and consequently 1 chloride ion for the solution simulation and 4 chloride ions for the crystal simulation were added, by replacing water molecules at the water oxygen positions where the ions would experience the most favorable electrostatic interactions. The solvent simulation then had 5049 remaining water molecules, whereas the crystal simulation had 688 water molecules.

Non-bonded interactions were treated with a double or triple-range method [32], where all interactions on a charge group based 0.8 nm cut-off radius pair list are evaluated every time step of 2 fs, whereas interactions within a 1.4 nm radius are updated every 5 time steps (10 fs), concurrent with updating the non-bonded pair list. Electrostatic interactions with everything outside the 1.4 nm sphere were either treated as a dielectric continuum reaction field model [42], as described above, with a self-consistent relative dielectric constant $\varepsilon_{rf} = 54.0$ as found for SPC water [43] and $\kappa_{rf} = 0$, or were neglected, as described.

All covalent bond lengths were constrained to their ideal lengths with the SHAKE algorithm [3] using a geometric accuracy of 10^{-4}.

Simulation temperature was controlled by coupling solute and solvent to separate heat baths [6], with coupling constants of 0.1ps. In the solution simulation without added counterions, pressure was also (inadvertently) kept at 1 atm by coupling to an isotropic pressure bath [6] with a coupling constant of 0.5ps, assuming an isothermal compressibility of 4.575×10^{-4} $(\text{kJ mol}^{-1}\,\text{nm}^{-3})^{-1}$. The simulation box side length changed only marginally during this MD run, with a final value after 2.0ns of 6.915nm.

In preparation for the molecular dynamics simulations, the protein atoms were restrained to their initial positions with harmonic potentials $V = \frac{K}{2}(r - r_0)^2$ with force constants K of 2.5×10^4 kJ $\text{mol}^{-1}\,\text{nm}^{-2}$, and the system was subjected to 20 cycles of steepest descent minimization. The minimized structure was pre-equilibrated by consecutive 1ps runs of MD simulation at 300K, with restraining force constants of 2.5×10^4, 2.0×10^4, 1.5×10^4, 1.0×10^4, 5.0×10^3, 1.0×10^3, 3.0×10^2, 1.0×10^2, 30, 10, 3.0 and 1.0 kJ $\text{mol}^{-1}\,\text{nm}^{-2}$.

After this, each run was continued without restraining, collecting trajectory data every 500 time steps (i.e. 1 frame every picosecond), for 2.0ns. Analysis of the resulting trajectories was done with programs from the GROMOS96 package.

For graphics modelling, we used Rasmol (Roger Sayle, Glaxo, UK).

5.4 Results

5.4.1 Simulation modelling

We performed molecular dynamics simulations using the GROMOS96 program package [2, 64] of ubiquitin in five different systems: three in a solvated crystal environment and two free in aqueous solution. Details of the simulation procedure are given in the Computational methods section above.

Starting crystallographic coordinates were acquired from the Protein Data Bank [83], structure 1UBQ [30], which has a resolution of 1.8 Å. The crystal is $P2_12_12_1$ orthorhombic, with 1 protein molecule in the asymmetric unit [84]. This leads to 4 protein molecules being present in

the unit cell, which is a rectangular prism (all angles 90°) of dimensions 5.084nm × 4.277nm × 2.895nm.

In the crystal simulations, the crystallographic unit cell was used as the simulation periodic cell, and the simulations comprised four protein chains and 692 water molecules. For the solvent simulations, a single protein chain was surrounded by 5050 waters using truncated octahedron periodic boundary conditions. The net charge of a protein chain was $+1e$, so one chloride counterion per protein chain was added in two of the simulations. Chloride ions were initially placed to replace water molecules at the water oxygen positions where a chloride ion would experience the most favourable electrostatic potential. The crystal simulation then had 4 chloride ions and 688 water molecules, while the solution simulation had 1 chloride ion and 5049 water molecules.

In all simulations, the simulation temperature was kept at 300K by heat bath coupling [6]. The crystal simulations were carried out at constant volume conditions, as was the solution simulation with an added counterion. The solution simulation without counterions was performed at constant atmospheric pressure, with a pressure bath coupling [6]. The periodic box size changed only marginally during this simulation.

Each of the systems was briefly energy minimized and simulated in several consecutive short dynamics runs at 300K with decreasing harmonic potentials restraining the protein atoms to their initial positions. After this, all five systems were simulated without restraints at 300K for 2.0ns, collecting one trajectory data point every picosecond.

5.4.2 The effect of a reaction field on crystal simulations

Figure 5.1 shows the root-mean-square atom-positional deviation (RMSD) from the crystallographic coordinates of the four chains, in the three crystal simulations, after fitting of the alpha carbons to remove overall translation and rotation. In 5.1a, we see the RMSD of the C_α atoms for the simulation without a reaction field, which reaches values around 0.20–0.30nm. This would not be very remarkable values for a solution simulation starting from a crystallographically determined structure, but for a crystal simulation where one expects to reproduce the environment of the protein molecules very accurately, the deviations are too large. It is also obvious that the simulation cannot be said to be converged, even after 2.0ns.

Figure 5.1b depicts the RMSD of the C_α atoms for the crystal simulation with a reaction field treatment of longer-ranged electrostatics, whereas Figure 5.1c refers to the simulation with both reaction field and added counterions. We note that the values reach stable plateaux at around 0.12nm in both cases. The ion simulation seems more stable, leveling off early, but has an episode of slightly irregular behaviour in the period 1.1–1.5ns.

The trend is the same for the root-mean-square atom-positional deviation of all protein atoms in the three simulations, with values around 0.30–0.35nm, 0.18nm and 0.18nm, respectively (not shown).

These results show clearly that some treatment of longer-ranged electrostatics, e.g. with the reaction field method, is necessary to reproduce effects of the crystal environment.

Analysis of the secondary structure further underlines this impression. In Figure 5.2a we see the development over simulation time of the secondary structure assignments using PROCHECK [85, 86] for one representative chain from the simulation without a reaction field. Figure 5.2b gives the analogous result for the same chain in the simulation with reaction field and counterions. We note the much higher stability of the secondary structure elements in the simulations with a

reaction field. The reaction field results with and without counterions do not show any obvious differences (not shown).

5.4.3 The effect of counterions on crystal simulations

As noted above, both atom-positional root-mean-square deviations from the initial structure and secondary structure elements are much stabilized by a reaction field treatment of longer-ranged electrostatics. There are however no clear differences in these properties between the simulations with and without added counterions.

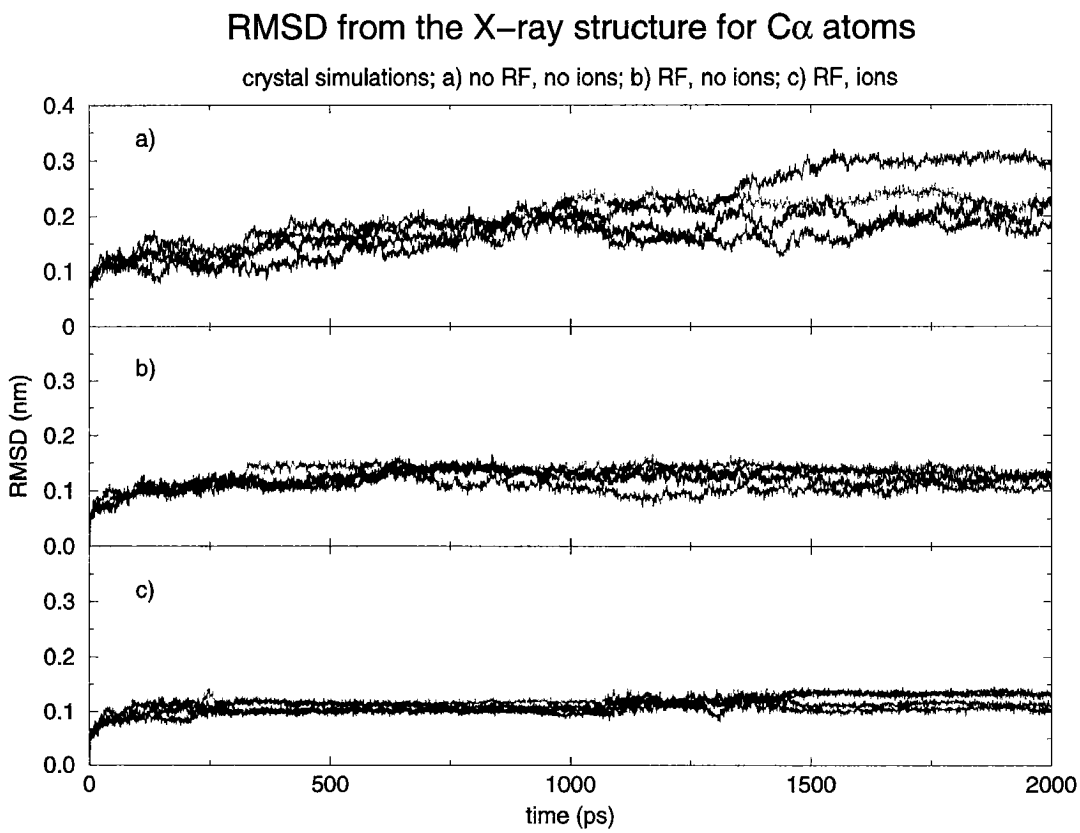

Figure 5.1: *Atom-positional root-mean-square deviation (RMSD) of the C_α atoms in nm from the initial crystallographic structure for the crystal simulations, as a function of simulation time. Each chain was least-squares fit to the initial structure using the C_α atoms of residues 1–72, to remove overall translation or rotation. a) The simulation without a reaction field treatment of longer-ranged electrostatic interactions and without counterions. b) The simulation with a reaction field, but without neutralizing counterions. c) The simulation with a reaction field and 4 neutralizing counterions.*

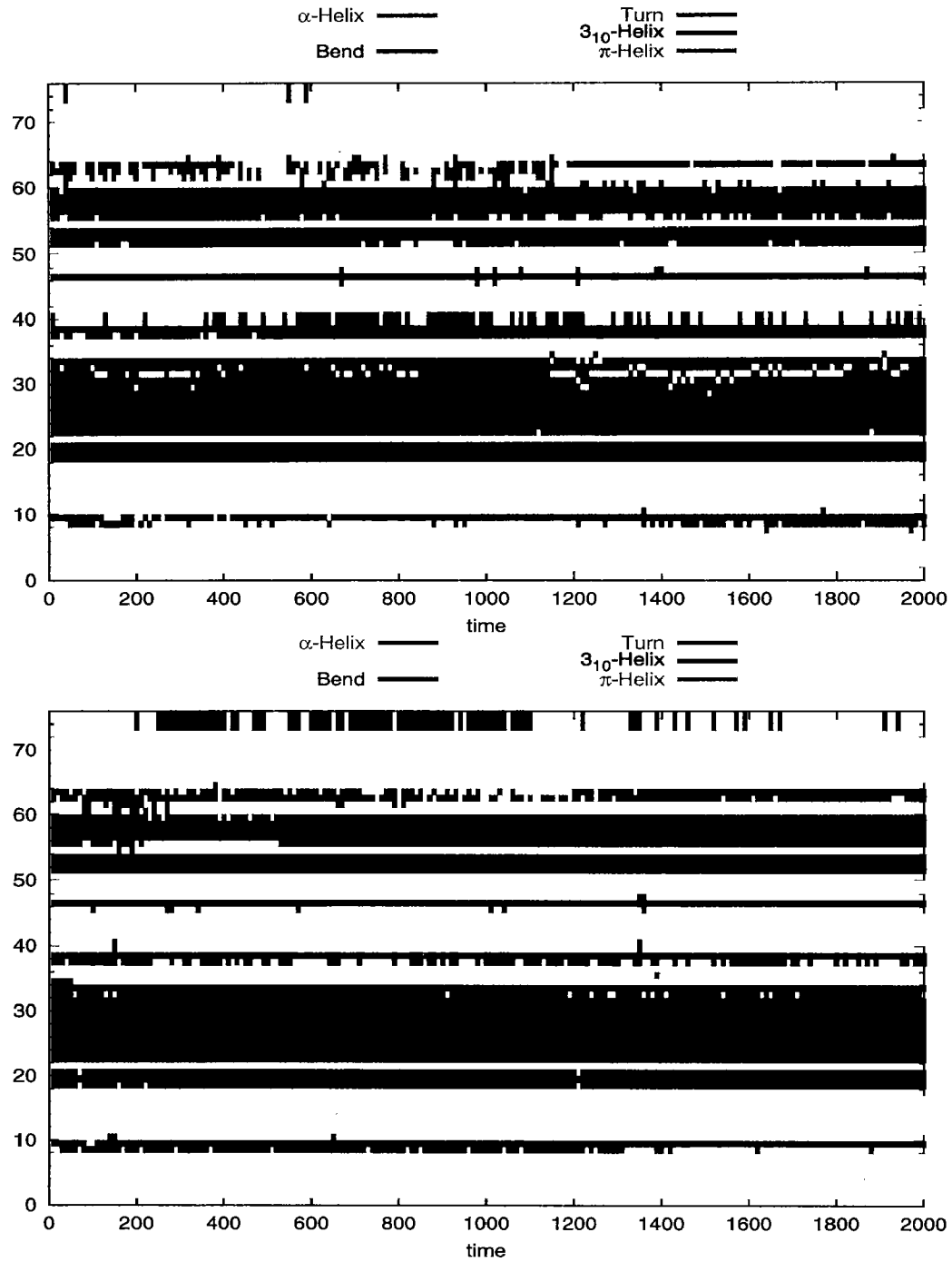

Figure 5.2: *Secondary structure assignment according to PROCHECK [85, 86] for one representative protein chain in the crystal simulations. In the top figure, secondary structure for the crystal simulation without a reaction field treatment of longer-ranged electrostatic interactions and without counterions is displayed, in the bottom figure, secondary structure for the crystal simulation with a reaction field and 4 neutralizing counterions is shown. Not shown is the crystal simulation with a reaction field but without counterions, where the result is similar to case b).*

However, when dihedral transitions observed in the protein are counted over simulation time (Table 5.1), we observe a strong effect of the counterions on the frequency of these transitions,

simulation	60°	120°	with H
cry1	4357	724	1841
cry2	3937	711	1933
cry3	5389	769	1916
cry4	4950	689	1689
cry_rf1	3022	447	1162
cry_rf2	2947	512	1095
cry_rf3	3568	526	1164
cry_rf4	3714	483	1445
cry_rf_ions1	1171	168	534
cry_rf_ions2	1098	176	546
cry_rf_ions3	1394	194	488
cry_rf_ions4	1061	170	536
sol_rf	7543	1165	1583
sol_rf_ions	6296	874	1727

Table 5.1: *Number of dihedral angle transitions during the simulations exceeding* 60° *for dihedrals of periodicity 6 or exceeding* 120° *for dihedrals of periodicity 3. Dihedral angle transitions for rotating polar hydrogen atoms are listed separately. The dihedrals were counted during the final 1500ps of simulation time, except for the values for the crystal simulation with a reaction field but with no counterions, where the final 1520ps were analyzed. In the table, "cry" refers to simulation in the crystal environment, and the four protein chains in the unit cell are counted separately, whereas "sol" indicates simulation in a solution environment. The use of a reaction field treatment of longer-ranged electrostatic interactions is shown by "rf", and "ions" implies the addition of neutralizing counterions.*

which are indications of mobility and fluctuations in the protein. Over the final 1.5ns of the simulations, the frequency of such transitions (here given as events per ns per protein chain) was 4818 for the simulation without reaction field, 3303 for the simulation with reaction field but without counterions, and 1256 for the simulation with both reaction field and counterions. The counterions thus seem to significantly reduce the fluctuations in the protein chains.

Figure 5.3 shows the effect on the radius of gyration of the protein chains, as an indication of shape changes of the protein. The instability of the simulation without a reaction field is marked (Figure 5.3a). In the reaction field simulation without counterions, we see a more stable behaviour (Figure 5.3b). When counterions are added (Figure 5.3c), much less fluctuations occur, except in the episode between 1.1 and 1.5ns, after which the simulation attains a stable state where all four protein chains have almost exactly the same radii of gyration. Again, the stabilizing influence of the added counterions can be seen.

5.4.4 Results of solution simulations

The solution simulations show, as expected, large deviations in atom-positional root-mean-square deviation from the crystallographic starting structure. Interestingly, here the simulation without

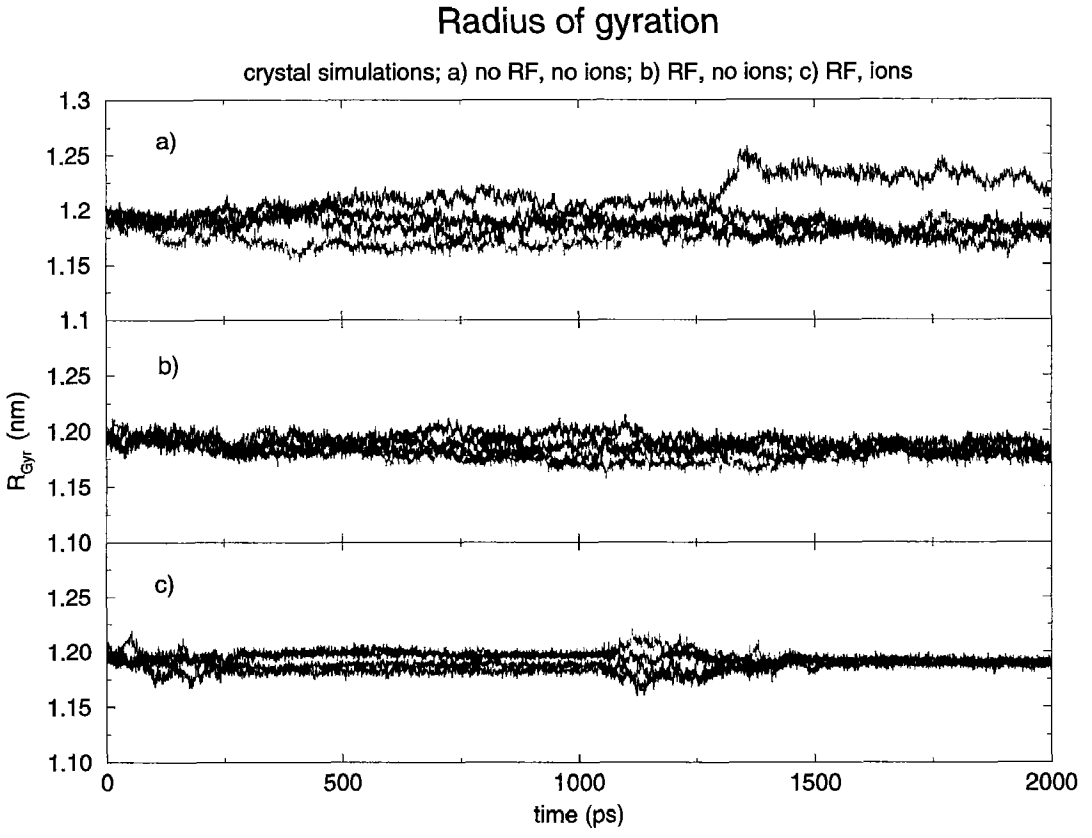

Figure 5.3: *Protein chain radius of gyration (R_{Gyr}) for the crystal simulations, in nanometers, as a function of simulation time. In each case, all four chains are plotted. a) The crystal simulation without a reaction field treatment of longer-ranged electrostatics and without counterions. b) The crystal simulation with a reaction field, but without counterions. c) The crystal simulation with a reaction field and with 4 neutralizing counterions.*

a counterion (Figure 5.4a) deviates less from the initial structure than the simulation with a counterion (Figure 5.4b), with values around 0.15nm and 0.20nm, respectively.

Furthermore, the fluctuations in the molecular structure are much larger than in the crystal simulations, as shown by large fluctuations in the molecular radius of gyration (Figure 5.5). For the radius of gyration, however, the trend found in the crystal simulations reappears, with larger fluctuations without counterion (Figure 5.5a) than with a counterion (Figure 5.5b) reappears.

We see the same tendency in the dihedral-angle transitions (Table 5.1) as found for the crystal simulations, namely that counterions decrease the fluctuations of the protein. Here, the overall mobility of the protein is much larger, with 6861 and 5931 dihedral transition events per ns per protein chain for the simulations without a counterion and with a counterion, respectively.

One would like to compare available NMR data, e.g. nuclear Overhauser effects, J-coupling constants and order parameters, with these solution simulations. Just as X-ray crystallographic data are best compared with crystal simulations, solution structure data should be compared to solution simulations. This will be addressed in later additions to this work.

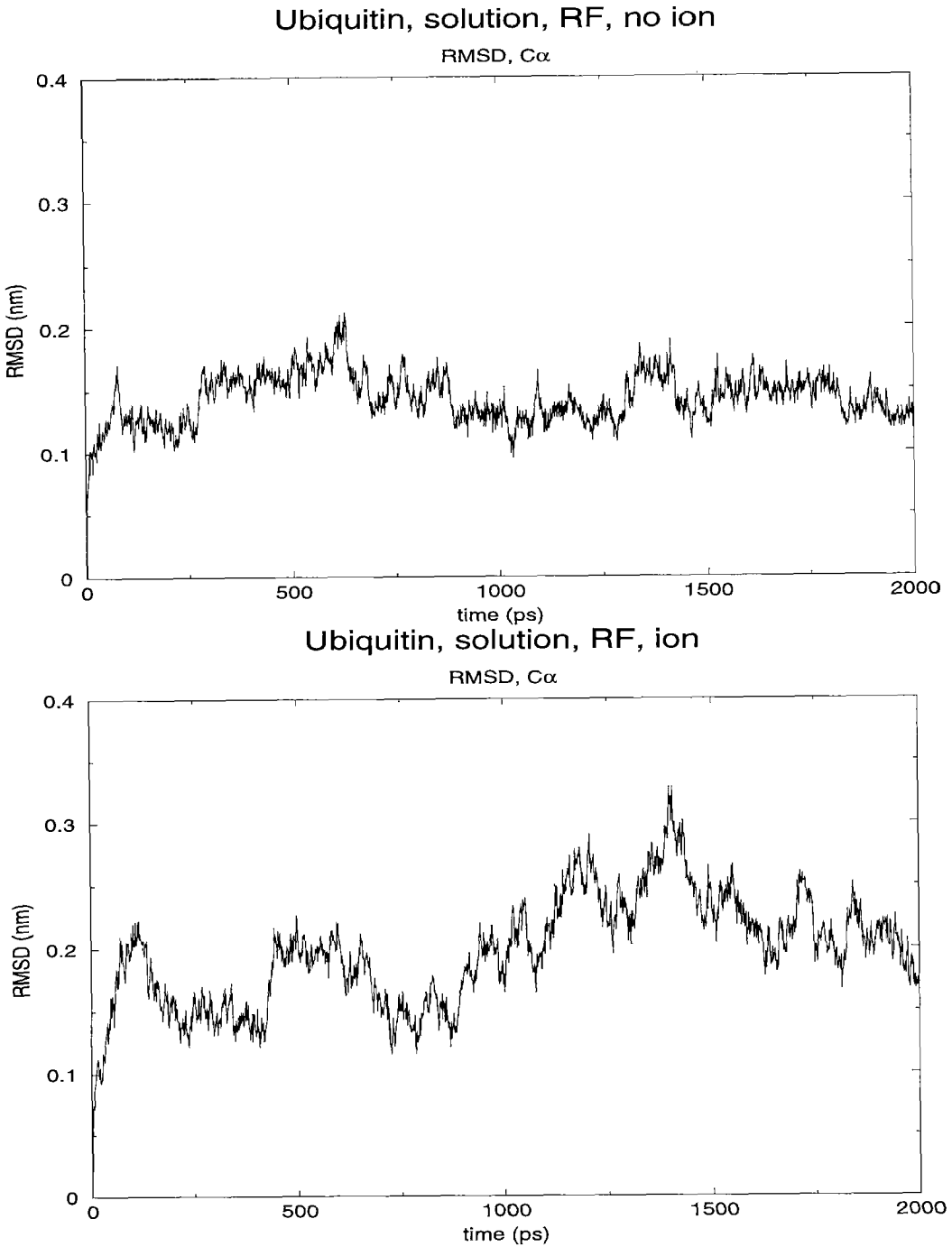

Figure 5.4: *Atom-positional root-mean-square deviation (RMSD) of the C_α atoms in nm from the initial crystallographic structure for the solution simulations, as a function of simulation time. Each chain was least-squares fit to the initial structure using the C_α atoms of residues 1–72, to remove overall translation or rotation. The top figure shows the simulation with a reaction field treatment of longer-ranged electrostatic interactions, but without a neutralizing counterion. In the bottom figure, the simulation with a reaction field and a neutralizing counterion is displayed.*

5.4.5 What are the counterions doing?

It would be interesting to follow what the counterions are doing, in the simulations where they occur. This will be addressed in later additions to this work.

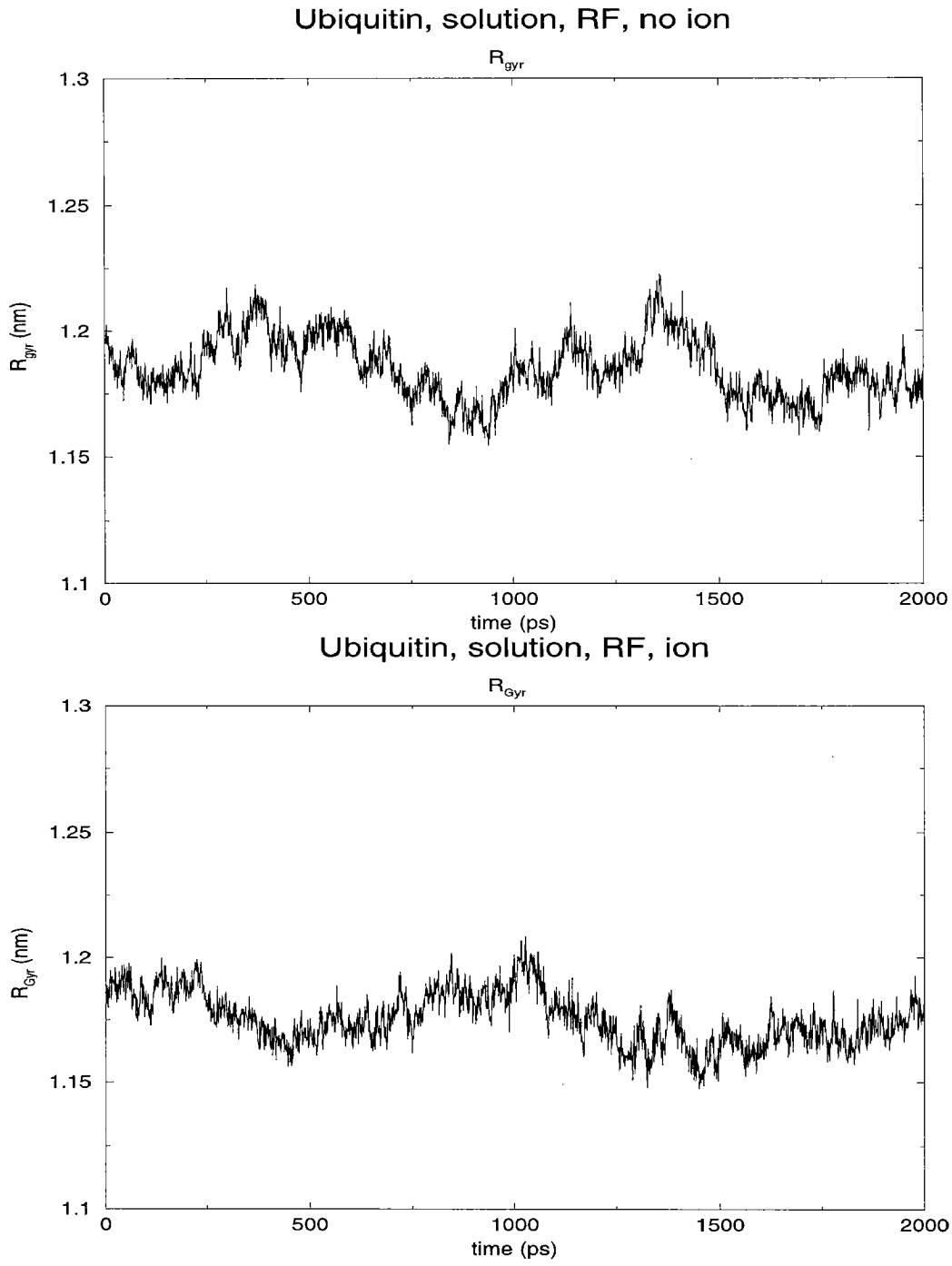

Figure 5.5: *Protein chain radius of gyration (R_{Gyr}) for the solution simulations, in nanometers, as a function of simulation time. In the top figure, the simulation with a reaction field treatment of longer-ranged electrostatics, but without counterion is shown, in the bottom figure, the simulation with a reaction field and a neutralizing counterion.*

5.5 Discussion

We have simulated ubiquitin in crystal and in solution, under various conditions. Without a reaction field treatment of longer-ranged electrostatic forces, the crystal simulations cannot reproduce crystallographic experimental data very well, but with a reaction field, the results are

good. This is true both with and without the addition of neutralizing counterions, but adding counterions reduces the fluctuations and mobility in the protein. In solution, the deviations from the initial structure are larger and the proteins fluctuate more, as expected. There is again some stabilization by the addition of a neutralizing counterion.

There are very few crystal simulations of proteins reported in the literature. This may be because the force-fields are usually parameterized for a solution environment, or because many proteins are so large, that treating several protein chains in a crystalline environment has been considered unfeasible.

One must distinguish between crystal simulations and periodic boundary simulations. Periodic boundaries (toroidal boundary conditions) are a scheme for removing boundary effects by removing all boundaries. This scheme is used also with solution simulations, where the periodic boxes used must be large enough that the protein molecule has pure solvent surroundings and experiences minimal influence from the periodically repeated copies of itself. The minimum image convention is applied to remove pseudo-crystalline effects, and is usually implemented by choosing a short enough cut-off radius for the non-bonded interactions. In a simulation of a protein crystal, one certainly wants protein molecules to interact with their crystal neighbors. But one still does not want a single protein chain to interact directly with its periodically repeated copies. The motions of periodic copies are by definition perfectly correlated, and therefore much more correlated than what would be expected for motions of crystal neighbors. To simulate a protein crystal, one should instead include several copies of the protein in the periodically repeated box, and ensure that any given protein molecule does not interact with periodic copies of itself.

Simulations of protein crystals provide good tests of the simulation methodology. Firstly, the most accurate experimental data available on protein structure is derived from crystallographic experiments. One can thus directly compare a crystal simulation with the X-ray data, without the systematic error of assuming that a solution structure is exactly the same as the structure in the crystal. Granted that this systematic error has been shown to generally be small in cases where NMR solution structures and X-ray crystal structures have been compared, it is still a more satisfying situation when the comparison can be done without worries whether a certain discrepancy is a "crystal contact". If there is an effect of a crystal contact, this should indeed be reproduced by the crystal simulation. Secondly, the problem of treating electrostatics becomes more challenging in a crystal environment, due to the much higher charge density in the system. In a solvent surrounding, there is a large screening bulk of water which dampens electrostatic interactions. While there is still quite some water present in protein crystals – 34 volume percent in the ubiquitin crystals [84], they are dense enough that charged protein groups will interact appreciably with charges on other protein molecules, and problems regarding electrostatics are highlighted.

It is often stated that the Ewald summation treatment automatically gives the only correct treatment of long-range electrostatics [79, 80, 87, 88]. However, since the summation involved is only conditionally convergent, the choice of summation order becomes vital. This is in the context of Ewald summation equivalent to the assumption of certain boundary conditions [1]. Åqvist & Hansson [89], for instance, have shown that using or not using a charge group based treatment of an aqueous solvent is equivalent to using different boundary conditions for the Ewald summation, and there arises the need of further decisions as to which exact version of Ewald summation should be taken as the truth.

The basis of the effects of counterions is still not clear. Is it the overall neutralization of the system, making the reaction field method work better, which stabilizes the system? Or is it the

counterions sitting relatively fixed close to explicit charges on the protein surface, thereby turning charges into dipoles and lowering electrostatic noise? Interestingly in this context, Ibragimova and Wade [90] found a large effect of ionic strength in particle-mesh Ewald simulations. They note that in order to maintain the conformation of a peptide, the YAP-WW domain, which is experimentally known to have a fixed conformation but is easy to melt, it was necessary to add further pairs of ions, corresponding to an ionic strength of 0.2M, in addition to the counterions needed to neutralize the system.

In summary, we have found that long range electrostatics must be taken into account, for instance by a reaction field treatment, and that counterions further stabilize protein simulations.

5.6 Acknowledgements

Financial support from the Schweizerischer Nationalfonds, project number 21-41875.94 is gratefully acknowledged, and so is support from the Swedish Foundation for International Cooperation in Research and Higher Education (TH & WFvG).

Chapter 6

Increasing the Time Step and Efficiency of Molecular Dynamics Simulations: Optimal Solutions for Equilibrium Simulations or Structure Refinement of Large Biomolecules

6.1 Abstract

In biomolecular simulation, the different force field terms determine the motional frequencies found in a particular simulation. By eliminating the high frequencies, longer time steps can be achieved in equilibrium and refinement (non-equilibrium) simulations, sometimes at the expense of decreasing physical correctness of the simulated system. The different force field terms used in molecular simulations can be constrained or softened leading to a maximum time step of 5fs for a simulation, in which static equilibrium properties are not affected by the changes in the force field parameters introduced, but dynamical quantities are. Using a very smoothed, soft interaction function and increased hydrogen masses, a time step of 14fs can be reached. However, it should only be used in refinement simulations, in which the unphysical properties of the force field and dynamics are of minor importance.

6.2 Introduction

Since its introduction in the sixties [91] the computer simulation method of molecular dynamics (MD) has become a well-known method to investigate the equilibrium properties of a variety of molecular systems [1], in particular biomolecular systems [32, 92, 93]. In the eighties, MD simulation was introduced as a technique to efficiently search configurational space in the process of structure determination and refinement of proteins based on experimental NMR data [8, 9]. This application of MD simulation has since then become a standard tool in structure determination and refinement using NMR and spectroscopic or X-ray diffraction data [20, 64, 94]. Although in both applications of MD simulation techniques, (i) study of equilibrium properties and (ii)

structure refinement, the same methodology is used, their characteristic features and choice of optimal parameters are rather different:

Ad (i). By simulating the real atomic motions over a sufficiently long time period the equilibrium properties of a biomolecular system can be calculated from the simulated trajectory. Newton's or Lagrange's equations of motion are integrated using time steps of the order of 1fs for as long as possible a period, nowadays in the order of nanoseconds (or 10^6 steps). The goal is to produce a Boltzmann-weighted ensemble of structures, which are, moreover, properly dynamically related or correlated. So, the use of proper equilibrium MD is mandatory to obtaining meaningful results.

Ad (ii). In structure refinement, MD simulation is only used as a technique to search conformational space for structures that satisfy best a large number of spatial restraints or conditions, which originate from NMR spectroscopic or X-ray diffraction measurements: NOE (Nuclear Overhauser enhancement) atom-atom distance bounds, J-coupling constants, X-ray diffraction intensities, etc.. We write "best", since there will generally not exist a single molecular structure which satisfies all spatial restraints, because the measurement involves averaging over many different structures (in time and space) leading to average restraints. The goal is to produce one or a handful (10, 20, 50, depending on the flavour of the particular research group) of structures that satisfy as good as possible a given set of spatial restraints. Here, the use of MD simulation techniques is validated only by the quality of the final structures generated.

Although both applications put different demands on the MD simulation techniques used, they share a need for extensive sampling of conformational space, that is, as long simulation periods as possible at a given accuracy and computational cost. In other words, increasing the time step and efficiency of MD simulations is of interest to both applications.

The four basic limitations of classical MD simulation in describing a biomolecular system are the following.

1. The use of classical-dynamical equations of motion to simulate motion along the chosen degree of freedom, thereby neglecting quantum effects.

2. The finite size of the system or number of degrees of freedom that can be simulated, currently typically 10^4 - 10^5 atoms or degrees of freedom.

3. The finite accuracy of the energy function or force field representing the energy of the system as function of the degrees of freedom along which the motion is simulated.

4. The finite length of or time period covered by a simulation, which limits the sampling and time scale of processes that can be simulated, currently typically 10^{-8} seconds.

In order to lengthen the time period covered by an MD simulation a variety of approaches has been tried in the past.

1. Reduction of the number of degrees of freedom to be sampled to the minimum that is essential to the particular properties of the system one is interested in. Examples of removing of, hopefully, non-essential degrees of freedom are the following

i. In structure refinement of proteins, solvent degrees of freedom are generally omitted [95, 96], although it is experimentally known that protein structure is sensitive to solvent composition.

ii. Aliphatic hydrogen-atom degrees of freedom can be removed by treating CH_n-groups as united (effective) atoms, an approximation which reduces the number of atom-atom interactions in proteins by a factor of 4, in membranes by a factor of up to 9, without significant loss of accuracy [97].

iii. If hydrogen atoms are to be kept as interaction sites, but their positions need not be sampled, they can be treated as so-called virtual atoms or interaction sites, which have zero mass (no inertia) and a position that is completely determined by (generally neighbour) non-virtual atoms (which have non-zero mass) [98–101].

iv. Bond-length or bond-angle degrees of freedom can be eliminated from the system by treating them as constrained [95, 96, 102, 103].

2. Simplification of the energy function or force field leading to a reduction in the number and complexity of atom-atom interactions to be evaluated, and so to a reduced computational cost. Since the calculation of non-bonded interactions dominates the cost of a molecular simulation, a considerable speed-up may be obtained at the expense of loss of force field accuracy. Examples are the following.

   i. Neglect of (long-range) electrostatic interactions [95, 96].

   ii. Use of a very short (smaller than $0.8nm$) cut-off radius for non-bonded interactions or neglect of attractive van der Waals interactions [95, 104].

3. Use of multiple-time-step (MTS) algorithms to integrate the equations of motion. When integrating the forces arising from different terms in a force field that are characterized by different relaxation times, it can be efficient to use different integration time steps Δt for the different force field terms. If for the more costly force field terms a longer time step can be used, the efficiency of the integration of the equations of motion is enhanced. MTS algorithms have been used for integrating bond-stretching and bond-angle-bending forces separately from the remaining forces [105–107], and for integrating long-range Coulomb and van der Waals forces separately from the remaining ones in the so-called twin-range method [32].

4. Reduction of the frequencies of the fastest motions in the system will allow for a longer time step Δt to be used: Algorithms to integrate equations of motion forward in time are based on the approximation that the forces or their derivatives are constant during the integration time step Δt, which approximation is of course more exact the smaller Δt is. This approximation is exact in the special case that the force is constant or in other words, if the derivative of the force or the second derivative of the energy function V, viz., the curvature of V equals zero. So the length of the integration time step Δt is limited by the size of the local curvature of V or in physical language, by the highest-frequency (ν_{max}) motions in the molecular system:

$$\Delta t \ll \nu_{max}^{-1} = \tau \tag{6.1}$$

For example, for the 1-dimensional classical harmonic oscillator with potential energy function

$$V(x) = \frac{1}{2}Kx^2 \tag{6.2}$$

and mass m, one has

$$\nu = \frac{(K/m)^{\frac{1}{2}}}{2\pi} \tag{6.3}$$

and the curvature equals the force constant K. The larger the curvature K of V(x), the higher the frequency ν of the motion and the smaller time step Δt must be used. Expressions (6.1-6.3) show different ways the time step in an MD simulation can be lengthened.

A. Redistribution of the total mass of a molecular system over the atoms such that the range of motional frequencies is reduced. When the inertia of the fastest motions is increased, ν_{max} is reduced thereby allowing a longer time step [108]. Of course, dynamical properties are modified when modifying atomic masses non-uniformly, but thermodynamic properties remain unchanged. For flexible molecules with internal spatial constraints such as bond lengths or angles this is not true due to metric-tensor effects [109]. By making the atomic masses more equal, that is by increasing the masses of hydrogens and reducing the masses of non-hydrogen atoms, a lengthening of the time step by 50% can be obtained for proteins [110] and for water [111].

B. Softening or smoothing of the high-frequency (most strongly curved) interaction terms in the potential energy function V will also allow the use of longer integration time steps Δt. This technique is often used in the earlier stages of structure refinement in order to enlarge the radius of convergence of the restraining simulation and in order to avoid a reduction of Δt when simulating at high temperature in simulated annealing refinement.

C. High frequency motions can also be eliminated from a molecular system by the application of atom-atom distance constraints. Two types of distance constraints can be distinguished.

   1. Holonomic (time-independent) constraints can be implemented in two ways: (i) by formulating Lagrange equations of motion in generalized (e.g. torsional) coordinates [112, 113], or (ii) by formulating these equations in Cartesian coordinates (i.e. using Newton's equations) and then using Lagrange multipliers to satisfy the constraints at each MD time step [3, 114, 115].
   2. Soft or adiabatic (variable length) distance constraints can only be implemented using the latter technique. The difference with respect to hard or holonomic constraints is that the length of a constrained distance is not a constant throughout the simulation, but varies per integration time step without involving kinetic energy. Reich [116] applies a smoothing operator to the Hamiltonian of a non-constrained system, which suppresses high-frequency oscillations and formulates its effect in terms of a soft-constrained Hamiltonian. An alternative is to adjust the length of a constrained distance adiabatically at each time point such that no strain is built up along the constrained degree of freedom.

How the various techniques to reduce the frequency of the fastest motions can be applied and combined most effectively to lengthen the MD time step depends on the hierarchy of frequencies

originating in different types of interatomic interactions governing the degrees of freedom of a particular molecular system. It also depends on the goal of a simulation. In standard structure refinement [95, 96], solvent degrees of freedom are omitted, and only solute torsional-angle degrees of freedom are sampled using a simplified and smoothed energy function, with the goal of obtaining a model structure of a protein. The severe approximations allow for fast refinement at the cost of loss of physical content, which is no problem as long as the quality of the obtained model structure can be independently assessed. On the other hand, in equilibrium simulations to study the dynamics of a protein only a few approximations can be accommodated without significant loss of accuracy, e.g. elimination of the bond-length degrees of freedom and aliphatic hydrogen degrees of freedom.

The literature on the subject is rather confusing. One reason is that different techniques are often not compared to each other under identical circumstances. Another reason is that confusing nomenclature is used. Recently, it has been suggested, that by integrating Lagrange equations of motions for torsional-angle degrees of freedom using an implementation in generalized coordinates, conformational space is more efficiently sampled than by using Cartesian coordinates [95, 96]. This conveys a misunderstanding. The coordinate system used will not determine the efficiency of sampling, the choice of degree of freedom along which the sampling is done may, of course, affect the efficiency. Here, coordinate system and degree of freedom are confused. In fact, one can integrate Lagrange equations of motion for torsional degrees of freedom using Cartesian coordinates and Lagrange multipliers as efficient as, if not more efficient than, when integrating Lagrange equations of motion formulated in torsional, generalized coordinates. A third cause of confusion is that in the literature structure refinement protocols are compared, which differ in more than one aspect [95, 96]. This makes it difficult, if not impossible, to analyze which of the many differences between the methods that are compared is responsible for the observed difference in efficiency.

In order to clear the field and to enable a rational choice of approximations aimed at speeding up biomolecular simulations, we review the topic and list the various aspects and choices involved in Section 6.3. In Sections 6.4 and 6.5 we illustrate the findings with examples: a simple model chain of 100 CH_n atoms and a small protein, bovine pancreatic trypsin inhibitor (BPTI), both using simulations in vacuo. Section 6.6 contains a number of practical conclusions.

6.3 Methods

In this section, we briefly review and list a number of aspects and concepts of molecular simulation that are of importance when analysing the computational efficiency and accuracy of equilibrium or structure refinement simulation.

6.3.1 Simulation and sampling methods

Since biomolecular conformational space is too large to be exhaustively sampled, one generally has to use in biomolecular modeling or structure refinement heuristic methods for sampling and searching for low-energy conformers. An overview of types of methods to search and sample conformational space can be found in [117]. Only a subset of the great variety of methods has been tried in structure refinement based on spectroscopic or diffraction data. The most widely used sampling methods are the following.

1. Non-Boltzmann sampling, such as
   a. repeated distance geometry calculations
   b. structural database sampling
   c. random or gradient-driven variation of torsional angles sampling
2. Boltzmann sampling, such as
   a. conventional or configuration-bias Monte Carlo (MC) simulation [118]
   b. molecular dynamics (MD) simulation [1]
   c. stochastic dynamics (SD) simulation [119]

The efficiency of the sampling is generally restricted by the general nature of the energy surface or the function V. The occurrence of high-energy barriers between local minima may inhibit proper sampling. Therefore, techniques have been developed to enhance the sampling power of the methods [117]. Only methods of type 1.c and of type 2 involve stepping through conformational space. In gradient-driven step-methods force-biased MC, MD, or SD simulations, the maximum possible step size is determined by the local curvature of the energy function, and in the dynamics simulation methods by the mass distribution along the degrees of freedom that are sampled.

6.3.2 Hierarchy of motional frequencies in biomolecular systems

In biomolecular systems a hierarchy of motional frequencies originating in different types of interatomic interactions can be distinguished [64, 111, 120]. In order of decreasing frequency or increasing smoothness of the corresponding atom-atom interaction term in the biomolecular force field we have (Table 6.2):

I. Bond-stretching vibrations with an approximate oscillation or relaxation time $\tau_I \approx 10fs$ for bonds involving a hydrogen atom and $\tau_I \approx 20fs$ for bonds involving only carbon or heavier atoms.

II. Bond-angle bending vibrations with $\tau_{II} \approx 20fs$ for bond angles involving hydrogen atoms and $\tau_{II} >\approx 40fs$ for bond angles involving only carbon or heavier atoms.

III. Improper dihedral angle vibrations due to force field terms used to impose the proper chirality at chiral CH_1 united atoms or to impose planarity on ring structures with conjugated double bonds with $\tau_{III} >\approx 30fs$.

IV. Torsional-angle vibrations around double bonds (e.g. peptide bonds) with $\tau_{IV} \approx 30fs$.

V. Water or solvent librational vibrations with $\tau_V \approx 30fs$.

VI. Torsional-angle vibrations around single bonds with $\tau_{VI} \approx 40fs$ for torsional angles involving a hydrogen atom and $\tau_{VI} >\approx 80fs$ for torsional angles involving only carbon or heavier atoms.

VII. Motions dominated by (not softened) van der Waals contacts and short-range (e.g. hydrogen bonding) Coulomb interactions with $\tau_{VII} \approx 150fs$.

VIII. Motions dominated by long-range Coulomb (ionic, dielectric) interactions with relaxation time $\tau_{VIII} > \approx 2000 fs$.

It is this hierarchy that is exploited to enhance the efficiency of a simulation through the use of longer time steps Δt [121].

6.3.3 Choice of molecular model and force field

In its most simple form a biomolecular force field contains the following terms, where we have used the functional form of the GROMOS96 force field [2, 44] as example.

1. Bond-stretching terms

$$V^{bond}(\vec{r}(t)) = \frac{1}{4} \sum_{n=1}^{N_b} K_{b_n} \left[b_n^2(t) - b_{0_n}^2 \right]^2 \tag{6.4}$$

in which b_{0_n} is the ideal bond length and $b_n(t)$ is the actual (at time point (t)) bond length of the bond with sequence number n in the list of N_b bonds in the molecular system, and K_{b_n} determines the strength of the n-th bond, which determines the curvature or smoothness of term (6.4).

2. Bond-angle bending terms

$$V^{angle}(\vec{r}(t)) = \frac{1}{2} \sum_{n=1}^{N_\theta} K_{\theta_n} \left[cos(\theta(t)) - cos(\theta_{0_n}) \right]^2 \tag{6.5}$$

in which θ_{0_n} is the ideal bond angle and $\theta_n(t)$ the actual bond angle value of the n-th bond angle in the list of N_θ bond angles, and K_{θ_n} is the force constant determining the curvature of term (6.5)

3. Improper dihedral angle terms

$$V^{har}(\vec{r}(t)) = \frac{1}{2} \sum_{n=1}^{N_\xi} K_{\xi_n} \left[\xi_n(t) - \xi_{0_n} \right]^2 \tag{6.6}$$

in which ξ_{0_n} is the ideal improper dihedral, $\xi(t)$ the actual improper dihedral angle value, and K_{ξ_n} the force constant of the n-th improper dihedral.

4. Torsional (or proper) dihedral angle terms

$$V^{trig}(\vec{r}(t)) = \sum_{n=1}^{N_\varphi} K_{\varphi_n} \left[1 + cos(\delta_n) cos(m_n \varphi_n(t)) \right] \tag{6.7}$$

in which δ_n is the phase shift (0 or π), m_n the multiplicity, $\varphi_n(t)$ the actual value and K_{φ_n} the force constant of the n-th torsional angle.

5. Non-bonded (Lennard-Jones, Coulomb) interaction terms

$$V^{LJ}(\vec{r}(t)) = \sum_{non-bonded pairs(i,j)} \frac{1}{\alpha_{LJ}(i,j)C_{126}(i,j) + (r_{ij}(t))^6} \cdot \left[\frac{C_{12}(i,j)}{\alpha_{LJ}(i,j)C_{126}(i,j) + (r_{ij}(t))^6} - C_6(i,j)\right] \tag{6.8}$$

and

$$V^{CRF}(\vec{r}(t)) = \sum_{non-bonded pairs(i,j)} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_1} \left[\frac{1}{\left[\alpha_C(i,j) + (r_{ij}(t))^2\right]^{\frac{1}{2}}} - \frac{\frac{1}{2}C_{rf}(r_{ij}(t))^2}{\left[\alpha_C(i,j) + R_{rf}^2\right]^{\frac{3}{2}}} - \frac{1-\frac{1}{2}C_{rf}}{R_{rf}}\right] \tag{6.9}$$

in which r_{ij} is the actual distance between atoms i and j. $\alpha_{LJ}(i,j)$ and $\alpha_v(i,j)$ are so-called soft-core parameters for the Lennard-Jones and charge interactions [122], respectively, which can be used to convert the hard-core singularity at $r_{ij} = 0$ for $\alpha_{LJ}(i,j) = \alpha_C(i,j) = 0$ into a smooth soft core for $\alpha_{LJ}, \alpha_C > 0$. The r^{-6} and r^{-12} Lennard-Jones parameters for atom pair (i,j) are $C_6(i,j)$ and $C_{12}(i,j)$, respectively and

$$C_{126}(i,j) = \begin{cases} \frac{C_{12}(i,j)}{C_6(i,j)} & if \quad C_6(i,j) \neq 0 \\ 0 & if \quad C_6(i,j) = 0 \end{cases} \tag{6.10}$$

The partial charge of atom i is denoted by q_i and the constant

$$C_{rf} = \frac{(2\varepsilon_1 - 2\varepsilon_2)(1+\kappa R_{rf}) - \varepsilon_2(\kappa R_{rf})^2}{(\varepsilon_1 + 2\varepsilon_2)(1+\kappa R_{rf}) + \varepsilon_2(\kappa R_{rf})^2} \tag{6.11}$$

determines the Poisson-Boltzmann reaction field [42] contribution due to a medium with relative dielectric permittivity ε_2 and inverse Debye screening length κ outside the cut-off sphere with radius R_{rf}, and ε_1 is the relative permittivity inside this sphere. The dielectric permittivity of vacuum is ε_0.

The smoothness of the energy function consisting of terms (6.4-6.9) can be increased by decreasing the force constants K_{b_n}, K_{θ_n}, K_{ξ_n}, and K_{φ_n} in terms (6.4-6.7) and by increasing the values of the soft core parameters α_{LJ} and α_C in terms (6.8-6.9).

6.3.4 Choice of degrees of freedom to be simulated or sampled

The hierarchy of motional frequencies in a molecular system (as determined physically or artificially by the force field terms used) will determine which degrees of freedom should be omitted in order to allow for a longer time step Δt.

1. Omission of aliphatic hydrogen degrees of freedom has two advantages. The number of non-bonded atom-atom interactions in the summations of (6.8-6.9) is more than halved for proteins and lipids. Secondly, the high-frequency ($\tau_I \approx 10fs$) hydrogen motions need not be integrated.

2. If the aliphatic hydrogen atoms are kept as virtual interaction sites (with zero mass) only the second advantage is retained.

3. This is also true, when the bonds or in addition the bond angles involving hydrogen atoms are constrained in a simulation.

4. Omitting water or solvent degrees of freedom essentially yields only the first advantage since their typical motional frequencies differ not much from those in proteins. However, since solute-solvent and solvent-solvent interactions easily account for 80-99% of the computational effort in explicit solvent simulations, omitting solvent degrees of freedom yields by far the largest gain in computational efficiency.

When omitting degrees of freedom, their (mean) effect on the remaining explicitly simulated degrees of freedom should be as much as possible incorporated into the energy function for these degrees of freedom. Mean solvation force terms have been reviewed in [123].

6.3.5 Equations of motion

The classical equations of motion have been given by Lagrange in a most general form.

$$\frac{d}{dt}\left(\frac{\partial L(q,\dot{q})}{\partial \dot{q}_i}\right) = \frac{\partial L(q,\dot{q})}{\partial q_i} \qquad \text{i=1,2,...,} N_{df} \tag{6.12}$$

where q_i denote the generalized coordinates, $\dot{q}_i$ their time derivatives and the Lagrangean $L(q,\dot{q})$ is the kinetic energy $K(\dot{q})$ minus the potential energy $V(q)$ of the system which contains N_{df} degrees of freedom. When using Cartesian coordinates $q \equiv x$, one has $K(\dot{x}) = \frac{1}{2}m\dot{x}^2$ and equations (6.12) reduce to Newton's equations of motion (for N_{df} degrees of freedom).

$$m_i \frac{d^2 x_i}{dt^2} = \frac{-\partial V(x_1, x_2, \ldots, x_{N_{df}})}{\partial x_i} \qquad \text{i=1,2,...,} N_{df} \tag{6.13}$$

When considering branched polymers, the choice of internal coordinates, bond lengths, bond angles, and torsional angles seems to be natural. However, the equations of classical dynamics (6.12) expressed in the N_{df} internal, generalized coordinates $q_i \equiv \theta_i$,

$$\begin{aligned} \sum_{j=1}^{N_{df}} a_{ij} \frac{d^2\theta_j}{dt^2} &= \frac{-\partial V(\theta_1, \theta_2, \ldots, \theta_{N_{df}})}{\partial \theta_i} \\ &- \sum_{j=1}^{N_{df}} b_{ij} \left(\frac{d\theta_j}{dt}\right)^2 - \sum_{j=1}^{N_{df}} \sum_{k=1}^{N_{df}} c_{ijk} \left(\frac{d\theta_j}{dt}\right)\left(\frac{d\theta_k}{dt}\right) \qquad 1,2,\ldots,N_{df} \end{aligned} \tag{6.14}$$

are considerably more complex than when expressed in Cartesian coordinates. They contain two additional summations over the number of degrees of freedom and two additional quadratic (i. e. non-linear) terms in the generalized velocities. Equations (6.14) have been presented in different forms [103, 112, 113, 124–129], and the coefficients a_{ij}, b_{ij}, and c_{ij} depend on the atomic masses and the molecular topology of the polymer considered.

6.3.6 Choice of coordinate system

Since the Cartesian (Newtonian) equations of motion (6.13) do not involve the explicit coupling of the equations through the summation over the index j in (6.14) and lack the two non-linear terms depending on the generalized velocities in (6.14), Cartesian equations of motion have been the method of choice in MD and SD simulations of biomolecular systems for obvious efficiency reasons. Even in the presence of spatial constraints, such as bond-length and bond-angle constraints, it is possible to use a Cartesian coordinate system, while imposing the constraint conditions at every time step through the Lagrange multiplier technique [130]. When only bond lengths are constrained, it takes for a protein only a few percent of the total simulation effort to determine the Lagrange multipliers [131]. This is done by iteratively solving the constraint equations [3, 115]. When bond-angle degrees of freedom are to be constrained in addition, the simple SHAKE algorithm [3] becomes inefficient [102]. However, using techniques proposed in [114] the convergence of solving the constraint equations can be improved considerably. In this case it takes for a protein at most 10-20% of the total simulation effort to determine the Lagrange multipliers.

Simulation of a protein through the equations of motion in generalized, torsional coordinates (6.14) requires a larger computational effort, since at each time step a set of N_{df} non-linear equations is to be solved. One iteration to this end may take as much computational effort as the calculation of all forces and energies, thereby roughly doubling the overall computational effort [95]. Secondly, it is practically impossible in the framework of generalized, torsional coordinates to treat chain closure of a chain of atom-atom distance constraints while maintaining flexibility within the closed chain. As a consequence, disulfide bridges or other polypeptide chain closures are treated as unconstrained degrees of freedom in simulations based on equations (6.14). This implies that their high-frequency motions are still present in the system, unless they are removed using other techniques (e.g. potential-energy smoothing).

In summary, the use of equations of motion in non-Cartesian coordinates should be avoided if the computational efficiency is to be optimized.

6.3.7 Choice of numerical algorithm to integrate Newton's equations of motion

A great variety of time integration algorithms has been tried over the years in MD simulation [1, 132]. Since Newton's equations of motion (6.13) are time-invariant and do not contain terms dependent on the atomic velocities, simple algorithms that are time-invariant are most efficient to integrate them [132, 133]. The leap-frog algorithm [5] and the equivalent Verlet algorithm [134] are most efficient.

6.3.8 Use of multiple-time-step algorithms

One way to exploit the hierarchy of motional frequencies in biomolecular systems is to use multiple-time-step (MTS) algorithms, in which different time steps Δt_I, Δt_{II-VII} or Δt_{VIII} are used, each satisfying condition (6.1) with respect to the oscillation or relaxation times τ_I, τ_{II-VII}, or τ_{VIII} of the various atom-atom interaction terms discussed in Section 6.3.2, when integrating the contributions of the different forces. Since evaluating the non-bonded force terms (6.8-6.9) is by far the most costly, it is efficient to integrate high-frequency bond-stretching and bond-angle

bending forces (τ_{I-II}) [106, 107] and the very low-frequency non-bonded forces (τ_{VIII}) [32] with time steps shorter and longer than the standard time step by which the remaining forces are integrated ($\tau_{III-VII}$).

6.3.9 Redistribution of atomic mass along degrees of freedom

Another way to exploit the hierarchy of motional frequencies is to change atomic masses such that the fastest motions are slowed down and the slowest motions are sped up. A simple approach is to increase the masses of hydrogens to those of carbons or nitrogens, which allows the MD time step to be lengthened, however, at the cost of distortion of the dynamics of the system. Thermodynamic properties, however, are not influenced by mass changes. This technique is also used in structure refinement using MD along torsional angle degrees of freedom [95].

We note that a scaling of the total mass of a molecular system makes not much sense: At constant temperature, the scaling of all atomic masses with a factor λ (or scaling the unit of mass by a factor of λ^{-1}) is equivalent to scaling the time with a factor $\lambda^{-\frac{1}{2}}$. The increase in size of the time step due to the mass increase will be precisely offset by an equally big reduction of the time unit.

6.3.10 Softening or smoothing the highest-frequency interaction terms

The oscillation or relaxation time of the highest-frequency motions can be reduced by decreasing the force constant of the corresponding force field terms. For example, a reduction of the strength K_{b_n} of the bond stretching term (6.4) by a factor 10 will allow a 3 times longer time step, as indicated by relation (6.3). The price to be paid is an increase of the bond-length fluctuations by a factor of 3. Similar conditions apply to the bond-angle term (6.5) and the improper dihedral angle term (6.6).

A peculiar situation may occur when using very soft bond-angle forces with the torsional energy term (6.7) switched on. If the bond-angle i-j-k defined by atoms i, j, and k increases towards 180° and beyond, the torsional dihedral angle i-j-k-l will make a step of 180°, which may induce a large sudden change in torsional-angle energy. This effect could be avoided by introducing a coupling between bond-angle and torsional angle forces.

6.3.11 Use of hard, soft or adiabatic atom-atom distance constraints

High-frequency motions can be eliminated from a molecular system by constraining the corresponding degrees of freedom. This can be implemented either by formulating the equations of motion in generalized coordinates or by sticking to Cartesian coordinates and using the Lagrange multiplier technique to impose the constraints. When using soft or adiabatic (variable length) constraints the latter method is most practical.

6.3.12 Consequences of the use of the various techniques and approximations

When applying constraints, so-called metric tensor correction terms may have to be added to the interaction function V, depending on the type of constraints used [109, 135]. Moreover, a physical force field calibrated for use in unconstrained simulations may have to be recalibrated

for use in conjunction with hard constraints, which will rigidify the molecules [102], or with soft or adiabatic constraints, which will mollify the molecules.

Another relevant aspect is the quantum-mechanical nature of particular intramolecular vibrations. Bond-stretching vibrations have frequencies ω or wave numbers in the range 2000-4000cm^{-1}. At room temperature, k_BT corresponds to a wave number of 200cm^{-1}. Thus we have (k_B is Boltzmann's constant and h is Planck's constant)

$$\frac{h}{2\pi}\omega \gg k_BT \tag{6.15}$$

which implies that the bond-stretching vibrations are essentially of quantum-mechanical nature and that only the ground state will be populated. Treating the bonds as hard constraints is most likely to be a more correct approximation of the quantum dynamics than adiabatic dynamics or classical (harmonic oscillator) dynamics, with its different energy distribution, would be. On the other hand, zero-point energy is neglected.

Another unpleasant aspect of fully flexible molecular models is the presence of weakly coupled modes of different frequencies, which makes the energy distribution over these modes in a molecular simulation a slow process. This is a technical problem, which can be avoided by coupling the different degrees of freedom (high-frequency ones versus the rest) separately to a heat bath [2]. This problem is likely to occur when MTS algorithms are used for higher-frequency motions.

An obvious, non-desirable effect of the softening or smoothing of the highest-frequency interaction terms in V is an enlargement of the structural fluctuations of the molecules along the degrees of freedom for which the interaction has been smoothed. This may influence thermodynamic properties that depend on fluctuations.

The unpleasant consequences of using equations of motion in non-Cartesian coordinates have been discussed in section 6.3.6.

6.3.13 Relative merits of the various time-saving techniques and approximations

We now reconsider the hierarchy of motional frequencies and the corresponding force field terms given before, and discuss the relative merits of applying the different time-saving techniques to lengthen the integration time step Δt to a particular motion or force field term.

I. Bond-stretching motion (τ_I=10-20fs). The most appropriate treatment of the bond-stretching degree of freedom is the use of hard constraints. It is a good approximation of their quantum-mechanical nature, it avoids the energy redistribution problem, and it can be very simply carried out using Cartesian coordinates (Newton's equations of motion) with the SHAKE method to determine the Lagrange multipliers [3]. Metric tensor effects are negligible [135] and force-field corrections are not necessary: the dynamics of the molecular system is not affected by the bond-length constraints [102]. Use of any of the the time-saving techniques (constraining of the degree of freedom, mass redistribution along the degree of freedom, smoothing of the force field term, MTS integration) may reduce the computational effort by up to a factor of 4 through a lengthening of the time step.

II. Bond-angle bending motion involving hydrogen atoms ($\tau_{II} \approx 20fs$). Constraining the bond-angle degrees of freedom involving hydrogen atoms would allow for a small gain

in time step. Metric tensor effects are probably small, and the heavy atom dynamics probably not much influenced. Alternatives are to use larger hydrogen masses or to reduce the force constants for the bond-angle terms involving hydrogens slightly.

II-VI Bond-angle bending motion of non-hydrogen atoms, torsional motion around double bonds, improper dihedral vibrations, and water librational motion ($\tau_{II-VI} \approx 30fs$). In simulations of biomuolecules including explicit water molecules, little gain in computational efficiency can be obtained by use of one of the time-saving techniques due to the presence of the high-frequency water molecule librations governed by the non-bonded (van der Waals, Coulomb) interactions. These limit the time step to about $\Delta t=2fs$. Using heavier hydrogens would allow for a somewhat larger $\Delta t=3fs$, at the cost of distorting dynamical properties. When simulating a macromolecule in vacuo, the bond-angle and stiff (double bond character) torsional-angle degrees of freedom could be treated using one of the time-saving techniques: constraining, mass redistribution, or smoothing of the force field terms. This may reduce the computational effort by up to a factor of 4. Typically, the time step is limited to about $\Delta t=8fs$ when using physically realistic force fields. By combining constraints, hydrogen mass scaling, and softening of the various interaction terms a time step $\Delta t=14fs$ can be reached. The use of bond-angle constraints or so-called torsion-angle dynamics [95, 96, 103] introduces two problems: (i) metric-tensor effects are not small, so should not be ignored [135], and (ii) the dynamics of the macromolecule is severely altered due to the rigidification of the molecular model, structural fluctuations and torsional-angle transitions are quenched [102]. To restore the proper physical behaviour of the molecular system a metric-tensor term should be added to the energy function V, and the force field should be recalibrated for use with bond-angle constraints.

VII. Motions dominated by van der Waals contacts, single-bond torsional interactions, and short-range Coulomb interactions, ($\tau_{VII} \approx 150fs$). These interactions dominate the essential degrees of freedom of molecular systems and should therefore not be approximated using time-saving techniques. They limit the time step using physically realistic force fields to $\Delta t=8fs$ for macromolecules and to $\Delta t=2fs$ when water is used as solvent. As before, softening of torsional and non-bonded interaction terms allows for larger Δt, but at a loss of physical accuracy of the molecular model.

VIII. Motions dominated by long-range Coulomb interactions ($\tau_{VIII} > \approx 2000fs$). Long-range electrostatic interactions can be computationally very demanding due to their dependence on a large number of atoms. Since these interactions are only changing slowly during a simulation, the use of multiple-time-step algorithms when integrating these forces can reduce the computational effort by up to a factor of ten [2, 32]. On the other hand, application of so-called particle-particle-particle-mesh (P3M) methods to evaluate electrostatic interactions [5] may reduce the computational effort by a factor of 100 over conventional Ewald summation techniques [136].

6.3.14 Testing the precision and efficiency of the various simulation techniques

To compare the various techniques aimed at lengthening the time step Δt in MD simulations, one should compare MD simulations (i.e. trajectories) that differ only in one aspect, all other

parameters, force field terms, set-up procedures being the same. Considering the variety of techniques discussed, a comprehensive comparison would require too much effort. Therefore, we have limited ourselves to a comparison of the following properties.

In equilibrium MD simulations without coupling to temperature or pressure baths, the total energy

$$E_{tot}(t) = E_{kin}(t) + E_{pot}(t) \tag{6.16}$$

should be conserved over the trajectory. This implies that the root-mean-square fluctuation of the total energy

$$\Delta E_{tot} \equiv < [E_{tot} - < E_{tot} >]^2 >^{\frac{1}{2}} \tag{6.17}$$

and the drift in E_{tot} should be ideally zero. Here a trajectory average is denoted by $< >$. Due to numerical and integration errors, ΔE_{tot} will be non-zero and could be used as an indication of the precision with which the equations of motion are integrated forward in time. One would expect that for longer integration time steps Δt larger values for ΔE_{tot} will be obtained. For very small time steps Δt the value of ΔE_{tot} should become independent of Δt, since in that limit, numerical errors are expected to dominate integration errors [131]. It makes little sense to calculate the relative fluctuation of the total energy $\Delta E_{tot} / < E_{tot} >$ as was done by Güntert et al. [95], since the zero-point of the potential energy E_{pot} and therefore of the total energy E_{tot} can be chosen at will. This means that the relative fluctuation $\Delta E_{tot} / < E_{tot} >$ can be made arbitrarily large or small depending on the choice of zero on the potential energy scale. A more useful quantity to measure the precision of the time integration is the ratio of the fluctuations in the total energy to those in the kinetic energy,

$$\Delta E_{tot} / \Delta E_{kin}, \tag{6.18}$$

since the latter should be independent of the time step and of the zero-point of the potential energy. This ratio measures how well the total energy is conserved given the kinetic and potential energy fluctuations characteristic for the molecular system at the chosen temperature [131]. For a more detailed analysis of energy conservation in MD simulations we refer to Mazur [133].

In non-equilibrium MD simulations as used in structure refinement, conservation of energy or correct dynamics is of no importance. In this case the efficiency of an algorithm can be measured by determining the computational effort required to obtain a correctly folded structure starting from an arbitrary or extended structure and using spatial restraints derived from NMR or X-ray data to guide the refinement.

6.3.15 Molecular dynamics simulations

Equilibrium MD simulations were carried out for two test systems in order to determine the degree of energy conservation as a function of MD time step size Δt and the maximum possible Δt value in dependence of the setting of the following force field and simulation parameters:

a. application of bond-length constraints (bond-constraints),

b. hydrogen mass m_H,

c. force constant K_{b_n} of bond-stretching terms,

d. force constant K_{θ_n} of bond-angle bending terms,

e. force constant K_{ξ_n} of improper-dihedral angle terms,

f. force constant K_{φ_n} of proper torsional-angle terms,

g. van der Waals and Coulomb soft-core parameters α_{LJ} and α_C,

h. atom partial charges.

If not explicitly stated otherwise, the force field parameter values are those of the GROMOS96 force field, and the hydrogen masses are m_H=1.008u. The first test system consisted of a linear C_{100} chain built of 98 united CH_2 atoms and 2 united CH_3 atoms as end groups. In the united atom approach, aliphatic hydrogen atoms are incorporated into the carbon atoms and not treated explicitly. So, some of the highest frequencies were not present: (i) there are no hydrogen atoms and thus the fastest bond-stretching vibrations and the fastest bond-angle bending vibrations are not present; (ii) the system has no improper dihedral angles which would be the fastest motions in the system after bond-length vibrations and bond-angle bending vibrations. The second test system was the small (58 amino acids) protein bovine pancreatic trypsin inhibitor (BPTI) including a variety of internal degrees of freedom. All simulations were carried out using the following conditions.

(i) Vacuum boundary condition.

(ii) An infinite cut-off radius, because the noise arising from the use of a cut-off scheme would heat the system and spoil the energy conservation. The fluctuation of the total energy (6.18) would be determined by the cut-off noise rather than by the time step used in the integration.

(iii) When SHAKE [3] was applied, a relative geometric precision of 10^{-6} was used. In molecular dynamics of proteins using a cut-off radius a lower relative SHAKE tolerance of 10^{-4} can be used, since the cut-off noise will dominate the fluctuation in the total energy.

In order to obtain an equilibrated starting structure in the C_{100} system, an extended chain was simulated for 1ns using standard conditions (2fs time step, weak coupling to a temperature bath of 300K using a relaxation time τ_T of 0.1ps; bonds kept constrained using the SHAKE algorithm with a relative geometric tolerance of 10^{-4}, and twin-range cut-off radii of 0.8nm and 1.4nm). This leads to a compact, almost spherically shaped molecule. This new structure was then used as a starting structure for a simulation of 20ps using the various force field and simulation parameter sets for which the total energy fluctuation was to be determined, still using temperature coupling to eliminate the energetic effect of a change in parameter settings used. Each equilibrated system was then simulated for 80ps without temperature coupling. These 80ps were used to determine energy conservation for the selected parameter combination as a function of Δt. For the second test system, the protein BPTI, the structure 1BPI of the Brookhaven protein databank [31] was used as a starting structure. The cysteine bridges were reduced, cysteine residues protonated. The following residues were used in their protonated form: Lysine (Lys) and arginine (Arg). The solvent water molecules were removed except for the four internal water molecules in BPTI. The GROMOS96 vacuum force field [2], code 43B1, was used to simulate the protein for 1ns at standard conditions. As for the C_{100} test system, parameters were then changed to various combinations and another 20ps were simulated using temperature coupling to absorb the energetic effects of changing the parameters. Afterwards, 80ps without temperature coupling were used for analysis. For determining the different frequencies and relaxation times, another 1ps was appended to these simulations. A time step of 0.5fs was used and every trajectory

configuration was saved for analysis. All simulations were carried out using the GROMOS96 package and force field [2].

6.4 Results

6.4.1 C_{100} test system

When simulating the C_{100} test system using a fully flexible model (without SHAKE to constrain bonds), a time step of 6fs can be achieved (Fig. 6.1). Using larger time steps leads to a loss of energy conservation due to integration errors. ΔE_{tot} is larger than ΔE_{kin}. The oscillation time of the vibrations of a C-C bond or a C-C-C bond-angle is around 30fs (table 6.1). These are the

	force field term					
	no constraints, fully flexible	bonds constrained				
		with dihedral forces, $s_{dih} = 1.0$		no dihedral angle forces, $s_{dih} = 0.0$		
		scaling factor s_{ba} for bond angle forces				
		s_{ba}=1.0	s_{ba}=0.3	s_{ba}=0.2	s_{ba}=0.1	s_{ba}=0.0
single bond	30	-	-	-	-	-
bond angle	30	85	128	171	171	-
torsional angle	128	256	171	-	-	-
max step size (fs)	6	17	17	25	25	22
average $\Delta\varphi$ (°) per step	2.0	6.5	5.3	10.5	9.4	16.3

Table 6.1: *Oscillation or relaxation times (in fs) of various degrees of freedom or force field terms in the C_{100} simulations at 300K. Deviations of force field parameters from GROMOS96 values is indicated by scaling factors. The maximum step size Δt is indicated and the corresponding average change of torsional-angle value per time step is given.*

fastest motions in the system and thus limit the time step. For a physically reasonable simulation ($\Delta E_{tot}/\Delta E_{kin} < 5 \cdot 10^{-2}$, figure 6.1), a time step of 1fs should be used. The average torsional angle change $\Delta\varphi$ at the largest possible time step of 6fs is 2°. Removing the bond-stretching vibrations from the system using the SHAKE algorithm leads to an increase of the maximal time step by almost a factor of three. Use of even larger time steps causes SHAKE to fail because the distortions from ideal bond geometry become too large. The average $\Delta\varphi$ is also increased by a factor 3 indicating that the motion along torsional degrees of freedom is not inhibited by constraining bonds. Bond-angle vibrations, however, become slower compared to the flexible model, with an oscillation time of about 85ps on average. A physically meaningful time step would lie around 4fs for such a system. Reducing the highest frequencies of the system by scaling all bond-angle force constants K_θ by a factor of 0.3 does not affect the energy conservation very much. The maximum time step that can be used stays the same. The vibrational frequencies are slightly differently distributed between the bond-angle and torsional-angle degrees of freedom. Bond-angle vibrations become slower and torsional-angle motions faster. The average $\Delta\varphi$ per step stays about the same. Further reduction of the bond-angle force constant was not possible in the presence of torsional-angle interactions. The very soft bond-angle forces allow the bond-angles to become larger than 180°, causing a jump of 180° for the torsional angles defined by the

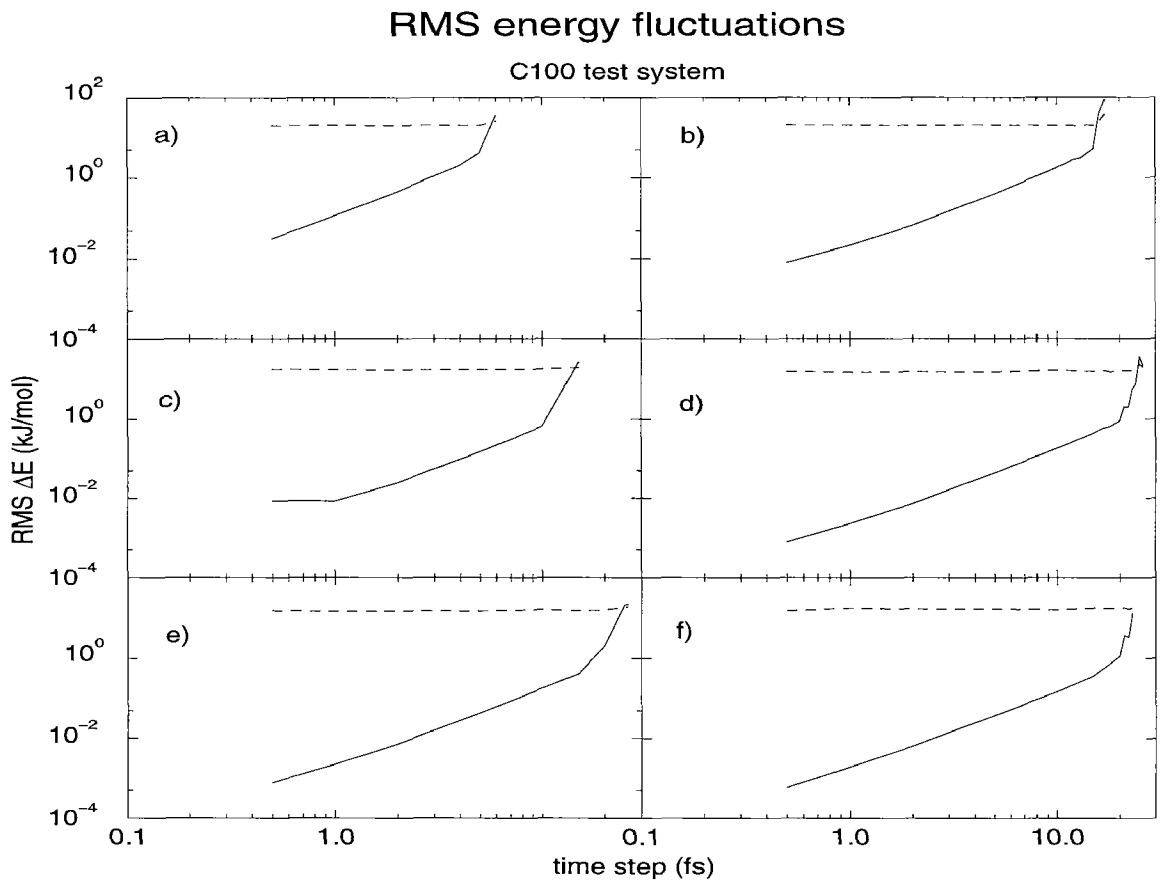

Figure 6.1: *Root-mean-square fluctuations of total (E_{tot}, solid line) and kinetic (E_{kin}, dashed line) energy for the C_{100} test system for different parameter setting as function of the time step at 300K; a) fully flexible model, b) bonds constrained, c) bonds constrained, bond-angle force constants K_{θ_n} scaled by a factor 0.3, d) bonds constrained, bond-angle force constants K_{θ_n} scaled by a factor 0.2, no torsional-angle forces, e) bonds constrained, bond-angle force constants K_{θ_n} scaled by a factor 0.1, no torsional-angle forces, f) bonds constrained, no bond-angle forces, no torsional-angle forces.*

same three atoms as define the bond-angle. These discontinuities in the torsional angle trajectories imply discontinuities in the torsional angle forces, which cause the system energy to diverge. So, when using even softer bond-angles, bond-angle force constants K_θ scaled by s_{ba}=0.2 and 0.1, the torsional-angle interaction term of the force field was turned off. Both these simulations led to nearly identical results. The largest achievable time step was 25fs, since bond-angle vibrations have an oscillation time of about 170fs. The average torsional-angle change at Δt=25fs was around 10°. Completely removing bond-angle forces (s_{ba}=0.0) reduced the maximum time step to 22fs. However, the large freedom of the system (now mainly a covalently bound chain of van der Waals atoms) led to a large average Δφ of 16.3°. Energy conservation was very good for all the simulations using softened or smoothed forces. At a time step of 10fs, the criterion for a physically correct simulation, $\Delta E_{tot}/\Delta E_{kin} < 5 \cdot 10^{-2}$, was fulfilled. But, since the force field used is very unphysical, this time step cannot be used to simulate equilibrium conditions of a real system.

6.4.2 BPTI test system

Simulating the protein bovine pancreatic trypsin inhibitor (BPTI) involves a wider spread of and also higher frequencies than the simple C_{100} test system. The fastest motions in a biomolecular simulation are the vibrations of bonds involving hydrogen atoms. The period of one oscillation of such a bond is around 10fs (table 6.2). Double bonds with their higher force constant have

	force field term						
	T=300K					T=1000K	
	no constraints	bonds constrained					
	m_H=1.008u	m_H=14.027u				m_H=1.008u	m_H=14.027u
				$s_{ba} = 0.5$	$s_{ba} = s_{imp}$ $=s_{dih} = 0.4$		$s_{ba} = 0.4$ $s_{imp} =$ $s_{dih} = 0.25$
single bond H	10	-	-	-	-	-	-
double bond	20	-	-	-	-	-	-
single bond	30	-	-	-	-	-	-
bond-angle 2H	20	20	57	102	102	20	102
bond-angle 1H	32	22	64	73	85	32	85
bond-angle	21	57	73	85	102	73	102
improper H	51	51	51	73	85	51	128
improper planar	51	102	64	73	128	102	85
improper tetrahedral	73	57	47	73	85	43	85
torsion H	34	32	73	64	171	34	171
torsion	85	47	85	85	73	64	85
double bond torsion	73	64	85	64	128	73	128
max step size (fs)	2	4	6	6	14	3	12
average $\Delta\varphi$ (°) per step	0.7	1.3	2.0	2.0	5.1	1.0	6.1

Table 6.2: *Oscillation or relaxation times (in fs) of various degrees of freedom or force field terms in the BPTI simulations. Deviations of force field parameters from GROMOS96 values is indicated by scaling factors. The maximum step size Δt is indicated and the corresponding average change $\Delta\varphi$ of torsion angle value per time step is given.*

a period of about 20fs compared to around 30fs for a single bond involving only heavy atoms. Using the flexible bond model, one can reach time steps of 2fs until the total energy starts to diverge. At this time step, the average torsional-angle change $\Delta\varphi$ is 0.7° per time step. For a physically meaningful simulation with vacuum boundary conditions ($\Delta E_{tot}/\Delta E_{kin} < 5 \cdot 10^{-2}$, figure 6.2), a time step of 1fs can be used. Removal of the highest-frequency motions by applying constraints using the SHAKE algorithm increases the largest possible (until geometric distortions get too large for SHAKE to correct them) time step to 4fs. The energy conservation criterion limits the time step to 2-3fs. The fastest motions under these conditions which are usually applied in standard biomolecular simulations, are the vibrations of bond angles with hydrogen atoms. The oscillation periods of these bond-angle vibrations are around 20fs. By increasing the masses of hydrogen atoms to a value similar to those of the other simulated atoms (in this simulation 14.027u, the mass of a nitrogen atom) one can increase the time step further to 6fs. Since static

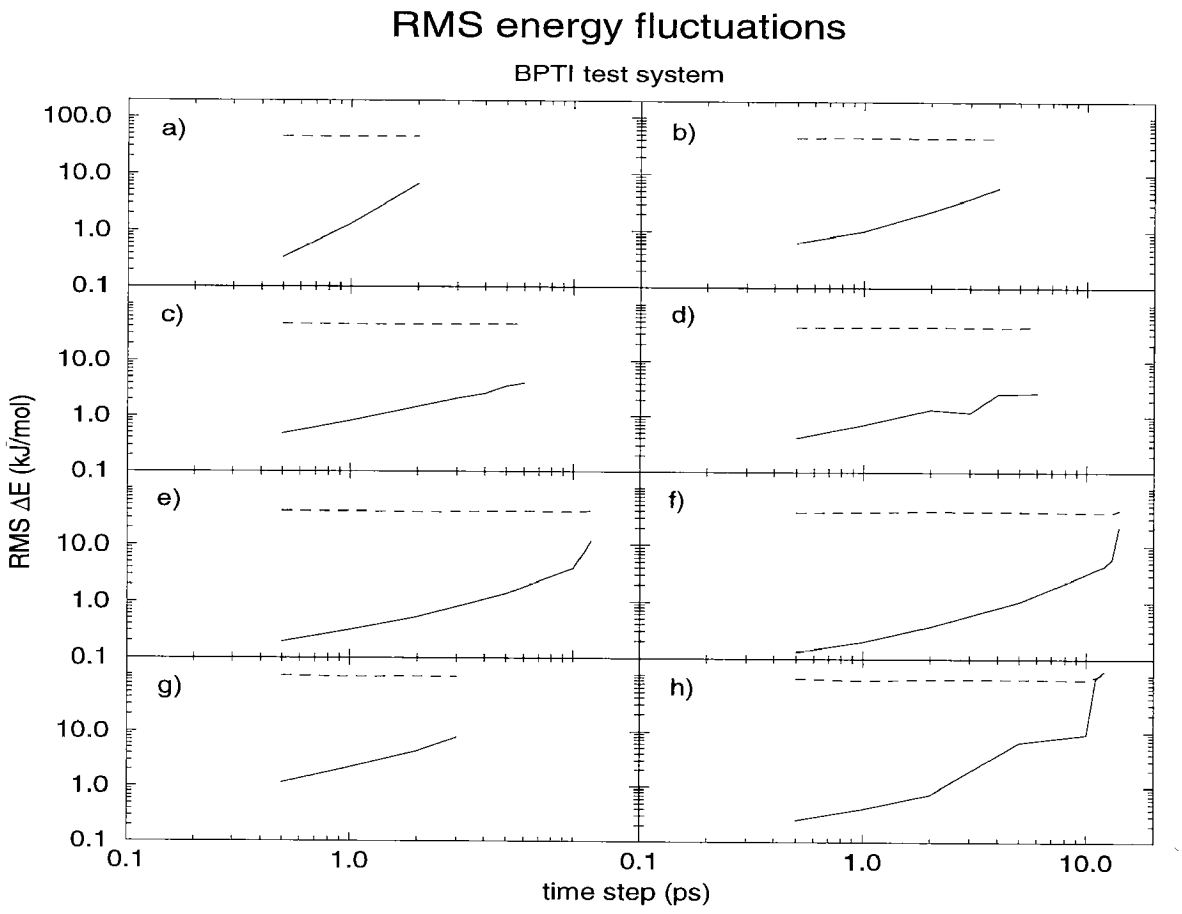

Figure 6.2: *Root-mean-square fluctuations of total (solid line) and kinetic (dashed line) energy for the protein bovine pancreatic trypsin inhibitor (BPTI) for different parameter settings as function of the time step at 300K if not mentioned otherwise; a) flexible bonds, b) bonds constrained, c) bonds constrained, hydrogen masses increased to 14.027u, d) bonds constrained, bond-angle force constants scaled by a factor of 0.5, hydrogen masses increased to 14.027u, e) bonds constrained, bond-angle force constants scaled by a factor of 0.4, improper-dihedral force constants scaled by a factor of 0.4, torsional force constants scaled by a factor of 0.4, hydrogen masses increased to 14.027u, charges turned off, f) bonds constrained, bond-angle force constants scaled by a factor of 0.4, improper-dihedral force constants scaled by a factor of 0.25, torsional force constants scaled by a factor of 0.25, hydrogen masses increased to 14.027u, charges turned off, g) bonds constrained, T=1000K, h) bonds constrained, bond-angle force constants scaled by a factor of 0.4, improper-dihedral force constants scaled by a factor of 0.25, torsional force constants scaled by a factor of 0.25, hydrogen masses increased to 14.027u, charges turned off, T=1000K.*

equilibrium properties do not depend on masses, physical simulations can still be carried out under these conditions as long as dynamical properties are not of interest. The high-frequency bond-angle vibrations are no longer present in the system. The fastest motions are now improper-dihedral-angle motions, although other degrees of freedom show similar frequencies. A time step of 5fs can be applied without the energy conservation being violated to an untolerable extent. Decreasing the bond-angle force constants in addition to the scaling of hydrogen masses brings no further improvement in time step. Under both conditions, using a time step of 6fs, the average torsional-angle change is 2° per step. Since the frequencies corresponding to the various force

field terms are very similar, further improvement could only be achieved by reducing bond-angle-, improper dihedral-, and torsional-angle force constants simultaneously. To do so, charges were set to zero, because otherwise, hydrogen atoms (which do not have a van der Waals interaction) turned out to be attracted too strongly by negatively charged neighboring atoms (e.g. in the side chain of the amino acid Arg), causing an electrostatic collapse.

The largest time step for the simulation of BPTI in vacuum was possible using the following parameter set:

- hydrogen masses (m_H) = 14.027u,
- bond-angle force constants K_{θ_n} scaled by a factor of 0.4,
- improper-dihedral force constants K_{ξ_n} scaled by a factor of 0.4,
- torsional-angle force constants K_{φ_n} scaled by a factor of 0.4,
- all charges set to zero.

The largest possible time step was 14fs with an average $\Delta\varphi$ per step of 5.1°. A further reduction of force constants (e.g. scaling K_{ξ_n} and K_{φ_n} by a factor of 0.25) did not lead to larger time steps. Like in the C_{100} test system, physically meaningful equilibrium simulations can no longer be performed using this parameter set, but it is useful for sampling conformational space. Increasing the temperature to 1000K did not change the results very much.

6.5 Discussion

Molecular dynamics using Cartesian coordinates and Newton's equations of motion is an efficient method to integrate the classical equations of motion. When water molecules solvating a biomolecule are explicitly treated, their fast librational motions limit the integration time step to 2fs, if physically realistic dynamics is to be obtained. In the absence of explicit water, the time step can be lengthened to 6fs when using bond-length constraints and by applying hydrogen mass scaling. The latter will not change significantly the static equilibrium properties, but will modify the dynamics. When using in addition a smoothed, very soft atomic interaction function or force field and neglecting electrostatic interactions at the expense of loosing physical correctness, a time step of 14fs can be reached. The use of another algorithm than SHAKE to constrain bonds would probably enlarge this maximum time step even further. The time step that can be used in Cartesian dynamics is of a size similar to the ones employed in torsional-angle molecular dynamics. The use of Newton's equations of motion instead of Lagrange's equations of motion, however, has a number of advantages:

i) Newton's equations of motion in Cartesian space are much simpler than Lagrange's equations of motion using generalized, torsional, coordinates, so one can use simpler algorithms to integrate the equations of motion. This leads to higher computational efficiency, less computational noise and a reduced chance of software errors. In Cartesian coordinates, the computational cost for one step of simulation is roughly half of the computational cost using generalized coordinates.

ii) Metric tensor effects play a non-negligible role when using generalized, non-Cartesian, coordinates and have to be corrected if a Boltzmann-weighted ensemble is to be generated.

iii) The complete freezing of bond-angle degrees of freedom without adjustment of the physical force field to these conditions reduces atomic motions and thus sampling of conformational space.

iv) The transition from a physically correct force field to a smoothed, very soft force field and back is continuous and more straight-forward than when using dynamics in torsional coordinates

v) Using generalized, torsional coordinates, it is virtually impossible to treat chain closure of a chain of particles using bond constraints while maintaining flexibility within the ring. So, disulfide bridges have to be treated as unconstrained degrees of freedom and the corresponding high-frequency motions are still present in the simulated system.

Considering all these points, for equilibrium simulations, Newton's equations of motion in Cartesian coordinates are to be preferred to Lagrange's equations of motion using generalized, non-Cartesian, coordinates.

6.6 Acknowledgments

Financial support was obtained from the Schweizerischer Nationalfonds, project number 21-57069.99, which is gratefully acknowledged.

Seite Leer /
Blank leaf

Chapter 7

Anisotropic Pressure Coupling Applied to Small, Highly Charged Systems

7.1 Abstract

Use of anisotropic pressure coupling allows for a deformation of the simulated system. The effect of anisotropic pressure coupling on small systems containing highly charged particles is investigated. Finite and infinite periodic boundary conditions were used. 27 charged particles (with equal charges) were simulated in water and argon using various charge sizes to investigate the dampening effect of the different dielectric permittivities. A deformation of the initial box is observed in almost all the simulations. Only the most dilute system at 300K, 27 particles with charges of +1 electron charge in water, and the two simulations at 400K did not deform. The simulations in argon, using finite periodic boundary conditions, converge very slowly. The simulations in water need shorter time to find the optimal box shape. The simulations using infinite periodic boundary conditions converge even quicker and lead to a different box shape than the simulations using finite periodic boundary conditions.

7.2 Introduction

7.2.1 Periodic boundary conditions

In molecular dynamics simulation (MD), due to the small size of the system, boundary effects can not be neglected. There are several ways to treat boundaries, of which the following two are nowadays most commonly used [1]:

(i) Finite periodic boundary conditions, i.e. minimum image periodic boundary conditions using a cut-off radius to truncate long-range interactions. A space-filling shape, like a cubic, rectangular, or truncated octahedral box represents the whole space. An atom leaving the box on one side re-enters it at the opposite side. Non-bonded interactions are only evaluated if the two atoms selected lie within a given cut-off radius R_c. The simulated box is chosen such that a selected atom can interact with at most one periodic copy of any other atom (the minimum image criterion). Thus, the shortest edge of the box has to be at least twice as long as the cut-off radius. To minimise the effect of truncating the electrostatic interactions, the space outside the cut-off radius is often treated as a dielectric continuum

with an adjustable relative dielectric permittivity (the reaction field approach). These finite periodic boundary conditions impose an artificial periodicity onto the simulated system, since normally solutions are to be simulated.

(ii) Infinite periodic boundary conditions, i.e. without any truncation. The effect of truncation of interactions at the cut-off radius is removed. Each atom interacts with each of the other atoms of the system and infinitely many periodic copies of itself and of the other atoms. These boundary conditions induce periodicity artifacts for simulations of proteins in solution and of liquids. These methods are generally computationally more demanding and the implicitly included infinite periodicity introduces artifacts for non-periodic systems. Infinite periodic boundary conditions, however, are well suited to simulate real periodic systems, like (protein) crystals.

7.2.2 Pressure scaling

MD simulation can be carried out at constant pressure using a weak coupling scheme, the so-called Berendsen barostat [6]. The pressure in a MD simulation can be calculated using the virial theorem,

$$PV = \frac{2}{3}[E_{kin} - W], \tag{7.1}$$

where P is the pressure and V the volume of the simulated system. E_{kin} is the kinetic energy of particles defined as

$$E_{kin} = \sum_{i=1}^{N} \frac{1}{2} m_i \vec{v}_i^2 \tag{7.2}$$

and W is the virial of the particles in the periodic box, which is for pair forces

$$W = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j>i}^{N} \vec{r}_{ij} \cdot \vec{f}_{ij} \tag{7.3}$$

with $\vec{r}_{ij} = \vec{r}_i - \vec{r}_j$ and $\vec{f}_{ij}$ the force on particle i due to particle j. According to (7.3), an attractive force between two particles reduces the kinetic pressure of the system. In general, the pressure is a 3-dimensional tensor,

$$P_{xy} = \frac{1}{V} 2 \left[\frac{1}{2} \sum_{i=1}^{N} m_i v_x v_y - \frac{1}{2} \sum_{i=1}^{N} \sum_{j>i}^{N} x_{ij} f_{y,ij} \right] \tag{7.4}$$

or for the diagonal components along the x-, y-, and z-axis:

$$P_{xx} = \frac{1}{V} 2 \left[\frac{1}{2} \sum_{i=1}^{N} m_i v_x^2 - \frac{1}{2} \sum_{i=1}^{N} \sum_{j>i}^{N} x_{ij} f_{x,ij} \right] \tag{7.5}$$

The pressure can also be defined in terms of the molecular (center of mass) kinetic energy and the intermolecular virial. The formulae used are very similar to the ones presented for pressure defined on the atomic level. The major difference is that intramolecular forces do not contribute to the molecular pressure. This is correct, if the molecules simulated are small enough, so that

they cannot see parts of their own periodic copies, which is usually the case in biomolecular simulations in solution. When simulating crystals of proteins, on the other hand, it is well possible, that a protein molecule might see itself through the periodic boundary.

For scaling the pressure in MD, one can use a simple weak coupling scheme [6]. The equations of motion are modified such that the pressure P of the system relaxes towards a preset reference pressure P_0

$$\frac{dP(t)}{dt} = \frac{1}{\tau_P}[P_0 - P(t)] \tag{7.6}$$

The isothermal compressibility κ_T relates a pressure change ΔP to a volume change ΔV,

$$\Delta P(t) = -\frac{\Delta V(t)}{\kappa_T V(t)} \tag{7.7}$$

and the volume change is obtained by scaling the atomic coordinates and the edges of the box by a factor μ,

$$\Delta V(t) = \left[(\mu(t))^3 - 1\right] V(t) \tag{7.8}$$

solving for the pressure scaling factor $\mu(t)$ one gets

$$\mu(t) = \left[1 - \kappa_T \frac{\Delta t}{\tau_P}[P_0 - P(t)]\right]^{\frac{1}{3}} \tag{7.9}$$

for isotropic (all axes are scaled by the same factor) pressure coupling, and

$$\mu_x(t) = \left[1 - \kappa_T \frac{\Delta t}{\tau_P}[P_0 - P_{xx}(t)]\right]^{\frac{1}{3}} \tag{7.10}$$

and corresponding equations for μ_y and μ_z for anisotropic pressure coupling where the the three axes are scaled by the different factors allowing for a deformation of the computational box. The compressibility κ_T is usually not known for the simulated system, so κ_T and the pressure relaxation time τ_P are adjustable parameters in a simulation to obtain the desired coupling strength.

7.2.3 Box deformations due to box multipoles

If a solute with a charge multipole is simulated using anisotropic pressure coupling and infinite periodic boundary conditions, the following effects are to be expected. A charge monopole, such as an ion will prefer a cubic over a rectangular box. A dipole, on the other hand, will change a cubic starting box to a rectangular one, shrinking along the axis the dipole is aligned with and expanding isotropically in the other directions. The total volume of the computational box or cell should not change, since condensed-phase systems, as for example liquid water, the most commonly used solvent in biomolecular simulations, are rather incompressible. The effect of changing box shape will be stronger in non-polar solvents, as for example Argon, compared to a polar solvent such as water, since the latter reduces the electrostatic interaction between the solute and its periodic images.

When using finite periodic boundary conditions, there is no direct interaction between the charged species and its periodic images, so the effect is reduced compared to infinite periodic boundary conditions. Yet, there should still be an effect observable for highly packed or immobile charges or charge distributions, such as, for example, protein crystals. An effect should as well be observed for high charge concentrations, such as many ions or dipoles in solution.

7.3 Methods

In order to analyse the effects of anisotropic pressure coupling on small, highly charged systems, a series of simulations was carried out for the following two systems. 27 ions were simulated either in 1968 simple point charge (SPC) water [29] molecules or in 49 Argon atoms as a solvent. The following parameters were the same for all simulations:

- A time step of 2fs.
- Weak coupling of the solutes and the solvent to two separate temperature baths using a temperature relaxation time τ_T of 0.1ps.
- Anisotropic pressure coupling to a pressure bath of 1atm using a pressure relaxation time τ_P of 0.5ps and an isothermal compresibility κ_T of 4.575 10^{-5}(kJmol^{-1}nm^{-3})$^{-1}$.
- When applying finite periodic boundary conditions, a twin range cut-off radius scheme was used with 0.8nm and 1.4nm for the short- and long-range interactions, respectively. The pairlist and the long-range interactions were updated every 5 steps. A reaction field correction was applied using a self-consistent relative dielectric permittivity of 54 [101].
- Infinite periodic boundary conditions were applied using the particle-particle particle-mesh scheme(P3M) of [5]. A cut-off for real-space interactions of 0.8nm was used. The width of the Gaussian charge shaping function was 0.6nm and 64 grid points in every direction were used.

The differences between the various simulations are summarized in table 7.1. The standard

name	charge per ion (e)	starting geometry (nm) of the box	solvent	temperature (K)	cut-off radii (nm)	boundary method
wa2.0c	2	4x4x4	water	300	0.8/1.4	finite
wa2.0r	2	3x4x5	water	300	0.8/1.4	finite
wa1.5c	1.5	4x4x4	water	300	0.8/1.4	finite
wa1.0c	1	4x4x4	water	300	0.8/1.4	finite
wa2.0c_400	2	4x4x4	water	400	0.8/1.4	finite
wa2.0r_400	2	3x4x5	water	400	0.8/1.4	finite
wa2.0c_p3m	2	4x4x4	water	300	0.8	infinite
ar2.0r	2	3x4x5	argon	300	0.8/1.4	finite
ar1.5r	1.5	3x4x5	argon	300	0.8/1.4	finite
ar1.0r	1	3x4x5	argon	300	0.8/1.4	finite
ar2.0r_small	2	3x4x5	argon	300	0.8/1.2	finite

Table 7.1: *Differences between the various simulations used to analyse the effect of anisotropic pressure coupling. The names of the simulations indicate the solvent used (wa for water; ar for argon), the charge of the solutes in e, the starting shape of the simulated box (c for cubic 4nm edge length; r for rectangular (3x4x5nm)). Particular features of a simulation are indicated at the end of the name, 400 for a higher temperature (400K), p3m for infinite boundary conditions, and small for a smaller long-range cut-off radius.*

simulation ($27Mg^{2+}$ in water using a cubic starting box, a twin range cut-off, and reaction field in conjunction with finite periodic boundary conditions at 300K, wa2.0c, was slightly changed to obtain several slightly different simulations. In a first series, a rectangular starting box with box edges 3nm, 4nm, and 5nm (wa2.0r) was used instead of a cubic box. The ions were either placed randomly in the box or on a regular grid, which turned out not to make a difference for the results (not shown). Therefore, here only simulations with randomly placed ions are reported. Secondly, the charges were changed to $1.5e$ (wa1.5c) and 1e (wa1.0c), leaving all other simulation conditions the same. The decrease in charge leads to a dilution in charge density in the system, so a smaller effect is expected compared to the standard simulation. A dilution of charge density can also be expected using a higher temperature. The volume would be slightly higher, and the particles more mobile thus minimizing the effect of static order, which could influence the box shape. So, the third series involves two simulations at 400K, starting from a cubic (wa2.0c_400) or from a rectangular box (wa2.0r_400). The influence of electrostatic interactions is expected to be much larger using infinite periodic boundary conditions, since here, particles do not only interact with at most one copy of another atom but with all periodic images of all atoms. To observe this effect, a simulation using the P3M method (wa2.0c_p3m) was performed.

To minimize the screening effect of the polar solvent water, some of the simulations were also carried out using argon as a solvent. $27Mg^{2+}$ were simulated in 49 argon atoms (ar2.0r), which leads to a comparable box size as for the simulations in water. A rectangular box with edges 3nm, 4nm, and 5nm was used as a starting point. Again, charges were decreased to see the effect of charge dilution, charges of $1.5e$ (ar1.5r) and $1.0e$ (ar1.0r) were used. The last system simulated used standard charges ($2.0e$), argon as a solvent and a shorter cut-off radius of 1.2nm for evaluating electrostatic interactions (ar2.0_small).

7.4 Results

In Table 7.2, the results for the different simulations are summarized. The total simulation time and the time needed for the box to reach its final shape are shown. The average box edge lengths and the average volume are calculated using the trajectory from the time the box shape has converged until the end of the simulation.

The standard simulation, wa2.0c takes one nanosecond to change its shape from initially cubic to a quadratic prism, where it stays for the remaining 3ns of the simulation. The simulation starting from a rectangular box converges slightly faster and finds the same optimal box shape with the shorter two axes being equal in length, 3.4nm and 3.5nm, and the longer axes being 5.4nm and 5.2nm in the two simulations, respectively. The wa1.5c simulation, having a lower charge of $1.5e$ on the solute atoms compared to $2e$ for the standard simulations, needs a much longer time to converge, 4ns, but ends up with a comparable box shape, where the three axes are 3.4nm, 3.6nm and 5.0nm. The volume of the simulated box is slightly lower than for the doubly charged ions which is not unexpected since the electrostatic repulsion between like charges tends to increase the pressure through the virial. The same trend in volume can be observed, when the charges are further decreased to $1e$ (wa1.0c). The volume is again slightly lower. The computational box, on the other hand, does not deform in this system. The density of charges is low enough that the effect disappears. Increasing the temperature also leads to a decrease of charge density. The volume of the simulated box becomes considerably larger, by more than a factor 2, leading to a charge density in the same order of magnitude for the two 400K simulations (wa2.0c_400, wa2.0r_400) and the wa1.0c simulation. Although in the simulation starting

simulation name	simulation time (ns)	convergence time (ns)	edge length (nm) x	y	z	volume (nm^3)
wa2.0c	4	1	3.5	3.5	5.2	63.8
wa2.0r	2	0.75	3.4	3.4	5.4	64.0
wa1.5c	8	4	3.6	5.0	3.4	61.3
wa1.0c	2	0	4.0	3.9	3.9	60.5
wa2.0c_400	5	1	5.0	5.3	5.6	148.0
wa2.0r_400	5	1	4.2	5.4	6.5	148.5
wa2.0c_p3m	1	0.2	4.2	3.4	4.2	60.0
ar2.0r	200	30	2.8	5.1	5.1	74.4
ar1.5r	200	30	2.8	4.1	6.8	79.9
ar1.0r	40*	35	2.8	5.2	5.2	75.6
ar2.0r_small	200	25	2.5	3.6	6.0	55.3

Table 7.2: *The convergence time in ns of the box shape is shown together with the value (in nm) to which the different axes converge. The trajectory reaching from the time of convergence to the end of the simulation was used to calculate the average edge lengths and volume of the box. The simulation names are explained in table 7.1. In the simulation ar1.0r, the box gets smaller than twice the cut-off radius, which causes the simulation to fail after 40ns.*

from a cubic box, the axes diverge a little, this effect is not significant, and there is also no clear trend (Fig. 7.1). The simulation starting from a rectangular box expands isotropically to adopt to the higher temperature and thus lower density. The ratio of the different axes to one another, the box shape, does not change. Using infinite periodic boundary conditions (wa2.0c_p3m), the box deforms an order of magnitude faster than using a finite periodic boundary cut-off scheme. This is expected, as much more (periodic) interactions are calculated. The final shape of the box is already reached after 200ps, and differs from the shapes found with finite periodic boundary conditions. The shorter axis is again 3.4nm, but in contrast to the other simulations, the base of the quadratic prism found as optimal shape is built by the longer axes.

The simulations in argon all need a much longer time to converge. After having found an ideal box shape, the axes stay much more stable than when water is the solvent. The root-mean-square fluctuations of the axes in the wa2.0c simulation are 0.09nm, 0.19nm, and 0.25nm for the x-, y-, and z-axis respectively, in the ar2.0r simulation they are 0.01nm, 0.03nm, and 0.03nm. The shortest box edge found in the three standard argon simulations, ar2.0r, ar1.5r, and ar1.0r is always around 2.8nm, which is twice the cut-off radius for the long-range interactions. In the case of ar1.0r, marked with an asterisk (*) in table 7.2, the shortest axis get even shorter than the cut-off radius, causing the simulation to fail after 40ns. For the longer axes, no trend can be observed. In the two simulations with integral charge, again a quadratic prism is found, whereas the simulation ar1.5r, where the solutes are charged $1.5e$, is the only one that shows the box deformation effect but does not find a quadratic prism as final box shape. Shortening the long-range cut-off radius in the argon simulation, ar2.0r_small, the shortest axis is again slightly bigger than twice the cut-off radius, and the shape is as in the previous example a non-regular rectangular box. The volume decreases considerably compared to the other argon simulations. This is to be expected since the loss in electrostatic repulsion can not be compensated by the loss in attraction arising from the van der Waals interaction, which is much shorter ranged.

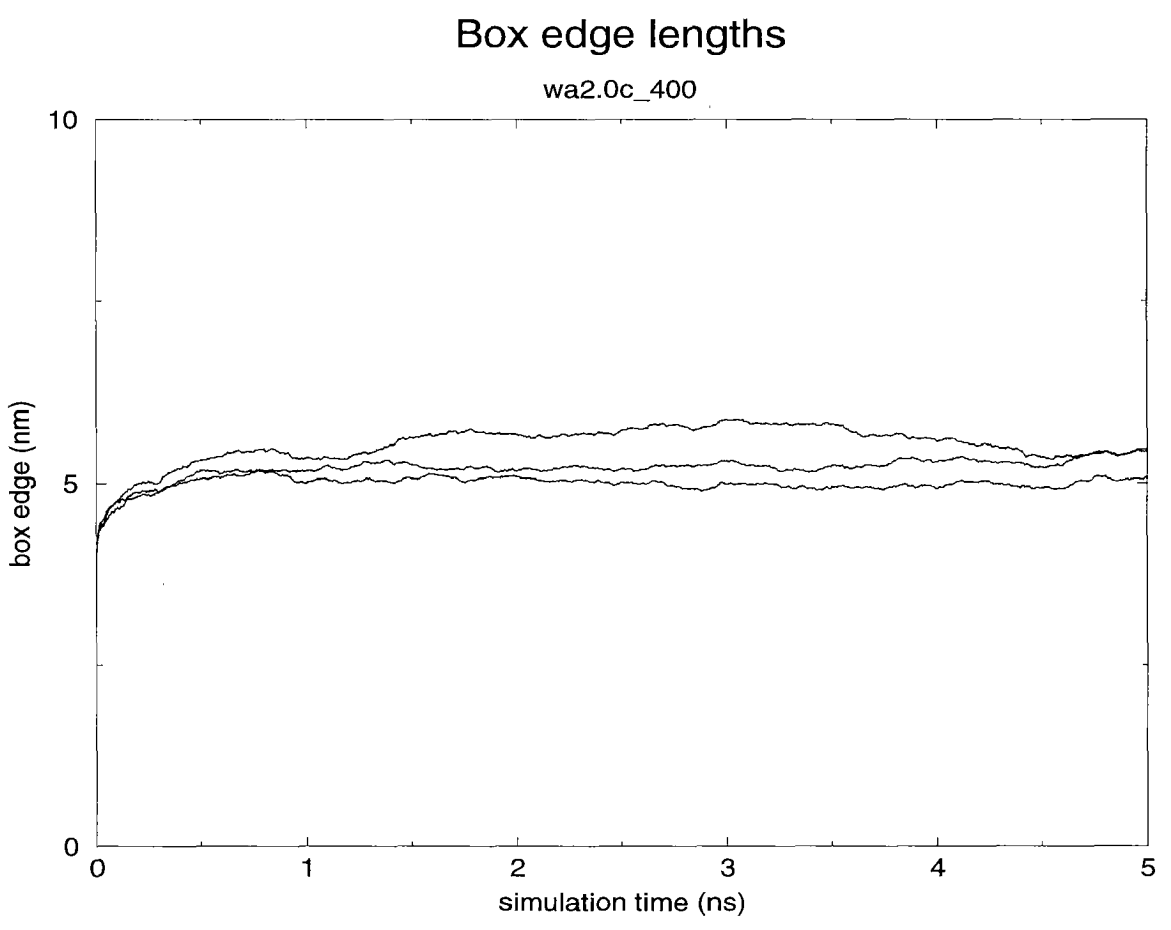

Figure 7.1: *Edge lengths of the computational box of the wa2.0c_400 simulation. 27 Mg^{2+} ions in water, starting from a cubic box at T=400K.*

7.5 Discussion

The presented results show a strong influence of the density of simulated charges on the box shape using anisotropic pressure coupling. If the charge density is low, the edges of the simulated box deform rather isotropically, as in the wa1.0c simulation, and in the high-temperature simulations. If the critical charge density is exceeded, the box deforms towards an ideal shape. An explanation could be that the simulated ions, together with a dynamic water shell behave as large, isotropically repulsive spheres in a smaller number of solvent molecules. These spheres or balls then try to minimize interactions thus ending in a kind of closest (or, in this case, the furthest away from each other as possible) packing. The radial distribution functions magnesium-water oxygen and magnesium-magnesium all show clear structure. With the charge being decreased, these balls would become smaller and softer, so the effect is dampened or even disappears. For the high-temperature runs, molecules move more, the balls become softer, and the volume accessible is much higher, so the effect also disappears. Using infinite periodic boundary conditions, where one takes many more interactions into account, the effect gets stronger and the box shape converges faster. In argon, the convergence is much slower. On the one hand, the electrostatic interactions are less screened by the completely apolar solvent, and the ions can feel each other much more strongly. On the other hand, the argon atoms do not cluster around the ions, so the simulated spheres are effectively smaller. A possible test system to evaluate the assumption that one effectively simulates repulsive spheres would be a system consisting of 27 Lennard-Jones

particles. There, like in the real system, one has isotropic spheres, trying to position in a closest packing. The problem with such a test system is the very small size, the cut-off radius has to be chosen extremely small, and the temperature has to be adjusted to a value which corresponds to the ionic system. The very small cut-off radius has the disadvantage that cut-off artifacts become non-negligible compared to the system energy and thus could mask the desired effect. Finding the right pseudo-temperature to maintain a liquid system is also not a trivial task. The test system would be very sensitive to small changes leading to either a frozen system or to vaporization, which cases both have to be avoided. A simulation of 27 argon-like particles using a cut-off of 0.4nm at a pseudo temperature of 4K showed a similar box-deformation effect. A rectangular box is found after several microseconds of simulation. After this box shape has been reached, the noise in energy is extremely high with the fluctuation of potential energy being about ten times the value of the potential energy.

Bibliography

[1] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Clarendon, Oxford (1987).

[2] W. F. van Gunsteren, S. R. 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*. Vdf Hochschulverlag, Zürich, Switzerland (1996).

[3] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. *J. Comput. Phys.*, 23:327 (1977).

[4] I. G. Tironi, R. M. Brunne, and W. F. van Gunsteren. *Chem. Phys. Letters*, 250:19 (1996).

[5] R. W. Hockney and J. W. Eastwood. *Computer Simulation using Particles*. McGraw-Hill, New York (1981).

[6] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. *J. Chem. Phys.*, 81:3684 (1984).

[7] C. A. Schiffer, R. Huber, K. Wüthrich, and W. F. van Gunsteren. *J. Mol. Biol.*, 241:588 (1994).

[8] W. F. van Gunsteren, R. Kaptein, and E. R. P. Zuiderweg. In: W. K. Olson, ed., *Proceedings NATO/CECAM workshop on nucleic acid conformation and dynamics*, pp. 79–97. CECAM, France, Orsay (1984).

[9] R. Kaptein, E. R. P. Zuiderweg, R. M. Scheek, R. Boelens, and W. F. van Gunsteren. *J. Mol. Biol.*, 182:179 (1985).

[10] A. E. Torda, R. M. Brunne, T. Huber, H. Kessler, and W. F. van Gunsteren. *J. Biomol. NMR*, 3:55 (1993).

[11] T. S. Harvey and W. F. van Gunsteren. In: *Techniques in protein Chemistry IV*, pp. 615–622. Academic press, New, York (1993).

[12] A. T. Brünger, J. Kuriyan, and M. Karplus. *Science*, 235:458 (1987).

[13] M. Fujinaga, P. Gros, and W. F. van Gunsteren. *J. Appl. Cryst.*, 22:1 (1989).

[14] A. E. Torda, R. M. Scheek, and W. F. van Gunsteren. *J. Mol. Biol.*, 214:223 (1990).

[15] P. Gros, W. F. van Gunsteren, and W. G. J. Hol. *Science*, 249:1149 (1990).

[16] P. Gros and W. F. van Gunsteren. *Molecular Simulation*, 10:377 (1993).

[17] C. A. Schiffer, P. Gros, and W. F. van Gunsteren. *Acta Cryst.*, D51:85 (1995).

[18] R. M. Scheek, A. E. Torda, J. Kemmink, and W. F. van Gunsteren. In: J. C. Hoch, F. M. Poulsen, and C. Redfield, eds., *Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy*, vol. A225 of *NATO ASI Series*, pp. 209–217. Plenum Press, New York (1991).

[19] J. Fennen, A. E. Torda, and W. F. van Gunsteren. *J. Biomol. NMR*, 6:163 (1995).

[20] W. F. van Gunsteren, R. M. Brunne, P. Gros, R. C. van Schaik, C. A. Schiffer, and A. E. Torda. In: T. L. James and N. J. Oppenheimer, eds., *Methods in Enzymology: Nuclear Magnetic Resonance*, vol. 239, pp. 619–654. Academic press, New York (1994).

[21] W. F. van Gunsteren, A. M. J. J. Bonvin, X. Daura, and L. J. Smith. Aspects of modelling biomolecular structures on the basis of spectroscopic or diffraction data (1997). To appear in "Modern techniques in Protein NMR", Krishna and Berliner eds, Plenum Publ. Corp.

[22] W. F. van Gunsteren and M. Karplus. *Nature*, 293:677 (1981).

[23] W. F. van Gunsteren and M. Karplus. *Biochemistry*, 21:2259 (1982).

[24] W. F. van Gunsteren, H. J. C. Berendsen, J. Hermans, W. G. J. Hol, and J. P. M. Postma. *Proc. Natl. Acad. Sci. USA*, 80:4315 (1983).

[25] H. J. C. Berendsen, W. F. van Gunsteren, H. R. J. Zwinderman, and R. Geurtsen. *Ann. New York Acad. Sci*, 482:269 (1986).

[26] S. Yun-yu, Y. Ru-huai, and W. F. van Gunsteren. *J. Mol. Biol.*, 200:571 (1988).

[27] A. P. Heiner, H. J. C. Berendsen, and W. F. van Gunsteren. *Proteins*, 14:451 (1992).

[28] W. F. van Gunsteren, S. R. 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*. Vdf Hochschulverlag, Zürich, Switzerland (1996).

[29] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans. In: B. Pullman, ed., *Intermolecular Forces*, pp. 331–342. Reidel, Dordrecht (1981).

[30] S. Vijay-Kumar, C. E. Bugg, and W. J. Cook. *J. Mol. Biol.*, 194:531 (1987).

[31] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. *J. Mol. Biol.*, 112:535 (1977).

[32] W. F. van Gunsteren and H. J. C. Berendsen. *Angew. Chem. Int. Ed. Engl.*, 29:992 (1990).

[33] M. Levitt, M. Hirshberg, R. Sharon, and V. Daggett. *Comp. Phys. Commun.*, 91:215 (1995).

[34] P. E. Smith and W. F. van Gunsteren. *Molecular Simulation*, 15:233 (1995).

[35] W. F. van Gunsteren, X. Daura, and A. E. Mark. *Encyclopaedia of Computational Chemistry*, 2:1211 (1998).

[36] L. J. Smith, A. E. Mark, C. M. Dobson, and W. F. van Gunsteren. *Biochemistry*, 34:10918 (1995).

[37] W. F. van Gunsteren and H. J. C. Berendsen. *Groningen Molecular Simulation (GROMOS) library Manual.* Biomos, Nijenborgh 4, 9747 AG Groningen, The Netherlands, (1987).

[38] P. E. Smith, R. C. van Schaik, T. Szyperski, K. Wüthrich, and W. F. van Gunsteren. *J. Mol. Biol.*, 246:356 (1995).

[39] X. Daura, A. E. Mark, and W. F. van Gunsteren. *J. Comput. Chem.*, 19:535 (1998).

[40] D. Carter, J. He, J. R. Ruble, and B. Wright. The structure of the orthorhombic form of hen egg-white lysozyme at 1.5 Ångstrøms resolution (1997). Protein Data Bank, entry 1AKI.

[41] P. J. Artymiuk, C. C. F. Blake, D. W. Rice, and K. S. Wilson. *Acta Cryst, Sect. B*, 38:778 (1982).

[42] I. G. Tironi, R. Sperb, P. E. Smith, and W. F. van Gunsteren. *J. Chem. Phys.*, 102:5451 (1995).

[43] P. E. Smith and W. F. van Gunsteren. *J. Chem. Phys*, 100:3169 (1994).

[44] W. R. P. Scott, P. H. Hünenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Krüger, and W. F. van Gunsteren. *J. Phys. Chem. A*, 103:3596 (1999).

[45] L. J. Smith, M. J. Sutcliffe, C. Redfield, and C. M. Dobson. *Biochemistry*, 30:986 (1991).

[46] L. J. Smith, M. J. Sutcliffe, C. Redfield, and C. M. Dobson. *J. Mol. Biol.*, 229:930 (1993).

[47] M. Buck, J. Boyd, C. Redfield, D. A. MacKenzie, D. J. Jeenes, D. B. Archer, and C. M. Dobson. *Biochemistry*, 34:4041 (1995).

[48] J. C. Hoch, C. M. Dobson, and M. Karplus. *Biochemistry*, 24:3831 (1985).

[49] M. Billeter, A. D. Kline, W. Braun, R. Huber, and K. Wüthrich. *J. Mol. Biol.*, 206:677 (1989).

[50] E. T. Baldwin, I. T. Weber, R. S. Charles, J.-C. Xuan, E. Appella, M. Yamada, K. Matsushima, B. F. P. Edwards, G. M. Clore, A. M. Gronenborn, and A. Wlodawer. *Proc. Natl. Acad. Sci. USA*, 88:502 (1991).

[51] J. M. Moore, C. A. Lepre, G. P. Gibbert, W. J. Chazin, D. A. Case, and P. E. Wright. *J. Mol. Biol.*, 221:533 (1991).

[52] D. Neri, M. Billeter, and K. Wüthrich. *J. Mol. Biol.*, 223:743 (1992).

[53] J. Kallen, C. Spitzfaden, M. G. M. Zurini, G. Wider, H. Widmer, K. Wüthrich, and M. D. Walkinshaw. *Nature*, 353:276 (1991).

[54] M. Billeter, J. Vendrell, G. Wider, F. X. Avilès, M. Coll, A. Gausch, R. Huber, and K. Wütrich. *J. Biomol. NMR*, 2:1 (1992).

[55] K. D. Berndt, P. Güntert, L. P. M. Orbons, and K. Wüthrich. *J. Mol. Biol.*, 227:757 (1992).

[56] W. Braun, M. Vašák, A. H. Robbins, C. D. Stout, G. Wagner, J. H. R. Kägi, and K. Wüthrich. *Proc. Natl. Acad. Sci. USA*, 89:10124 (1992).

[57] A. Fede, M. Billeter, W. Leupin, and K. Wüthrich. *Structure*, 1:177 (1993).

[58] U. Dornberger, J. Flemming, and H. Fritzsche. *J. Mol. Biol.*, 284:1453 (1998).

[59] J. Lu, C.-L. Lin, C. Tang, J. W. Ponder, J. L. F. Kao, D. P. Cistola, and E. Li. *J. Mol. Biol.*, 286:1179 (1999).

[60] F. Yang, C. A. Bewley, J. M. Louis, K. R. Gustafson, M. R. Boyd, A. M. Gronenborn, G. M. Clore, and A. Wlodawer. *J. Mol. Biol.*, 288:403 (1999).

[61] W. F. van Gunsteren and H. J. C. Berendsen. *J. Mol. Biol.*, 176:559 (1984).

[62] J. Evenäs, S. Forsén, A. Malmendal, and M. Akke. *J. Mol. Biol.*, 289:603 (1999).

[63] P. H. Hünenberger, A. E. Mark, and W. F. van Gunsteren. *J. Mol. Biol.*, 252:492 (1995).

[64] W. F. van Gunsteren, A. M. J. J. Bonvin, X. Daura, and L. J. Smith. In: *"Structure Computation and Dynamics in Protein NMR", Biol. Magnetic Resonance*, vol. Vol. 17, pp. 3–35. Plenum Publishers, New York (1999).

[65] B. A. Luty, I. G. Tironi, and W. F. van Gunsteren. *J. Chem. Phys*, 8:3014 (1995).

[66] T. Darden, D. York, and L. Pedersen. *J. Chem. Phys*, 98:10089 (1993).

[67] G. Goldstein, M. Scheid, U. Hammerling, E. A. Boyse, D. H. Schlesinger, and H. D. Niall. *Proc. Natl. Acad. Sci. USA*, 72:11 (1975).

[68] D. H. Schlesinger, G. Goldstein, and H. D. Niall. *Biochemistry*, 14:2214 (1975).

[69] A. Hershko, H. Heller, E. Eytan, G. Kaklij, and I. A. Rose. *Proc. Natl. Acad. Sci. USA*, 81:7021 (1984).

[70] D. Finley and A. Varshavsky. *Trends Biochem. Sci.*, 10:343 (1985).

[71] A. Hershko and A. Ciechanover. *Prog. Nucl. Acid. Res. Mol. Biol.*, 33:19 (1986).

[72] A. Hershko. *J. Mol. Biol.*, 263:15237 (1988).

[73] B. P. Monia, J. Ecker, and S. T. Crooke. *Biotechnology*, 8:209 (1990).

[74] R. E. Lenkinski, D. M. Chen, J. D. Glickson, and G. Goldstein. *Biochim. Biophys. Acta*, 494:126 (1977).

[75] D. M. Schneider, M. J. Dellwo, and A. J. Wand. *Biochemistry*, 31:3645 (1992).

[76] J. A. Braatz, M. D. Paulsen, and R. L. Ornstein. *J. Biomol. Struct. Dyn.*, 9:935 (1992).

[77] D. O. W. Alonso and V. Daggett. *J. Mol. Biol.*, 247:501 (1995).

[78] R. Abseher, S. Lüdemann, H. Schreiber, and O. Steinhauser. *J. Mol. Biol.*, 249:604 (1995).

[79] H. Schreiber and O. Steinhauser. *Biochemistry*, 31:5856 (1992).

[80] H. Schreiber and O. Steinhauser. *Chem. Phys.*, 168:75 (1992).

[81] S. F. Lienin, T. Bremi, B. Brutscher, R. Brüschweiler, and R. R. Ernst. *J. Am. Chem. Soc.*, 120:9870 (1998).

[82] N. Juranic, V. A. Guileic, F. G. Prendergast, and S. Macura. *J. Amer. Chem. Soc.*, 118:7859 (1996).

[83] H. M. Berman, J. Wesstbrook, Z. Feng, G. Gilliand, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. *Nucl. Acid Res.*, 28:235 (2000).

[84] W. J. Cook, F. L. Suddath, C. E. Bugg, and G. Goldstein. *J. Mol. Biol.*, 130:353 (1979).

[85] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton. *J. Appl. Cryst.*, 26:283 (1993).

[86] J. A. C. Rullmann. Aqua, computer program (1996). Department of NMR Spectroscopy, Bijvoet Center for Biomolecular Research, Utrecht University, The Netherlands.

[87] S. W. Rick and B. J. Berne. *J. Am. Chem. Soc.*, 116:3949 (1994).

[88] G. Hummer, L. R. Pratt, A. E. Garcia, B. J. Berne, and S. W. Rick. *J. Phys. Chem. B*, 101:3017 (1997).

[89] J. Åqvist and T. Hansson. *J. Phys. Chem. B*, 102:3837 (1998).

[90] G. T. Ibragimova and R. Wade. *Biophys. J.*, 74:2906 (1998).

[91] A. Rahman. *Phys. Rev.*, 136A:405 (1964).

[92] J. A. McCammon, B. R. Gelin, and M. Karplus. *Nature*, 267:585 (1977).

[93] D. P. Tieleman, S. J. Marrink, and H. J. C. Berendsen. *Biochim. Biophys. Acta*, 1331:235 (1997).

[94] A. T. Brünger and M. Nilges. *Quart. Rev. Biophys.*, 26:49 (1993).

[95] P. Güntert, C. Mumenthaler, and K. Wüthrich. *J. Mol. Biol.*, 273:283 (1997).

[96] E. G. Stein, L. M. Rice, and A. T. Brünger. *J. Magn. Res.*, 124:154 (1997).

[97] L. D. Schuler and W. F. van Gunsteren. *Mol. Simulation*, in press (2000).

[98] A. Rahman and F. H. Stillinger. *J. Chem. Phys.*, 55:3336 (1971).

[99] H. J. C. Berendsen and W. F. van Gunsteren. In: A. J. Barnes and *al*, eds., *Molecular Liquids - Dynamics and Interactions*, vol. C135 of *NATO ASI Series*, pp. 561–564. Reidel, Dordrecht (1984).

[100] W. F. van Gunsteren, R. Boelens, R. Kaptein, R. M. Scheek, and E. R. P. Zuiderweg. In: *Molecular Dynamics and Protein Structure*, pp. 92–99. Polycrystal Book Service, P.O. Box 27, Western Springs, Ill. 60558, USA (1985).

[101] W. F. van Gunsteren, S. R. 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*. Vdf Hochschulverlag, Zürich, Switzerland (1996).

[102] W. F. van Gunsteren and M. Karplus. *Macromolecules*, 15:1528 (1982).

[103] L. M. Rice and A. T. Brünger. *Proteins*, 19:277 (1994).

[104] M. Levitt, M. Hirshberg, R. Sharon, K. E. Laidig, and V. Daggett. *J. Phys. Chem. B*, 101:5051 (1997).

[105] M. Tuckerman, B. J. Berne, and G. J. Martyna. *J. Chem. Phys.*, 97:1990 (1992).

[106] M. Watanabe and M. Karplus. *J. Chem. Phys.*, 99:8063 (1993).

[107] J. A. Izaguirre, S. Reich, and R. D. Skeel. *J. Chem. Phys.*, 110:9853 (1999).

[108] C. H. Bennett. *J. Comput. Phys.*, 19:267 (1975).

[109] J.-P. Ryckaert. In: M. Meyer and V. Pontikis, eds., *Computer Simulations in Material Science*, vol. E205 of *NATO ASI Series*, pp. 43–66. Kluwer, Dordrecht (1991).

[110] R. Pomès and J. A. McCammon. *Chem. Phys. Lett.*, 166:425 (1990).

[111] K. Feenstra, B. Hess, and H. J. C. Berendsen. *J. Comput. Chem.*, 20:786 (1999).

[112] H. Katz, R. Walter, and R. L. Somorjay. *Comput. Chem.*, 3:25 (1979).

[113] A. K. Mazur, V. E. Dorofeev, and R. A. Abagyan. *J. Comput. Phys.*, 92:261 (1991).

[114] G. Ciccotti, M. Ferrario, and J.-P. Ryckaert. *Mol. Phys.*, 47:1253 (1982).

[115] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. *J Comput. Chem.*, 18:1463 (1997).

[116] S. Reich. *Physica D*, 89:28 (1995).

[117] W. F. van Gunsteren, T. Huber, and A. E. Torda. In: *Conference Proceedings*, vol. 330, pp. 253–268. American Institute of Physics (1995). European Conference on Computational Chemistry (E.C.C.C 1).

[118] D. Frenkel. In: *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*, vol. 2, pp. 37–66. Escom, Leiden, The Netherlands (1993).

[119] W. F. van Gunsteren. In: *Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications*, vol. 2, pp. 3–36. Escom, Leiden, The Netherlands (1993).

[120] A. K. Mazur. *J. Phys. Chem.*, B102:473 (1998).

[121] W. F. van Gunsteren. In: *Advances in Biomolecular Simulations*, vol. 239, pp. 131–146. American Inst. of Physics (1991).

[122] T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber, and W. F. van Gunsteren. *Chem. Phys. Letters*, 222:529 (1994).

[123] W. F. van Gunsteren, F. J. Luque, D. Timms, and A. E. Torda. *Ann. Rev. Biophys. Biomol. Structure*, 23:847 (1994).

[124] J. Wittenburg. *Dynamics of Systems of Rigid Bodies*. Teubner, Stuttgart (1977).

[125] D.-S. Bae and E. J. Haug. *Mech. Struct. Mach.*, 15:359 (1987).

[126] D.-S. Bae and E. J. Haug. *Mech. Struct. Mach.*, 15:481 (1988).

[127] A. Jain, N. Vaidehi, and G. Rodriguez. *J. Comput. Phys.*, 106:258 (1993).

[128] J. D. Turner, P. K. Weiner, H. M. Chun, V. Lupi, S. Gallion, and U. C. Singh. In: *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*, vol. 2, pp. 535–555. Escom, Leiden, The Netherlands (1993).

[129] A. M. Mathiowetz, A. Jain, N. Karasawa, and W. A. Goddard III. *Proteins*, 20:227 (1994).

[130] H. Goldstein. *Classical mechanics*. Addison-Wesley, Reading MA, USA, 2nd ed. (1980).

[131] W. F. van Gunsteren and H. J. C. Berendsen. *Mol. Phys.*, 34:1311 (1977).

[132] H. J. C. Berendsen and W. F. van Gunsteren. In: G. Ciccotti and W. G. Hoover, eds., *Molecular-Dynamics Simulation of Statistical-Mechanical Systems*, pp. 43–65. Proceedings of the International School of Physics "Enrico Fermi", course 97, North-Holland, Amsterdam (1986).

[133] A. K. Mazur. *J. Comput. Phys.*, 136:354 (1997).

[134] L. Verlet. *Phys. Rev.*, 159:98 (1967).

[135] W. F. van Gunsteren. *Mol. Phys.*, 40:1015 (1980).

[136] B. A. Luty, M. E. Davis, I. G. Tironi, and W. F. van Gunsteren. *Mol. Simulation*, 14:11 (1994).

Curriculum Vitae

Personal Information

Name	Urs Stocker
Date of birth	June 11, 1971
Place of birth	Winterthur, Switzerland
Citizenship	Neftenbach and Schönenberg, Zürich, Switzerland

Education

1976–1977	Primary School in Neuhausen a. d. F, Germany
1977–1982	Primary School in Estoril and Lisboa, Portugal
1982–1984	Primary School in Embrach
1984–1986	Secondary School in Embrach
1986–1990	Mathematisch-Naturwissenschaftliches Gymnasium in Bülach 1990 Matura Typus C
1991–1995	Study of Chemistry at ETH Zürich 1995 Dipl. Chem. ETH
1996–2000	Ph. D. research at the Laboratory of Physical Chemistry at ETH Zürich, Prof. Wilfred F. van Gunsteren