

Toward an Efficient Simulation of Biomineralization: A Computational Study of the Apatite/Collagen System

Vom Fachbereich Chemie
der Technischen Universität Darmstadt

zur Erlangung des akademischen Grades eines
Doktor rerum naturalium (Dr. rer. nat.)

genehmigte
Dissertation

vorgelegt von

Dipl.-Chem. Thorsten Schepers
aus Lengerich

Berichterstatter:	Prof. Dr. Jürgen Brickmann
Mitberichterstatter:	Prof. Dr. Florian Müller-Plathe
Tag der Einreichung:	27.3.2006
Tag der mündlichen Prüfung:	8.5.2006

Darmstadt 2006

D17

Diese Arbeit wurde am Institut für Physikalische Chemie der Technischen Universität Darmstadt in der Zeit von Juli 2002 bis März 2006 unter der Leitung von Prof. Dr. J. Brickmann angefertigt.

Ich danke Prof. Brickmann für die finanzielle und wissenschaftliche Unterstützung und für die weitreichende Freiheit bei der Bearbeitung des Projekts. Allen Mitgliedern der Gruppen Brickmann und Kast danke ich für hilfreiche Diskussionen, Austausch von Wissen, gemeinsame Computeradministration und die angenehme Atmosphäre im Büro.

1	Einleitung und Problemstellung	1
2	Introduction	3
2.1	Composites	3
2.2	The apatite/collagen composite system	4
2.3	Objectives of this work	5
3	Theory	7
3.1	Molecular dynamics simulation	7
3.1.1	Equations of motion	7
3.1.2	Force field	9
3.1.3	Statistical interpretation	10
3.2	Potential of mean force	11
3.2.1	Aim of the method	11
3.2.2	Reversible work theorem	11
3.2.3	Simulation technique	12
3.2.4	Error estimation	13
3.3	Analysis techniques	16
3.3.1	Mobility analyses	16
3.3.1.1	<i>RMS values</i>	16
3.3.2	Structural analyses	19
3.3.2.1	<i>Hydrogen bonds</i>	19
3.3.3	Radial distribution function	20
3.3.4	Analysis of kinetics	21
3.4	Collagen	22
4	Computational details and results	26
4.1	Sequence analysis	26
4.2	Structural analyses	30
4.2.1	Collagen telopeptide	30
4.2.1.1	<i>Model preparation</i>	30
4.2.1.2	<i>Long time simulation</i>	32
4.2.1.3	<i>Protein flexibility</i>	34
4.2.1.4	<i>Hydrogen bonds</i>	39
4.2.2	Collagen triple helix	42

4.2.2.1	<i>Model preparation</i>	42
4.2.2.2	<i>Long time simulations</i>	44
4.2.2.3	<i>Protein flexibility</i>	44
4.2.2.4	<i>Ramachandran analysis</i>	48
4.2.2.5	<i>Hydrogen bonds</i>	50
4.2.3	Discussion	51
4.3	Charged groups distances	52
4.3.1	Telopeptide	52
4.3.2	Triple helix	54
4.3.3	Discussion	55
4.4	Ion attachment simulations	55
4.4.1	Analysis	59
4.5	Microscopic computational experiments with complexes	65
4.6	PMF calculations	70
4.6.1	Adsorption site model structures	71
4.6.1.1	<i>Model preparation</i>	71
4.6.1.2	<i>Simulations</i>	73
4.6.1.3	<i>Analyses</i>	74
4.6.1.4	<i>Discussion</i>	82
4.6.2	Telopeptide	93
5	Summary, conclusions and outlook	96
6	Zusammenfassung und Ausblick	102
7	Literatur	109
8	Anhang	112

1 Einleitung und Problemstellung

Die Natur bringt durch die Kristallisation von anorganischem Material in einer organischen Matrix eine enorme Vielfalt von interessanten Strukturen hervor, deren atomare Ordnungsprinzipien noch größtenteils unbekannt sind. Diese sogenannten Kompositmaterialien eröffnen ein weites Feld an Eigenschaften, die ein großes Potential für industrielle und technische Anwendungen in vielen Bereichen bieten. Ein für den medizinischen Bereich interessantes Komposit stellt das System Apatit/Kollagen dar, aus dem Knochen und Zähne aufgebaut sind. Für die Behandlung von Knochenbrüchen, Karies etc. wäre eine Kontrolle über die Bildung dieser Materialien oder eine Beeinflussung der Eigenschaften äußerst attraktiv. Dazu ist das Verständnis der Strukturen und der Bildung dieser Kompositmaterialien notwendig.

Busch et al. stellten durch Doppeldiffusionsexperimente zur Kristallisation des Apatit/Kollagen-Systems faszinierende morphologische Strukturen her, deren Verständnis ebenfalls noch aussteht. Insgesamt ist über die Kristallisation von anorganischem Material an organischen Matrices mechanistisch kaum etwas bekannt. Modellvorstellungen wurden weitgehend von Experimentatoren geäußert und bleiben auf supramolekularen Größenordnungen, so wurde zwar die Kalzifizierung von Kollagen-Fibrillen diskutiert, ohne jedoch die Feinstruktur der Fibrillen zu berücksichtigen. Im Rahmen dieser Arbeit soll die Erforschung der Kompositbildung mit theoretischen Methoden von der Seite der atomaren Größenskala her angegangen werden, um neue Erkenntnisse zu erlangen, die dem Experiment nur schwer oder gar nicht zugänglich sind. Durch die vereinten Anstrengungen mit beiden Ansätzen können vielleicht in Zukunft die beobachteten Strukturen erklärt werden.

Für das Verständnis der mikroskopischen Strukturbildung der Komposite bis hin zu ihrer makroskopischen Morphologie ist für die Zukunft die Etablierung von Simulationsmethoden auf Basis vergrößerter Modelle geplant. Hierfür ist die Kenntnis der elementaren Prozesse notwendig, die der Kristallisation zugrundeliegenden. Generell wurden die Mechanismen der Anlagerung von Ionen an Proteine mit theoretischen Methoden bisher wenig untersucht. Gegenstand dieser Arbeit ist es, die grundlegenden Mechanismen, Energien und Reaktionsraten der Anlagerungen einzelner Ionen an das Protein zu untersuchen, um die

gewonnenen Daten als Basis für die Parametrisierung vergrößerter Modelle verwenden zu können.

Eine Besonderheit des Kollagens sind die endständigen Telopeptide, deren Peptidsequenz zwar bekannt, aber deren Struktur bis heute nicht geklärt ist. Im Fall des Apatit/Kollagen-Systems wird eine besondere Rolle der Telopeptide bei der Nukleation von Apatit diskutiert. Um dies näher zu untersuchen ist es wichtig, die Struktur der Telopeptide zu kennen. Experimente von verschiedenen Gruppen am N-terminalen Telopeptid wurden unterschiedlich gedeutet, ein Konsens ist bisher nicht erreicht. Im Wesentlichen teilen sich die Meinungen in zwei Lager, nach denen die Struktur globulär oder ausgestreckt ist. Scheraga et al. haben eine Modellstudie angefertigt, deren Ergebnisse auf eine kompakte Struktur hindeuten. Es wurden aber energetische Minima durch Geometrieoptimierung im dielektrischen Kontinuum berechnet, die sich formal auf eine Temperatur von 0 K beziehen und damit für die Struktur in wässriger Lösung nur begrenzt aussagekräftig sind. Die Flexibilität des Telopeptids kann für die Anlagerung von Ionen von großer Wichtigkeit sein, da bei hoher Flexibilität möglicherweise Relaxationen erfolgen, die eine Bindung von Ionen stabilisieren. Es ist ein weiteres Ziel dieser Arbeit, die Struktur des N-terminalen Telopeptids und seine Flexibilität mithilfe der Molekulardynamik (MD) zu untersuchen.

Eine hohe Flexibilität des Proteins bedeutet für die Untersuchung mit den Methoden der MD, dass der relevante Phasenraum sehr groß wird. Ionenanlagerungen bei verschiedenen Konformationen des Proteins können nicht in genügender Anzahl beobachtet werden, als dass daraus die statistische Verteilung dieser Ereignisse erhältlich wäre. Ein alternativer Ansatz ist die genauere Untersuchung von Ionenanlagerungen als einzelne Ereignisse. Aus diesem Grund ist es wichtig abzuschätzen, wie groß der Einfluss der chemischen Umgebung einer Bindestelle für Ionen für die Anlagerung ist. Denn aus der Kombination vieler Adsorptionsplätze im Protein und einer sich ständig ändernden chemischen Umgebung durch Relaxationen im Protein ergibt sich prinzipiell eine unüberschaubare Zahl an Szenarien für die Ionenanlagerung. Wünschenswert wäre eine Systematisierung der möglichen Reaktionen bzw. Adsorptionsplätze in Kategorien. Für die Entwicklung vergrößerter Modelle wird detailliertes Wissen über die Energetik und Kinetik der Anlagerungsreaktionen benötigt. Im Rahmen der vorliegenden Arbeit soll eine theoretische Behandlung dieser Fragestellungen unternommen werden.

2 Introduction

2.1 Composites

Composite materials or composites for short, are materials made of two or more components that are combined to take advantage of the favorable qualities of each individual component. In that sense, composites were already used by humans as they built their early dwellings with mud bricks. These were a composite of mud, which when dried was firm and solid, and straw, which gave some flexibility and tensile strength to the otherwise rather crumbly bricks. Nowadays, a composite is understood as a material whose components are arranged on a microscopic or nanoscopic level. The relatively new technical exploration and utilization of composite materials began about in the early 20th century. Fiberglass, engineered in the 1930s, was one of the first newly designed composites. Today, the engineering of composites is an own established field of scientific research.

As in many other cases, the technology is copied from nature. Wood, bone, teeth, nacre and many more materials built by living organisms are composites, called biocomposites. In all these materials, a component providing hardness and rigidity is mixed with a component providing flexibility and tensile strength. The hard material alone would be brittle, the elastic material would be too soft. Only the composite provides the needed combination of qualities. In biocomposites, the organic matrix is provided by a protein or saccharide polymer, of which the most widespread species is collagen¹. The inorganic material is a calcium based salt like carbonate, oxalate, silicate or phosphate in almost all composites found¹. In experiments, other polymers and minerals have been successfully utilized as matrices for mineralization^{2,3}.

Research has revealed that composites are often micro- or nanostructured in an ordered manner. One material functions as a matrix in which the other component is embedded. While technical composites may have a structure ordered on an atomic scale, e.g. the crystal structures of alloys, that are fairly easily determined experimentally, biocomposites are structured on a microscopic or nanoscopic level which is more difficult to study. The principles governing the formation and structure of biocomposites are subject of intensive research. Promising ideas for medical applications exist which makes the research are

attractive for the industry as well. The components of bone and tooth enamel are a collagen matrix and embedded apatite crystallites, which is a system of special interest. The vision of the control over bone and teeth formation is driving many researchers, for the reward could be the end of painful dentist visits and the expansion of possibilities in prosthesis and implantation techniques and bone surgery.

Experimental research of biological composite materials has just in the last decade explored new dimensions from a milli- to micro- to nanometer scale resolution. For an understanding of the formation mechanism of a composite, insight on a molecular scale is needed. The structure and dynamics of the organic matrix as well as the kinetics, mechanisms and energetics of ion adsorption processes are vital information of need. However, little is known so far on these subjects. The structure of collagen is not yet completely clarified, especially the three-dimensional structures of the telopeptides are largely unknown. The very telopeptides are suspected to play a possible key role for the nucleation of minerals, however. Ion adsorption processes are the initial and fundamental processes leading to the nucleation of crystals on an organic matrix. Several studies on the interactions of amino acid side chains exist (see chapter 4.6), but no results for apatite ion-protein interactions have been reported yet.

2.2 The apatite/collagen composite system

Experiments with the apatite/collagen composite system have yielded highly interesting morphologies. Busch et al. obtained spherical aggregates from calcium and phosphate solutions diffusing into a gelatin layer in a double diffusion setup⁴. They have conducted intensive research to enlighten the structure and growth of these aggregates^{5,6,7} by means of transmission electron microscopy (TEM) with resolutions down to a few nanometers. An understanding at a chemical level of detail could not be gathered yet, however.

Collagen is the single most abundant protein in the animal kingdom. The experiments by Busch et al. were carried out with gelatin from calf and pig. There are different types of collagen of which type I is the most common. A detailed description of collagen will be given in chapter 3.4.

The biocomposites bone and teeth are based on calcium phosphate. The closest mineral species is hydroxyapatite of an ideal composition $\text{Ca}_5(\text{PO}_4)_3\text{OH}$ with a small and varying proportion of hydroxide substituted by fluoride. Also, minor substitutions by trace element cations and other halogenide anions are found⁸. The solubility of these species is very low, in saturated solutions, ion concentrations are in the range of 10^{-4} to 10^{-1} mmol/l^{9,10}.

2.3 Objectives of this work

Experimental research on composite formation mechanisms is limited in terms of spatial resolution. Experiments on a molecular or atomic scale are expensive and extremely difficult to realize. It was intended in this work to overcome this situation by employing computational methods to tackle the problem. As will be discussed in detail in this work, this is not a trivial task either. Little theoretical work has been done so far in this field, so that at first models have to be established and the proper functioning of computational methods has to be verified.

It was intended with this work to investigate the fundamentals of the calcification of collagen. The structure of the collagen, especially that of the terminal telopeptide, and the energetics of ion adsorption mechanisms are considered essential knowledge for the understanding of composite formation and shall be studied in this work. Modeling, simulation and theory are used as techniques to gain information on an atomic level.

Hereby, the ongoing experimental research on the formation of apatite/collagen biocomposites conducted on scales down to nanometer resolution can be supported with research on the atomic scale. Results are intended to provide a rationale for the development of mesoscale methods. The aim is to eventually get a consistent interpretation of experimental findings. It is hoped that this will ultimately enable the rational development of new materials.

Several questions should be explicitly addressed in this work, all of which are important for a coarse model development. The structures and particularly the dynamics of the human type I collagen telopeptides are still largely unknown. The size of the N-terminal telopeptide allows

2.3 Objectives of this work

a study with molecular dynamics (MD) simulation. This is the method of choice to assess dynamical properties of a structure. For the C-terminal telopeptide, preparatory work would have to be done to investigate the folding of the protein, which exceeds the scope of this work. A comparison of the dynamics of the triple helix and the telopeptide of collagen is needed for a meaningful evaluation of the results. The reaction energy and stability of ion binding reactions has to be elucidated. The electrostatic properties of the collagen telopeptide and triple helix are to be compared to check for a possible preference of apatite nucleation at either of these. Any hint to the possibility of complexes with multiple ions could be pointing to a preferred nucleation site.

3 Theory

3.1 Molecular dynamics simulation

Molecular dynamics (MD) simulation is a field of constantly expanding applications in chemistry. New methods to simulate rare events or to extract new information are continuously being developed, and ever larger systems can be studied as computer power is incessantly growing. Several program packages exist that are designed for the simulation of polymers, crystals, biomolecules etc.

MD simulation is a technique that is based on classical physics. Matter is described on the basis of atoms as soft bodies, i.e. inertial point mass particles interacting through smooth electrostatic potentials. Two forms of intermolecular potentials are generally considered, the Coulomb type and Van-der-Waals type interactions. This is the basic model that underlies practically all MD simulations, and the assumptions made should be kept in mind. More sophisticated models exist like polarizable atom force fields or quantum mechanical MD, but these are more expensive in terms of computing power and are usually used for smaller systems only. An outline of the treatment of MD is given in the following.

3.1.1 Equations of motion

The physics of classical systems can be described by Newton's equations of motion:

$$\dot{\vec{p}}_i = m_i \ddot{\vec{r}}_i = \vec{F}(\vec{r}_i) = -\vec{\nabla} V_i(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N) \quad (3.1)$$

where i is the particle index, r is the position, p the momentum, m the mass, F the force and V the potential. For N particles, a set of N differential equations of type (3.1) have to be solved simultaneously, where V is a function of *all* particle coordinates in the ensemble. The equations cannot be solved analytically but have to be integrated numerically. This is done by algorithms specifically developed for the solution of motion equations, popular implementations are the Leapfrog and Verlet-Leapfrog schemes.

The discretization demands a resolution in time that describes the fastest motion in the system with adequate precision. Insufficient resolution will result in an unstable simulation trajectory. Bonds between hydrogen and heavy atoms are commonly held fixed by a special algorithm (Shake or Rattle) because they show the fastest vibrations but contribute very little to large scale dynamics significantly. The time step for integration can then be chosen as 1 or 2 fs. This predefines the timescales that are accessible to simulation, since 10^{15} integrations have to be solved to simulate a real-time interval of a second. On a current CPU, the calculation of one integration step for a system of 10^4 to 10^5 atoms takes between one and a few seconds. Depending on the simulation system and on available computing power, up to about 10^2 nanoseconds of simulated real-time are currently the limit.

Each particle in a simulation system interacts with every other particle in the system. Since the Coulomb and Van-der-Waals potentials diminish quickly with increasing r , the potential is only computed for atoms that are closer than a certain cutoff distance r_{cut} . The cutoff can be chosen independently for Coulomb and Van-der-Waals potentials. A shifting function is applied to the potential rather than simply truncating it to minimize artifacts.

To minimize artificial surface effects in a small simulation system, periodic boundary conditions are applied. This means that the box is surrounded by identical copies of itself that are periodically continued to infinity. Together with the cutoff scheme this results in the minimum image convention. It demands that a particle only interacts with the nearest of all copies of the other particles. This also implies that the simulation box must be greater than $2 r_{cut}$ to avoid self-interacting particles.

Electrostatic interactions are not calculated on a pair basis but by the more efficient Ewald algorithm which transforms a part of the computation to the Fourier transformed domain.

Algorithms to simulate a system in contact with a heat bath and a pressure piston have been developed to permit constant temperature and constant pressure simulations. These are far more realistic representations of lab conditions compared to the constant volume and constant energy conditions opposed by the simple simulation. Popular algorithms are Berendsen, Nosé-Hoover and Langevin.

3.1.2 Force field

While integration algorithms are a key part of MD programs and different implementations have their peculiar advantages and disadvantages, the core of applied MD simulations is the formulation of the potential energy function V used in the equations of motion. The general approach is to ignore many-body forces (which is realized through the initial assumption of non polarizable point mass particles) so that the potential energy can be broken down to pair potentials. These are described by a combination of a Coulomb term and a Van-der-Waals term usually in form of a Lennard-Jones (12-6), a Buckingham or a Born-Huggins-Meyer type potential. In addition, chemical knowledge is incorporated by classifying selected types of pairs which are described by specialized interaction terms. Atoms located in 1-2, 1-3 and 1-4 topological order are classified as bonds, angles and dihedrals, respectively. These are typically described by harmonic potential terms and consequently excluded from the calculation of the normal electrostatic pair potentials. A simple general form of the potential energy function V is given in eq. 3.2. The bond, angle and dihedral terms are intramolecular, the Coulomb and Van-der-Waals terms are intermolecular potentials.

$$\begin{aligned}
 V(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N) = & \sum_{i_{bond}=1}^{N_{bond}} V_{bond}(i_{bond}, \vec{r}_a, \vec{r}_b) \\
 & + \sum_{i_{angle}=1}^{N_{angle}} V_{angle}(i_{angle}, \vec{r}_a, \vec{r}_b, \vec{r}_c) \\
 & + \sum_{i_{dihedral}=1}^{N_{dihedral}} V_{dihedral}(i_{dihedral}, \vec{r}_a, \vec{r}_b, \vec{r}_c, \vec{r}_d) \\
 & + \sum_{i_{Coulomb}=1}^{N_{Coulomb}} V_{Coulomb}(i_{Coulomb}, \vec{r}_a, \vec{r}_b) \\
 & + \sum_{i_{vdW}=1}^{N_{vdW}} V_{vdW}(i_{vdW}, \vec{r}_a, \vec{r}_b)
 \end{aligned} \tag{3.2}$$

The whole of the potential energy function and the parameter values in it are usually referred to as the force field. Not all individual atoms or atom pairs, respectively, are assigned individual parameters, but categories of chemically similar atoms are defined. The potential energy function for a specific molecule is then built up by summing over all applicable terms and assigning parameters taken from a look-up table of force field parameters. Several

program packages each include a proprietary force field. They use more or less the same potential energy functional form and essentially differ only in the classification of chemically similar atoms and the parameter values assigned. Parameter values are generated in a process of fitting energies to experimental spectra or quantum mechanical results. The partial charges of atoms are usually calculated with quantum mechanical methods, on a per residue basis for proteins. Modern simulation programs can handle multiple force fields so the user is not restricted to the use of any specific force field.

3.1.3 Statistical interpretation

The theoretical background to extract information from the simulation is provided by statistical mechanics and thermodynamics. Average values for structural properties can be calculated straightforward from a simulation run. The calculation of thermodynamic properties is more complicated. An important prerequisite is the ergodic hypothesis. Statistical mechanics calculations are based on the idea of an ensemble of copies of a system in identical macroscopic states but different equiprobable microscopic states. The ergodic hypothesis formulated by Boltzmann states that N statistical copies of a system and N snapshots of the same system evolving in time, each in the same macroscopic state, assume the same distribution over microscopic states as N approaches infinity. Thus, expectation values of observables can be computed as averages from a simulation trajectory. Effectively, the averaging of ensembles is replaced by averaging over time. From the few quantities that can be calculated for a single system snapshot (coordinates, momenta and energies), virtually any thermodynamic quantity can be calculated using relations that can be deduced by statistical thermodynamics.

3.2 Potential of mean force

3.2.1 Aim of the method

The potential of mean force (PMF) is the profile of the thermodynamical function G (or A), which controls the progression of chemical reactions in an NpT (or NVT) environment, along an arbitrary reaction coordinate. The PMF allows a detailed characterization and comparison of the energetics of different reactions of the same type. From the PMF, key data like the reaction free energy and the activation barrier can be extracted, from which in turn estimates of the reaction rate can be made. With these data, the stability of chemical products can be assessed, as well as the rate at which they are produced.

3.2.2 Reversible work theorem

The theoretical deduction of the formalism as described by Kapral and Ciccotti¹¹ is reproduced here in a schematic description for a system in which the constraint distance is defined by the designated particles 1 and 2. All other particles are referred to as solvent.

An observable A in an ensemble with the constraint ξ is given by

$$\langle A \rangle_c = \frac{\int e^{-\beta H} A \delta(\xi) \delta(\dot{\xi}) d\Gamma}{\int e^{-\beta H} \delta(\xi) \delta(\dot{\xi}) d\Gamma} \quad (3.3)$$

where Γ includes the coordinates and momenta of all particles in the ensemble and δ is the delta function.

The mathematical form of the constraint is given by

$$\xi(\vec{r}_1, \vec{r}_2) = |\vec{r}_1 - \vec{r}_2| - d \quad (3.4)$$

$$\dot{\xi}(\vec{r}_1, \vec{r}_2) = \vec{r}_{12} \cdot \vec{u}_{12} = \vec{r}_{12} \cdot (\vec{u}_2 - \vec{u}_1) \quad (3.5)$$

where r is the position, d is the specified distance of the constraint and \vec{u} is the velocity. The distance between the particles 1 and 2 is fixed (eq. 3.4), thus there can be no relative velocity between them (eq. 3.5).

The force acting on the particles is split into a contribution F_d by the potentials of the bare particles and a contribution F_{PS} by the particle-solvent interaction.

The direct force F_d is constant and can be calculated from the force field without simulation. The instantaneous force F_{PS} along the constraint induced by the particle-solvent interaction can be calculated from the interaction potential of particle and solvent V_{PS} by

$$F_{1|_{12}} = \nabla V_{PS} \cdot \frac{\vec{r}_{12}}{|\vec{r}_{12}|} \quad (3.6)$$

The expectation value must be obtained from simulation according to

$$\langle F_{1|_{12}} \rangle_c = \frac{\int \nabla V_{PS} \cdot \frac{\vec{r}_{12}}{|\vec{r}_{12}|} \cdot e^{-\beta H} \cdot \delta(\xi) \delta(\dot{\xi}) d\Gamma}{\int e^{-\beta H} \cdot \delta(\xi) \delta(\dot{\xi}) d\Gamma} \quad (3.7)$$

This is the average value of the projection along the inter-particle axis of the force exerted on particle 1 by the solvent in the constrained ensemble. The total mean force can be calculated by computing the average of the force on particle 1 and 2 and adding the direct force F_d

$$F(r_{12}) = F_d(r_{12}) + \frac{1}{2} \langle F_{2|_{12}} - F_{1|_{12}} \rangle_c \quad (3.8)$$

Known as the reversible work theorem, the mean force can also be written as

$$F(r_{12}) = \frac{dW(r_{12})}{dr_{12}} \quad (3.9)$$

where W is the reversible work resulting from the mean force. Integration of the mean force hence yields the potential of mean force W .

3.2.3 Simulation technique

For the calculation of a PMF via MD simulation, different techniques exist like umbrella sampling or the coupling parameter approach. The coupling parameter can be an external variable λ that controls the mixing of a two state hybrid potential function, or an internal

coordinate ξ that is identified as a reaction coordinate. The latter technique is well suitable for systems in which a distance can be identified as the reaction coordinate, which is the case for the investigated ion adsorption processes.

For the MD simulation, a constraint is defined that describes the reaction coordinate ξ , called the pmf constraint. In praxi, multiple simulations are carried out with the pmf constraint held fixed to different distance values. For each simulation, the average of the force which acts upon the pmf constraint, called the constraint force, is calculated. By interpolation between the calculated points, the constraint force is obtained as a continuous function of the reaction coordinate. Integration of this function yields the potential of mean force.

The described method is well known as thermodynamic integration (TI) when used with an external variable λ as the coupling parameter. In our case an internal coordinate ξ is used as the coupling parameter. The term thermodynamic integration is less often used in this conjunction and the method shall be referred to as constraint force averaging.

3.2.4 Error estimation

Large errors can occur in PMF calculations which is discussed in the literature¹¹. This necessitates a thorough error estimation. Apart from the usual sources of errors in MD simulations, two parameters directly control the precision of the constraint force profile: the chosen number of points along the reaction coordinate, and the duration of each simulation.

The number of points on the ξ -axis is best decided heuristically by starting with a fairly coarse resolution and increasing it in regions where large changes of the mean force occur or decreasing it where little changes are observed. The total number of points should be as small as possible to keep the calculations economic but as large as needed to cover all important features of the profile.

The duration of each simulation controls the statistical error of the calculated mean. Mathematically, recording the mean force during the simulation represents the acquisition of a sample from a population. The mean force is a statistic used as a point estimate for the

corresponding parameter of the distribution function. The mean M and the standard deviation σ of a sample are given by

$$M = \frac{1}{N} \sum_{i=1}^N x_i \quad (3.10)$$

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - M)^2} \quad (3.11)$$

The standard error of the mean depends on the sample size and is given by

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \quad (3.12)$$

The standard deviation of the constraint force is controlled by physics and cannot be influenced by parameter choice. The sample size is the variable that can be controlled to reach the desired precision in the calculation. However, the application of this formalism requires that all samples are independent or random. This is not the case for the constraint force which obviously shows autocorrelation as it is governed by Newton's equations of motion. Thus, an effective sample size N_{eff} has to be determined which represents the number of independent scores in the sample. This is done by determining a decorrelation time (in samples) for the constraint force. One technique is to calculate it from the autocorrelation function (acf) which is the correlation of the quantity with itself at a different time (eq. 3.13). The acf C is plotted against the sample lag τ and ranges from 0 for a random process to 1 for completely correlated events.

$$C(\tau) = \frac{\sum_{i=1}^{N-\tau} (y_i - \bar{y})(y_{i+\tau} - \bar{y})}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (3.13)$$

Even for a completely random process, C is not exactly zero but fluctuates on a level dependent on the number of samples. For a given number of samples, a confidence interval can be calculated within which the acf may fluctuate¹². The equation above is the biased acf which fades to zero for $\tau = N$. This is in contrast to the unbiased acf for which statistical noise increases to a maximum for $\tau = N$ as it is normalized to the number of data samples which becomes zero. Both the biased and unbiased acf of a finite number of samples are estimates for the true acf of an infinite number of samples^{13,14,15}. They are therefore also called sample autocorrelations, an analysis of which is recommended to be restricted to the first 10 to 25 %

at maximum^{14,15}. The biased formulation of the acf is often preferred over the unbiased acf in statistics, because the latter one can lead to autocorrelation sequences which are not positive semidefinite, possibly causing problems with spectral estimation algorithms. The confidence band cb for a biased acf is

$$cb = \frac{z\sqrt{N-\tau}}{N}, \quad z = \text{ppf}(1 - \frac{\alpha}{2}) \quad (3.14)$$

where α is the significance level and z is the numerically computed percent point function (ppf) of the standard normal distribution. Typical confidence levels are 95 % ($\alpha = 0.05$, $z = 1.96$) and 99 % ($\alpha = 0.01$, $z = 2.58$). By reducing N to an effective N_{eff} , the confidence band is raised. When lowered until the acf fluctuates within the confidence bands, N_{eff} can be regarded as the number of statistically independent samples.

The mean force can be described as an interpolated continuous function of the constraint distance. This function can be integrated to obtain the potential of mean force. The data points will most practically be interpolated by spline functions and integrated with standard numerical methods. An error estimate for the integrated function can be made when approximating the integral by a Riemann sum.

$$F = \int_a^b f(x)dx \approx \sum_{k=1}^n f(x_k)\Delta x_k \quad (3.15)$$

where n is the number of sampling points x_k and Δx_k are the sample widths. The absolute error of the integral F can then be calculated by simple error propagation laws

$$\Delta F = \sum_{k=1}^n \Delta f(x_k)\Delta x_k \quad (3.16)$$

Note that Δx here refers to the discretization resolution of the Riemann sum, $\Delta f(x)$ refers to the absolute error of $f(x)$. A drawback of the integration of experimental data with measurement uncertainty is the accumulation of the absolute error from left to right. If the function F does not constantly rise but rather fluctuate around zero, the relative error can become tremendous, just in analogy to the small difference of two great quantities with measurement errors.

3.3 Analysis techniques

A premise for the analysis of molecular dynamics trajectories is that the system is in an equilibrium state and that the corresponding phase space is thoroughly sampled. Sufficient snapshots have to be taken to let the approximation be reasonable that the time average equals the average over an infinite number of fictitious ensembles. If these requirements are met, a large variety of properties can be calculated. For example, distribution functions, expectation values of thermodynamical variables and other observables, structural properties, diffusion coefficients and various time correlation functions can be obtained. The calculation of properties that are evaluated in this work are described in detail in the following.

3.3.1 Mobility analyses

Information about the dynamic behavior of a protein is the primary advantage of molecular dynamics over static methods like quantum mechanics or others. Motions can have very different orders of magnitude in time and space, ranging from femtosecond vibrations of single bonds to large domain protein folding that can take seconds. A standard means to assess the motions of a protein in simulations is the so called root mean square displacement (RMSD).

3.3.1.1 RMS values

RMSD_{t0}

The RMSD is an often calculated quantity in MD analysis. The mean quadratic distance of the atoms positions at time step t and at time step t_0 is calculated as a function of time.

$$RMSD(t) = \sqrt{\frac{1}{N} \sum_{i=1}^N (\vec{r}_i(t) - \vec{r}_i(t_0))^2} \quad (3.17)$$

where the sum is over N atoms in the system. The RMSD provides a measure of how far an atom moves from its initial position during the simulation. In the case that atoms vibrate about

a minimum structure, the RMSD is comparable to the temperature factors measured in X-ray experiments. Since the RMSD as given in eq. 3.17 is referenced to the time step t_0 , it will be referred to as RMSDt0.

According to the type of information needed, the quadratic averaging can also be done over all time steps so that the RMSD is a function of the atom index i . This information reveals which atoms are more mobile than others.

RMSDm

The RMSDt0 is however primarily useful to judge whether a more or less artificial starting structure has equilibrated to a stationary geometry. During relaxation the RMSD as a function of time usually rises until it reaches a roughly constant value. This is caused by the fact that e.g. a protein changes from its initial artificial and unfavorable geometry into a stable equilibrium state. The reference to the structure at t_0 is only as meaningful as this structure itself. If the starting structure is modeled and therefore not more than an intelligent guess, the reference to this structure is not reasonable. To assess the motions and flexibility in a protein, the RMSD with respect to the mean position of the atoms is more meaningful. This equals the standard deviation of the coordinates of an atom at all time steps. According to the reference structure used, the abbreviated terms RMSDt0 and RMSDm will be used in the following for distinction of these two.

At a first glance, the whole trajectory has to be processed two times to calculate the RMSDm. In a first pass, the mean has to be determined, in the second pass, the deviations from the mean can be summed up. With some quick rewritings, the RMSDm can be calculated more conveniently in one pass:

The second moment or standard deviation of a sample is

$$\sigma = \sqrt{\langle (x - \langle x \rangle)^2 \rangle} = \sqrt{\langle x^2 \rangle - \langle x \rangle^2} \quad (3.18)$$

This equation holds for vectors as well, since they satisfy the distributive and associative laws as scalars do.

Both terms are squared so that they become scalar products in the vector formulation. Substituting a vector and simplifying yields the RMSDm as

$$RMSDm = \sigma = \sqrt{\frac{1}{N} \sum_{t=1}^N (x_t^2 + y_t^2 + z_t^2) - \frac{1}{N^2} \left[\left(\sum_{t=1}^N x_t \right)^2 + \left(\sum_{t=1}^N y_t \right)^2 + \left(\sum_{t=1}^N z_t \right)^2 \right]} \quad (3.19)$$

RMSmob

As a measure for mobility, the distribution of an atom in space over the simulation time would be the most detailed information. For visualization purposes however, the information has to be reduced to a single number that can be assigned to an atom. It is clear that not the whole characteristics of the movements of an atom can be distilled into one number, so not all possible distributions can be distinguished. The main focus of the measure is to distinguish between atoms that move little or that move amply. Firstly, a grid approach is employed. The distribution of an atom in space is thus sampled as a three dimensional histogram of residence times in volume elements of the simulation box. Because the simulation time is equal for all atoms, the integral over the volume elements multiplied by their occupation must be the same for all atoms, as well as the mean residence time in a volume element. The residence times of an atom in volume elements are related to the number of volume elements occupied. An analysis of the occurring residence times thus also accounts for the number of volume elements occupied. The quadratic mean (RMS) of the residence times is thus an appropriate measure. If N time steps are simulated, the RMS of the residence time (RMSrt) can in principle reach the extreme values of 1 for an atom that is distributed to a maximum spatial range (as it moves to a new grid cell at each time step) and N for a fixed atom (as it resides in the same grid cell during all N time steps). Since the RMS residence time in a volume element is contrariwise to the volume populated by an atom, the reciprocal of RMSrt can be interpreted as a measure for mobility. Since the RMSrt values spread over a large range, they are scaled to the square root. This is then titled RMSmob. It is well comparable with other measures like RMSDm.

The program *spacedist* was written in FORTRAN that calculates the RMSDt0, RMSDm and RMSmob as measures for the distribution of atoms in space. For the grid based analysis of the movements of an atom, of course the orientation and movement of the whole protein influence the histogram. The protein structure was therefore reoriented for each time step so

that the location of the triple helical domain, which presumably shows least internal deformations, remained stationary in the laboratory system. This was achieved by applying a matching algorithm by Ferro and Hermans¹⁶, which was implemented and kindly made available by Müller¹⁷. An algorithm was developed that uses a grid as a three-dimensional array of counters. Each volume element in the grid is assigned to a counter that accumulates the number of time steps an atom resides in this volume element. The program analyzes the whole trajectory and updates all counters for every time step. Once this is completed, the discussed mobility measuring quantities are calculated with the respective formula.

Inherent to any grid based approach is the direct dependence of memory requirements on the size of the grid, which is in this case dependent on resolution and simulation box size. The grid is a three-dimensional four byte integer array. A box size of 100 x 100 x 100 Å³ with 0.5 Å resolution requires 30 MB for one grid. The RMSmob value is a per-atom measure, thus each atom needs its own grid array. For the given example, the roughly 400 heavy atoms of the telopeptide require 12 GB of memory which is feasible on modern machines.

An RMSD value (root mean square deviation) makes reference to a mean. The mean does not necessarily represent a meaningful information. If an atom more or less jumps between two locations where it resides mostly, the mean position would be a useless information, for example. For the question of mobility and flexibility, only the real movements of the atoms are of interest. Thus a grid approach is better suited and leads to more meaningful results because it is independent of the actual form of the distribution of an atom over space.

3.3.2 Structural analyses

3.3.2.1 Hydrogen bonds

Hydrogen bonds contribute an important effect to protein structure, they are essential for the formation of secondary structure motives. The basic form of a hydrogen bond is constituted by one donor and one acceptor atom. Fig. 3.1 depicts a common hydrogen bond formed by a nitrogen atom as donor and an amide group oxygen as acceptor, where β is the angle between

the N-H bond and the axis through the acceptor O and hydrogen, r is the distance between donor and acceptor atom.

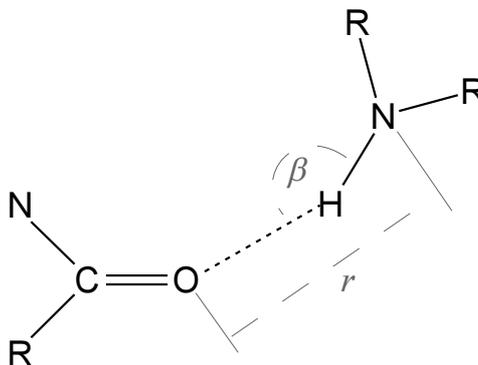


Fig. 3.1: Hydrogen bond definition.

Hydrogen bonds can also be mediated through water molecules but these are much weaker. For the automated recognition of hydrogen bonds, a simple geometric definition is used based on the distance between r donor and acceptor and the angle β formed by donor, hydrogen and acceptor. A hydrogen bond is detected when the distance r is less than r_{cut} and the angle β deviates less than β_{dev} from the ideal value of 180 degrees. The parameters r_{cut} and β_{dev} can be adjusted to a value proving to be reasonable for the system examined. The applied values will be stated with each calculation.

3.3.3 Radial distribution function

The radial distribution function $g(r)$, also simply called g -function, reflects the relative probability of finding two atoms at a distance r apart. The function is normalized to the probability at infinite distance. The function can also be interpreted as the probability to find atom 2 at a distance r to atom 1. This illustrates the radial character of the function, since atom 2 may be located on a spherical shell around atom 1. When the g -function is calculated for atom types, the function reflects the bulk density at infinite r and the short-range order at small values of r . In the case that one atom type is a solute species and the other type is a solvent species, the g -function contains information about the solvation coordination sphere.

The scalar g -function is formulated as

$$g(r) = \frac{V}{N^2} \left\langle \sum_i \sum_{j \neq i} \delta(|\vec{r}_i - \vec{r}_j| - r) \right\rangle \quad (3.20)$$

where δ is the delta function that is ∞ when the distance between particle i and j is equal to r , and zero otherwise. For the calculation of g from a simulation trajectory, the delta function is replaced by a Boxcar function, so that effectively a histogram is calculated. The integrated distribution function G (eq. 3.21) gives the mean number of atoms 2 within a sphere around atom 1. This is a useful representation to examine e.g. solvation spheres.

$$G(r) = 4\pi \frac{N}{V} \int_0^r g(s) s^2 ds \quad (3.21)$$

where s is a radius as an argument of g to be distinguishable from r as the argument of G . G has the same unit as N .

3.3.4 Analysis of kinetics

The Eyring transition state theory (TST) provides a basis to derive kinetic data from the activation free energy of a reaction. In our case, the activation free energy is obtained from simulation. To keep the application of the theory simple and efficient, some additional approximations are made. The resulting kinetic data should therefore be interpreted as estimates.

The Eyring theory assumes an equilibrium between the reactants and an activated complex with an equilibrium constant K^\ddagger . According to the theory the rate constant k of the reaction is then

$$k = \kappa \frac{k_B T}{h} K^\ddagger \quad (3.22)$$

or

$$k = \kappa \frac{k_B T}{h} (C^\ominus)^{1-n} \exp\left(\frac{-\Delta^\ddagger G^\ominus}{RT}\right) \quad (3.23)$$

where k_B is the Boltzmann constant, h is the Planck constant, κ is the transmission coefficient, C^\ominus is the standard concentration, n is the molecularity of the reaction and $\Delta^\ddagger G^\ominus$ is the Gibbs energy of activation defined for a single reaction coordinate. The transmission coefficient is a measure of the fraction of trajectories that, once they have reached the activated complex, proceed to the products state rather than return to the reactants state. The transmission coefficient depends on the friction and diffusion properties of reactants and solvent. The determination of κ by means of MD simulation is a complex and very costly procedure. Since reactants, solvent and reaction mechanism are supposedly very similar for all reactions investigated in this work, and the results are primarily subject to comparative studies, we refrain from the calculation of κ . For the reasons stated it is assumed that κ will be similar for all reactions and that qualitative trends of k will be correctly reflected. Comparative analyses of the rates for the investigated reactions should be reliable, comparisons with other studies should be made with caution.

3.4 Collagen

Collagen is the family of proteins constituting the basis for the connective tissue (extracellular matrix) in all multicellular organisms, and it is the most common protein. The main structural unit is a coiled coil of three helical molecules forming a thin long rod of approximately 15 Å diameter, the triple helix. There are at least 19 types of collagens varying in sequence and folding¹⁸. All of them contain triple-helical segments of different length that are interrupted by non-helical domains to induce different kinds of three-dimensional structures¹⁹. The collagen types I, II, III, V and XI feature a triple helix of 3000 Å length. Type I collagen is the most common one and it is found in bone and teeth. The gelatin used by Busch et al. for their experiments is also based on type I collagen²⁰. In vivo, it is synthesized within cells as soluble procollagen. After removal of the C- and N- terminal propeptides by enzymes, collagen is excreted from the cell²¹ (fig. 3.2). The molecule then spans three individual chains of about 1050 amino acids each.

3.4 Collagen

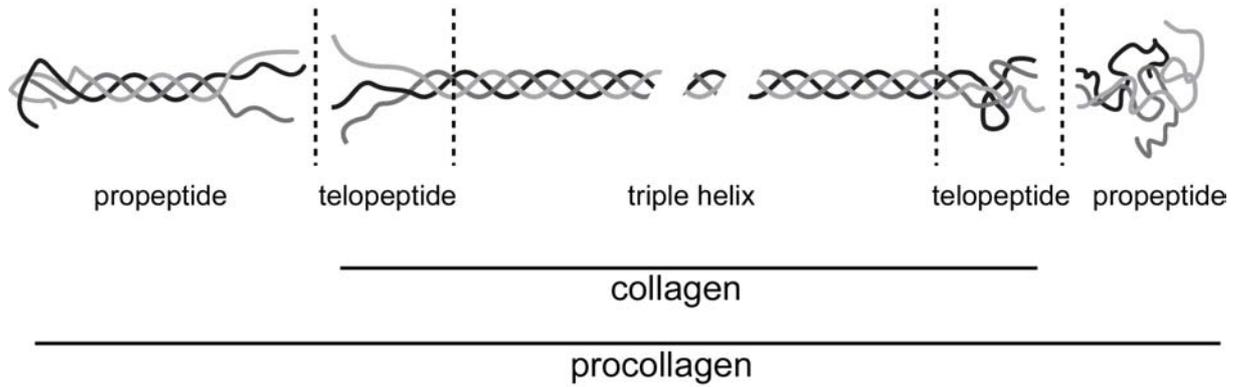


Fig. 3.2: Collagen domains, illustration following Kadler et al.²¹

The triple helices assemble to large fibrils of 10 to 300 nm diameter and up to many μm length. These fibrils finally group to larger bundles, the collagen fibers. The fibrils show a highly regular packing of collagen molecules. Adjacent molecules are aligned with a displacement of 67 nm in the fibril direction, while molecules in a row are separated by a 35 nm gap region²². Hence the terminal domains of each collagen molecule, the telopeptides, are in contact with a large amount of solvent, which is also obviously true for the telopeptides at the ends of the fibril, too. Thus a possible dedicated function of the telopeptides as the sites of initial mineral nucleation can be discussed.

Since the telopeptides are located at gap regions and at fibril ends, they all reside in large volumes of solvent compared to the triple-helical domains. This provides the possibility for ions to reach the telopeptides, and for the structure to be flexible, since they are not as much restricted by the packing of molecules in the fibril as the triple helices.

Owing to the size of the protein, an experimental X-ray structure could not be obtained yet. However, X-ray structures have been measured for small synthetic or natural fragments of collagen^{23,24,25}. Based on these data, the triple helix model was refined^{26,27}. At the ends of the chains, the N- and C-terminal telopeptides are located. The secondary structure of these domains is not known. Despite years of research, there is still a debate among experimentalists about the three-dimensional structure of the N-terminal telopeptide. Extensive experimental studies using various methods were carried out on isolated telopeptides in solution^{28,29,30}. The results were interpreted as hints on a mainly extended structure, with a likely beta turn stabilized by a hydrogen bond in both the $\alpha 1$ and $\alpha 2$ chains.

However, the isolated telopeptide chain molecules presumably adopt a structure very different from the telopeptides *in situ*, since the interactions of the individual chains in the trimer and the interactions between different molecules in a fibril are obviously not existent for isolated molecules. The relevance of the results for the telopeptide trimer is thus questionable. Newer experimental and theoretical studies focus on the consideration of the telopeptide trimer interacting with its environment in a fibril or in the fibril forming process. Based on the results of an X-ray structure on an intact fibril, Miller et al. proposed a contracted structure with a sharp hairpin turn³¹. Scheraga et al. confirmed a contracted structure in a modeling study³². Based on an energy minimization technique with a continuous dielectric environment, they found multiple minimum structures and a distinct global minimum within their results. All structures found were of a contracted nature.

In a recent theoretical modeling study, Veis et al. found that the free telopeptide is essentially unstructured, but adopts a folded structure in interaction with other helices³³. The shortcomings of the studies to date are an insufficient resolution in experimental works and the static nature of the modeling techniques which only search for an energy minimum rather than modeling the dynamic structure at experimental temperature conditions. Veis et al. applied short dynamic runs of a few hundred picoseconds, but did not treat the solvent explicitly. The investigations trying to treat the telopeptide in its fibril environment mainly suffer from the lack of knowledge about the fibril packing in *atomic* detail. It is also accepted that only a fraction of the potential cross-link sites actually form cross-links.

Information needed to unambiguously consolidate the experimentally available information into a three-dimensional atomistic model is missing, such as the rotational orientation of collagen molecules relative to each other, the exact triple helix chain registration, the geometry of side chains and the identification of chains involved in cross-link sites. Veis et al. provide some results regarding these questions, but assumptions were made so that the results are not generally applicable but rather hypothetical. Simulation of the telopeptide in a section of the fibril environment thus appears premature, more proven results are needed to build a reliable model system.

The dynamics of the free telopeptide in solution is still a valuable source of information. Dynamical simulation can reveal whether a structure is folded or flexible, and can quantify

the flexibility of the structure. Since the gap region in the fibrils provide some free space for the telopeptides, the flexibility of the structure is an important feature. Also the question may be answered whether the telopeptide has a folded structure in solution or adopts this structure only when interacting with other helix molecules. Of course, the final goal for future work still remains the simulation of a complete fibril environment.

For the C-terminal telopeptide, neither modeling studies nor experimental data with sufficient resolution exist to give a workable idea of the structure. Since with 64 residues it is a fairly large protein, an MD study does not make sense until a reasonable model is established with other techniques tackling the folding problem. Homology modeling is not a promising approach since no structural data exist for any homologous proteins. For this reason, the present work is focused on the N-terminal telopeptide.

4 Computational details and results

For all simulations presented in this work, the charmm22 force field³⁴ was used with the TIP3P water model. The simulations were carried out with the DLPOLY³⁵ and NAMD³⁶ program packages. Parameters for calcium, hydrogen phosphate and fluoride ions were taken in part from Hauptmann et al. who published parameters for calcium, fluoride and phosphate ions. Building upon these results, Zahn et al. derived parameters for hydrogen phosphate by calculating new charges and the rotational barrier for the hydrogen atom using quantum mechanical methods³⁷.

Further supplementations to the force field are described in the chapters where they are first used.

Analyses and visualizations are carried out with the VMD³⁸ and MOLCAD³⁹ programs as well as special, personally written programs. Thanks are due to Dr. B. Schilling for providing an adapted version of the byte swapping program *histool* for DLPOLY trajectories.

Structure manipulations were done with the SYBYL⁴⁰ program.

4.1 Sequence analysis

The sequence of human type I collagen has been completely determined and is available from the Protein Information Resource International Protein Sequence Database (PIR-PSD)⁴¹ (PIR codes CGHU1S, CGHU2S). The sequences from other organisms are only available in fragments (bovine, rat, chicken). An alignment of the $\alpha 1$ and $\alpha 2$ chains shows that approximately 64 % of the total sequences are identical and 70 % of the charged amino acids are identical.

4.1 Sequence analysis

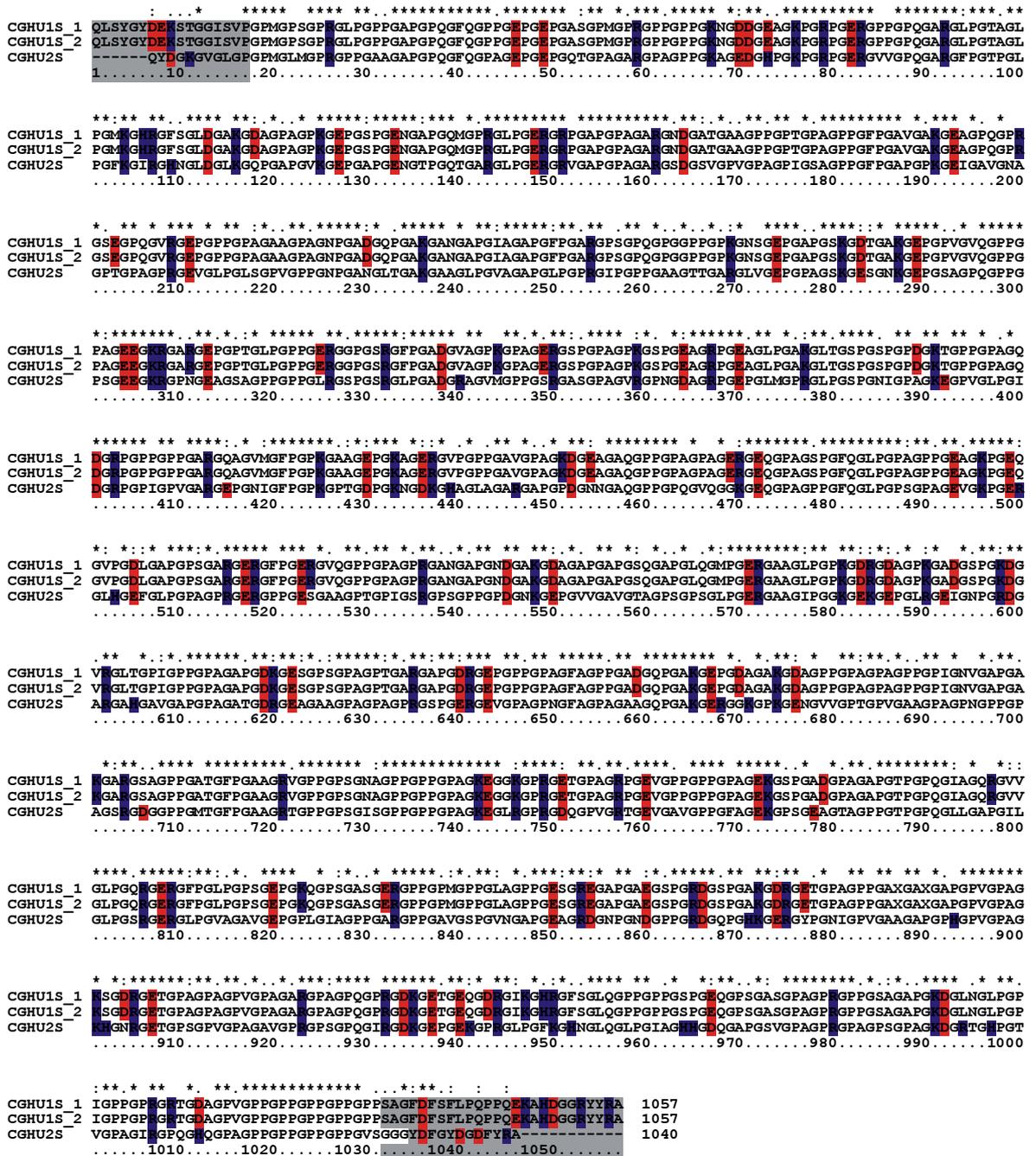


Fig. 4.1: Human type I collagen $\alpha 1$, $\alpha 2$ chain alignment. (gray = telopeptides, red = negatively charged, blue = positively charged residues)

The N-terminal telopeptides extend over 17 and 11 residues in the $\alpha 1$ and $\alpha 2$ chain, respectively, and 26 and 12 residues in the C-terminal telopeptides. Charged residues are often located in groups of two or more in direct neighborhood. Statistics on the amino acid composition of the complete collagen trimer (after cleavage of the propeptides) shows that charged residues are fairly equally distributed among triple helix and telopeptides. Only a slight accumulation of charged residues in the C-terminal telopeptides can be observed. It is

4.1 Sequence analysis

concluded that a special role of the telopeptides for crystallite nucleation cannot originate from a higher density of charged residues.

domain	total res.	charged res.	(%) charged
collagen	3154	503	15,9
N-telopeptide	45	8	17,8
triple helix	3045	479	15,7
C-telopeptide	64	16	25,0

Table 4.1: occurrence of charged amino acids (HIS assumed neutral).

triple helix	total res.	Pro/Hyp/Gly	other res.
residues	3154	1724	1430
%	100	54,7	45,3

Table 4.2: Pro/Hyp/Gly content of triple helix.

The triple helix motif demands a Gly residue in every third position, because in the helix interior, no space is available for a greater side chain. A nonstandard amino acid found in collagen but only few other proteins is 4-hydroxyproline. It is produced from proline through hydroxylation after protein expression and stabilizes the triple helical structure through hydrogen bonds. About 42 mass percent of proline is hydroxylated in collagen. The three-letter-code for hydroxyproline recommended by the IUPAC⁴² is Hyp, while no recommendation for a one-letter symbol is given except X for any nonstandard amino acid. The sequence data from PIR-PSD does not distinguish between Hyp and Pro. The amino acids Pro, Hyp and Gly make up approximately half of the protein, and the PXG triplet is often stated as an ideal collagen prototype. Of the 338 helical triplets per chain (1015 total), however, only 111 (10,9 %) are pure PXG triplets.

A comparison of the N-terminal collagen telopeptides from different organisms shows high homology in the $\alpha 1$ chain, but only little identity in the $\alpha 2$ chain. If the telopeptides are to fulfill a special function, it does not seem to depend on an exact sequence motif.

4.1 Sequence analysis

```

      *:*****:* ::*
CGCH1S QMSYGYDEK SAG-VAVPG
CGRT1S QMSYGYDEK SAG-VSVPG
CGBO1S QLSYGYDEK STG-ISVPG
CGHU1S QLSYGYDEK STGGISVP-
      1.....10.....

      *: * * . . **
CGHU2S QY-D-GKGVGLGP-
CGBO2S QF-D-AK G-G-GPG
CGRT2S QYSD--K GVSAGP-
CGCH2S QY-DPSK AADF GP-
      1.....10....

```

Fig. 4.2: Alignment of collagen telopeptides from different organisms (red = negatively charged, blue = positively charged, HU human, BO bovine, CH chicken, RT rat, 1S = $\alpha 1$, 2S = $\alpha 2$).

No distinct features of the mere primary structure of the telopeptide could be found that indicate a special function. A more detailed analysis of the secondary structure and the dynamics and chemical and physical properties of the telopeptide are required to reveal possible functionalities.

The collagen molecule consists of 3000 amino acids and is much too large to be simulated as a whole. The selection of subsections is required to generate systems that can be simulated. An obvious approach is to firstly distinguish the telopeptides and the triple helix as the primary secondary structure domains. A model for the structure of the triple helix exists which was gained from crystallographic data of small fragment molecules, modeling and statistical analyses. The structures of both the N- and C-terminal telopeptides are unknown. For the C-terminal, very few detailed data exist. The N-terminal telopeptide has been subject of quite a few studies, which did not lead to a consensus model yet but rather to an ongoing controversy about whether the structure is extended or contracted, flexible or rigid. To clarify the structure of the N-telopeptide, a detailed study is necessary. A good atomistic protein model structure is needed to further investigate the nucleation of apatite. The protein's flexibility is an extremely important feature to be clarified to choose the best methods for further in-depth studies. A highly flexible and yet large protein cannot be simulated with full MD for the timescales needed to observe crystallization events.

The N-terminal telopeptide is small enough so that its dynamics can be simulated for several nanoseconds. The C-terminal telopeptide is too large, protein folding methods have to be applied first which cannot be done in the scope of this work. The complete triple helix is also far too large to enable the simulation of nanosecond durations. A subsystem as meaningful as possible has to be chosen as a simulation system. Therefore, the N-terminal telopeptide and a reasonable selection of the triple helix will serve as two model systems that will be studied in

detail in this work. The selection and buildup of the simulation systems will be described in the following subchapters, together with simulation results and structural analyses.

4.2 Structural analyses

4.2.1 Collagen telopeptide

4.2.1.1 Model preparation

Buildup of the simulation system and force field

The global minimum structure of the N-telopeptide as modeled by Scheraga et al. is used as a starting model for the MD simulations. The telopeptides are connected to the triple helix, of which a six residue long segment is included. By this, the effect of the triple helix should be sufficiently accounted for and at the same time flexibility of the individual telopeptide chain arrangement is allowed. At the cut end of the triple helix, artificial bonds between the α -carbon atoms of the last residue of each chain are introduced to mimic the cohesion of the individual chains that would be present in a continuing helix. The cut end was also terminated with neutral charge by saturating the free valences on backbone carbon atoms with hydrogen. Standard protonation states for the neutral milieu were employed for all amino acids. This is a reasonable choice since mineralization experiments were reported with pH values between 5 and 10^{20,43}.

The telopeptide sequence contains a special and fairly rare residue. The N-terminus is formed by pyroglutamic acid (or 5-oxoproline).

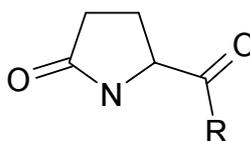


Fig. 4.3: non-standard amino acid pyroglutamic acid.

4.2 Structural analyses

The one-letter code Q is used in the PIR database entries, but Z is recommended by the IUPAC as one letter code, Glp as three-letter code. Pyroglutamic acid is always amino terminal and is created enzymatically from glutamic acid after expression of a protein. Glp is a neutral amino terminus under physiological conditions, pK_a-values for protonation at the pyrrolidinoyl oxygen are below zero and the nitrogen lone pair is delocalized in the amide bond⁴⁴.

As a non-standard amino acid, Glp is not included in most force fields. Ponder et al. have implemented Glp for the Amber98 force field⁴⁵. The atom type assignments were ported to the charmm22 force field by choosing chemically most similar types for which all needed parameters were available. Since the Amber98 force field has fewer atom types than the charmm22 force field, there was some freedom in the assignment. Charges were taken directly from Ponder. The atom types assigned are listed in table 4.3.

atom types	C	O	CA	HA	CB	HB	CG	HG	CD	OE	N	HN
Amber98	2	24	1	35	1	34	1	34	2	24	14	29
charmm22	C	O	CT1	HB	CT2	HA	CT2	HA	C	O	NH1	H

Table 4.3: Designation of atom types for Pyroglutamic acid.

To allow for a weak distance restraining that does not have unnecessarily high artificial influence on a system, a new bond type potential was implemented. The potential U_{conf} was created that is zero below a threshold d_{thr} and harmonic above the threshold (eq. 4.1).

$$U_{conf}(d_{12}) = \begin{cases} 0 & \text{for } d_{12} < d_{thr} \\ k(d_{12} - d_{thr})^2 & \text{for } d_{12} > d_{thr} \end{cases} \quad (4.1)$$

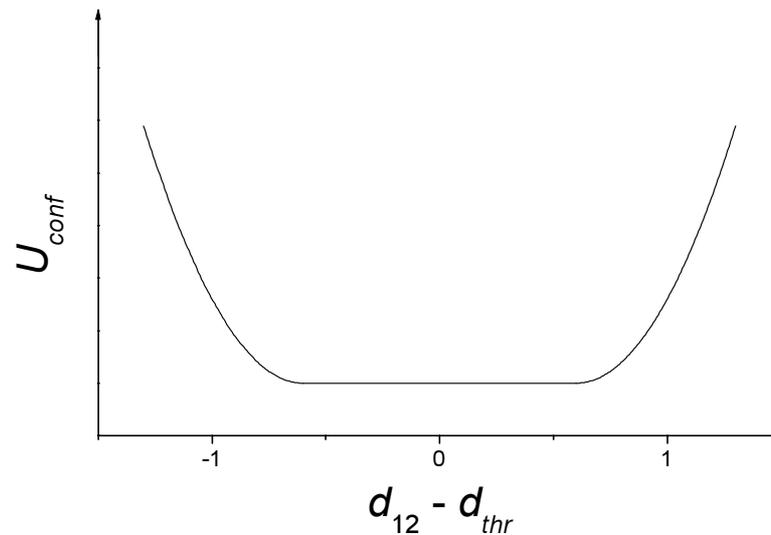


Fig. 4.4: "Confining" potential for distance restraints

It is termed "confining" potential and it is intended as a distance restraint which allows for a certain deviation before a restoring force applies. This way, the simulation is unrestrained as long as the distance varies only within a defined threshold. Only when the distance deviates more than the threshold from its desired value, the restraining potential actually applies. The potential was implemented in FORTRAN in the DLPOLY code and in Tcl in the NAMD code.

4.2.1.2 Long time simulation

Simulation protocol

A simulation system was set up consisting of three protein strands à 15, 21 and 20 residues (56 total) in a $60 \times 60 \times 60 \text{ \AA}^3$ water box. The protein has a total charge of -2. Six sodium and four chloride ions were added to simulate a 50 mmol/l NaCl solution as well as to keep the complete system neutral. A snapshot of the system is shown below (fig. 4.5).

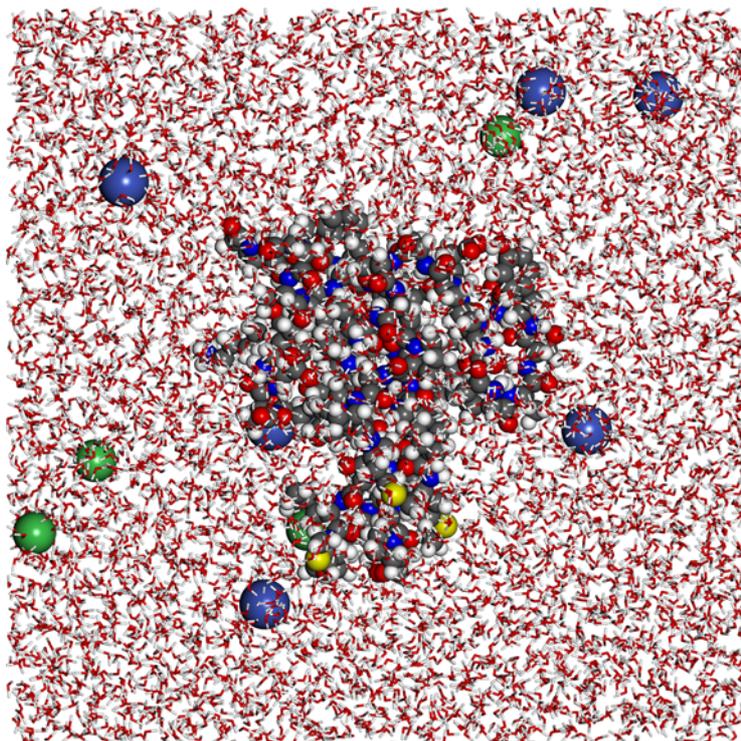


Fig. 4.5: Snapshot of the telopeptide simulation system 2. Protein in small CPK, ions in CPK, water in stick representation. Blue = sodium, green = chloride.

The telopeptide in solution was equilibrated for 0.5 ns and then simulated at constant temperature and constant pressure conditions for a duration of 20 ns with the NAMD program. The Langevin-Piston algorithm was used at 1 bar and 300 K, PME electrostatics were used with an approximately 1 Å grid resolution. All bonds involving hydrogen atoms were held fixed using the SHAKE algorithm (SETTLE for water molecules) with a convergence criterion of 10^{-8} . The simulation was run for 20 ns with an integration time step of 2 fs. The telopeptide was free except restraints at the end of the triple helix cutting site. The distances of the C-terminal residues of the three chains were restrained. The distances of the C atoms of each possible chain pair were restrained, as well as the distances between all possible C and C_{α} atom pairs, using the confining potential with a threshold of 0.7 Å and a force constant of 300 kcal/mol. The restraints were needed to prevent untangling of the triple helix, which was observed after a few nanoseconds in a completely unrestrained simulation. The distance restraints on the triple helix ends are a physically reasonable way to account for the missing triple helix that was cut off. Plots of some physical properties taken from the simulation and indicating equilibration are shown in figure 4.6.

4.2 Structural analyses

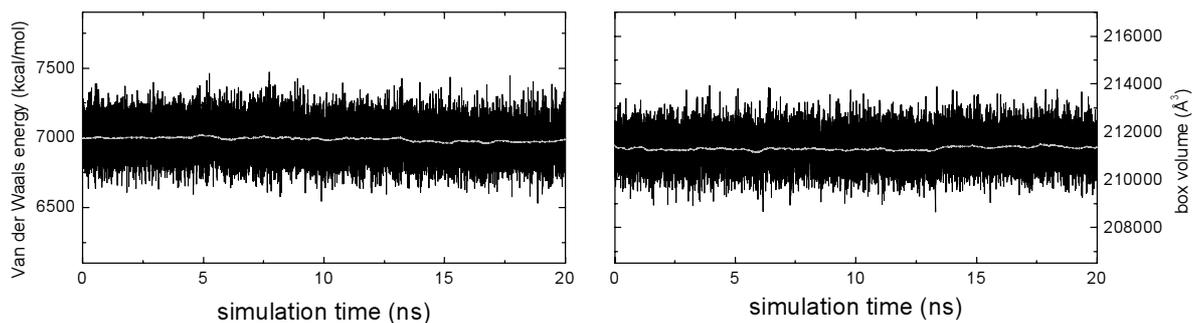


Fig. 4.6: Plots of selected physical properties of the system sampled during the 20 ns simulation. Left: Van-der-Waals energy, right: box volume. Black: sampled data, gray: running average.

4.2.1.3 Protein flexibility

To simplify descriptions, the telopeptide chains $\alpha 2(I)1$, $\alpha 1(I)1$ and $\alpha 1(I)2$ are termed A, B and C, respectively, in the following. Residues are referred to in capitalized three letter code, optionally with the chain letter prefixed and the residue number suffixed.

The first important subject to deal with is the structure of the telopeptide, as it was stated earlier that it is still a contentious issue. The flexibility of the protein has not been investigated to date. MD simulation is a very suitable approach here, since it can provide information about structure and dynamics in atomistic detail.

Three different values measuring the mobility of atoms were calculated for the whole trajectory: RMSDt0, RMSDm, RMSmob (see chapter 3.3.1 for details). The RMSmob was calculated at grid resolutions of 5.0, 2.0, 1.0 and 0.5 Å. For a histogram it is important that the best compromise is found between bin width and height. It was found that a resolution of 0.5 Å yielded optimal results. Results produced with this resolution are used for analysis.

At first, the RMSDt0 and RMSDm will be compared (fig. 4.7):

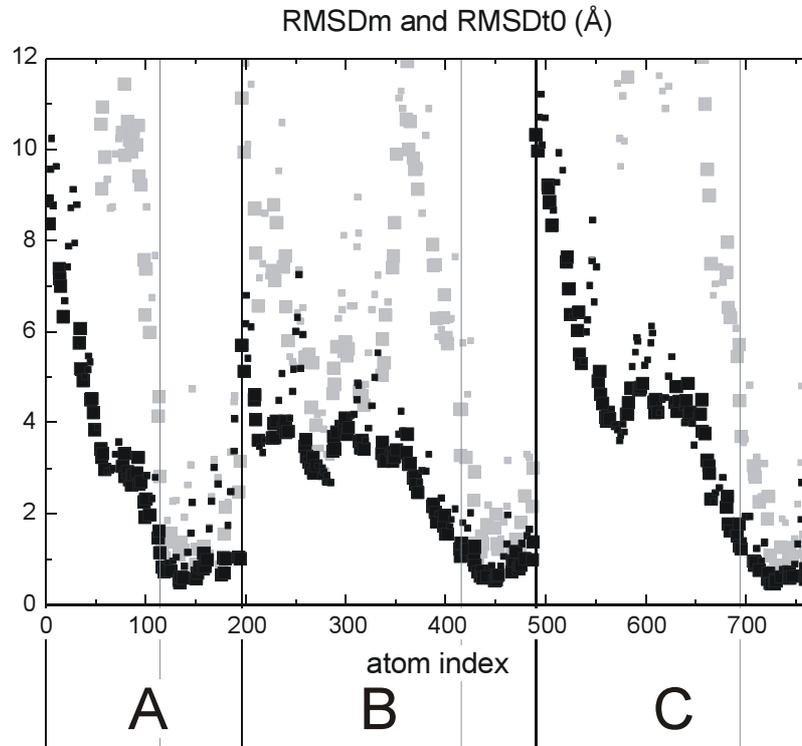


Fig. 4.7: Comparison of RMSDm (black) and RMSDt0 (gray). Scale fits RMSDm, RMSDt0 exceeds graph area. Protein chains separated by bold vertical lines, telopeptide and triple helix separated by thin lines. A, B and C chains are $\alpha 2(I)1$, $\alpha 1(I)1$ and $\alpha 1(I)2$, respectively. Backbone: large blocks, side chains: small blocks.

It can be clearly seen that the RMSD with respect to time step t_0 is always higher than the RMSD with respect to the mean position of an atom. The deviation is extreme for many atoms in the example. The RMSDt0 thus does not serve well as a measure for mobility and should not be used for this type of question. It can also be seen that the side chain atoms (small blocks) nearly always have higher values than the backbone atoms. This is easily understood as the side chains can move or rotate more or less freely while the backbone is a large entity in which atoms can only move conjointly.

The RMSDm and RMSmob are compared next (fig. 4.8). Both measures are shifted and normalized to fill the range from zero to one for better comparability. There are no units in this case.

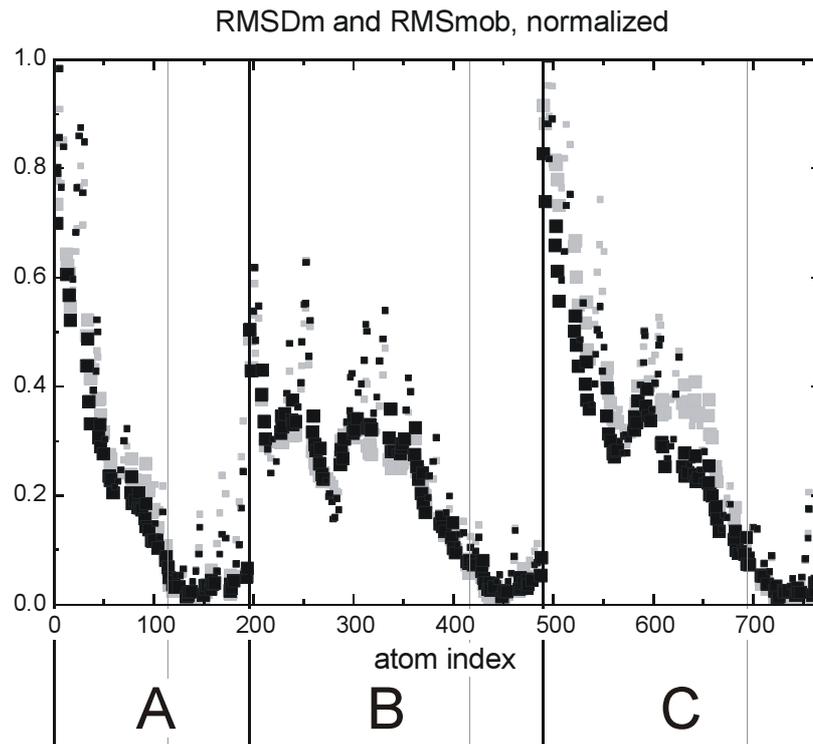


Fig. 4.8: Comparison of RMSDm (gray) and RMSmob (black). Protein chains separated by bold vertical lines, telopeptide and triple helix domains by thin lines. Backbone: large blocks, side chains: small blocks.

Judging from theory (chapter 3.3.1), RMSmob should be the most reliable and accurate measure for an atoms mobility. Nonetheless, both measures come out surprisingly similar. In the A and C chain, the RMSDm slightly overestimates the mobility. In the B chain, the curves are almost identical. Also in this plot, the higher mobility of the side chains can be recognized easily.

The plot shows homogeneous mobility characteristics for all three chains. The triple helical domains, separated by thin gray lines from the telopeptide domains, have a consistently low mobility. Side chain atoms rise above the backbone atoms in the plot, especially for longer side chains like Met. One Met in the B and C chains and two Met and one Leu in the A chain can be seen sticking out from the backbone in the plot.

The adjacent telopeptide domains are characterized by a small rising segment about five to seven amino acids long where the mobility rises roughly linearly. A region resembling somewhat of a plateau is reached, rather small in the A and C chains but pronounced in the B

chain. In the extensions of the telopeptides, especially in the A and C chains, the mobility again rises strongly to reach a maximum at the N-terminus.

An instructive visualization is obtained by mapping the calculated RMSmob values onto atoms as color. It is also possible to do this for the ribbon representation, which is less detailed but more clear. Both representations are shown in figure 4.9 (atoms) and 4.10 (ribbon). These are intended as intuitive graphic illustrations and are therefore printed without color bar legend here.

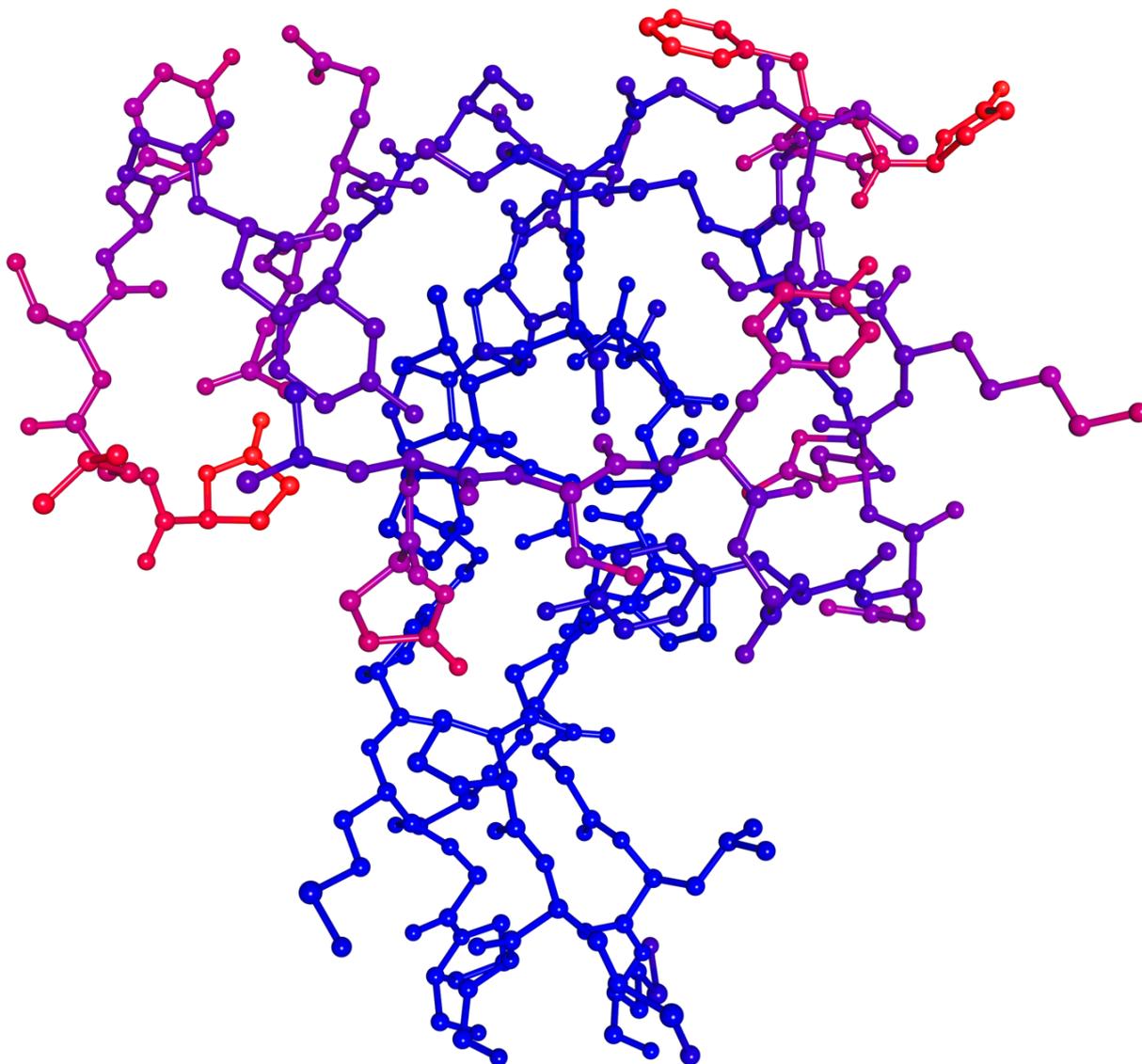


Fig. 4.9: RMSmob mapped on atoms as color. Red = high values, blue = low values.

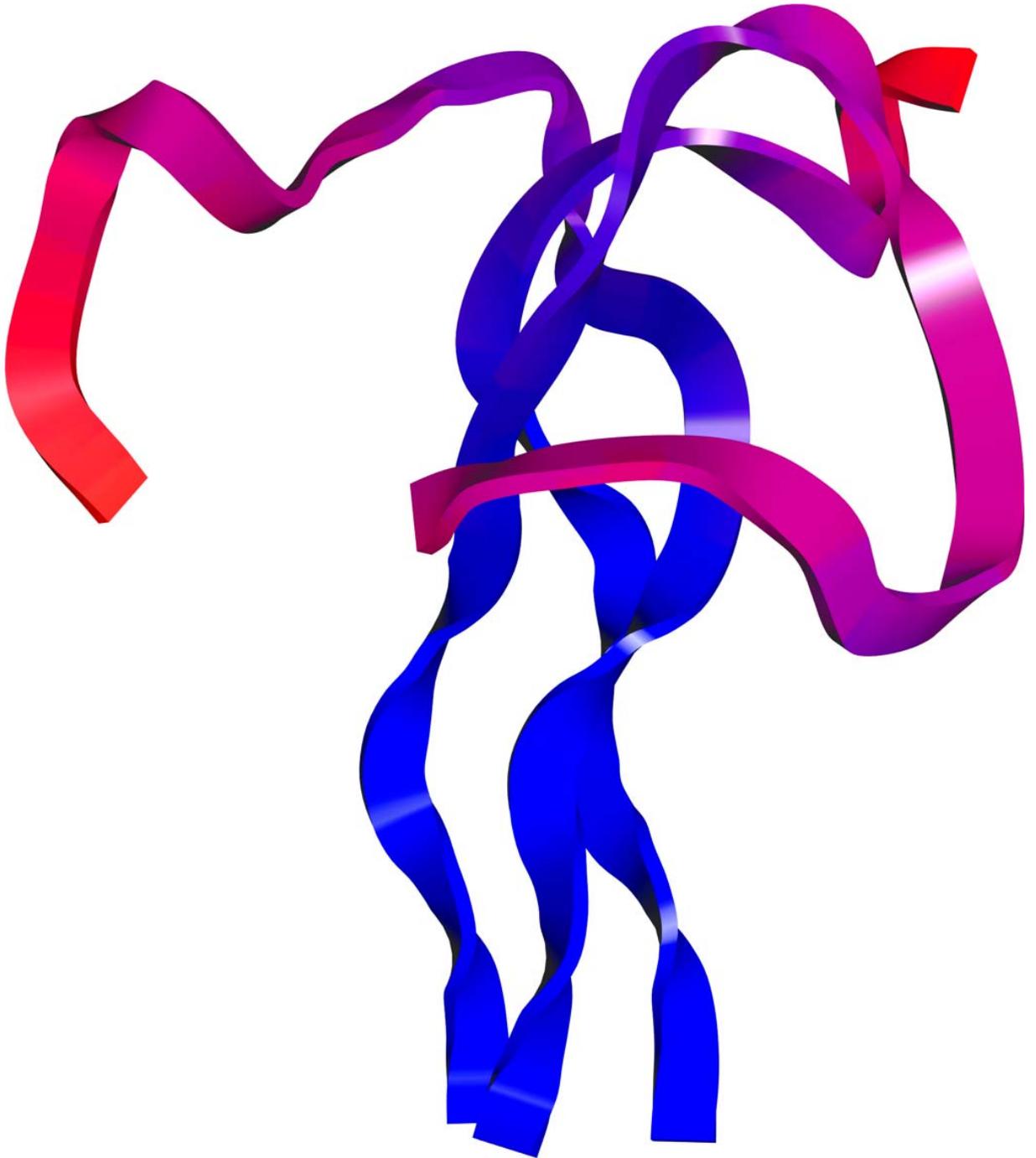


Fig. 4.10: RMSmob mapped on ribbon as color. Red = high values, blue = low values.

As a conclusion it can be stated that the RMSDt₀ should not be used to evaluate atom mobility. It always exceeds the true RMS fluctuation of an atoms position due to the reference to an arbitrary position t_0 and yields misleadingly high values. The most accurate measurement is given by the RMSmob. For the telopeptides, the RMSDm and RMSmob yield

almost identical results. Because the RMSDm is calculated with much less computational demand, the RMSDm is a good measure to assess atom mobility in this case.

4.2.1.4 Hydrogen bonds

The frequency of direct hydrogen bonds in the telopeptide chains during the simulation is shown in figure 4.11. Numerous water mediated hydrogen bonds were also observed, but those are not plotted.

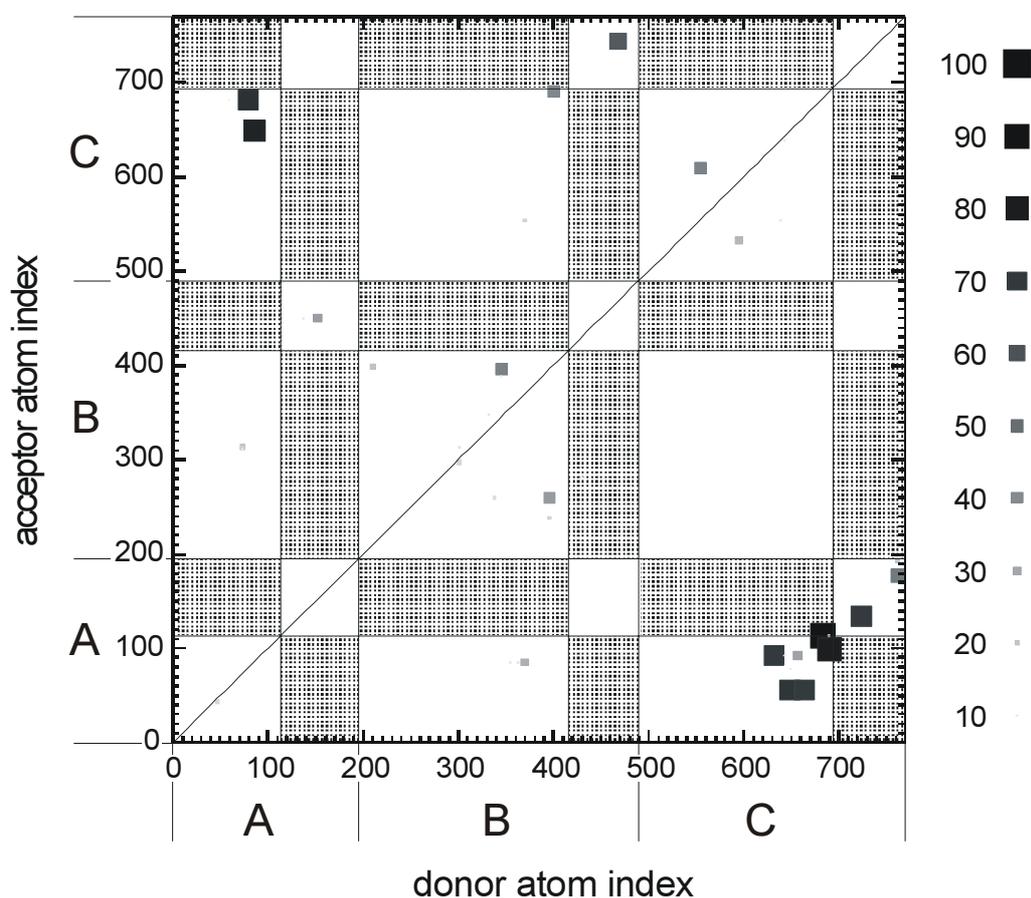


Fig. 4.11: Hydrogen bond frequency coded as square size and gray shading (white = 0 %, black = 100 % of simulation time, legend on the right). Chains marked A, B, and C. Bonds between triple helix and telopeptide residues in hatched areas.

The most multitudinous and stable hydrogen bonds are observed between chains A and C, both providing acceptor and donor residues. The typical triple helix hydrogen bonds can also be observed within the respective areas. Chain C and especially chain B also show some intra-chain hydrogen bonds which presumably stabilize the secondary structure of the individual chains. The most stable hydrogen bond observed between residues C_SER14 and A_PRO9 is intact for 85 % of the total 20 ns simulation time. This indicates a very stable structural feature in the telopeptide. No hydrogen bonds between the triple helix and telopeptide domains are observed. There are 33 hydrogen bonds which are intact for at least 10 % of the simulation time. These numbers were obtained with the relatively weak criterion parameters $r < 3.3 \text{ \AA}$ and $\Delta\beta < 40^\circ$ (see chapter 3.3.2.1). With more strict parameters of $r < 3.0 \text{ \AA}$ and $\Delta\beta < 20^\circ$ the exact same hydrogen bonds are detected, only with marginally less durations. It is remarkable that the six most stable hydrogen bonds are located within the telopeptide and not in the triple helix. The donor and acceptor atoms of the ten most stable hydrogen bonds are highlighted in figure 4.12.

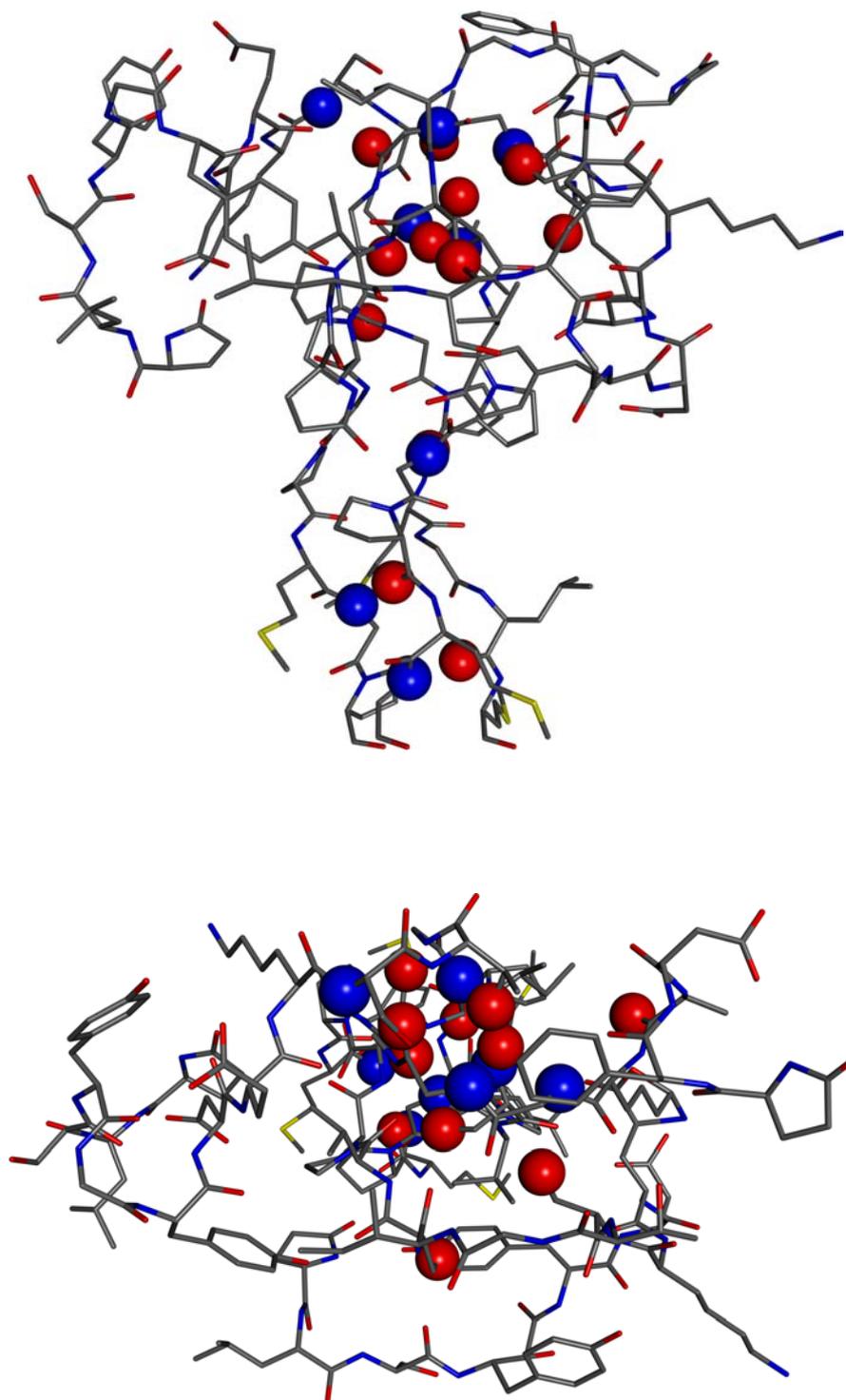


Fig. 4.12: Hydrogen bond donors (blue) and acceptors (red) of ten most stable hydrogen bonds in telopeptide. Protein in capped stick, donors and acceptors in CPK representation. Top: front view, bottom: top view.

Eight hydrogen bonds are active for 70 % or more of the simulation time. These can be regarded fairly stable. All hydrogen bonds are found in the triple helix and in the interior region of the telopeptides, where the individual chains are entangled. In the extensions of the

telo peptide chains, almost no hydrogen bonds are detected. The results are in perfect agreement with the flexibility analysis in chapter 4.2.1.3 where a high mobility was found for the extensions of the telopeptide, while the central region and the triple helix were found to be more rigid.

4.2.2 Collagen triple helix

4.2.2.1 Model preparation

The triple helical domain of human type I collagen has a secondary structure that is uniform throughout the three chains each 1014 amino acids long. A possible preference for mineral nucleation is thus very unlikely to depend on the secondary structure. From the primary structure analysis in chapter 4.1, no outstanding features in the sequence of collagen could be found that point to a special functionality for mineral nucleation. Charged residues are present throughout the complete molecule with a fairly random distribution. The high similarity of the $\alpha 1(I)$ and $\alpha 2(I)$ chain sequences also results in a very high identity of charged amino acid locations in all three chains. Charged residues often appear clustered to small groups of up to ten amino acids in close vicinity. Two charged residues are often observed in direct neighborhood, but not more than two. As a rare feature it is merely observed that two like charged residues in direct neighborhood are found only four times in the whole triple helix, plus once in the N-terminal telopeptide.

For simulation, a small section of approximately one period of the triple helix has to be chosen due to limitations of computational resources. As there is no other feature that can help single out a special portion of the triple helix, the choice is based on the charged residue pattern. A typical section, also including two like charged residues in direct neighborhood, is found at residues 217 to 246 of the human collagen $\alpha 1(I)$ chain and residues 129 to 158 of the $\alpha 2(I)$ chain, as found in the PIR database under codes CGHU1S and CGHU2S. Counting from the telopeptide N-termini (as in chapter 4.1) the residue numbers are 56 to 83 for $\alpha 1(I)$ and 50 to 77 for $\alpha 2(I)$. The difference in the residue numbers for the different chains is due to the different length of propeptides. The sequence in one letter code is given below:

```

      * .***.***** *:***.*****
CGHU1S_1 MGPRGPPGPPGKNGDDGEAGKPGKPGKRGK
CGHU1S_2 MGPRGPPGPPGKNGDDGEAGKPGKPGKRGK
CGHU2S_1 AGARGPAGPPGKAGEDGHPGKPGKPGKRGV
      1.....10.....20.....30

```

Fig. 4.13: 30 residue sequence section from human collagen used as one period triple helix model.

The triple helix contains 4-hydroxyproline (Hyp), which is a nonstandard residue and not contained in the charmm22 force field (figure 4.14).

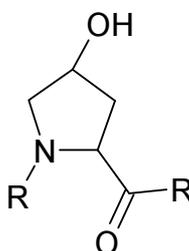


Fig. 4.14: Rare amino acid 4-hydroxyproline found in collagen

Klein et al. supplemented the Amber94 force field with the hydroxyproline residue by developing additional dihedral parameters and partial charges⁴⁶. The parameters were developed as an add-on to the existing proline parametrization. In the same manner, the extension was implemented to the Charmm22 force field. Chemically analogous atom types for the hydroxy group were assigned (Charmm types OH1 and H for Amber types 22 and 31, respectively), and required parameters were adopted from the Amber force field. The dihedral parameters and charges from Klein were implemented, dihedral parameters being taken from the paper, charges made available by Klein upon personal request.

A 30 residue long triple helical structure resembling one period of the triple helix was built using the designated application *TheBuScr* developed by Rainey et al⁴⁷ using the sequence described above. Side chains and hydrogen atoms were generated with Sybyl. The protein was put in a water box and ions added to resemble a 50 mmol/l NaCl solution. With the condition of electric neutrality and a net charge of +4 for the protein, three sodium and seven chloride ions were added to the box. The water was first equilibrated while holding the

protein fixed, then the whole system was allowed to relax. The box dimensions settled to approximately $117 \times 47 \times 47 \text{ \AA}^3$. A snapshot of the box is shown in figure 4.15.

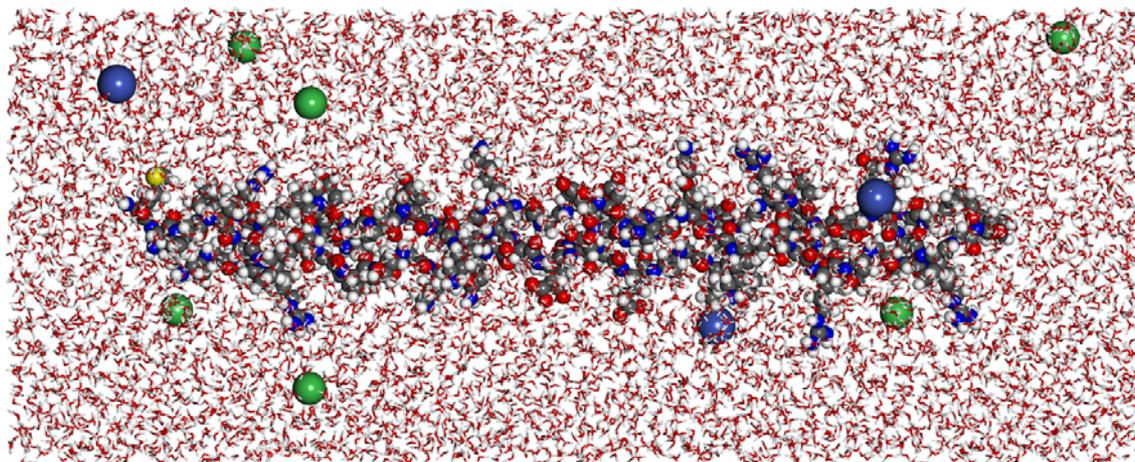


Fig. 4.15: Triple helix simulations system. Protein CPK, ions CPK, water sticks representation. Blue = sodium, green = chloride.

4.2.2.2 Long time simulations

The system was simulated using the same program and protocol as described for the telopeptide system long time simulation (chapter 4.2.1.2). Weak "confining" restraint potentials were applied to the ends of the triple helix with a threshold of 2.5 \AA and a low force constant to prevent unfolding of the molecule.

4.2.2.3 Protein flexibility

The RMSDt0, RMSDm and RMSmob were calculated for the complete trajectory of 20 ns. The RMSDt0 and RMSDm are compared in figure 4.16., RMSDm and RMSmob are plotted in figure 4.17.

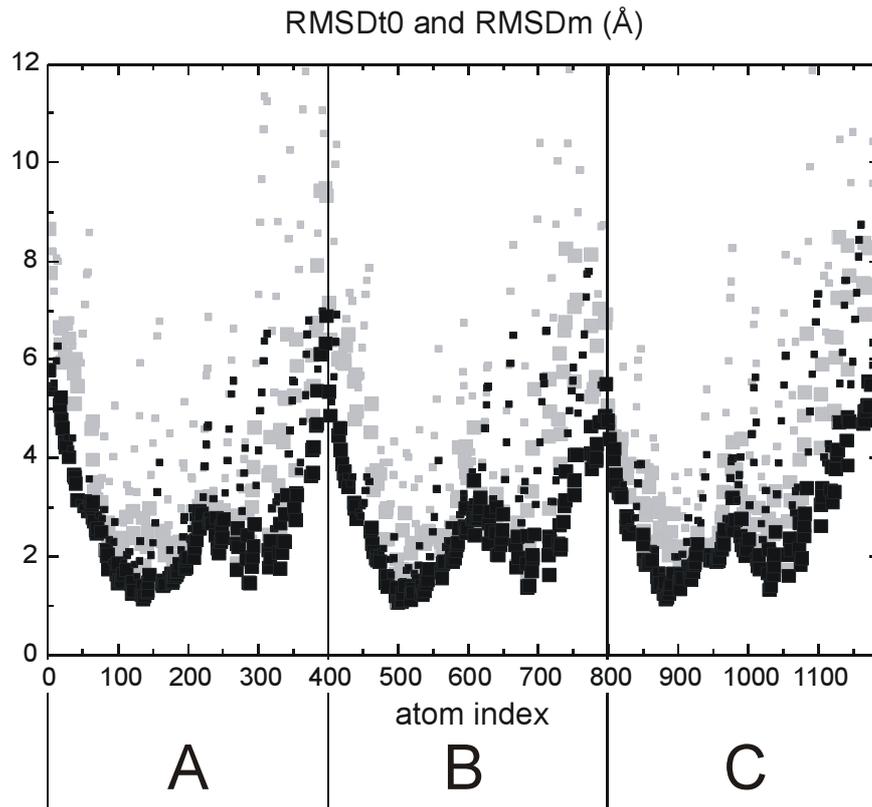


Fig. 4.16: RMSDt0 (gray) and RMSDm (black). Chains are separated by vertical lines. Small blocks: side chains, large blocks: backbone.

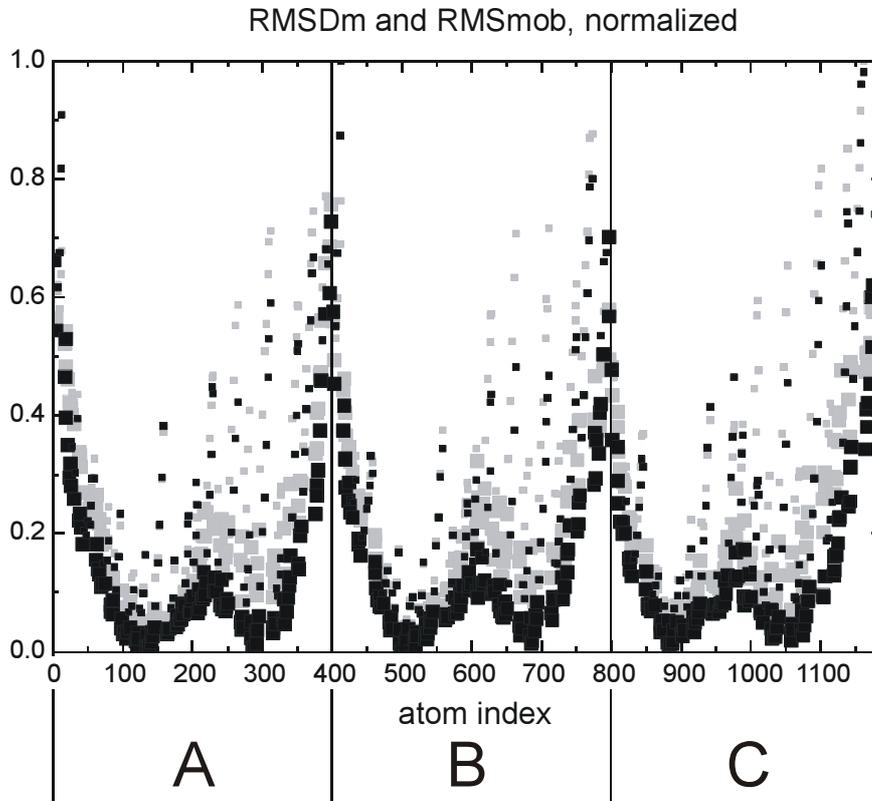


Fig. 4.17: RMSDm (gray) and RMSmob (black). Graph details see above figure.

The plot also visualizes the occurrence and length of side chains very well as the respective small squares are located higher than the backbone squares. It can be seen that at the N-terminus less and shorter side chains are located than in the rest of the molecule.

An instructive representation of the molecules mobility is obtained when plotting the chains superimposed, reflecting the actual arrangement of the chains. For increased clarity, only the RMSmob is plotted (fig. 4.18).

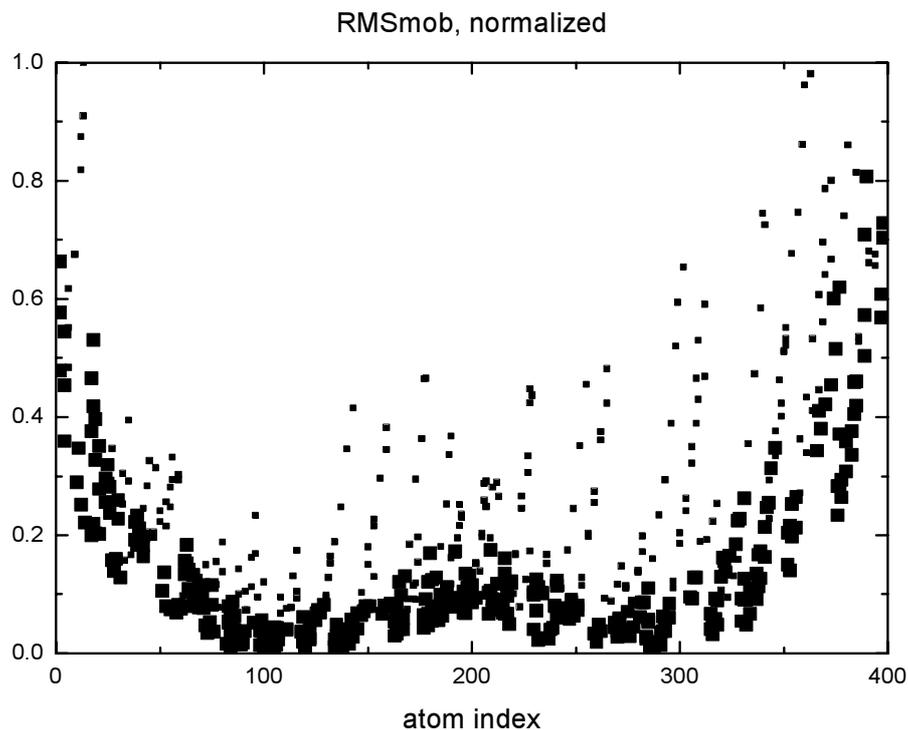


Fig. 4.18: Plot of RMSmob (normalized, without units) in actual chain registration.

The plot proves that all three chains show an identical mobility profile. A general trend is outstanding: the mobility is higher at the ends and in the middle of the molecule. This is exactly what was observed by experimentalists^{48,49}. Simulations with different molecules conducted in the initial preparing stage of this project, in part mutated sequences, showed the same behavior. Judging from the simulations conducted, the effect apparently does not depend on the exact sequence, charged residues or molecule length. It is, however, likely a finite chain effect that is probably not present in the real 1000 amino acid collagen triple helix. It is nevertheless certainly interesting for further in-depth research in the future.

The superior accuracy of the RMSmob in comparison to the RMSDm comes out clearly in this example. The RMSD type measures (see also fig. 4.16), show an asymmetry with higher values towards the C-termini (at higher atom numbers). The RMSmob in contrast gives a balanced picture of the mobility that is symmetric with respect to the molecule. The RMSmob represents a much more meaningful information compared to the other measures.

Comparison with telopeptide results

In figure 4.19, the results for the mobility of the triple helix and the telopeptide are compared.

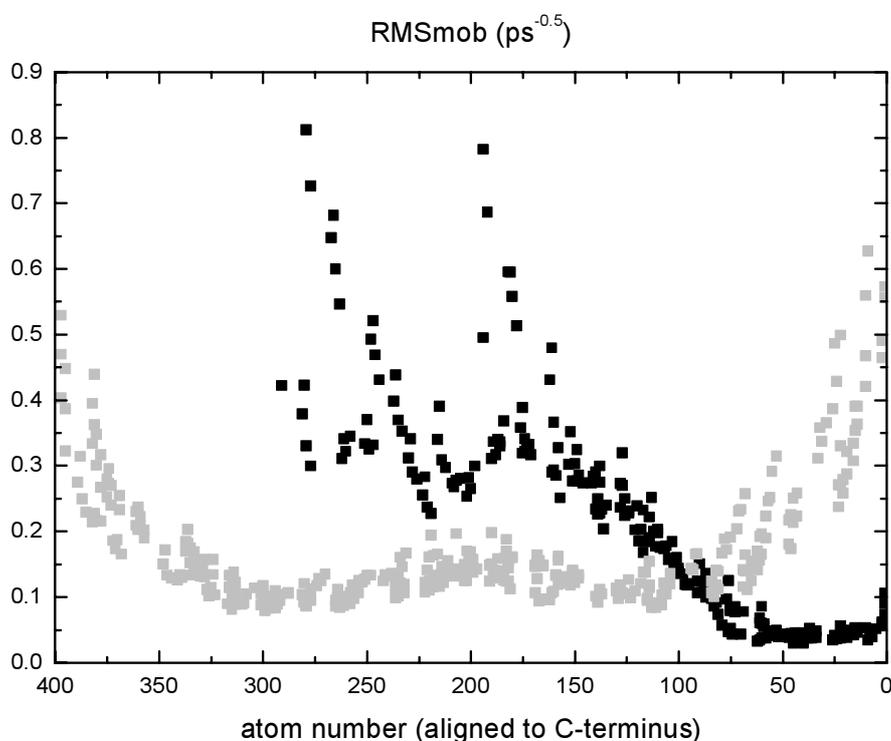


Fig. 4.19: Comparison of RMSmob for telopeptide (black) and triple helix (gray), all chains. Only backbone shown. Atom numbers aligned to C-termini. Units are ps^{-0.5}.

Apart from the ends of the triple helix, the mobility is distinctly higher in the telopeptide, especially in the extensions towards the N-termini. The lower mobility in the triple helical segment of the telopeptide is presumably dependent on the sequence. The dependence of the triple helix stability on sequence is also observed experimentally and was mentioned above. It has to be kept in mind that a short molecule is studied as a model for an extremely long

molecule. As seen in the mobility plot, the 15 central residues of the model molecule can be regarded to bear good resemblance with the real collagen triple helix.

4.2.2.4 Ramachandran analysis

The Ramachandran angle plot shows a concentration in the expected triple helix region. Also considerably high deviations from the ideal triple helix values is observed for some residues. This was reported by other groups as well^{33,48}. Below is a Ramachandran plot in which residues are plotted as circles (fig. 4.20). The size of the circles represents the mean fluctuation of the Φ and Ψ angles in the simulation.

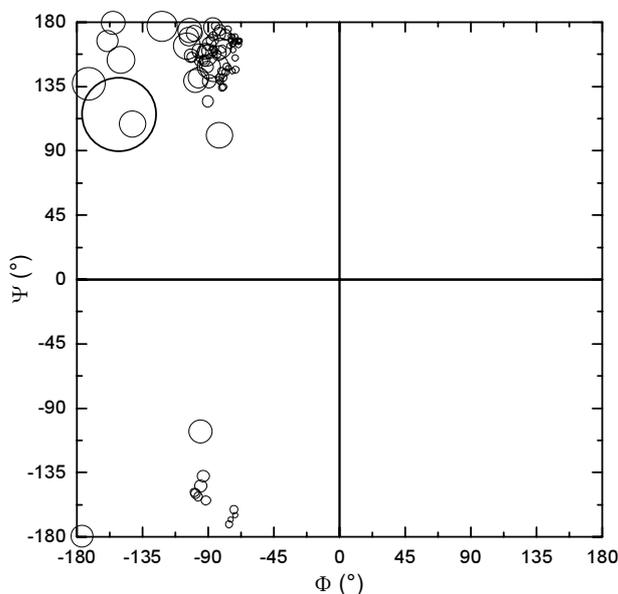


Fig. 4.20: Ramachandran plot of triple helix. Mean angle values obtained from 20 ns simulation are plotted with their mean standard deviation as circle size (symbolic scaling).

The by far largest circle belongs to the Gly residue at the very end of the C-terminus, which is basically free from interactions with the other chains and therefore has a low preference for any specific conformation.

The fluctuations of the backbone dihedral angles are plotted against the residue number in figure 4.21. The highest fluctuations are found at the ends and in the middle of the chain which agrees with the results found in the mobility analysis. This explains however, that the

4.2 Structural analyses

high mobility in the middle of the triple helix is not just due to a slight bending of the collagen molecule but due to a lack in secondary structure stability in the middle of the chain.

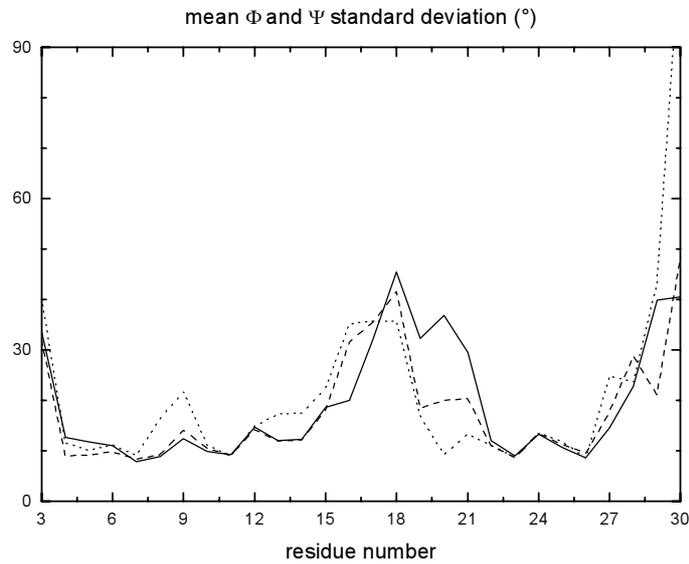


Fig. 4.21: Mean standard deviation of Φ and Ψ in triple helix during 20 ns of simulation. Chain A: solid, B: dashed, C: dotted.

In fact, visual inspection of the simulation trajectory showed that the triple helix frequently has the shape of two seemingly perfect triple helical segments connected by a kink. This is in agreement with experimental findings that the triple helix has regions of high and low stability, which is discussed in a recent review on collagen structure²⁷. A similar deformation of the triple helix was observed in simulations by Zahn et al. upon attachment of a phosphate ion⁵⁰.

4.2.2.5 Hydrogen bonds

The frequency of direct hydrogen bonds in the backbone of the triple helix during the simulation is plotted in figure 4.22. The regular hydrogen bonding pattern typical for the triple helix is clearly recognized in the residue vs. residue plot in the per chain numbered plot (fig. 4.22, right).

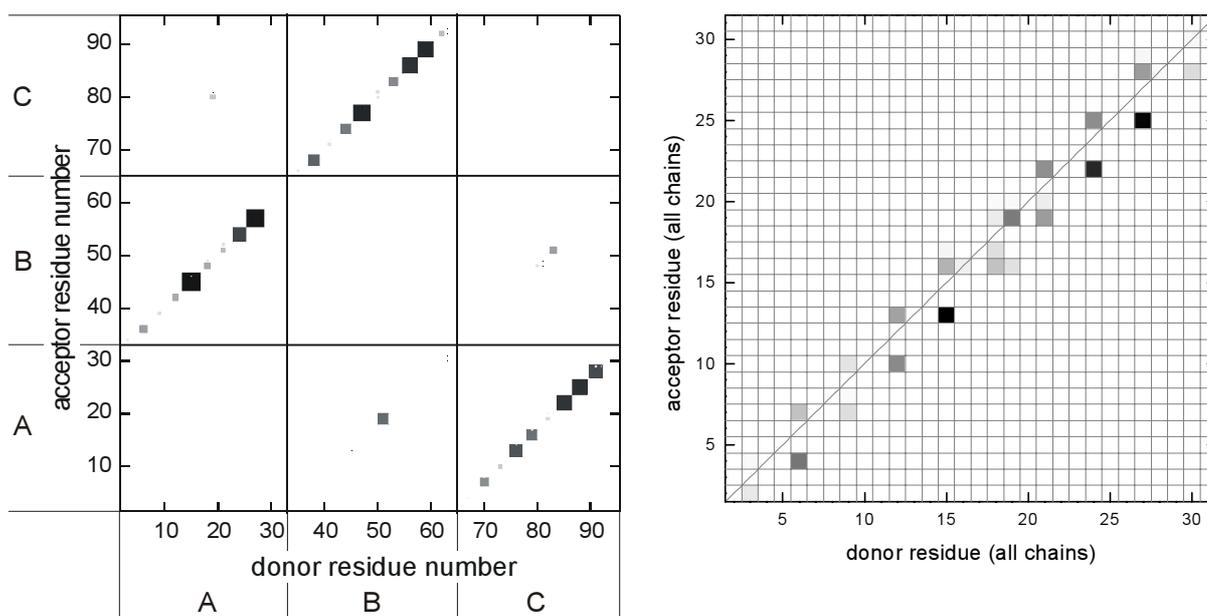


Fig. 4.22: Backbone hydrogen bonds counted on a per residue basis (left: continuous, right: per chain numbering). Frequency of bond indicated by gray shading (white = 0 %, black = 100 % of simulation time) and symbol size (left plot).

The well-known hydrogen bond pattern is $\text{Don}(n) \rightarrow \text{Acc}(n+1)$ and $\text{Don}(n) \rightarrow \text{Acc}(n-2)$ with $n = 3i+2$ ($i = 1, 2, \dots$). At the residues around 15 to 20, irregularities in the pattern are visible that agree with the results seen in the mobility analysis and in the Ramachandran analysis. This is additional evidence that the triple helical structure has a kink in this region.

In the left plot of figure 4.22, the occurrences of donor and acceptor residues in the individual chains is visible. There is a donor acceptor scheme $A \rightarrow B$, $B \rightarrow C$, $C \rightarrow A$ between the chains which is strictly followed. The only exception occurs in the kink region of the molecule.

More hydrogen bonds can be seen in the atom vs. atom plot of the complete molecule in figure 4.23. The donor-acceptor scheme between the chains is complemented in the opposite sense through the contribution of hydrogen bonds involving side chains.

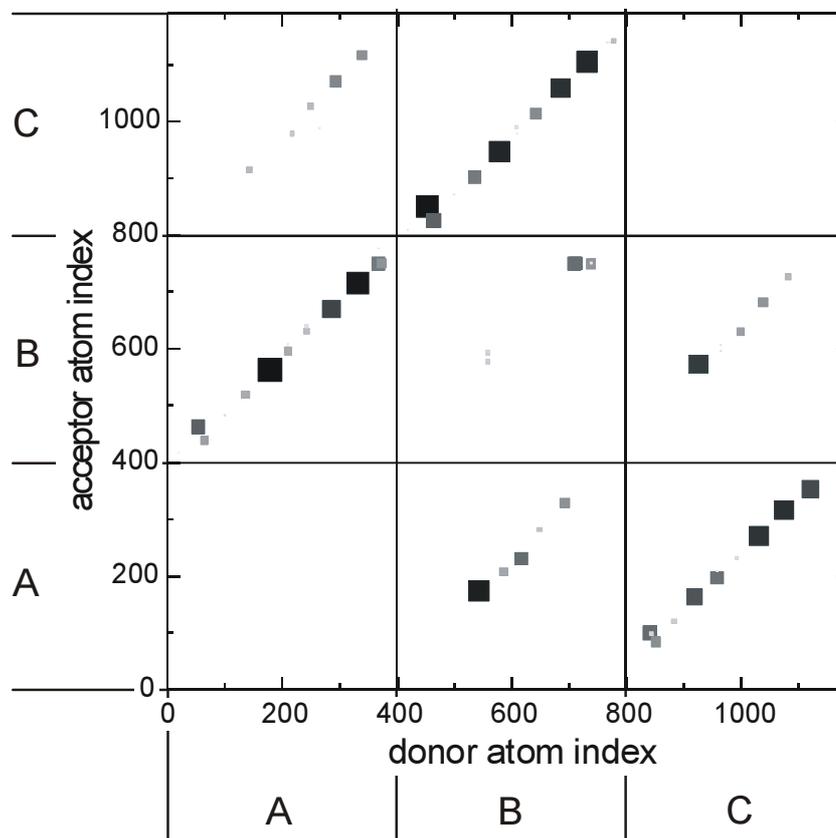


Fig. 4.23: Hydrogen bonds as in fig. 4.22, but for the complete molecule and counted on a per atom basis.

4.2.3 Discussion

The N-telopeptide structure shows two segments of different characteristics. The first segment is comprised by the nine residues adjacent to the triple helix. It is characterized by multiple inter-chain hydrogen bonding. The hydrogen bonds are not permanent but temporary, thus some flexibility of the structure is maintained. The structure is changing slowly over time, and no secondary structure motifs can be observed. The second segment is formed by the remaining N-terminal residues of the telopeptide. It is characterized by an outstanding

flexibility. No ordered or stable structure can be observed, the structure changes continuously. Also, no hydrogen bonds are observed. The first segment can be interpreted as a transition between the highly ordered triple helix and the completely unstructured and flexible N-termini.

The triple helix itself shows a stable secondary and tertiary structure. Local fluctuations from the ideal helix dihedral angles can be quite large without affecting the overall stability of the secondary and tertiary structure. A strong hydrogen bond pattern is observed as was described in literature^{51,52}. In the 30 amino acids long model protein, a kink in the middle could be observed, which was also observed by other groups in experiments and modeling studies^{27,50}.

4.3 Charged groups distances

It seems worth investigating whether measurable effects of the greater flexibility of the telopeptide can be detected which could affect ion binding. As a first approach the distances of charged amino acid side chains will be analyzed. Charged groups being significantly closer in the telopeptide or in the triple helix could give a first hint on the possibility of stable ion binding in chelate like conformations.

4.3.1 Telopeptide

The distances of the binding adsorption sites of the telopeptide (identification see chapter 4.4) that are conjectured to be most stable were monitored during the 20 ns simulation to inspect the arrangement of these groups and check on a possible ability to form chelate type complexes where an ion is coordinated by more than one charged group. The minimum distance between two groups observed during the whole simulation is plotted in figure 4.24 as gray shading. The telopeptide chains $\alpha 2(I)1$, $\alpha 1(I)1$ and $\alpha 1(I)2$ are termed A, B and C, respectively, in the following and will be printed with the residue name.

4.3 Charged groups distances

The lowest observed distance of groups of negative polarity is 2.7 Å for A_PYR1 and C_PYR1, for groups of positive polarity it is 3.0 Å for B_LYS9 and B_THR11. The lowest distance of like-charged groups with a negative net charge is 3.9 Å for B_ASP7 and B_GLU8, for a positive net charge it is 5.2 Å for A_LYS 5 and B_LYS 9. Several adsorption sites of like polarity come as close as about 5 Å, which means a separation by no more than one or two water molecules. An ion in the vicinity could possibly bind in a chelate type complex in such a situation. B_ASP7 and B_GLU8 are the only pair coming closer than 4.0 Å however, and no positively charged pair comes closer than 4.0 Å.

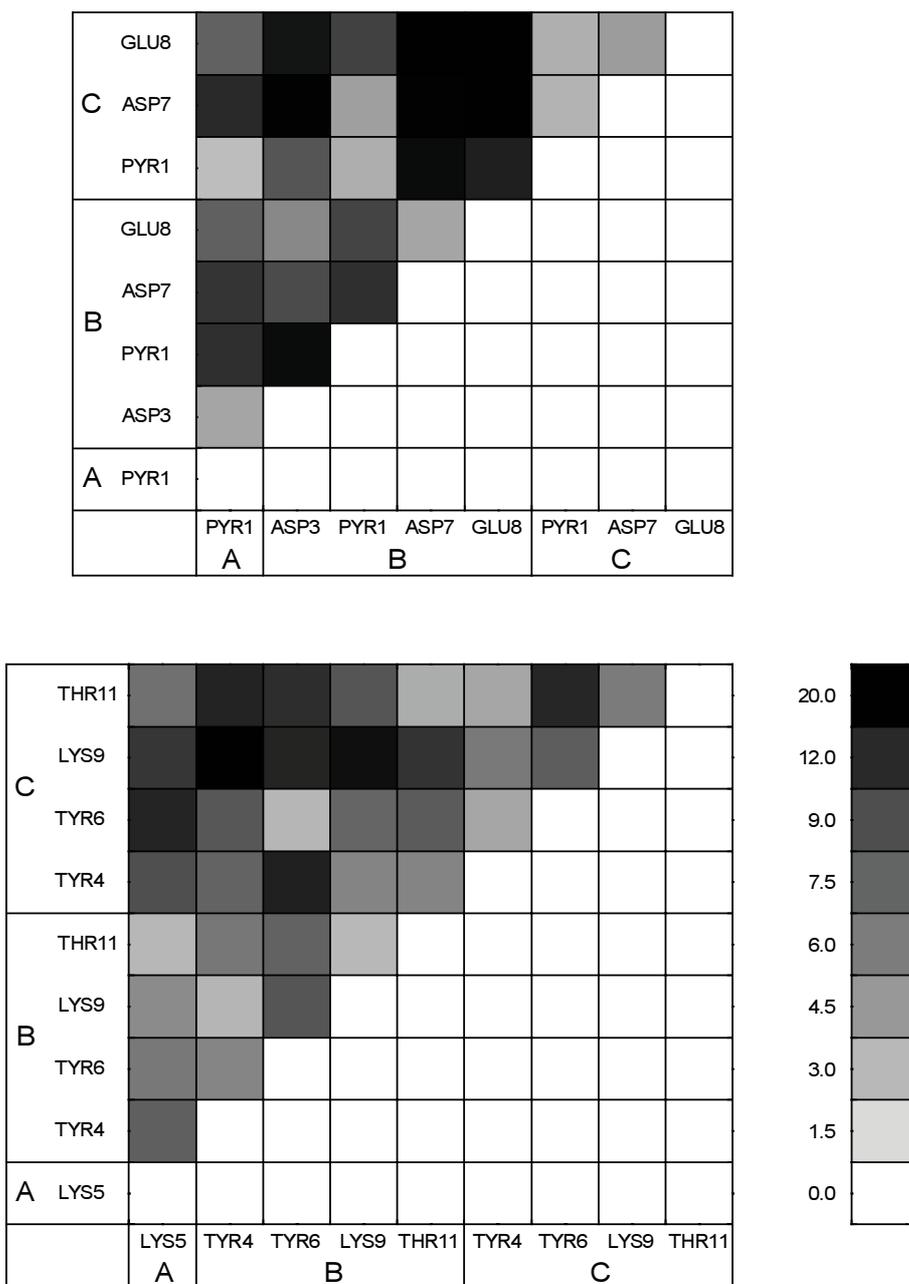


Fig. 4.24: Minimum distances (Å) of like-charged binding sites in 20 ns simulation. Shading scales with squared distance (see legend). Top: negatively charged sites, bottom: positively charged sites.

4.3.2 Triple helix

The smallest distances of possible adsorption sites of like polarity in the triple helix observed during the 20 ns simulation are shown in figure 4.25.

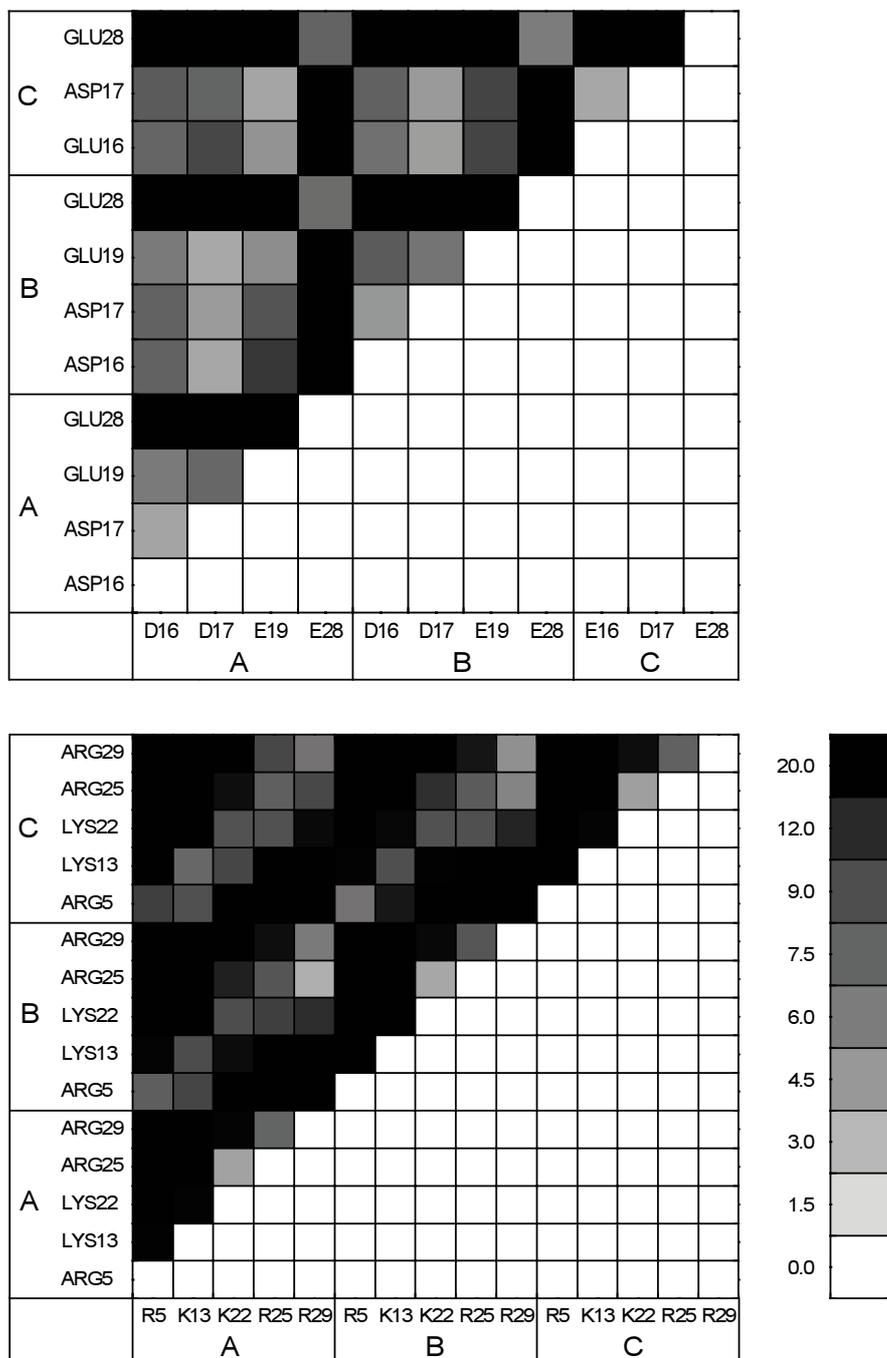


Fig. 4.25: Smallest distances of possible binding sites of like polarity in the triple helix observed during 20 ns simulation. Distance in Å is indicated by gray shading (see legend) for sites of negative (top) and positive polarity (bottom).

The nearest occurring distance of two positively charged residues is 3.5 Å for A_ARG29 and B_ARG25, for negatively charged residues it is 3.8 Å for A_ASP17 and B_GLU19. There are three positive and five negative amino acid pairs that reach a distance of less than 4 Å during the simulation. The results are comparable to those obtained for the telopeptide, in fact, in the triple helix even more close contacts of charged residues occur than in the telopeptide. It is concluded that, judging by the distances of charged groups, an equal ability to form chelate type complexes with ions must be expected for the telopeptide and the triple helix, if not a higher one for the triple helix.

4.3.3 Discussion

It was shown in the previous chapters that the telopeptide has a much greater flexibility and variability in its structure compared to the triple helix. The telopeptide thus in principle has the possibility to adapt its structure to optimally bind ions or even ion clusters. This is an important difference to the triple helix which is confined to a secondary and tertiary well defined structure. The analysis of the distances of charged groups in both molecules however did not reveal any significant differences.

The special structural features of the telopeptide found so far suggest more in-depth analyses concerning the ion binding. The focus of further studies is therefore concentrated on the telopeptide. The fact that in the literature much less is known about the properties of the telopeptide supports this decision. In the following chapters, the interactions of the telopeptide and ions will be studied in detail.

4.4 Ion attachment simulations

In addition to the simulation of the free telopeptide, simulations were setup to study the immediate reaction of the protein upon manual attachment of ions to possible binding sites. Ions were placed in close contact to the protein, and short relaxation simulations of fractions

of a nanosecond were performed to observe the immediate response of the protein and to identify possible rearrangements.

Possible binding sites were identified by analyzing the electrostatic potential (ESP) on the molecular surface of an equilibrated telopeptide structure. The structure snapshot is arbitrary since the structure of the telopeptide was changing throughout the long time simulation, but it is a first approach to look at the proteins behavior. The selected snapshot was taken after a short equilibration period of roughly a nanosecond that was still close to the minimum structure modeled by Scheraga.

The molecular surface (Connolly solvent accessible surface) was calculated with the program Fumee written by Keil⁵³. The partial charges for all atom were extracted from the charmm22 force field files used in the simulations. A molecular surface does not directly reflect a physical property, it is rather a concept that helps answer many chemically related questions. The surface is calculated based on a hard sphere model for the atoms, with average radii taken from many different compounds. The radius of the utilized probe sphere is an approximation the like. To generate the protein surface accessible to different ions, it is reasonable to set the probe radius to the radius of the respective ion. The radii of calcium and fluoride ions were approximated by the distance where the atom types Lennard-Jones (12-6) potential of the force field has its minimum. In ϵ, σ -form, the minimum is at $\frac{\sqrt{6}}{2} \sigma$. The calculated potential minima are at 2.7 and 3.1 Å for calcium and fluorine, respectively. For the hydrogen phosphate ion, the P-O bond distance plus the radius of the oxygen atom was chosen which was calculated to be 4.8 Å. Snapshots of the calculated ESP mapped on the molecular surfaces are shown in figure 4.26. The large radius of the probe sphere representing the hydrogen phosphate ion causes the sharp edges in the respective surface.

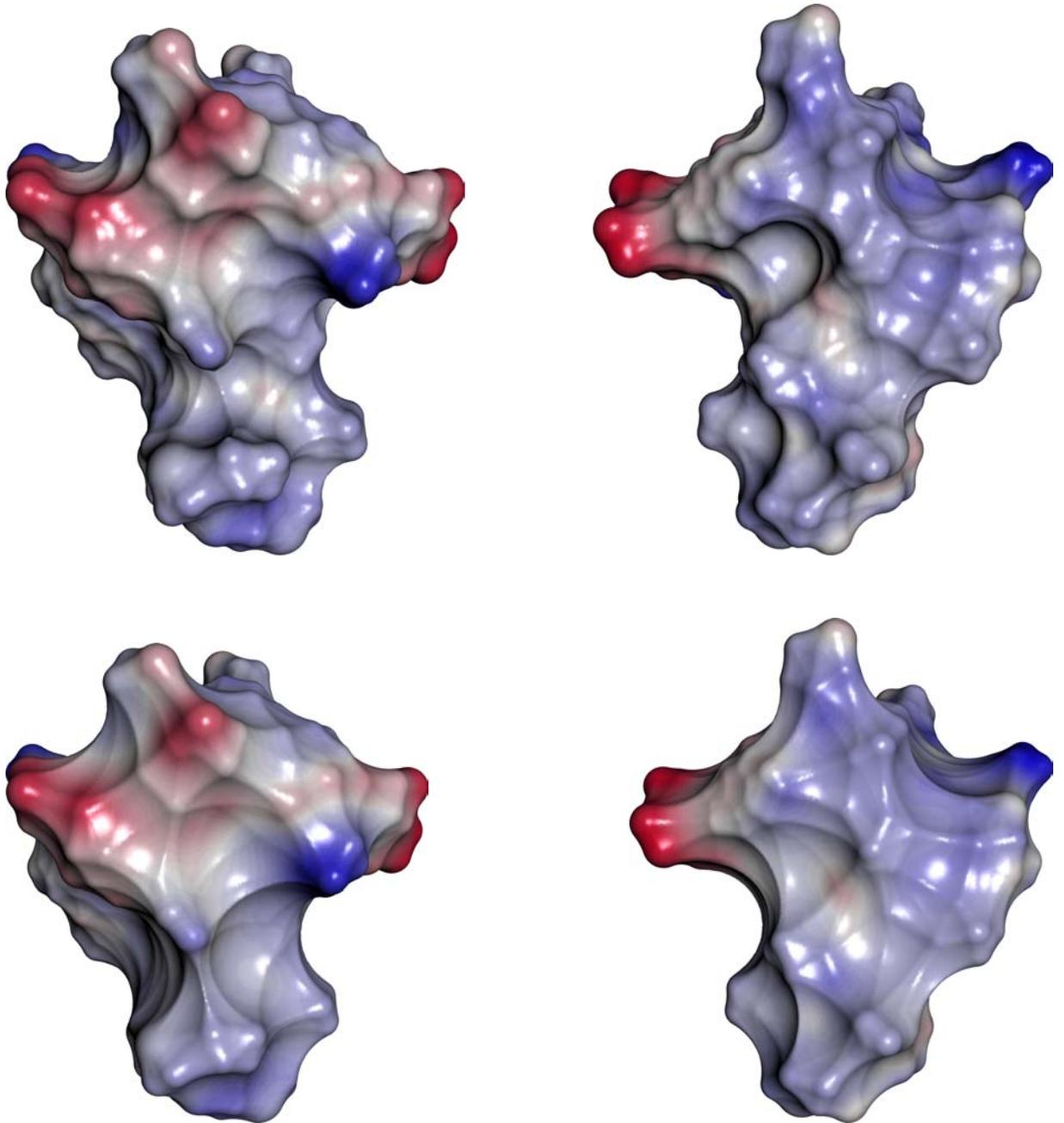


Fig. 4.26: Electrostatic potential mapped on the Connolly molecular surface of the telopeptide. Color is blue for minimum and red for maximum ESP for a negative probe charge (inverse for a positive probe charge). Top row: surface for calcium, bottom row: surface for hydrogen phosphate.

All charged sites in the protein appear exposed for both small and large ions. The surface for a large ion is smoother and partially charged spots visible on the surface for smaller ions are hidden in some cases. All sites with considerable negative ESP for calcium, fluoride or hydrogen phosphate were identified as possible binding sites. A list of the sites found is given in table 4.4. The chain nomenclature $A = \alpha 2(I)$; $B = \alpha 1(I), 1$; $C = \alpha 1(I), 2$ will be used from

4.4 Ion attachment simulations

here. All components of a simulation system will be referred to in capitalized letters from here, e.g. A_PYR1 (residue name and number, chain id prefixed), HPO4 (hydrogen phosphate ion), WAT (water).

residue	atom	ions
charged sites		
A_ASP3	CG44	CAL
A_LYS5	NZ75	HPO4, F
B_ASP7	CG297	CAL
B_GLU8	CD312	CAL
B_LYS9	NZ333	HPO4, F
C_ASP7	CG590	CAL
C_GLU8	CD605	CAL
C_LYS9	NZ626	HPO4, F
partially charged sites		
A_PYR1	CO13	CAL
B_GLY5	CO266	CAL
B_ASP7	CO296	CAL
C_TYR6	CO568	CAL
C_THR11	CO648	CAL
A_LYS5	HN58	F
B_TYR4	HH259	F
B_GLY12	HN363	F
C_PYR1	HT497	HPO4, F

Table 4.4: Possible adsorption sites in telopeptide identified by ESP. Chains: A = $\alpha 2(I)$; B = $\alpha 1(I), 1$; C = $\alpha 1(I), 2$.

A simulation system was setup consisting of the protein in a 54 x 54 x 54 Å water box. To simulate a very dilute apatite solution and simultaneously achieve electrical neutrality, two fluoride, four calcium and two hydrogen phosphate ions were added to the box (yielding approximately 10 mmol/l of $[\text{Ca}_4(\text{HPO}_4)_2\text{F}_2]^{2+}$).

The ions listed above were manually placed at the selected site, and a special relaxation protocol was applied. This included constrained relaxation for a few picoseconds where the ion was kept at the binding site but protein and water were allowed to equilibrate, and a subsequent free relaxation for a few picoseconds to let the binding distance between protein and ion equilibrate. The system was then simulated for 25 ps to record rdf functions, and subsequently for 100 picoseconds to monitor the protein relaxation. If significant conformational changes were observed, the simulation was continued for up to a nanosecond.

If the ion remained bound but no significant rearrangements in the protein could be observed, or the ion detached within this time, the simulation was not continued further.

4.4.1 Analysis

Stability of ion binding

The simulations showed that the attachment of a single ion to a suspected adsorption site in did not cause any significant reactions of the protein. Rather small local relaxations were observed which optimized the interactions between ions and polarities or charges in the protein. A list of all calculations performed is given in table 4.5.

#	res.	atom	ion	result
positively or negatively charged sites				
1)	A_ASP3	CG_44	CAL	ion remains bound
2)	A_LYS5	NZ_75	HPO4	ion remains bound
3)	A_LYS5	NZ_75	F	ion remains bound
4)	B_ASP7	CG_297	CAL	ion remains bound
5)	B_GLU8	CD_312	CAL	ion remains bound
6)	B_LYS9	NZ_333	HPO4	ion remains bound
7)	B_LYS9	NZ_333	F	ion remains bound
8)	C_ASP7	CG_590	CAL	ion remains bound
9)	C_GLU8	CD_605	CAL	ion remains bound
10)	C_LYS9	NZ_626	HPO4	ion remains bound
11)	C_LYS9	NZ_626	F	ion detaches after 65 ps
partially charged sites				
12)	A_PYR1	OC_13	CAL	see text
13)	B_GLY5	OC_266	CAL	see text
14)	B_ASP7	OC_296	CAL	see text
15)	C_TYR6	OC_568	CAL	see text
16)	C_THR11	OC_648	CAL	ion remains bound
17)	A_LYS5	HN_58	F	see text
18)	B_TYR4	HH_259	F	ion remains bound
19)	B_GLY12	HN_363	F	ion detaches after 62 ps
20)	C_PYR1	HT_497	HPO4	ion detaches after 12 ps
21)	C_PYR1	HT_497	F	ion detaches after 16 ps

Table 4.5: List of performed ion attachment simulations and result of the first 100 ps. Several calculations (marked "see text") are described in detail in the text.

4.4 Ion attachment simulations

Though no protein rearrangements were observed, yet from a number of examined cases several interesting findings about ion binding behavior can be learned. A short description of any noteworthy incidents or circumstances is given here. In simulation 2, A_LYS5 is in close vicinity to B_LYS9 which would allow chelate type complex, but B_LYS9 is strongly interacting with B_GLU8 and thus not available. In simulation 12, the CAL ion detaches from its initial place at A_PYR1 after 83 ps but is simultaneously approaching A_ASP3 (fig. 4.27).

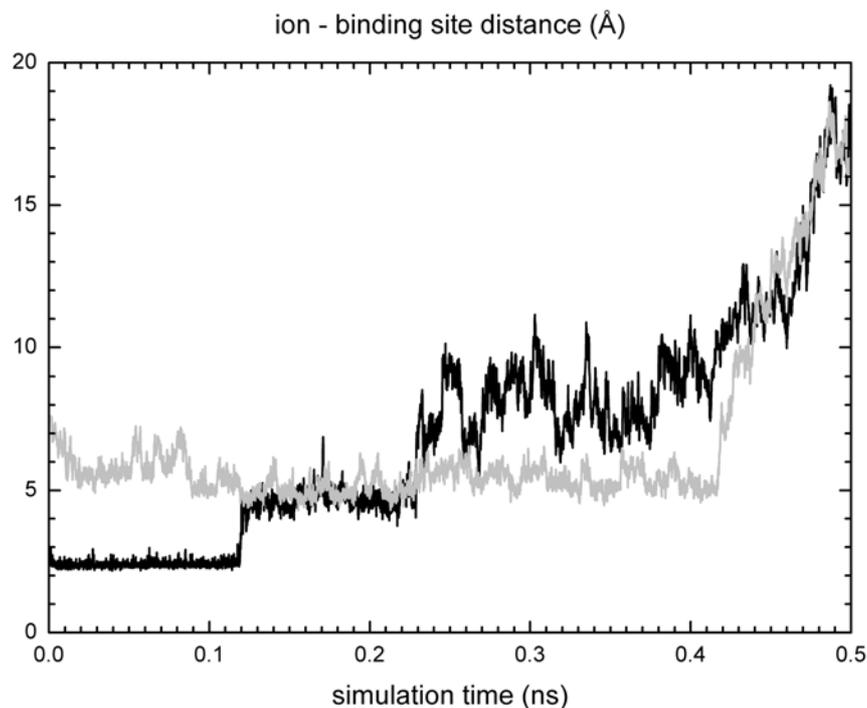


Fig. 4.27: Distance of calcium ion to carbonyl group of A_PYR1 (black) and carboxyl group of A_ASP3 (gray).

It seems to be "transferred" from one binding site to the next. For about a tenth of a nanosecond, the ion is coordinated by both the carbonyl and the carboxyl groups, each as a solvent separated ion pair. After about 0.4 ns the ion completely detaches from the protein. In simulation 13, the calcium ion is initially near to three carbonyl groups but detaches from all after 44 ps. In simulation 14, CAL detaches from its initial binding site B_ASP7 after 45 ps but approaches B_GLU8 simultaneously. It forms a solvent separated ion pair with the carboxyl group and carbonyl group existing temporarily during 1.1 ns, then CAL detaches completely (fig 4.28). The process looks like the ion is "handed over" from one binding site to the next. This kind of process could possibly help ions overcome high barriers. If a binding

4.4 Ion attachment simulations

site with a high barrier is located near other sites with smaller barriers, these sites could keep the ion in the area so that it diffuses to the high barrier site more often.

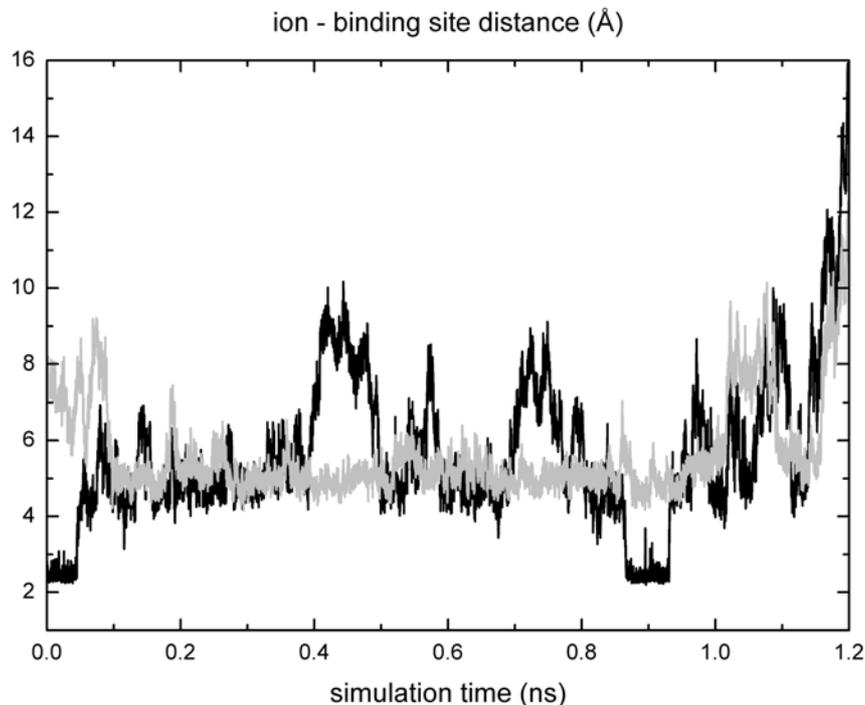


Fig. 4.28: Distances of calcium ion to B_ASP7 backbone carbonyl group (black) and B_GLU8 carboxyl group (gray). CAL forms a solvent separated pair with both groups approx. 5 Å distance each temporarily.

In simulation 15, CAL is initially coordinated by the backbone carbonyl groups of C_TYR6, C_ASP7 and C_GLU8. C_TYR6 leaves the arrangement after 55 ps, CAL stays bound to the remaining two sites. In simulation 17, fluoride is coordinated by the backbone amino groups of A_ASP3, A_ALA4 and A_LYS5. The ion moves between these sites and is mostly bound to one or two of them. After 50 ps it detaches from all three, but attaches to the charged amino group of A_LYS5 immediately. After 40 ps it also detaches from this group.

Three times, a process was observed where an ion was bound to partially charged groups with low barriers, detached from those to immediately attach to a charged group with a high barrier. This appeared like a "handing over" of the ion from one binding site to another. It seems to be a process that is occurring frequently with ions bound to partially charged sites. This could be a mechanism to bypass the high barrier for direct association to a charged group in solution. This is only conjecture at the moment, the necessary detailed mechanistic studies are not part of this work but can be performed in the future.

4.4 Ion attachment simulations

The simulations were also used to determine several mean binding distances of stable ion - binding site complexes.

binding site	ion	mean bond distance (Å)
<u>N</u> H ₃ ⁺	H <u>P</u> O ₄ ²⁻	3.62 ± 0.15
<u>N</u> H ₃ ⁺	F ⁻	2.67 ± 0.09
<u>C</u> OO ⁻	<u>C</u> AL ²⁺	2.73 ± 0.06

Table 4.6: Mean bond distances of ion - binding site complexes determined from 100 ps simulations. Underlined atoms used for distance measurement.

The measured binding distance for calcium to a carboxyl group agree well with distances found in the pdb database (1sra, 2cln, 1vfp, 1sn4). These are higher in some cases because calcium is coordinated by more than one carboxyl group, but distances in 2cln agree especially well. No compounds with a fluoride ion bound to an amino group were found in the pdb database. All complexes of fluoride involved iron or magnesium ions or multiple nitrogen coordination. Complexes of hydrogen phosphate and an amino group (1cnq, 1bup, 1fj6, 1ce8) were found with distances in very good agreement with the one found here.

Radial distribution functions

The *g*-functions and the integrated *G*-function were calculated for all atom pairs for every simulation. The coordination sphere of an ion can be investigated by analyzing the *G*-function. This serves also as a verification of the validity of the apatite force field which was developed for solid state crystals⁵⁴ and so far only used for crystals in water⁵⁵ but not single ions in water. In the following, the ion located at the binding site and those ions in solution will be distinguished as FB and F, CALB and CAL, PB and P, respectively. Oxygen atoms of water are termed OT.

Binding of calcium

The *g*- and *G*-functions of the solvated calcium ion show a distinct first coordination shell of eight water molecules. This is in good agreement with values found in the literature. The number of eight water molecules has been determined consistently with X-ray methods as well as ab initio (DFT) and combined QM/MM (HF, MD) methods^{56,57}.

The complexes of calcium bound to a carboxyl group show a straight replacement of two water molecules by the carboxyl group so that six water molecules remain in the solvation sphere. The bonding to partially charged carbonyl groups expels one water molecule from the shell. In two simulations CALB was coordinated by two (14) and three (15) carbonyl groups. This results in an ejection of two and three water molecules from the shell, respectively (see figures below).

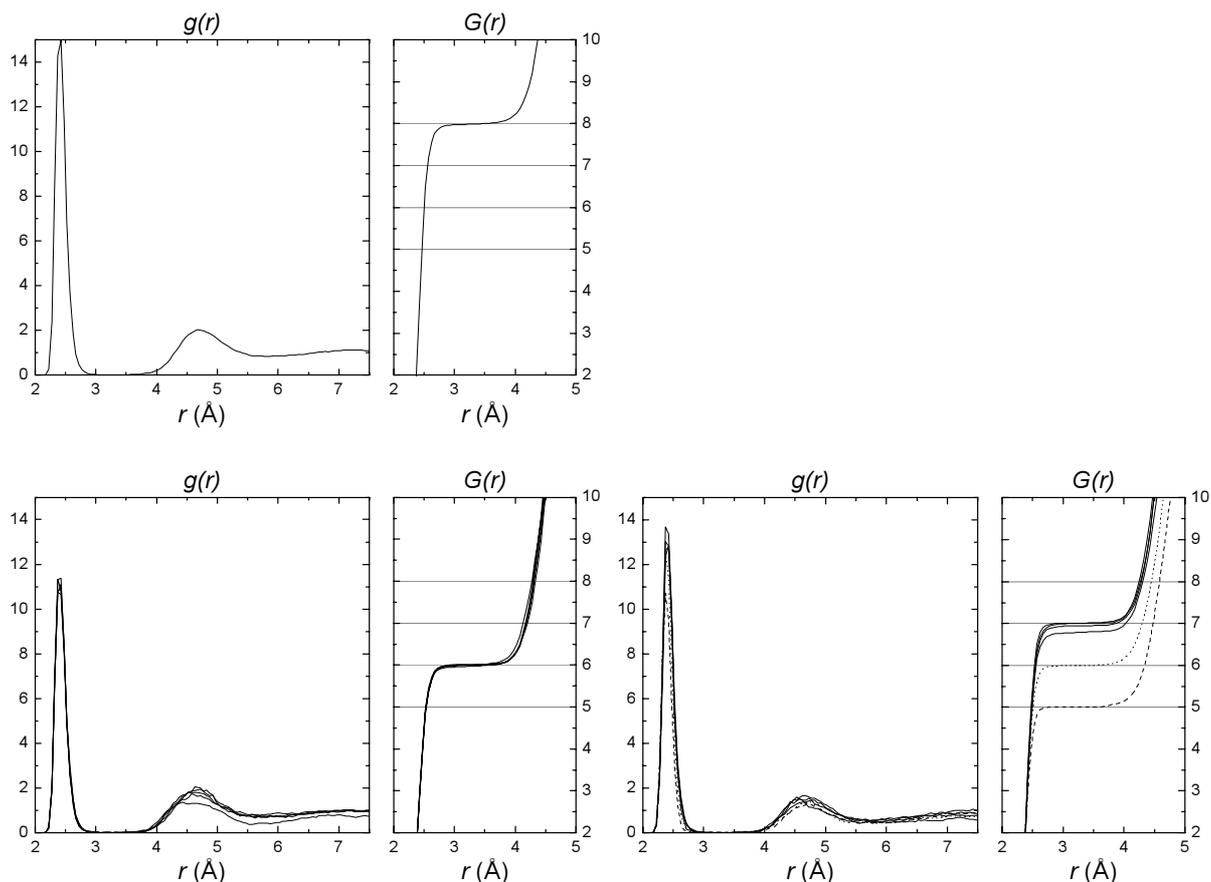


Fig. 4.29: g - and G -functions. Top: CAL / OT. Bottom: CALB / OT. Left: charged sites (1, 4, 5, 8, 9). Right: partially charged sites (12, 13, 16 solid; 14 dotted; 15 dashed). See table 3.5 for simulation numbers. Unit of G is number of particles.

The coordination number of fluoride in water has been determined with different results by various groups and is still controversial. Values of four to six have been published, also decimal numbers have been suggested^{58,59}. In the simulations conducted here, a coordination number of 6.7 was found (see figures below). The hydration shell is less sharply defined as that of the calcium ion, but it is still clearly pronounced.

4.4 Ion attachment simulations

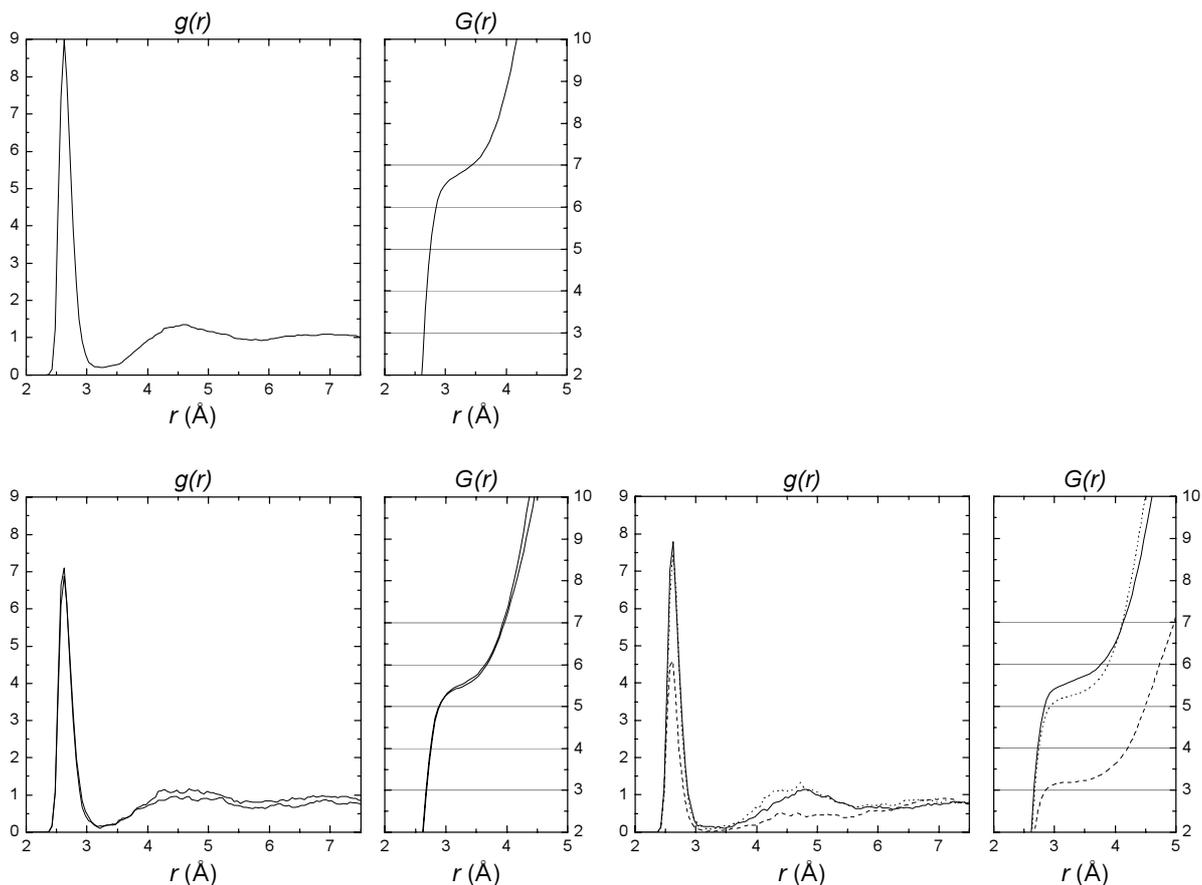


Fig. 4.30: g - and G -functions. Top: F / OT. Bottom FB / OT. Left: charged sites (3, 7, 11). Right: partially charged sites (17 dashed, 18 dotted, 19 solid). Unit of G is number of particles.

The charged amino group of lysine replaces one water in the hydration shell of fluoride. A backbone amino groups also replaces one water molecule in the shell (solid line). At the phenyl hydroxy group of tyrosine, the water content of the shell is a little less (dotted line). In simulation 17, the fluoride ion is actually coordinated by three backbone amino groups. This results in a reduction of the water shell to three (dashed line).

A coordination number of hydrogen phosphate could not be found in the literature. The hydration sphere is much less distinctly defined than for the other ions, the first and the second shell show some overlap. In our simulations, 14.4 water molecules are found within 4.3 Å of the hydrogen phosphate ion which is regarded as the first coordination shell.

The phosphate ion hydration shell was found very similar regardless of the nature of the binding site. One curve has a small bump above 4 Å which is attributed to a slightly different

geometry of the lysine side chain. The hydration shell of a bound hydrogen phosphate ion contains about one molecule less than the free ion (see figures below).

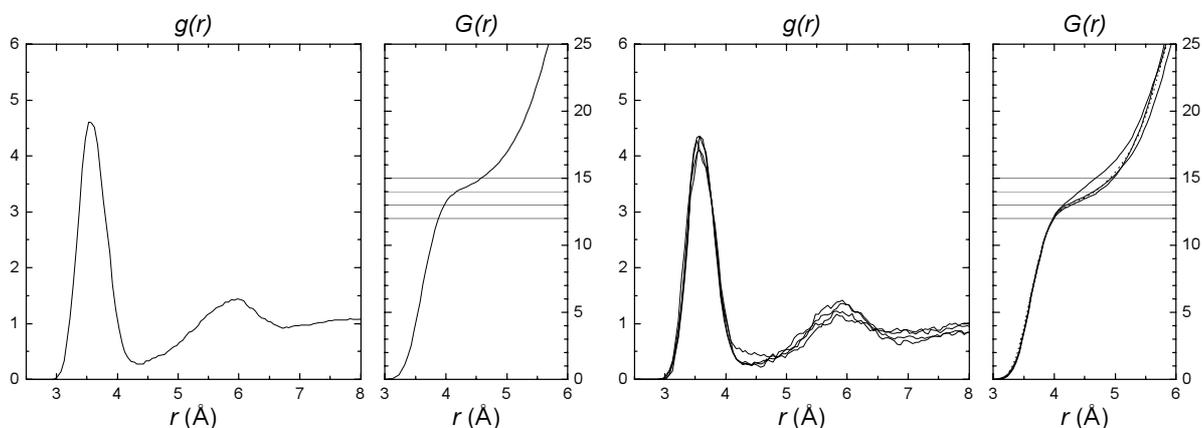


Fig. 4.31: g - and G -functions. Left: P / OT. Right: PB / OT (2,6,10,20). Unit of G is number of particles.

The findings are encouraging indications that the force field is valid and delivers reasonable results. As to the telopeptide, they show that single ion adsorptions do not induce visible conformational rearrangements in the protein. This supports the assumption that binding sites can be treated as independent sites within the protein. As a further look on the interaction of protein and ions, the stability of multiple ion complexes is tested next.

4.5 Microscopic computational experiments with complexes

The analysis of the longtime simulation showed that several groups identified as possible adsorption sites are sometimes in close proximity to each other. For a few cases it was attempted to place one or multiple counterions in direct contact to adsorption sites to create a complex. A short relaxation and a subsequent simulation was performed to test the stability of the aggregates generated in this way.

The configurations were simulated with the DLPOLY program in the NpT ensemble with the Nosé-Hoover barostat and thermostat at 1 bar and 300 K. Electrostatics were calculated using the SPME algorithm with a precision of 10^{-8} . All bonds involving hydrogen atoms were held

fixed using the SHAKE algorithm with a convergence criterion of 10^{-8} . Cubic periodic boundary conditions were applied. The integration time step was 1 fs.

The generated complexes of the telopeptide and one or more ions are tested on stability by monitoring the distances of the ions to the adsorption sites during the simulation and observing the protein relaxation. Three cases shall be discussed here. All simulations here represent single events and therefore have no statistical significance. Albeit, the attachment of ions to binding sites are naturally occurring events and therefore not unreasonable. The assumptions that are made are the selection of the binding sites and the *simultaneous* attachment of several ions. This is regarded as a measure to compress time and skip the explicit simulations of all single adsorption processes. The study of the complexes is not systematic but gives a good impression of the way the telopeptide behaves in interaction with multiple ions. Hints on the ability of the telopeptide to act as a center for mineral nucleation can be gained hereby which justifies the approach. The observations made in the three test cases are discussed here.

Case 1

A calcium ion and a hydrogen phosphate ion were placed in vicinity to the sites B_ASP7 and B_LYS9, respectively. A 3 ns simulation was conducted, during which the calcium ion detached from the protein. The hydrogen phosphate ion remained bound for the whole simulation, but no additional interactions to other groups of the protein were established. No observations were made worth mentioning.

Case 2

Three calcium ions and two hydrogen phosphate ions were attached to the sites A_ASP3, B_ASP7, B_GLU8, A_LYS5 and B_LYS9, respectively. A 2 ns simulation was conducted. The monitored distances of ions to the sites are plotted in figure 4.32.

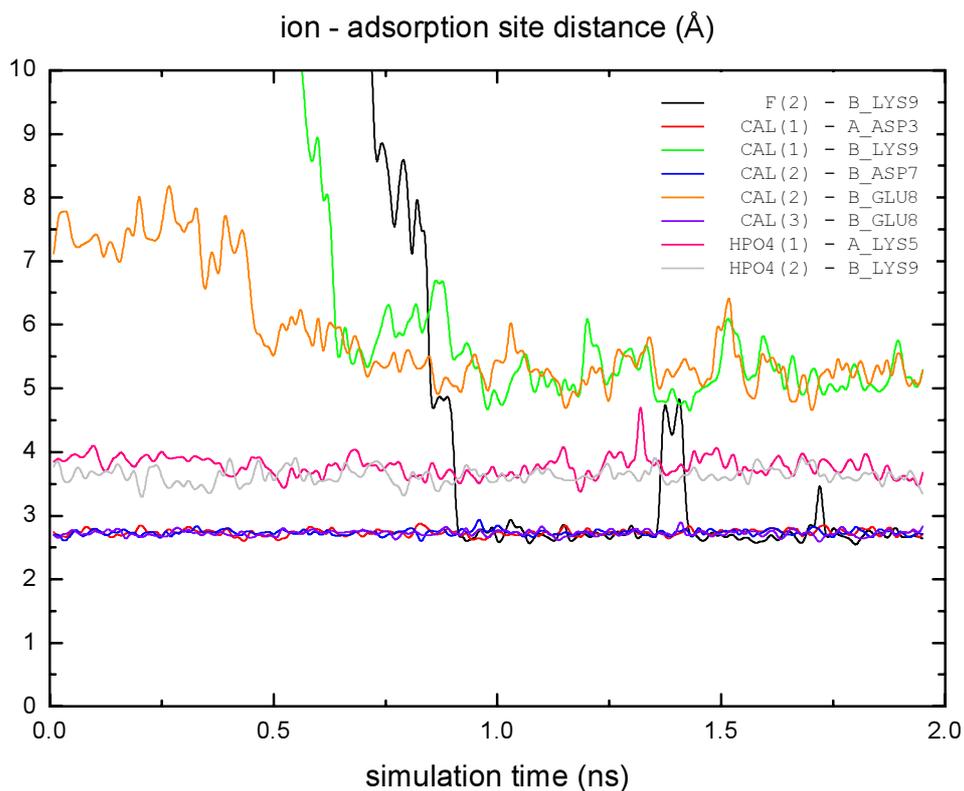


Fig. 4.32: Distances of ions to adsorption sites.

The five attached ions can be seen at the left end of the distance plot, calcium ions bound at approximately 2.7 Å distance, hydrogen phosphate ions at 3.8 Å. No interactions with other sites can be detected at the beginning of the simulation. All of the attached ions remain bound to their binding sites. During the course of the simulation, a rearrangement of the protein can be observed. The four groups A_ASP3, B_GLU8, B_ASP7 and B_LYS9 come near each other so that three more interactions are generated, one direct ion pair and two solvent separated ion pairs. In addition, a fluoride ion from solution spontaneously attaches to B_LYS9. It replaces a water molecule that was functioning as the solvent separation between B_LYS9 and a calcium ion before. Thereby, a triplet of ions in direct contact is formed that is in turn in direct contact to B_LYS9. Note that only pairs of ions attached to binding sites were present in the starting configuration, the agglomeration of these sites to such a cluster occurred spontaneously. The ionic interactions are schematically illustrated in the picture below, as well as snapshot view of the complex.

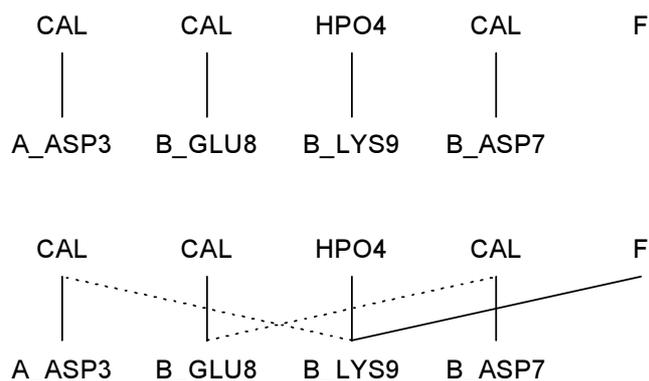


Fig. 4.33: Schematic illustration of ionic interactions in complex at the beginning of the simulation (top) and after 2 ns (bottom). Straight line: contact ion pair, dotted line: solvent separated ion pair.

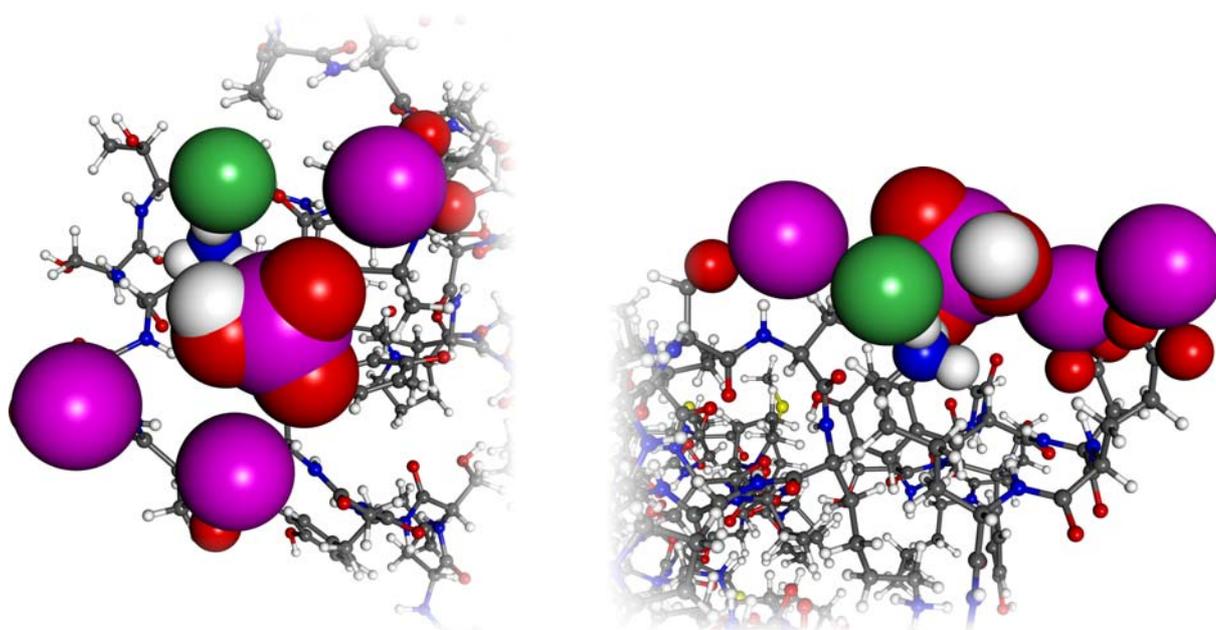


Fig. 4.34: Snapshots of the complex formed in the simulation. Protein in ball and stick, ions in CPK, binding sites in small CPK representation. Purple = calcium or phosphorus, green = fluorine. Left: top view, right: side view.

A detailed view of the structure at the beginning of the simulation shows that the second HPO4 ion is not only bound to the charged A_LYS5 residue. It is in addition stabilized by four hydrogen bonds located in a turn of the B chain which encloses the ion. The residues B_TYR6, B_ASP7 and B_SER10 act as H-donors, B_LYS9 as H-acceptors. During the simulation, this structure is opened in favor of more interactions of some residues to the other ions mentioned earlier.

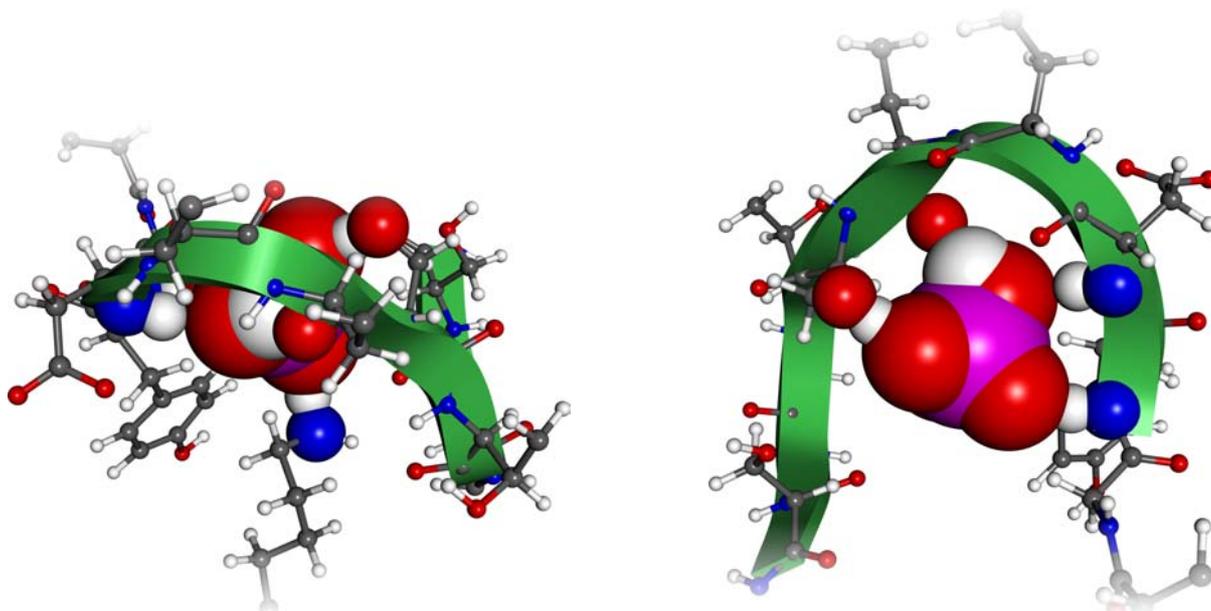


Fig. 4.35: Cutout showing the complexed hydrogen phosphate ion. Protein in ball and stick, HPO4 in CPK, binding sites in small CPK representation. B chain with green ribbon. Left: A_LYS5 rising from bottom of image, right: B chain enclosing the ion arcuately (A_LYS5 behind HPO4).

Case 3

A calcium ion was placed in proximity to C_ASP7 and a fluoride ion was attached to A_LYS5. A 3.5 ns simulation was conducted. The fluoride ion remained bound for 0.5 ns and detached afterwards. The calcium ion remained in its solvent separated state for 0.5 ns and likewise detached afterwards. The multiple ion complex was not stable in this case.

Discussion

The results show that stable complexes of the telopeptide with multiple ions are possible. A systematic study of complex formation or nucleation of apatite, respectively, is not possible with full MD simulation because the computational cost is too high. Coarse models are necessary to efficiently simulate large time scales needed for the study of the nucleation of inorganic crystallites. A model could for example describe the protein on a united residue basis, where a complete amino acid is represented by one interaction site. The inorganic component would still be described as single ions. Interaction potentials between the two are

needed for a parametrization of the model. This problem will be addressed by analyzing the process of single ions adsorptions in detail.

4.6 PMF calculations

As mentioned in the theory chapter, a PMF is the crucial information to analyze the energetics of a chemical reaction. Ion association reactions can be studied with classical molecular dynamics, since no covalent bonds are broken or formed. PMF profiles have been reported for amino acid side chain interactions by several sources^{60,61,62}. Likewise, PMF profiles for simple ion associations like NaCl⁶³ and also more complex ones like guanidinium-acetate⁶⁴ were calculated. The multiplicity of studies also revealed that free energy calculations are dependent on the treatment of long-range electrostatics, cutoff schemes, boundary conditions and water models used. Comparison of results with other studies should be therefore made with caution.

Coarse model development for proteins is also pursued by other groups to efficiently simulate protein folding^{65,66,67}. First Langevin dynamics simulations have been reported lately⁶⁸. A parametrization of the interaction of calcium with proteins was also reported⁶⁹. This was however done using Hartree-Fock methods, without considering solvent and without determining a reaction coordinate. The interaction potentials were obtained by simply integrating over all degrees of freedom that were lost in the coarse models. No PMF results exist, however, for the association of small ions to amino acid side chains, especially for apatite ions. This gap shall be closed here.

As is always the case in molecular dynamics, the simulated systems have to be designed with both chemical questions and feasibility and efficiency considerations in mind. PMF calculations require extensive sampling, due to the fact that tens of points on the general reaction coordinate axis have to be sampled hundreds of picoseconds to get results in acceptable precision, or statistical reliability, respectively. A typical PMF profile thus consumes a few nanoseconds of simulation time. It is therefore uneconomic to calculate a PMF profile for the attachment of all involved ion species to all possible adsorption sites of

the full telopeptide. To keep computational efforts in a feasible dimension, simplifications or approximations have to be made.

It is assumed that, in a first approximation, the potential of mean force is predominantly influenced by the ion species and by the chemical group or nearest environment of the adsorption site. If no particular situations like chelate type binding sites are considered, which will not be done with a one-dimensional internal PMF coordinate as in our case, this approximation will be fairly good. In that sense, all possible adsorption sites of the protein can be classified into a limited number of types by chemical similarity. PMF profiles can then be calculated for all appropriate combinations of ion species and adsorption site types.

4.6.1 Adsorption site model structures

4.6.1.1 Model preparation

System buildup

The telopeptide structure was analyzed as to possible adsorption sites, and a number of prototype sites were defined which can serve as approximations for all sites in the telopeptide. The sites found were individual amino acids with charged or polar groups, and also backbone elements. Where conformation was suspected to play an important role, cases with different conformations were distinguished. The collected prototypes are listed in table 4.7.

#	site ^a	description	geometry ^b	charge ^c	ion ^e
1	GLU	carboxyl group in glutamic acid	a.a.	-	CAL
2 a	CO	carbonyl oxygen in backbone amide group	$\Psi = -47^\circ$	δ^-	CAL
2 b			a.a.		
3	LYS	amino group of lysine	a.a.	+	F, HPO4
4 a	NH	amino hydrogen in backbone amide group	a.a.	δ^+	F, HPO4
4 b			$\Phi = -60^\circ$		
5	ASP	carboxyl group in aspartic acid	$\Phi = -90^\circ$ ^d $\Psi = -84^\circ$	-	CAL
6	PYR_H	amino hydrogen in pyroglutamic acid	$\Psi = 119^\circ$ ^d	δ^+	F, HPO4
7	PYR_O	carbonyl oxygen in pyroglutamic acid		δ^-	CAL
8	THR_H	hydroxy hydrogen in threonine	a.a.	δ^+	F
9	THR_O	hydroxy oxygen in threonine		δ^-	CAL
10	THR_OO	hydroxy oxygen in threonine and carbonyl oxygen in backbone	$\Phi = 134^\circ$ $\Psi = 171^\circ$	δ^-	CAL
11	TYR	hydroxy hydrogen in tyrosine	$\Phi = -108^\circ$ ^d $\Psi = 51^\circ$	δ^+	F, HPO4

Table 4.7: Model peptides for ion adsorption sites. ^aName of binding site amino acid or chemical group. ^bConformation of central residue, extended chains denoted a.a. (all anti). ^cCharge of binding site atom. ^dConformation chosen as modeled by Scheraga. ^eIons considered.

Prototype proteins were generated with SYBYL. A small protein was built from every amino acid or backbone group flanked by glycine residues, yielding two- or three-residue proteins. The terminal glycine residues were saturated with hydrogen atoms so that both N- and C-terminus were neutral. The backbone conformation was chosen extended (all-anti) if not otherwise stated. Where an extended conformation was unfavorable or uncommon, the conformation found in the telopeptide as modeled by Scheraga was taken. Side chains were generated in an extended conformation (all-anti where applicable) in an orientation pointing away from the backbone.

The PMF routine in the DLPOLY program was available for the *NVE* ensemble only. After communication with the authors, the routine was adapted to the *NVT* ensemble in a straightforward manner.

Equilibration

The generated proteins were put into water boxes of appropriate size so that the ions could move at least 13 Å away from the attachment site without entering the cutoff of the protein's periodic image. The boxes including protein, water and the ion were simulated in an NpT simulation to let the box size relax until no further drift was observed. In subsequent simulations of 0.5 to 2.5 ns, the mean values of the fluctuating box dimensions were determined and used as box sizes for the constant volume PMF simulations. The box volume sampling is shown below for model 1 as an example.

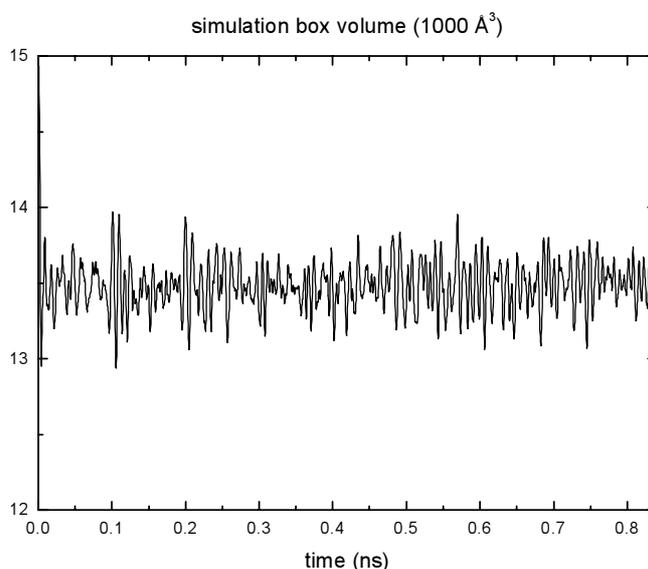


Fig. 4.36: Box volume during NpT sampling of equilibrium mean box dimensions for model 1.

4.6.1.2 Simulations

For the PMF calculations, the distance between the adsorption site atom and the ion was defined as the pmf constraint ξ . For an amino group the coordinates of nitrogen were used for the adsorption site, for a carboxyl group the coordinates of the carbon atom. A PMF on a coarse ξ coordinate grid was calculated first and subsequently refined where large changes in the PMF were observed. This way, the resolution on the ξ -axis was adaptive to the level of detail needed. The pmf constraint force was sampled at every time step of 1 ps. Tests showed very large fluctuations of the constraint force in an NpT ensemble. The simulations were

therefore performed in an NVT ensemble where fluctuations were smaller, which is a reasonable approximation in terms of the free energy calculation. The proteins were tethered at the terminal residues with the confining restraint potential to leave protein motions and vibrations minimally disturbed while still keeping them in place in the box. Due to the high fluctuations, the constraint force had to be sampled for 250 ps at each value of ξ to get satisfactory statistical confidence in the mean force. The PMF profiles were calculated in the range between 2 and 13 Å at about 35 different constraint distances. A full PMF profile thus required about 9 ns of simulation time.

4.6.1.3 Analyses

The mean of the sampled constraint force was calculated at each constraint distance and plotted in the mean force profile. Integration of the profile yields the potential of mean force. Care has to be taken to determine the effective number of samples N_{eff} . A physical quantity does not resemble a random variable due to the interdependence of samples owing to physical laws. Therefore, a decorrelation time has to be determined which defines the minimum time between samples to be interpreted as random events in a statistical sense. The mathematical background for the autocorrelation function (acf) was described in chapter 3.2.4. The decorrelation time is determined as the time lag where the acf has dropped to practically zero, which means the level on which the acf fluctuates for an equivalent number of random samples. Error bars are calculated based on the determined effective number of samples. Upon integration, the maximum error becomes the sum of the errors of all integrated points, which rises from left to right.

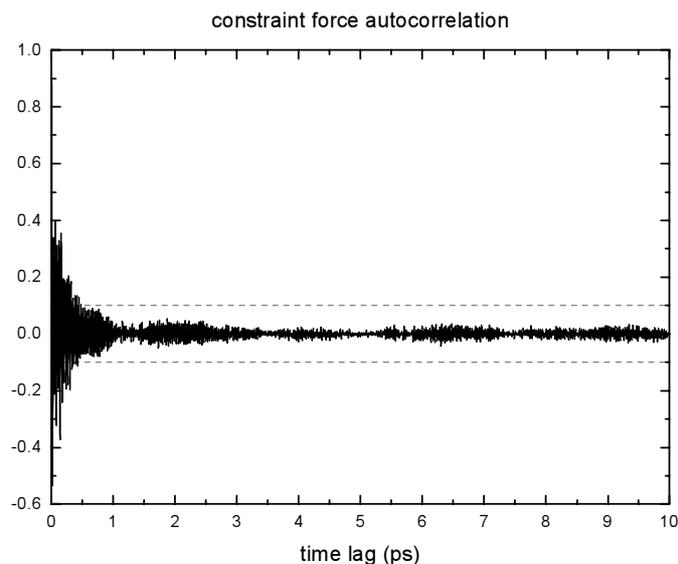


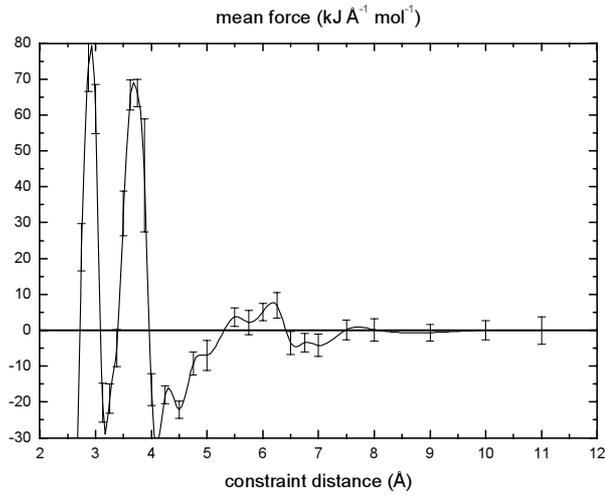
Fig. 4.37: Autocorrelation function of the constraint force for model 1 GLU at a constraint distance of $d = 5.00$. The threshold of 0.1 used as decorrelation criterion is drawn as gray dashed lines.

The acf was calculated for each point in the PMF. The decorrelation time τ_d was determined by finding the time lag where the acf fell below a threshold of 0.1 and remained stationary. A typical autocorrelation function of the constraint force taken from model 1 GLU at a constraint distance of $d = 5.00$ is shown as an example in figure 4.37. The acf exhibits a shape of a decay overlaid by a (high frequency) oscillation typical for partially autoregressive models. The determined decorrelation times are in the range of 0.5 to 3 ps for all models. Differing decorrelation times in this range were also observed by Hassan et al⁶².

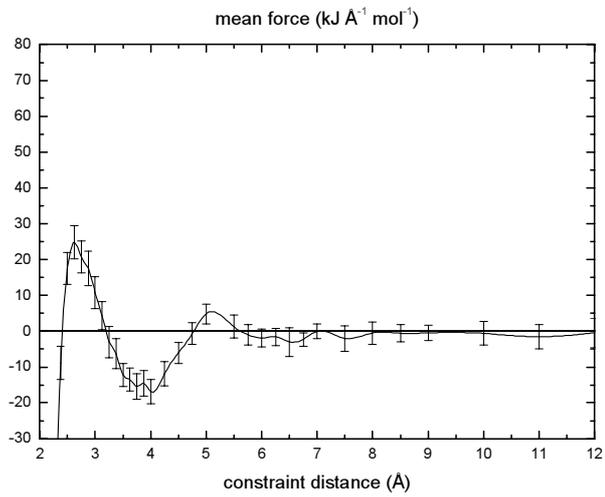
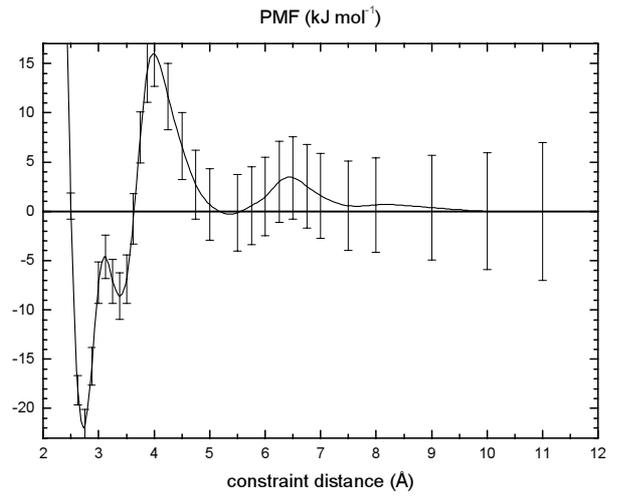
With the decorrelation time, the effective number of samples was calculated as $N_{eff} = N / \tau_d$. The threshold of 0.1 is a rather strict value, confidence bands for a 95 % confidence level for the determined N_{eff} are larger than 0.1 in all cases. The errors bars in the mean force plot are calculated as the standard error of the mean based on N_{eff} .

The calculated profiles of the mean force and the potential of mean force for all models are shown below.

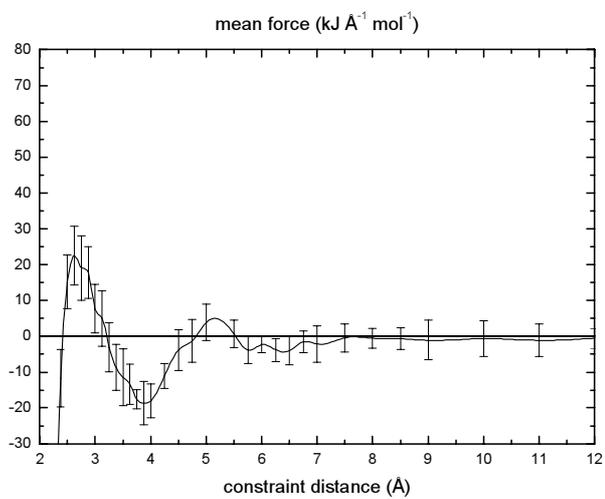
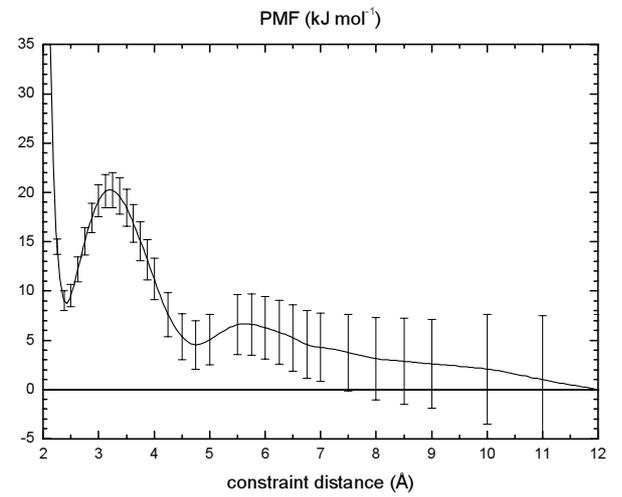
4.6 PMF calculations



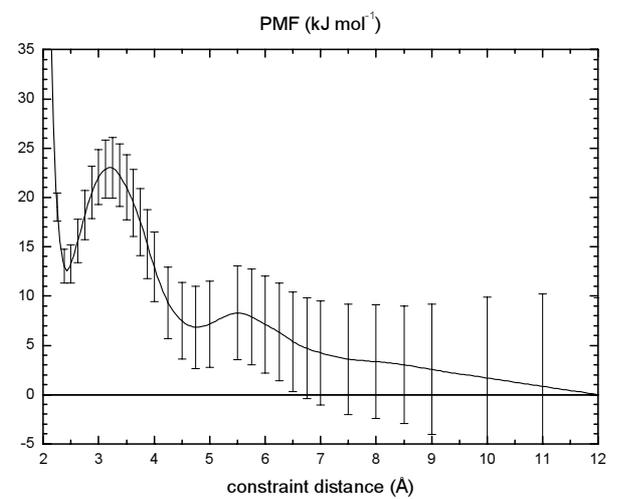
model 1 GLU, CAL



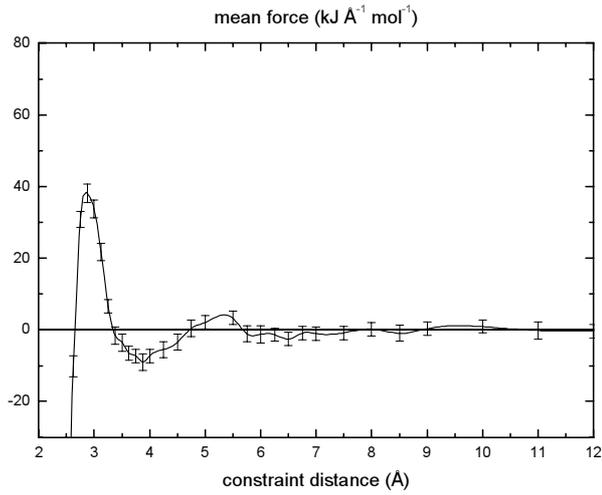
model 2a CO, CAL



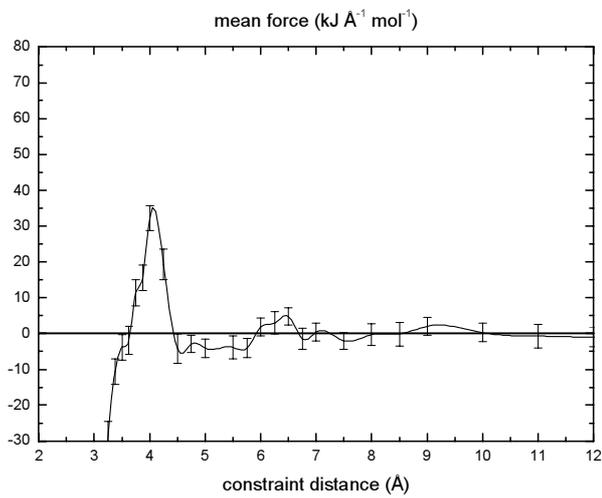
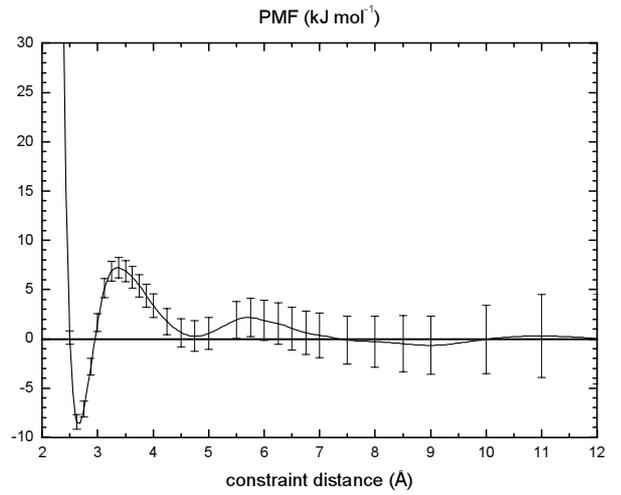
model 2b CO, CAL



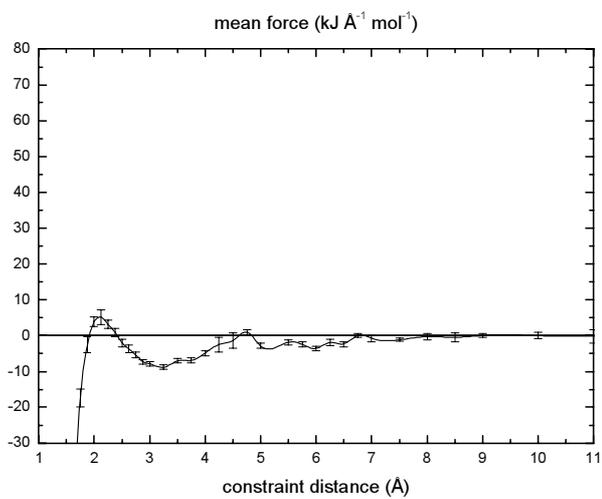
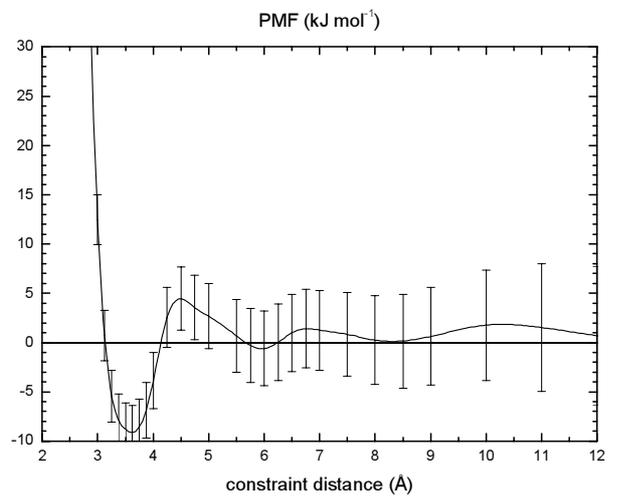
4.6 PMF calculations



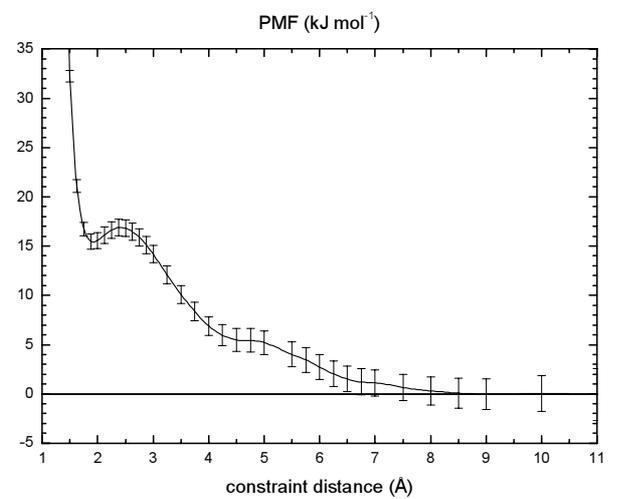
model 3 LYS, F



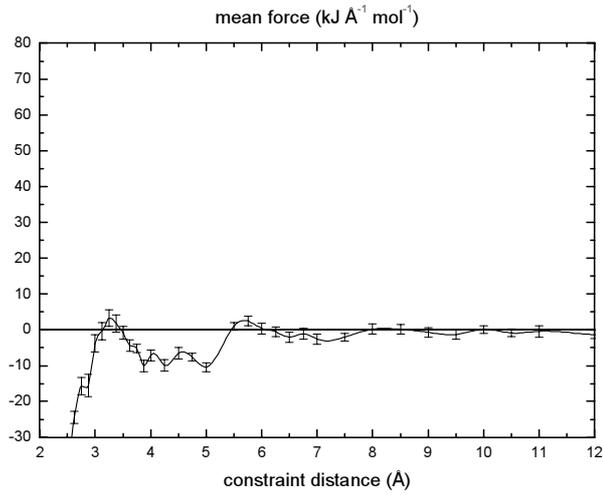
model 3 LYS, HPO4



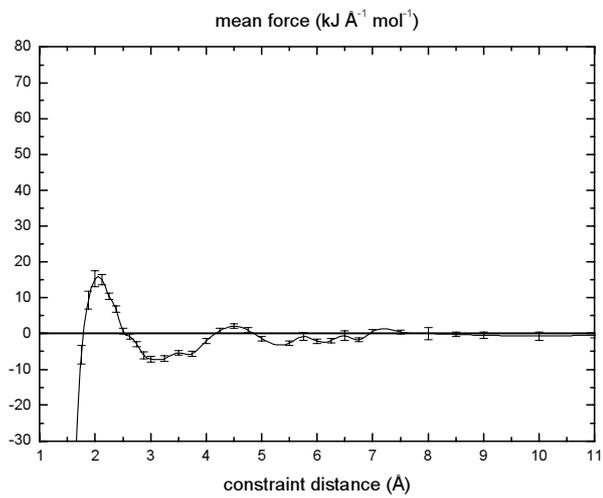
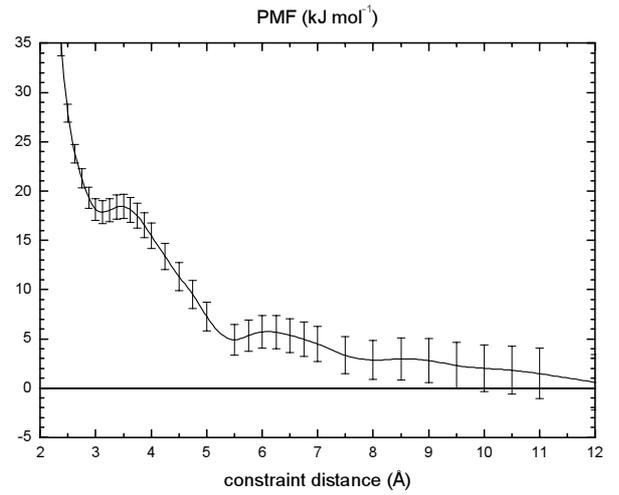
model 4a NH, F



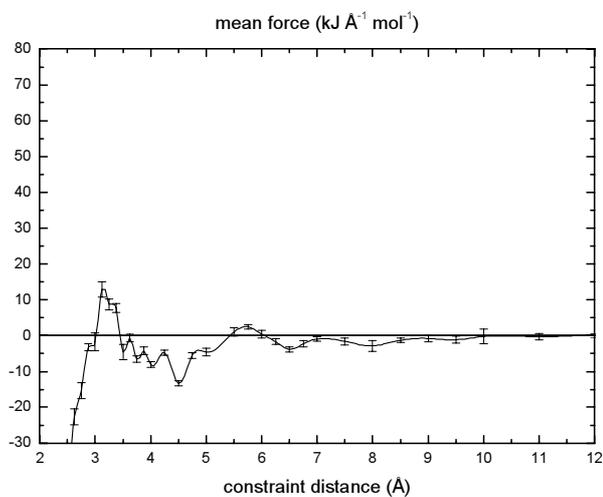
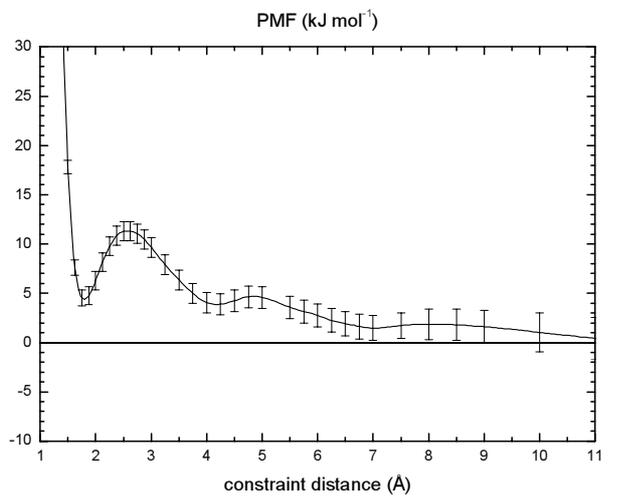
4.6 PMF calculations



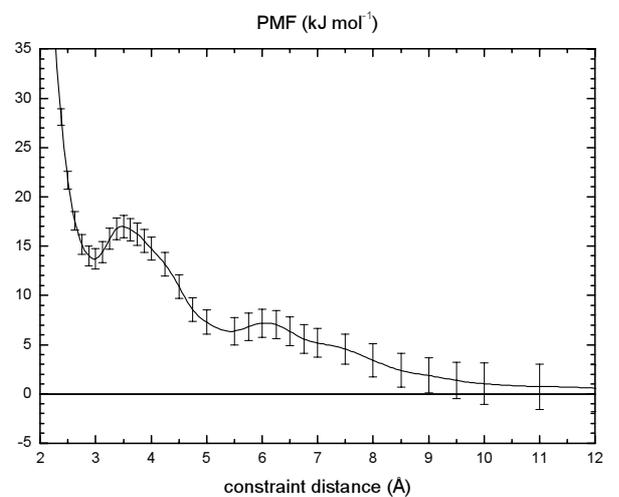
model 4a NH, HPO4



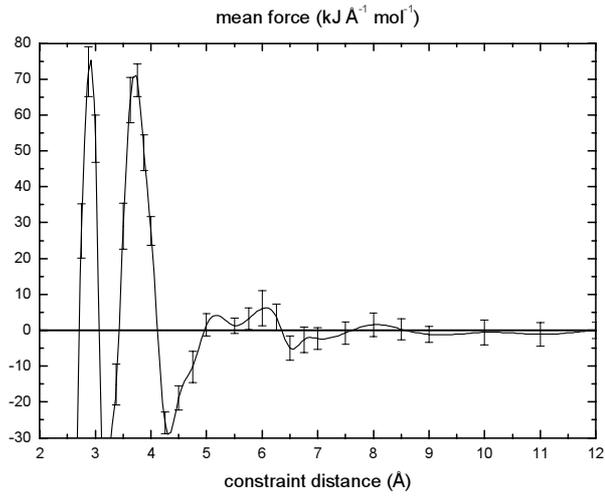
model 4b NH, F



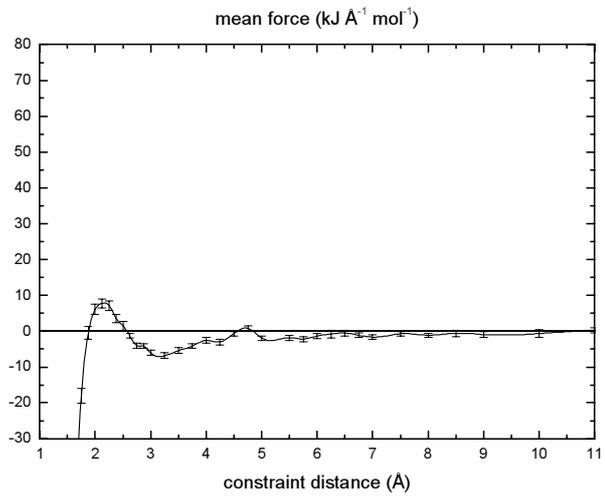
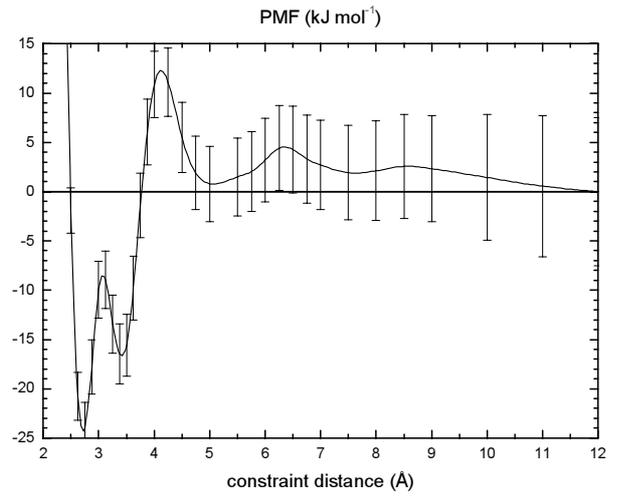
model 4b NH, HPO4



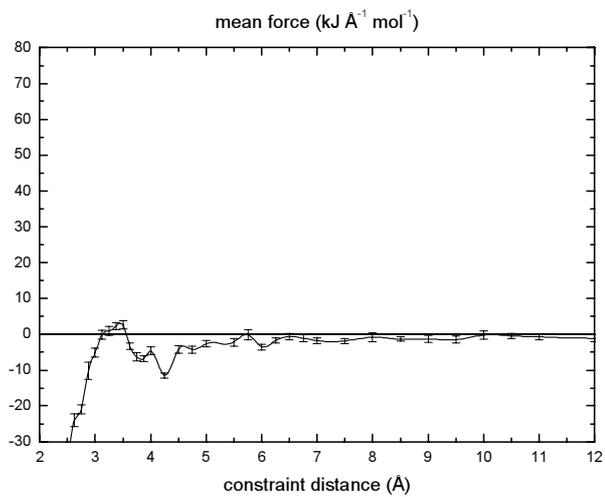
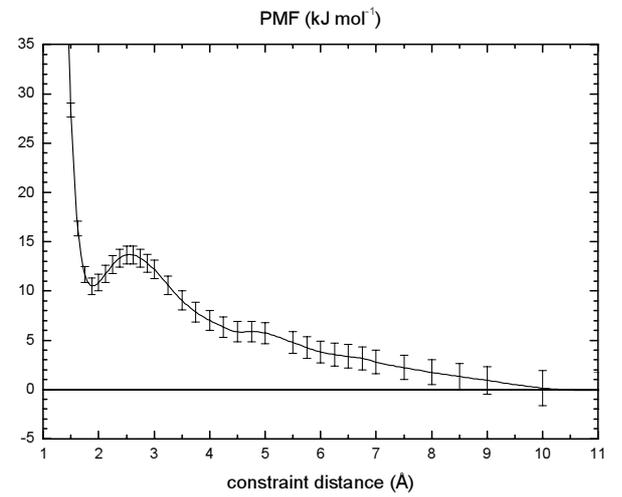
4.6 PMF calculations



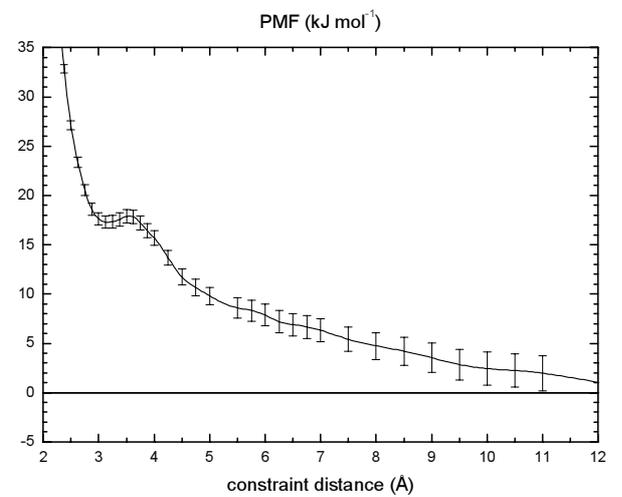
model 5 ASP, CAL



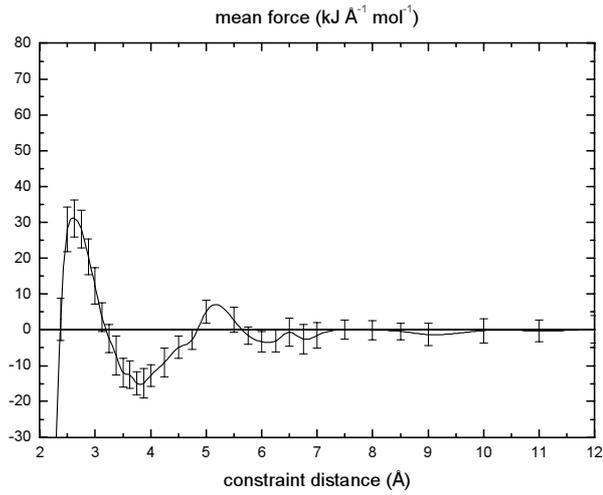
model 6 PYR_H, F



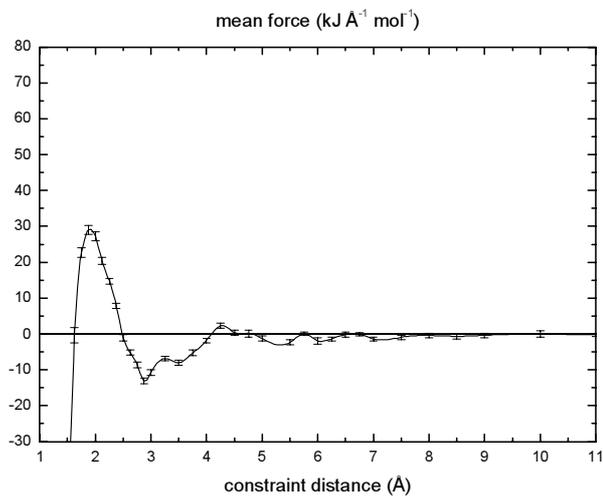
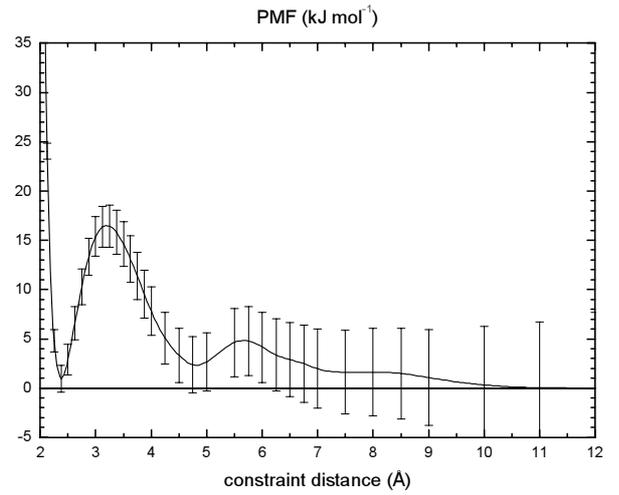
model 6 PYR_H, HPO4



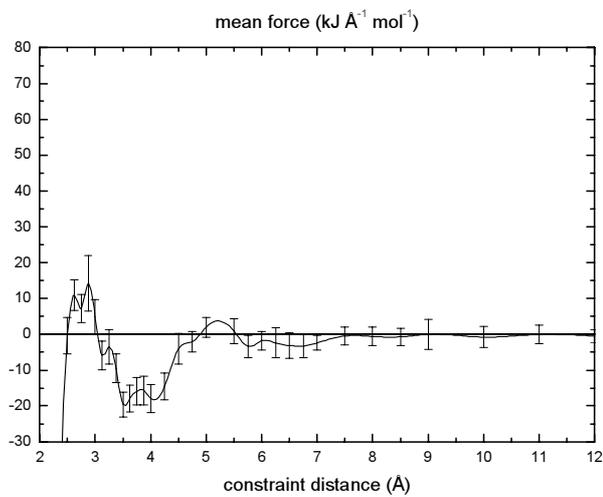
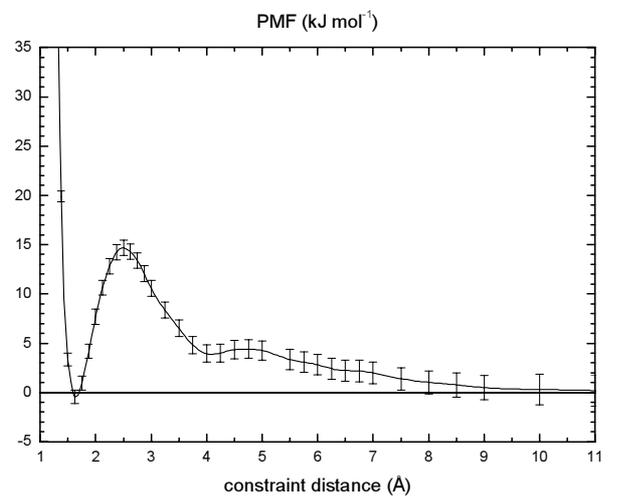
4.6 PMF calculations



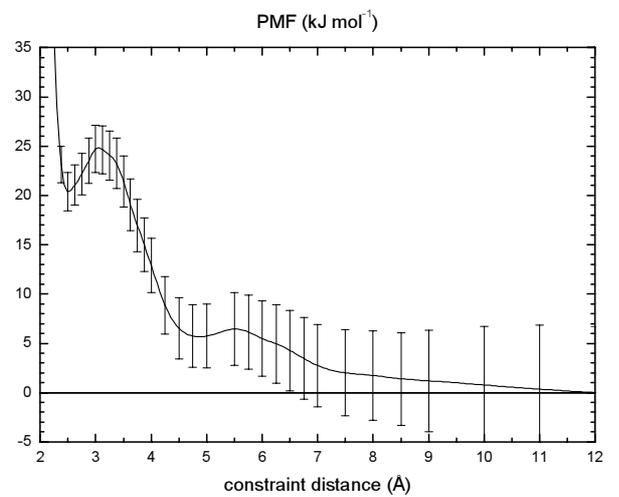
model 7 PYR_O, CAL



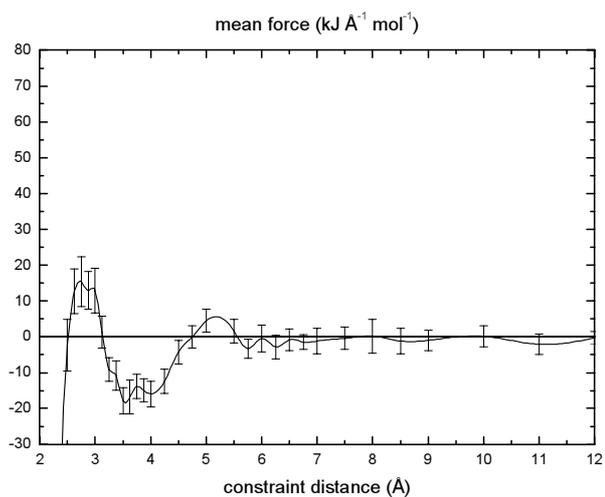
model 8 THR_H, F



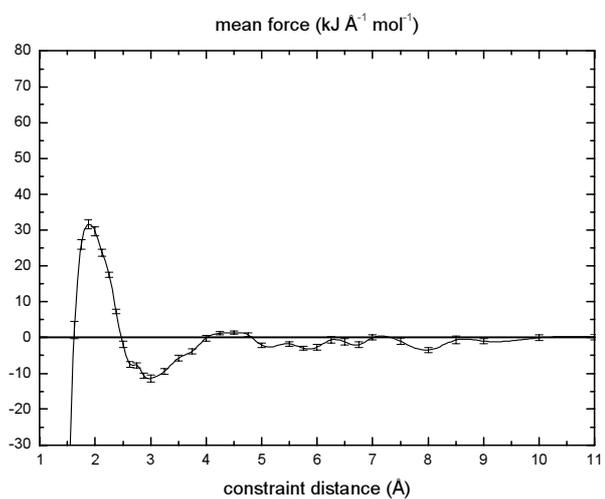
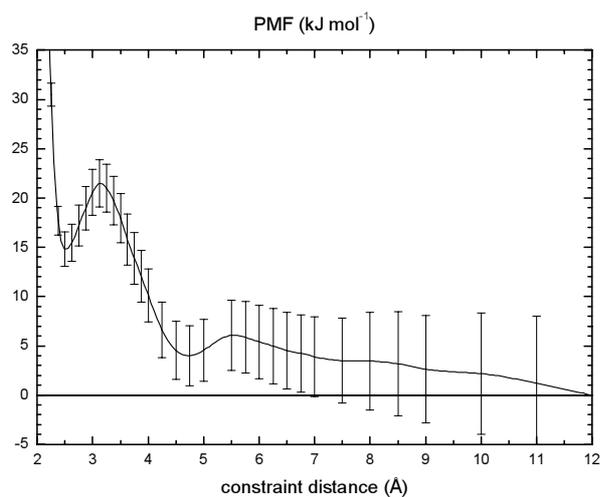
model 9 THR_O, CAL



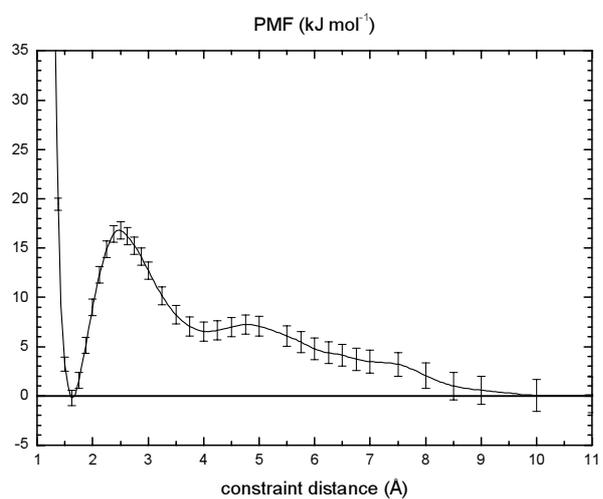
4.6 PMF calculations



model 10 THR_OO, CAL



model 11 TYR, F



4.6 PMF calculations

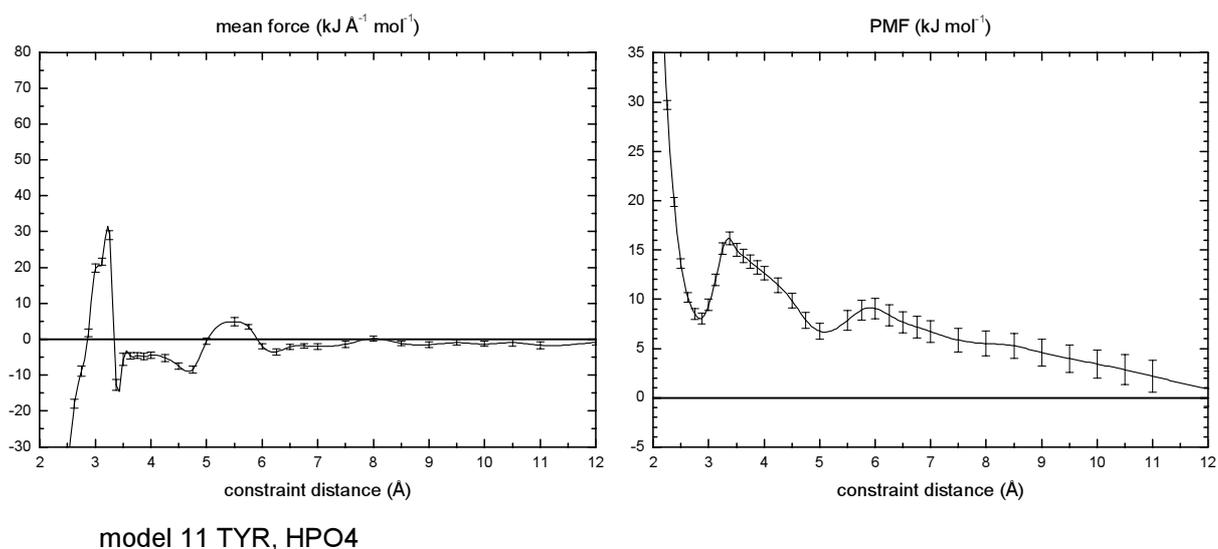


Fig. 4.38: Mean force (left) and PMF (right) of ion attachment to adsorption sites in model proteins.

Integration of the mean force accumulates errors from left to right in the PMF. The integral can be approximated as a Riemann sum, and consequently the Gaussian error propagation law can be applied which uses the geometrical mean of individual errors. Features of the PMF curve are thus most trustworthily analyzed for relatively small ranges of the constraint distance. For larger ranges and especially in the range of a high constraint distance, quantitative comparisons have to be made with caution. Unfortunately, the zero energy level must be determined at the highest constraint distance (as an approximation for infinite distance). In this sense, the calculation of e.g. a local barrier is more reliable than the calculation of the total reaction energy. This should be kept in mind when interpreting results.

4.6.1.4 Discussion

PMF curves

A few features of the calculated mean force curves immediately attract attention. The attachment of calcium to a carboxylic acid (models 1 and 5) is accompanied by strikingly great forces which have a pattern at low distances that is not observed in other models. The attachment of a hydrogenphosphate ion to a neutral backbone amino group (models 4 a and 4 b), a charged amino group in lysine (model 3) or other groups show an almost constant

mean force in the range between 3 and 5 Å (fluctuating in the range of the standard error) which converts into a uniquely asymmetric barrier in the PMF.

The results obtained give a good overview of the effective interaction potentials between apatite ions and possible binding sites in the N-terminal collagen telopeptide. All of the PMF curves feature a direct contact minimum and several solvent separated minima. The direct contact minimum is well pronounced in nearly all cases, only a few cases show a shallow direct contact minimum. A synopsis of all PMF curves leads to a classification into three energetically characteristic types of potentials. These are potentials that have a direct contact minimum I) below zero energy, II) at roughly zero energy or III) at higher energy, where zero energy is defined at infinite distance and measured at 12 to 14 Å distance as an approximation. It is found that only the association of oppositely charged species leads to a thermodynamically stable aggregate of type I. These are rather few sites in the telopeptide, namely the charged residues lysine, glutamic acid or aspartic acid, of which a total of eight are present in the telopeptides. It is readily understood that the attachment of a calcium ion to an uncharged carbonyl group leads to zero or positive net free energies. Water has to be expelled from the coordination sphere of calcium and is replaced by carbonyl groups. Since water has a higher polarization, a favorable interaction is exchanged with a less beneficial one, resulting in a disadvantageous overall process.

A majority of potentials exhibit a strong trend of rising energy with decreasing distance that is attributed to the direct component of the PMF, and clearly dominates the solvent induced component. For better comparison, the PMF curves are graphed in a combined plot per ion species in figure 4.39.

4.6 PMF calculations

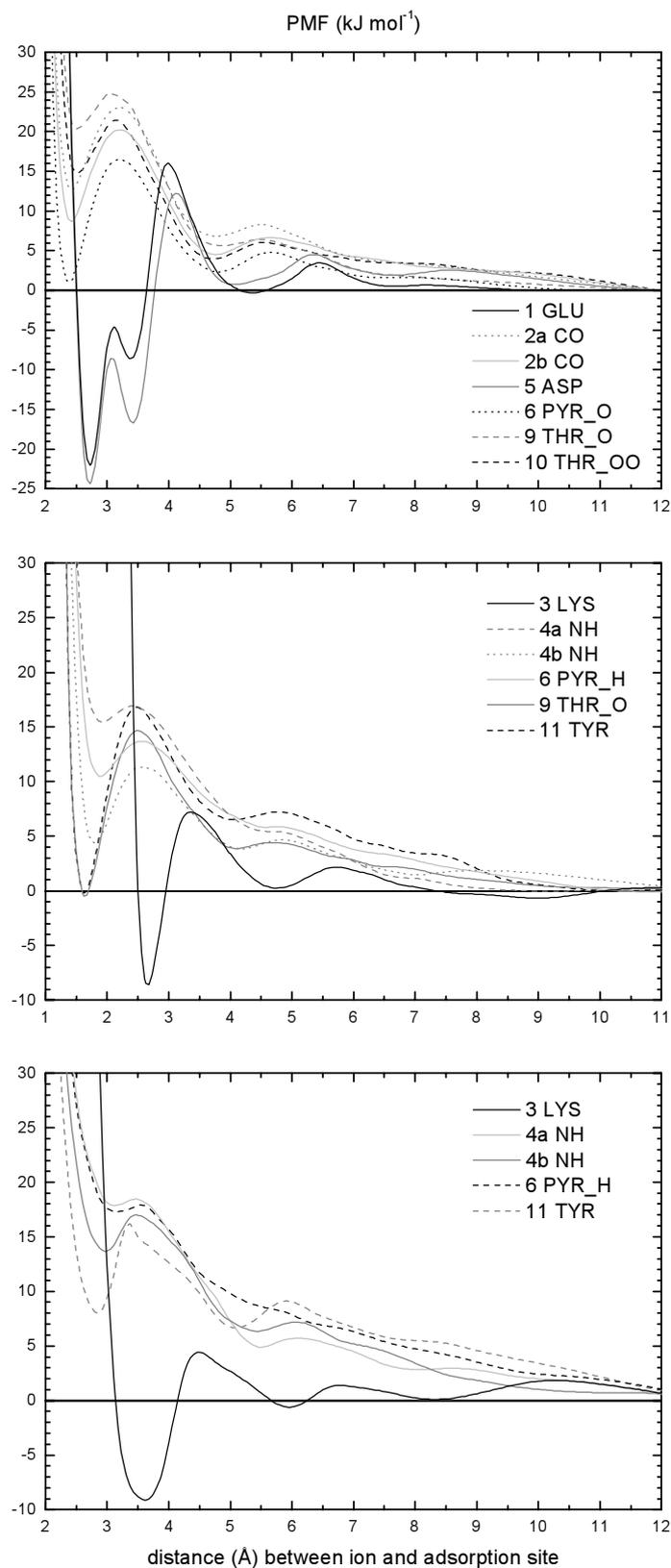


Fig. 4.39: Comparison of PMF profiles for the binding of calcium (top), fluoride (middle) and hydrogen phosphate (bottom) to model sites. Individual axes scaling in each plot. Error bars not shown for better clarity.

The carboxyl groups of glutamic acid and aspartic acid show a unique form of the PMF at small distances, where a local minimum very close to the global direct contact minimum can be observed. This feature will be studied in detail below. It is remarked that this local minimum was accurately reproduced only when working with a constraint distance resolution of at least $\frac{1}{8}$ Å. The two curves are sufficiently similar within error margins to state that both carboxylic acids can be modeled by one prototype protein in good approximation.

Also the PMFs of hydrogenphosphate show peculiar characteristics. The barriers are noticeably asymmetric here. The effect can most likely be ascribed to the symmetry of the hydrogenphosphate ion, which is C_s in eclipsed or staggered conformation of the hydroxy group or C_1 otherwise, in contrast to K_h for calcium and fluoride. In other words, the calcium and fluoride ions are perfect spheres while hydrogenphosphate is asymmetric, thus a changing orientation of the hydrogenphosphate ion while approaching the adsorption site will influence the PMF.

The attachment of calcium to a backbone carbonyl group turns out to be almost independent of the backbone conformation, both PMF curves are nearly identical. The attachment of a hydrogenphosphate ion and especially a fluoride ion to a backbone amino group (model 4), on the other hand, is highly dependent on backbone conformation. The carbonyl group of pyroglutamate binds calcium more firmly. This is presumably caused by the higher partial charges of this group in the force field. The partial charges are even higher in the TIP3P water model, this explains that the binding energy to a carbonyl group is positive, since the hydration shell is energetically favorable due to the Coulomb interactions.

Characteristic data of the PMF curves have been collected in table 4.8, namely the position and energy of the first and second minimum and maximum, counted from left to right. For three cases, a second minimum and maximum were not discernable. All PMF profiles have the same elementary features (see figure 4.40) so that order and interpretation of the local extrema are identical in all cases. The first minimum, first maximum, second minimum and second maximum are associated with the direct contact ion pair, a barrier, the first solvent separated ion pair and another barrier, respectively. The PMF curves are shifted so that the separated reactants in solution are at zero energy. Consequentially, the energy of the first minimum can be interpreted as the binding free energy, and the energy of the first maximum as the ion attachment reaction activation barrier.

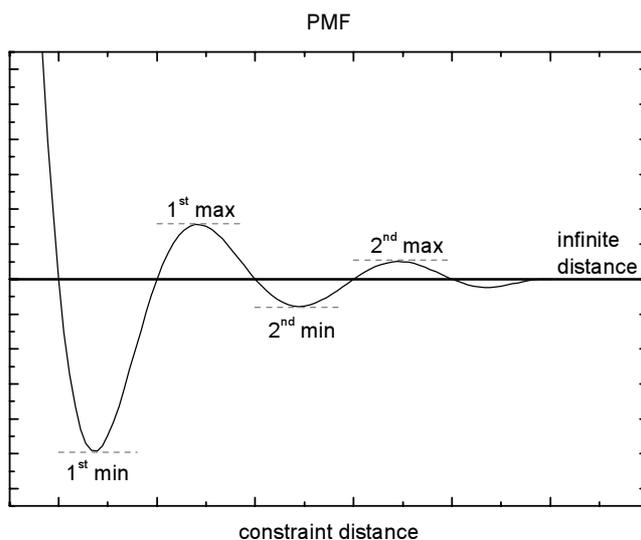


Fig. 4.40: Illustration of a common shape of PMF profiles and labeling of local extrema.

#	site	ion	1 st				2 nd			
			minimum		maximum		minimum		maximum	
			pos. ^a	ener. ^b	pos.	ener.	pos.	ener.	pos.	ener.
1	GLU	CAL	2.8	-21.8	4.0	16.0	5.5	-0.2	6.5	3.4
2 a	CO	CAL	2.4	9.0	3.3	20.2	4.8	4.5	5.8	6.6
2 b	CO	CAL	2.4	13.0	3.3	23.0	4.8	6.8	5.5	8.3
3	LYS	F	2.6	-8.4	3.4	7.2	4.8	0.2	5.8	2.2
3	LYS	HPO4	3.6	-9.2	4.5	4.4	6.0	-0.6	8.5	0.1
4 a	NH	F	1.9	15.5	2.4	16.9	-	-	-	-
4 a	NH	HPO4	3.1	17.9	3.5	18.4	5.5	4.9	6.5	5.3
4 b	NH	F	1.8	4.5	2.5	11.3	4.3	3.9	4.8	4.6
4 b	NH	HPO4	3.0	13.7	3.5	17.0	5.5	6.4	6.3	7.0
5	ASP	CAL	2.8	-24.0	4.3	11.1	5.0	0.8	6.3	4.4
6	PYR_H	F	1.9	10.5	2.6	13.7	-	-	-	-
6	PYR_H	HPO4	3.1	17.3	3.5	17.9	-	-	-	-
7	PYR_O	CAL	2.4	1.0	3.3	16.4	4.8	2.4	5.8	4.8
8	THR_H	F	1.6	-0.5	2.5	14.7	4.0	3.9	4.8	4.4
9	THR_O	CAL	2.5	20.4	3.0	24.7	5.0	5.7	5.8	6.1
10	THR_OO	CAL	2.5	14.8	3.1	21.5	4.8	4.0	5.5	6.1
11	TYR	F	1.6	-0.2	2.5	16.8	4.0	6.5	4.8	7.2
11	TYR	HPO4	2.9	8.1	3.4	16.2	5.0	6.7	6.0	9.1

Table 4.8: Characteristic data of the PMFs. ^aPosition (Å) and ^benergy (kJ/mol) for local extrema, counted ascending from small to large distances. Errors estimated ± 0.5 Å for positions and ± 3 kJ/mol for energies on average.

Thermodynamically stable minima are found for the attachment of calcium to aspartic acid or glutamic acid as well as fluoride or hydrogen phosphate to lysine. The binding free energies are -24.0, -21.8, -8.4 and -9.2 kJ/mol, respectively. The reaction barrier is 4 to 6 times kT for the attachment of calcium and 2 to 3 times kT for hydrogen phosphate and fluoride, respectively. In fact, the attachment of hydrogen phosphate to lysine is the only stable complex of hydrogen phosphate. This potentially entails a special role of lysine for the nucleation of apatite. The binding of calcium to the carboxylic function of aspartic acid or glutamic acid is the most stable complex found. This leads to the conclusion that calcium stays bound most durable and could thus be the initial step of apatite nucleation.

The attachment of fluoride to the hydroxy function of tyrosine or threonine is associated with a zero binding energy. Nevertheless these complexes are rather stable since the barrier for dissociation is around 15 - 17 kJ/mol. The same holds for the attachment of calcium to the carbonyl function of pyroglutamate. The binding energies of fluoride can be consistently categorized by the chemical nature of the binding site. It is approximately zero for all hydroxy groups, positive for all neutral amino groups and negative for all charged amino groups.

The PMF curves show anticipated similarities with the existing results for amino acid side chain interactions mentioned at the beginning of this chapter. The range of binding free energies determined are in good agreement with similar cases.

The geometry of a fluoride ion bound to an amino group is not symmetric as one could guess (C, N and F on a straight line), rather the fluoride ion is directly bound to exactly one of the hydrogen atoms (N, H and F on a straight line). In the case of a calcium ion bound to a carboxyl group, on the other hand, the complex is symmetric. This is discussed in detail below.

Reaction rates

Reaction rate constants were calculated from the obtained energies using transition state theory (TST). The ion attachment reaction was interpreted as the transition from the second to the first minimum in the PMF, with the first maximum as transition state. Reaction free energies and rate constants were calculated for the association and dissociation reaction. It

4.6 PMF calculations

was refrained from calculating the transmission coefficient κ as this would require a major effort. It is efficient to approximate κ as 1 and use the results as approximations for the real reaction rates. Comparison between the simulated reactions should be quite reasonable as all reactions are of the same type. Results are listed in table 4.9.

#	site	ion	$\Delta^\ddagger G_{ass}$	$\lg(k_{ass})$	$\Delta^\ddagger G_{diss}$	$\lg(k_{diss})$	$\lg(K_{ass})$
1	GLU	CAL	16.2	10.0	37.8	6.2	3.8
2 a	CO	CAL	15.7	10.1	11.2	10.8	-0.8
2 b	CO	CAL	16.2	10.0	10.0	11.1	-1.1
3	LYS	F	7.0	11.6	15.6	10.1	1.5
3	LYS	HPO4	5.0	11.9	13.6	10.4	1.5
4 a	NH	F	--	--	1.4	12.6	--
4 a	NH	HPO4	13.5	10.4	0.5	12.7	-2.3
4 b	NH	F	7.4	11.5	6.8	11.6	-0.1
4 b	NH	HPO4	10.6	11.0	3.3	12.2	-1.3
5	ASP	CAL	10.3	11.0	35.1	6.7	4.3
6	PYR_H	F	--	--	3.2	12.2	--
6	PYR_H	HPO4	--	--	0.6	12.7	--
7	PYR_O	CAL	14.0	10.4	15.4	10.1	0.2
8	THR_H	F	10.8	10.9	15.2	10.1	0.8
9	THR_O	CAL	19.0	9.5	4.3	12.0	-2.6
10	THR_OO	CAL	17.5	9.7	6.7	11.6	-1.9
11	TYR	F	10.3	11.0	17.0	9.8	1.2
11	TYR	HPO4	9.5	11.1	8.1	11.4	-0.2

Table 4.9: Free energy of activation $\Delta^\ddagger G$, TST reaction rate constants k and thermodynamic equilibrium constant K for association and dissociation reactions. Extremum not determinable where numbers are missing. Units are kJ/mol for energies, $l s^{-1} mol^{-1}$ for rate constants. Errors estimated 4.5 in above units on average for $\Delta^\ddagger G$ and $\lg(k)$.

The equilibrium constants for the simulated reactions are all fairly small with values ranging between 10^{-4} and 10^5 . Reaction rate constants for association and dissociation reactions are close in most cases, the absolute values of the rate constants are in the typical range for simple radical or ion combination reactions. The rate constant for the adsorption of a calcium ion to a carboxylic acid is also comparable to the other models for the association process but is strikingly low for the dissociation. This is reflected in the equilibrium constants which are between 10^{-3} and 10^2 for most of the reactions, but 10^4 to 10^5 for the calcium association to a carboxylic acid. All of the association reaction rate constants range between 10^9 and 10^{12} , so the adsorption of all ions should be roughly equally frequent for all ions. The adsorption of calcium ions to the telopeptide is however by far the most stable one as the dissociation is

very slow. This is at strong hint on the importance of calcium for the initiation of apatite nucleation. It is proposed that calcium is the species that initially attaches to the telopeptide, whereupon protein structure rearrangements and subsequent ion adsorptions follow.

Association of a calcium ion to a carboxylic acid group

As mentioned above, the attachment of a calcium ion to a carboxylic acid will be examined in detail. The PMF shows a unique small set of local maximum and minimum within the direct contact ion pair minimum. To fathom the origin of this feature, the integral of the g -functions for calcium (CAL) and the oxygen atoms of water (OT) and of the carboxyl group (OC) are calculated for each pmf constraint distance and plotted below. The function G_{12} marks the number of atoms of type 2 within a given distance of atoms of type 1.

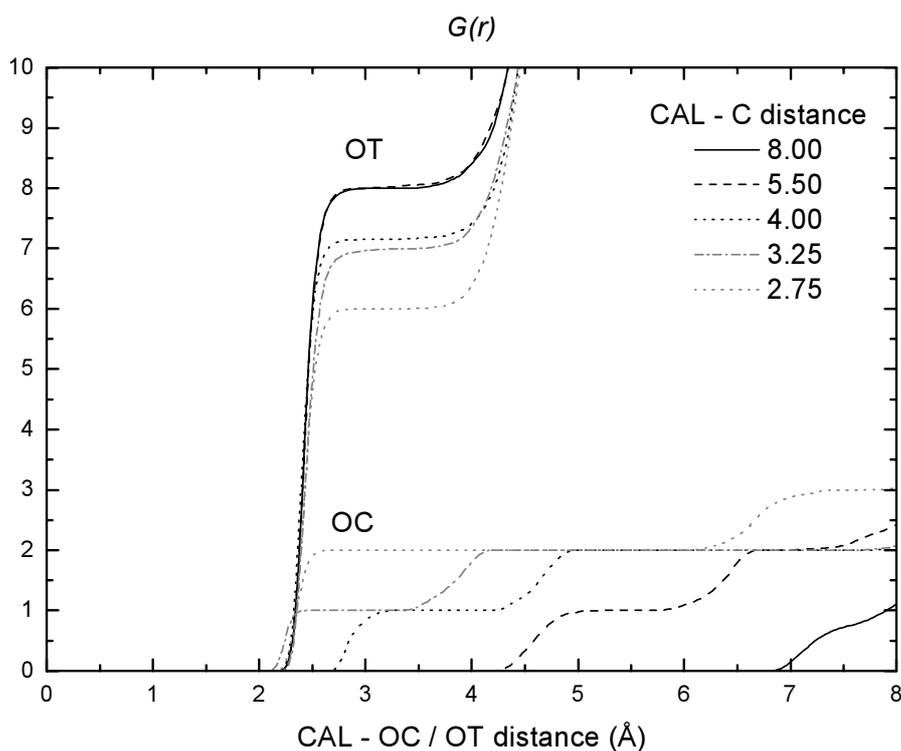


Fig. 4.41: Integral G of the g -functions for CAL and OC / OT at different pmf constraint distances (drawn in different shading / dashing patterns). Lower curves for CAL - OC, upper curves for CAL - OT.

The pmf constraint is defined between the calcium ion and the carbon atom of the carboxyl group. This leaves the orientation of the carboxyl group free, so that it is not introduced to the system as a predetermined constraint but can rather be obtained from the simulations. From

figure 4.41, the progression of the ion adsorption process can be rationalized. The "bond" length of associated calcium and oxygen is approximately 2.5 Å (the term "bond" is used in a loose definition here for any associated atoms). To judge which oxygen atoms are bonded to calcium the area around 2.5 to 3.0 Å of the G function is thus important. At 8.0 and 5.5 Å, eight OT are bonded to calcium resembling the hydration shell. At 4.0 Å an OT is fully expelled from the hydration shell, while at the same time the first OC is coming very close to CAL. At 3.25 Å, the OC is fully integrated into the coordination of CAL, and another OT is starting to be expelled from the hydration shell. At 2.75 Å finally, the second OT is completely replaced by the second OC. The chronologically successive steps of the rejection of a water molecule, the coordination of the first carboxyl oxygen, the rejection of the second water molecule and the coordination of the second carboxyl oxygen can be well observed. The described mechanism was observed in a free simulation as well and is illustrated in the figure below.

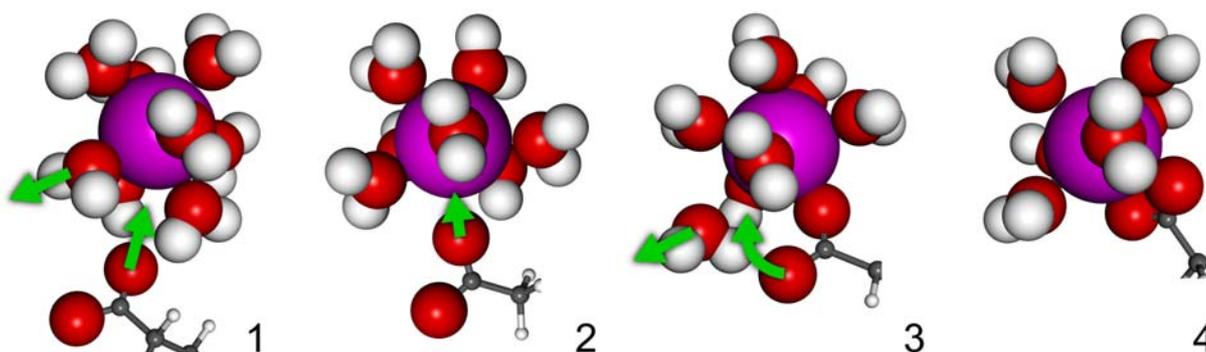


Fig. 4.42: Two-step mechanism of the attachment of a calcium ion (purple) to a carboxylic acid group.

The determined binding distance (global minimum in PMF) of approximately 2.75 Å is in good agreement with distances found in X-ray structures collected in the pdb database (see chapter 4.4.1). A multi-step mechanism for the association of calcium to carbonyl groups was also observed by Grubmüller et al.⁷⁰ in a similar scenario. The binding of calcium to phospholipids follows a similar mechanism where up to four carbonyl groups of different lipids are complexing the calcium ion one after another.

Monitoring the distances between the calcium ion and the individual OC atoms of the carboxyl group (fig. 4.43) illustrates the nature of the local PMF minimum at 3.5 Å. It can be

seen that the second OC atom is not yet bound to calcium at $d = 3.50 \text{ \AA}$ but is so at $d = 2.75 \text{ \AA}$.

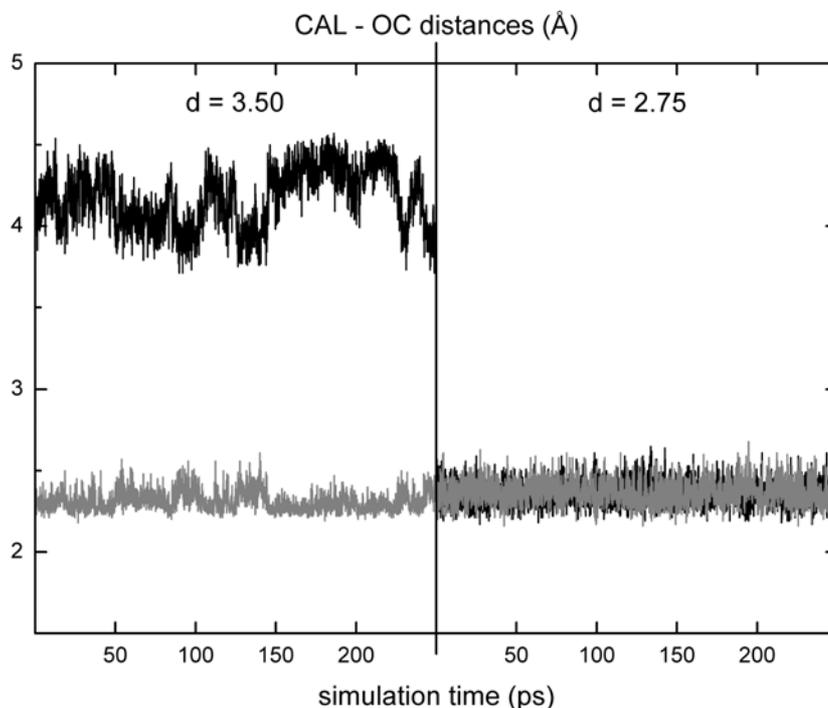


Fig. 4.43: Distance of CAL to OC1 and OC2 (black and gray) at constraint distances 3.50 \AA (local PMF minimum) and $d = 2.75 \text{ \AA}$ (global minimum). Note two independent simulations plotted on left and right.

It can also be noticed that the adsorption of calcium to a carboxylic acid has the deepest second minimum in the PMF with the highest dissociation barrier of 3.6 kJ/mol . This means that the solvent separated ion pair is most stable for these compounds compared with the other models. Free simulations perfectly support this finding as calcium ions are observed staying in a distance of approximately 5 \AA of a carboxyl group for as long as a nanosecond. To find out why the solvent separated ion pair is so stable, the respective simulation is analyzed. It is found that the calcium ion is bound to the carboxyl group via two stable water mediated hydrogen bonds. During 72 % of the total simulation time, two hydrogen bonds exist, 28 % of the time one hydrogen bond, and only 0.3 % of the time no hydrogen bond is detected. The mean number of hydrogen bonds during the simulation is 1.7. The timeline plot (fig. 4.44) shows that the hydrogen bonds are mediated by mainly three individual water molecules, one of which is almost constantly involved. The results prove a very stable solvent separated complex at a distance of 5 \AA with only minimal exchange of the involved water molecules.

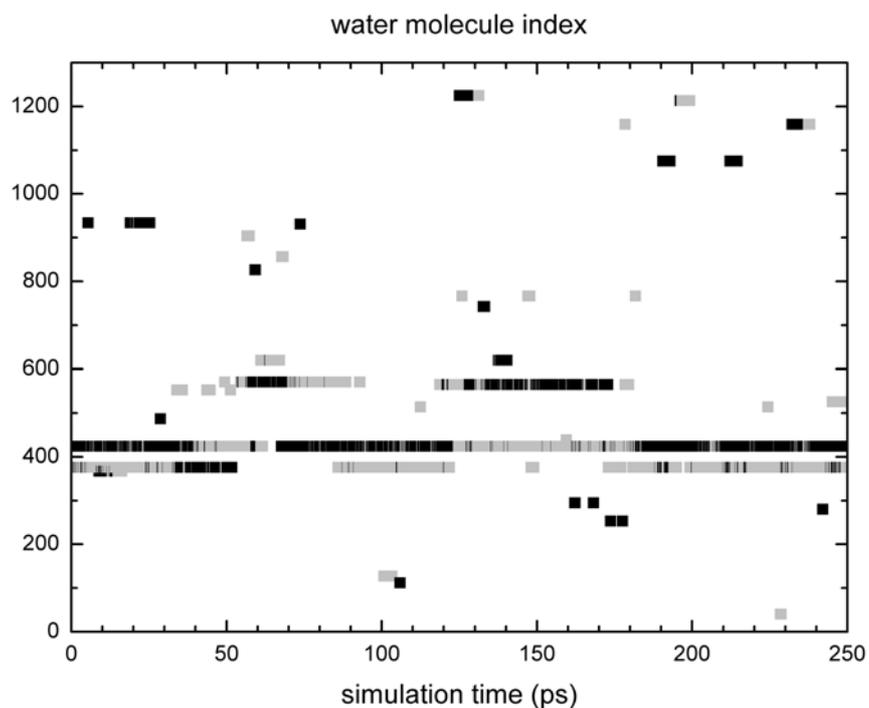


Fig. 4.44: Timeline plot of water mediated hydrogen bonds between carboxyl group and calcium ion. Index of hydrogen bond mediating water molecule plotted on ordinate (black: bond to OC1, gray: bond to OC2).

The calcium ion and its hydration shell can thus in good approximation be regarded as one particle, which is bound to the carboxyl group via two hydrogen bonds in a preserved orientation. A snapshot of this situation taken from the simulation is shown in figure 4.45.

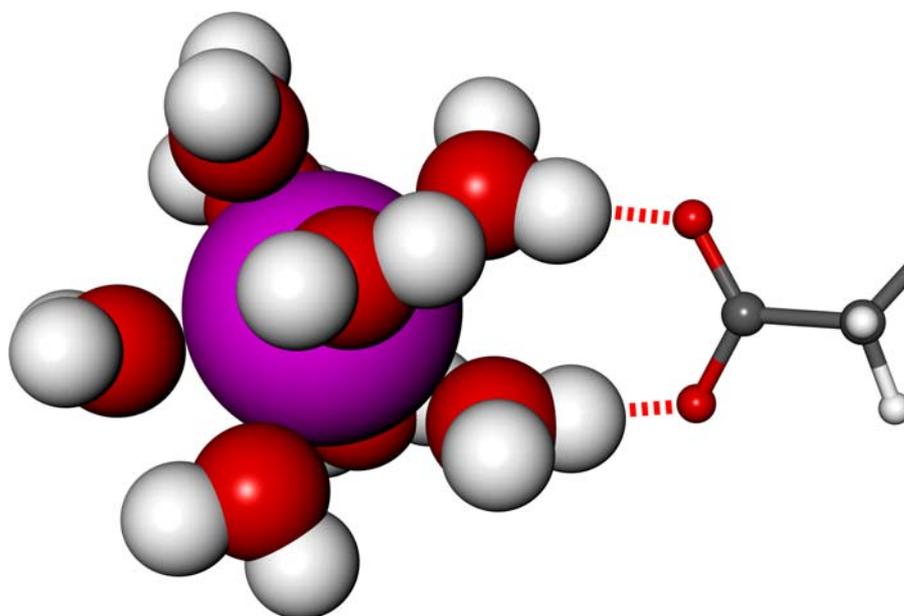


Fig. 4.45: Complex of calcium ion bound to carboxyl group via two water mediated hydrogen bonds. Complete first hydration shell of calcium is shown. Complex represents solvent separated minimum in the PMF.

4.6.2 Teloepptide

To test the validity of the initial assumption that the PMF is mainly dictated by the nearest chemical surroundings of a binding site, the PMF for a glutamic acid site of the teloepptide is calculated with the full teloepptide for comparison. This is an expensive calculation which is therefore only conducted for one test case. The residue B_GLU8 was chosen as it is located in a typical environment with several other charged residues in the near vicinity (fig. 4.46).

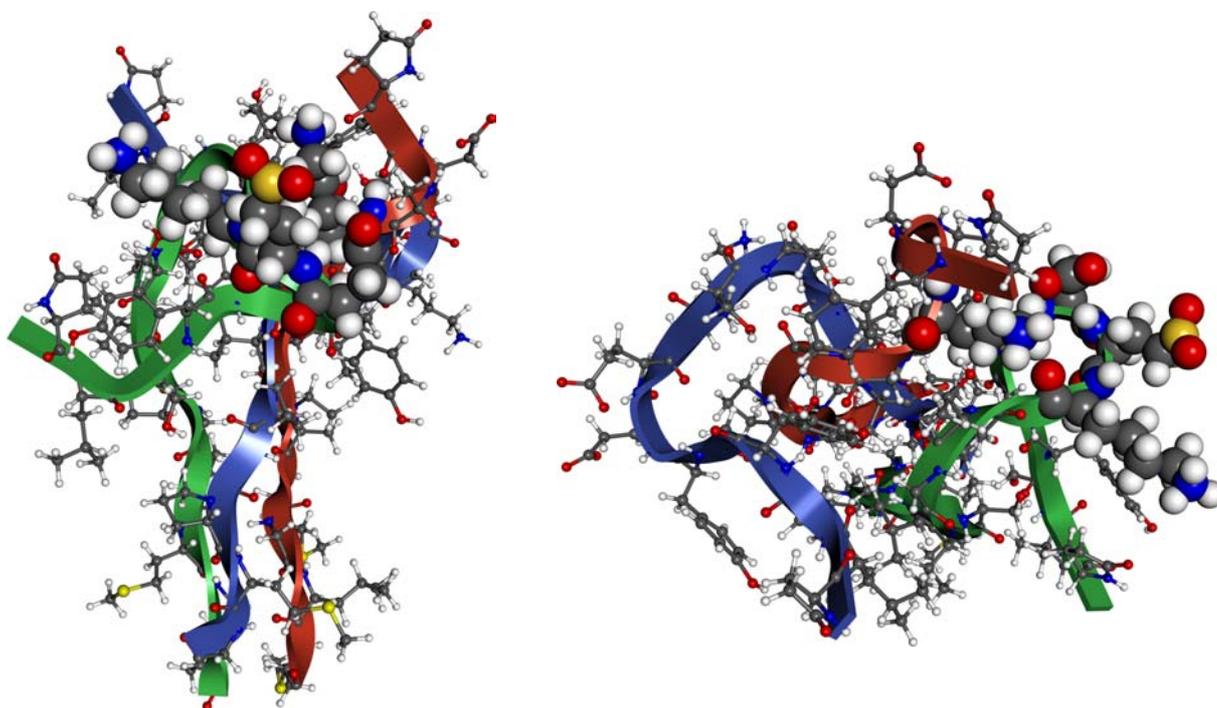


Fig. 4.46: View of teloepptide binding site B_GLU8 chosen as test case for PMF comparison with model proteins. Protein: ball and stick with ribbons, B_GLU8 and nearby charged residues: small CPK. The carboxylic carbon atom of B_GLU8 is marked yellow.

The same simulation system was used as for the ion attachment simulations in chapter 4.4. In a procedure analogous to that applied to the model proteins, the mean volume of the box was determined so the PMF simulations could be performed in an *NVT* ensemble where mean force fluctuations are smaller. Likewise, the PMF calculation procedure as described for the model proteins was used here as well. The results of the calculation are shown below.

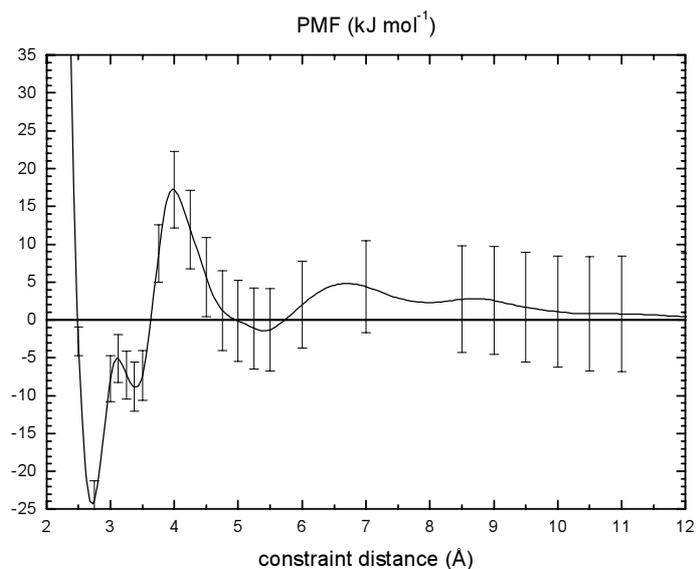


Fig. 4.47: PMF of the attachment of a calcium ion to B_GLU8 of the telopeptide with the full protein simulated.

Both PMFs (model 1 GLU and the full telopeptide) are shown in one plot below for comparison. Despite of the vicinity of B_GLU8 to one Asp and two Lys residues, the similarity of the PMF curves for the isolated Glu in the model protein and Glu in the telopeptide environment is remarkably high. Within the error margins, the PMFs can be regarded as identical, as was already found for the PMFs of glutamic acid and aspartic acid.

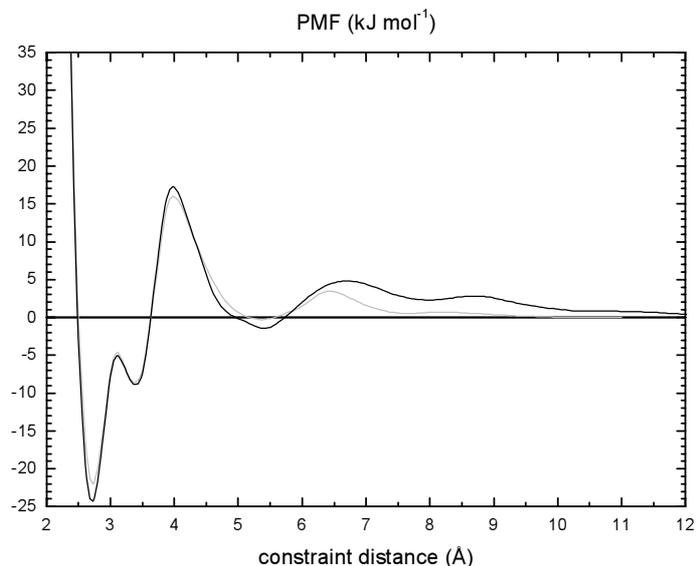


Fig. 4.48: PMF of calcium ion attachment to full telopeptide (black) and model 1 GLU (gray) in comparison. Error bars are omitted for better clarity.

Because of the high computational cost, only one such test case could be examined, and the results cannot be interpreted as a proof. Nevertheless, they are a strong hint that confirms the validity of the approximation that the attachment of ions to binding sites is in good approximation independent of the chemical environment in three or more Ångströms distance but mainly depends on the binding site itself.

The results are encouraging the pursuit of the coarse model development. The development of a full working model exceeds the possibilities of this work, however a good start for future efforts could be given here.

5 Summary, conclusions and outlook

Summary and conclusions

The intention of this work was to study the calcification of collagen with computational methods. The apatite/collagen system should serve as an example scenario of biomineralization, or more general, composite formation. The application of theoretical methods to this problem is a new approach that is expected to be capable of supporting the experimental studies that are on the edge of their possibilities. Simulation techniques can help to gain insight into processes at the atomic level. This can be very helpful to understand the mineralization mechanisms.

The apatite/collagen system was chosen as an example scenario because it is a very interesting example of biomineralization. Bone and teeth are made of these components, and many ideas exist in dentistry and surgery that could be realized when the composite formation is understood better and can be steered or controlled.

There were several questions that were of interest. What are the structures and dynamics of the collagen telopeptides, and what are the important differences to the triple helix? How can the calcification of collagen be simulated? What are the energies and reaction rates concerning the ion binding to the protein? Is there any evidence for a special function of the telopeptides for the nucleation of apatite?

To simulate the calcification of collagen, we first have to know the structure of collagen. Unfortunately it is not completely known. While good models exist for the triple helix, the structures of the N- and C-terminal telopeptides remain unknown. Several studies on the structure of the N-terminal collagen telopeptide have not led to a consensus model but are still in an ongoing controversy. Contracted and rigid as well as extended and flexible structures have been proposed. Particularly the dynamics of the protein dictate the theoretical methods that can be applied. The calcification at a folded protein might even be described using a static protein model. Proteins of high flexibility however need to be treated dynamically. Large proteins cannot be simulated with normal MD for very long timescales but have to be tackled with mesoscale methods. The examination of protein structures and dynamics of mediocre

size is most accurately done with atomistic MD simulation. Therefore, small subsections of collagen were simulated as model systems instead of the complete molecule. This way, a detailed analysis can be performed on important features of collagen.

First, a sequence analysis of the full collagen molecule was performed to possibly find distinct features of importance for mineral nucleation. Yet the analysis did not reveal any significant differences between the telopeptides and the triple helix except the well-known repetitive Pro and Gly pattern. The distribution of charged amino acids within the molecule, which could give a first hint on preferred ion binding sites, was found to be irregular, but similar throughout the molecule. Groups of two or more directly adjacent charged amino acids were found in the N-telopeptide as well as at four positions in the triple helix. Assuming the highest likelihood of ion binding in such a region, the sequence around one of these positions was chosen as a model system for simulation. An idealized backbone structure was generated which was then mutated into the section of the real collagen sequence. The N-terminal model system was built upon modeling results by Scheraga et al.. The force field was built on the charmm22 parameters. Missing amino acids such as hydroxyproline and pyroglutamic acid were taken from the literature and implemented into the charmm force field. The parameters for apatite ions were taken from coworkers and the OPLS force field. Simulations of 20 ns were performed with both systems to gather data for various analyses. These are the first simulations that were performed on the apatite/collagen system including both protein and ions. The binding distances of ions to charged amino acids and the coordination number of the ions in water were compared to literature values and proved the validity of the force field.

An analysis of the structure and dynamics of the proteins was performed. In the course of this task, a new measure for the flexibility or mobility of atoms was developed. The measure referred to as RMSmob proved to be superior over the normal root mean square calculation of the atoms position. It is a grid based approach which does not depend on a reference position for the RMS calculation but rather evaluates the distribution of an atom in space. It was found that the triple helix has a stable secondary and tertiary structure, which is stabilized through a well-known hydrogen bond pattern. This pattern could be reproduced in the current study. Nevertheless, relatively strong fluctuations of the structure were observed. The dihedral angles of the individual helices show fluctuations of several tens of degrees around their optimum value. The superhelix has significant flexibility in the center of the model molecule.

It was found that a kink can develop dynamically in the center of the structure, which through repeated formation and disappearance causes the observed flexibility. This phenomenon was observed by other groups in modeling and experimental works as well. The ends of the model molecule are naturally more flexible as the stabilization by a continuing triple helix is missing. This phenomenon is solely owed to the limited model structure, as there are no such cuts in the real collagen molecule. This was nevertheless also observed in experiments on short model triple helices.

The telopeptide is an entirely different structure. It can be divided into two domains. The segment directly adjacent to the triple helix is characterized by multiple hydrogen bond stabilization. The three collagen chains are still entangled here, but with less definition compared to the triple helix. The hydrogen bonds are fluctuating and the structure shows flexibility and no secondary structure motif. The very N-terminal segment is characterized by an extremely high flexibility. No salt bridges, hydrogen bonds or any secondary structure was observed, the individual chains are completely separated from one another and spread apart. A graphic illustration could be tentacles moving freely and continuously in space. The findings are an important contribution to the current debate in the literature. The high mobility of the telopeptide ends could possibly adapt to a scenario of multiple ion binding and cluster formation. This supports the idea of a special role of the telopeptides for mineral nucleation.

The distances of like charged groups were monitored in the trajectories to evaluate whether an accumulation of them occurs somewhere just by random motions of the protein. This could be a hint on a stable chelate type ion binding site, which in turn could be an initial point of mineral nucleation. The analysis revealed, however, that telopeptide and triple helix do not differ in this respect. Of course, the event of accumulating like charged amino acids to bind to an oppositely charged ion could also be induced by an approaching ion itself, which would not be observed here.

The results up to this point indicated strongly that a static description of the nucleation of inorganic crystallites at collagen would not make sense. Such large systems, however, cannot be simulated with normal MD on timescales required for the nucleation of crystallites. As a consequence, the development of a mesoscale model is necessary. A united residue protein model with ion binding sites modeled with special interaction potentials could be a first

approach. With this idea in mind, several analyses were performed to explore the present chemical system and evaluate the applicability of such a mesoscale model.

At first, the possible ion binding sites had to be identified. The electrostatic potential was mapped on the molecular surface of the telopeptide, and all sites with a considerable potential were registered. A counterion was placed at each site and a relaxation simulation performed to test the reaction of the protein. In none of the cases a significant rearrangement of the protein structure could be observed. All structural changes remained within normal thermal motions. It was thus justified to assume independent binding sites that do not depend on protein conformation.

The energetics of ion adsorption reactions were studied using potential of mean force calculations. These are fairly expensive calculations, so a minimization of computing needs was sought. Therefore the determined binding sites were categorized according to chemical similarity. The simulations could not be performed with the complete telopeptide either, so model proteins of two to four amino acids size were created to simulate just the binding site and its close vicinity. The results of the PMF calculations showed that only the binding site with the nearest vicinity of two or three neighboring atoms and the ion dictate the PMF. Several of the chosen site types had very similar PMFs, and conformational variations had only little influence on the PMF. The effect of the chemical environment in three or more bind lengths distance seems negligible.

Positive, negative and near zero binding free energies were found. The by far strongest complex is formed by a calcium ion bound to a carboxyl group of glutamic or aspartic acid. This is readily understood with the Pearson HSAB concept⁷¹. Among the studied species, calcium is the hardest acid (high charge, small, low polarizability) and the carboxyl groups the hardest base, which combine to a stable complex according to Pearson. A unique two-step mechanism was found for the adsorption. It was also found that lysine is the only site in the telopeptide that binds hydrogen phosphate and fluoride stably. This might be a reason for a potential special importance of lysine for the nucleation process. The adsorption reaction of calcium and fluoride ions have the highest barriers, so that the adsorption of these ions will most likely be the rate-determining steps for the nucleation. Since the desorption of a calcium ion is the slowest reaction, it stays bound longest and allows the protein to rearrange optimally for stabilization and further ion adsorption.

To test the assumption that binding sites are independent of the farther chemical environment, a PMF with the full telopeptide was calculated with a randomly chosen adsorption site. The obtained PMF was equal to that of the corresponding model protein in good approximation. Deviations occurred mainly at the far distance end, the first minimum and barrier and thus the binding energies were equal within the error margins. This finding supports the assumption that a self-contained description of binding sites is valid.

The development of a coarse model was shown to be applicable, and a basis was created with PMF data which can be used for a parametrization.

Computational experiments were set up to get an impression of the behavior of the telopeptide in a complex with multiple ions. Starting configurations were prepared where several ions were placed near binding sites. The configurations were taken from the long time simulations by choosing frames from the trajectory where several charged amino acids were relatively close to each other. In one of the cases, a complex was formed that was completely stable during a 2 ns simulation. The complex was characterized by many ionic interactions and mutual stabilization. The principle possibility of stable larger complexes could be demonstrated at this example.

The results of this work encourage further studies. Biocomposites such as the studied apatite/collagen system are very complex featuring multitudinous degrees of freedom, which makes systematic studies very difficult. The experiments done here are not systematic and complete, but they are first approaches to the extremely complex nucleation problem. The basis for further, more efficient coarse model simulations was laid in this work. The actual development of the model goes beyond the scope of this work, but that will be a very promising project for future research in this field.

Outlook

As a start, the structure of the N-terminal collagen telopeptide was investigated. This has to be completed by the C-terminal telopeptide. More effort is necessary to obtain a reasonable model system here, since to date no detailed experimental or modeling approaches have been

performed. A folding analysis of the C-terminal telopeptide will be laborious due to its size. Simulations with coarse models parametrized based upon the results of this work are a sensible and needed continuation of the project in the future. Mineralization can then be studied on longer time scales and on larger molecular systems. For this, collagen can be described by a reduced fragment-based representation with active sites modeled with the help of the obtained potential of mean force functions. An explicit solvent would not be necessary since it is included in the PMFs.

With increasing computing power, a long time simulation with grid based analysis of ion population could be performed. Moreover, integral theory based methods could be applied to quickly assess ion distributions for snapshots of the dynamic telopeptide structure.

6 Zusammenfassung und Ausblick

Die Aufgabe in dieser Arbeit war es, die Kalzifizierung von Kollagen mit theoretischen Methoden zu untersuchen. Das Apatit/Kollagen-System sollte dabei als ein Beispielszenario für die Biomineralisation bzw. noch allgemeiner für Kompositbildung dienen. Insbesondere sollten die Untersuchungen darauf abzielen, Aspekte der initialen Nukleation von Apatit am Kollagen zu beleuchten.

Die Anwendung theoretischer Methoden auf diesem Gebiet soll neue Erkenntnisse bringen, um die experimentelle Forschung zu unterstützen. Simulationstechniken ermöglichen Einsichten in chemische Prozesse auf atomarer Ebene, welche helfen können, ein mechanistisches Verständnis der Prinzipien der Kompositbildung zu erlangen.

Das System Apatit/Kollagen ist ein sehr interessantes Beispiel von Biomineralisation, Knochen und Zähne bestehen aus einem solchen Kompositmaterial. Im Bereich der Zahnmedizin und Chirurgie existieren viele Ideen für Anwendungen, die durch eine angestrebte Kontrolle bzw. Steuerung der Biomineralisation ermöglicht würden.

Einige Fragen waren von direktem Interesse. Wie sieht die Struktur und Dynamik der Kollagen Telopeptide aus? Was sind wichtige Unterschiede zur Tripelhelix? Wie kann die Biomineralisation simuliert werden? Was sind die Reaktionsenergien und -geschwindigkeiten bei der Anlagerung von Ionen an Kollagen? Gibt es Hinweise für eine spezielle Funktion der Telopeptide für die Nukleation von Apatit?

Um die Kalzifizierung von Kollagen zu untersuchen, muss man natürlich zuerst die Struktur des Kollagens kennen. Leider ist diese nicht vollständig bekannt. Für die Tripelhelix existieren gute Modelle, die Struktur der N- und C-terminalen Telopeptide ist hingegen unbekannt. Nur sehr wenige experimentelle Daten liegen vor. Diverse Modellierungstudien zum N-Telopeptid haben nicht zu einem übereinstimmenden Bild geführt, sondern zu einer andauernden Kontroverse. Eine globuläre rigide Struktur wurde ebenso vorgeschlagen wie eine flexible ausgestreckte Struktur.

Insbesondere die Dynamik der Peptide diktieren die Auswahl an theoretischen Methoden, die sinnvoll angewendet werden können. Bei einer gefalteten Struktur wäre eine rein statische Methode unter Umständen ausreichend. Peptide von hoher Flexibilität müssen aber dynamisch beschrieben werden. Große Proteine können auf großen Zeitskalen nicht mit regulärer Molekulardynamik simuliert werden, hier müssen mesoskopische Methoden angewendet werden. Da die Untersuchung von Struktur und Dynamik von Proteinen mit der Molekulardynamik aber die genaueste Methode ist, wurden möglichst repräsentative Untereinheiten des Kollagens als Modellsysteme zur Simulation ausgewählt, die auf hinreichend großen Zeitskalen simuliert werden können. Auf diesem Weg konnte eine detaillierte Analyse der wichtigen Strukturmerkmale des Kollagens durchgeführt werden.

Als Erstes wurde eine Sequenzanalyse des vollständigen Kollagenmoleküls durchgeführt, um eventuelle besondere Merkmale in der Primärstruktur zu finden, die von Wichtigkeit für die Mineralisation sein könnten. Die Analyse zeigte jedoch keine signifikanten Unterschiede zwischen Telopeptiden und Tripelhelix auf, abgesehen von dem bekannten Pro-Gly-Muster in der Tripelhelix. Die Verteilung von geladenen Aminosäuren, die einen ersten Hinweis auf bevorzugte Bindungsplätze für Ionen hätte geben können, war unregelmäßig, aber ähnlich innerhalb des gesamte Moleküls. Gruppen von zwei oder mehr direkt benachbarten geladenen Aminosäuren fanden sich sowohl in den Telopeptiden als auch an vier Positionen in der Tripelhelix. Diese Bereiche wurden als potentielle Bindungsregionen angesehen und daher die Sequenz eines dieser Abschnitte als Modellsystem ausgewählt. Eine idealisierte Tripelhelix mit der entsprechenden Sequenz wurde generiert.

Ein Modell für das N-Telopeptid wurde anhand der Ergebnisse einer Modellierungstudie von Scheraga et al. erstellt. Das Kraftfeld wurde aufbauend auf den charmm22-Parametern generiert, wobei Parameter für die seltenen Aminosäuren Hydroxyprolin und Pyroglutaminsäure aus Literaturquellen entnommen wurden. Parameter für Apatitionen wurden von Hauptmann et al. und aus dem OPLS Kraftfeld übernommen. Simulationen von 20 ns wurden für beide Systeme durchgeführt. Dieses sind die ersten Simulationen für ein kombiniertes Apatit/Kollagen-Modellsystem, das sowohl Protein als auch Ionen berücksichtigt. Die Bindungsabstände von Ionen an geladenen Gruppen sowie die Koordinationszahlen der Ionen in Wasser wurden mit Literaturwerten verglichen (siehe Kapitel 4.4.1). Es wurde eine sehr gute Übereinstimmung gefunden, was die Güte des generierten Kraftfelds bestätigt.

Im Zuge der Analyse von Struktur und Dynamik der Proteine wurde ein neues Maß für die Mobilität von Atomen entwickelt, welches als RMSmob bezeichnet wird. Es handelt sich dabei um einen Grid-basierten Ansatz, der keine Referenzposition zur Berechnung der RMS-Abweichung benötigt. Die dreidimensionale Verteilung der Atome im Raum wird zunächst im Gitter aufgetragen und dann die quadratisch gemittelte Aufenthaltszeit in einer Gitterzelle berechnet. Aus dem Reziprokwert wird in bezug auf die Gesamtsimulationszeit eine Mobilität errechnet. Es zeigte sich, dass der RMSmob-Wert die Beweglichkeit bzw. Mobilität der Atome genauer quantifiziert als die üblicherweise berechneten RMS-Werte von Atompositionen.

Für die Tripelhelix wurde eine stabile sekundäre und tertiäre Struktur gefunden, die durch ein bekanntes Muster aus Wasserstoffbrückenbindungen stabilisiert wird. Das Muster konnte in den Simulationen reproduziert werden. Trotzdem wurden relativ starke Fluktuationen der Struktur beobachtet. Die Diederwinkel der einzelnen Helices variierten um einige zehn Grad um ihren optimalen Wert herum. Die Superhelix zeigte eine signifikante Flexibilität in der Mitte des Moleküls. Genauere Analysen zeigten, dass eine dynamische Knickstelle in der Mitte des Moleküls existiert. Durch wiederholtes Ausbilden und Verschwinden dieses Knicks entsteht die beobachtete Mobilität im mittleren Bereich des Moleküls. Dieses Phänomen wurde auch von anderen Gruppen experimentell beobachtet, konnte jedoch bisher nicht erklärt werden. Die Enden des Moleküls sind naturgemäß etwas mobiler als der Rest, da hier die Stabilisierung einer fortgesetzten Tripelhelix fehlt. Diese Erscheinung ist nur der Tatsache geschuldet, dass das Modellsystem aus einem ausgeschnittenen Teil der Tripelhelix besteht. Im realen Kollagenmolekül gibt es diese Schnitte selbstverständlich nicht. Nichtsdestotrotz wird auch dieses in experimentellen Studien an kurzen Modellpeptiden beobachtet, was die Güte der Ergebnisse stützt.

Das N-Telopeptid besitzt eine vollkommen unterschiedliche Struktur. Zwei Bereiche können unterschieden werden. Ein neun Aminosäuren langer Bereich, der sich direkt an die Tripelhelix anschließt, stellt das erste Segment dar. Die drei Kollagenstränge sind hier noch ineinander verschränkt, jedoch in viel weniger definierter Struktur als in der Tripelhelix. Wasserstoffbrückenbindungen sind hier dynamisch und die Beweglichkeit der Struktur steigt zum N-Terminus hin. Sekundärstrukturmotive gibt es nicht. Das zweite Segment wird aus den verbleibenden N-terminalen Resten gebildet. In diesem Bereich gibt es keinerlei geordnete Struktur und keine Wasserstoffbrückenbindungen. Die einzelnen Stränge des Kollagens sind

komplett voneinander getrennt und abgespreizt. Sie bewegen sich kontinuierlich frei im Raum. Für eine anschauliche Illustration könnte man sie als Tentakeln bezeichnen.

Die Befunde stellen einen wichtigen Beitrag zur Aufklärung der Struktur der Telopeptide dar, wobei der letzte Beweis eines Strukturvorschlags immer durch das Experiment gegeben werden muss. Durch die hohe Mobilität der N-terminalen Telopeptide besteht prinzipiell die Möglichkeit, dass sich die Struktur für eine Bindung von mehreren Ionen bzw. Apatit-Nanoclustern anpassen kann. Dies ermutigt zu weiteren Untersuchungen über eine spezielle Rolle der Telopeptide für die Nukleation von Mineralien.

Wenn zwei oder mehr gleichnamig geladene Gruppen in unmittelbarer Nähe zueinander sind, könnte mit einem anlagernden Ion eine Art Chelatkomplex entstehen, welcher besonders stabil ist und daher initiale Nukleationsstellen für eine Mineralisierung darstellen könnte. Darum wurden in den Simulationen die Abstände geladener Gruppen verfolgt, um Hinweise zu erlangen, ob schon durch die zufälligen thermischen Bewegungen des Proteins eine Ansammlung solcher Gruppen entstehen kann. Die Auswertung ergab aber, dass sich Telopeptid und Tripelhelix in dieser Hinsicht nicht unterscheiden. Die Möglichkeit, dass gleichnamig geladene Gruppen erst durch die Anwesenheit des passenden Gegenions zu einem Komplex zusammengelagert werden, wurde in dieser Untersuchung nicht untersucht.

Die Ergebnisse bis zu diesem Punkt haben bewiesen, dass eine statische Beschreibung der organischen Komponente bei der Kalzifizierung von Kollagen nicht sinnvoll ist. Systeme dieser Größe können aber nicht komplett mithilfe der Molekulardynamik auf einer Zeitskala beschrieben werden, die für die Nukleation von Kristalliten nötig ist. Die Konsequenz ist, dass die Entwicklung mesoskopischer Modelle notwendig ist. Ein vernünftiger Ansatz ist die Bildung eines "united-residue" Kraftfeldes, bei dem komplette Aminosäuren als einzelne Zentren modelliert werden und die Wechselwirkung mit Ionen durch spezielle Potentiale beschrieben werden. Es wurden zahlreiche Analysen durchgeführt, um auf ein derartiges Modell hinzuarbeiten und die Chemie des vorliegenden Systems zu charakterisieren und die Anwendbarkeit eines mesoskopischen Ansatzes zu bewerten.

Zuerst mussten mögliche Bindungsstellen für Apatitionen identifiziert werden. Dazu wurde das elektrostatische Potential auf die molekulare Oberfläche des Proteins projiziert und alle Plätze mit merklich positivem oder negativem Potential registriert. Diese Auswahl wurde

getestet, indem jeweils Gegenionen an die Plätze gebracht wurden und in Simulationen die "Reaktion" des Proteins beobachtet wurde. Zum Teil blieben die Ionen gebunden, zum Teil lösten sie sich recht schnell wieder ab. In keinem Fall konnte jedoch eine deutliche Umlagerung im Protein beobachtet werden, alle Strukturänderungen des Proteins blieben im Rahmen der normalen thermischen Bewegungen. Die Annahme von unabhängigen Bindungsplätzen, die in erster Näherung unabhängig von der Konformation des Proteins sind, scheint somit gerechtfertigt.

Die Energetik der Ionenbindungsreaktionen wurde mithilfe von "potential of mean force"-Rechnungen untersucht. Diese Rechnungen sind recht teuer in bezug auf den Verbrauch an Rechenleistung, daher wurde eine Minimierung der benötigten Ressourcen angestrebt. Die zuvor bestimmten Bindungsplätze wurden daher nach chemischer Ähnlichkeit kategorisiert. Eine PMF-Rechnung unter Simulation des gesamten Telozeptids wäre nicht möglich gewesen, daher wurden für jeden Bindungsplatz-Typ ein kleines Modellprotein generiert, um nur die Bindungsstelle und ihre nahe Umgebung zu simulieren. Die Ergebnisse der PMF-Rechnungen bestätigten, dass im Wesentlichen nur die nächste Umgebung, bestehend aus der Bindungsstelle und zwei oder drei benachbarten Atomen, und das Ion selbst das PMF bestimmen. Mehrere der Bindungsplätze zeigten sehr ähnliche PMFs, außerdem spielte die lokale Konformation nur eine geringe Rolle. Der Effekt der chemischen Umgebung in drei oder mehr Bindungslängen Entfernung zeigte sich vernachlässigbar gering.

Unter den gefundenen Bindungsenergien gab es positive, negative und solche nahe null. Der mit Abstand stabilste Komplex ist ein an eine Carboxylgruppe gebundenes Calciumion. Dies kann mit dem Pearsonkonzept⁷¹ nachvollzogen werden, nach dem Calcium eine harte Säure und die Carboxylgruppe eine harte Base ist, die nach dem Konzept einen stabilen Komplex ergeben. Für die Reaktion der Anlagerung des Calciumions an eine Carboxylgruppe wurde ein besonderer zweischrittiger Mechanismus gefunden. Im Telozeptid stellte Lysin die einzige Aminosäure dar, an die Hydrogenphosphat und Fluorid stabil gebunden waren. Aus diesem Grund könnte Lysin von besonderer Wichtigkeit für die Nukleation sein. Die Anlagerungen von Calcium- und Fluorid-Ionen haben die höchsten Barrieren, so dass diese Spezies wahrscheinlich geschwindigkeitsbestimmend für die Nukleation sein werden. Da Calcium am stabilsten gebunden werden kann, wird es somit am längsten gebunden bleiben und Umlagerungen des Proteins erlauben, ohne zu desorbieren. Aus diesem Grund wird ein angelagertes Calciumion wahrscheinlich das Zentrum für einen Nukleationsprozess sein.

Um die Annahme zu testen, dass die Bindungsplätze unabhängig von der entfernteren chemischen Umgebung sind, wurde eine PMF-Rechnung an einer zufällig gewählten Bindungsstelle mit dem vollen Telopeptid zum Vergleich mit den Modellproteinen durchgeführt. Das PMF war in guter Näherung identisch mit dem für das entsprechende Modellprotein berechneten. Abweichungen waren hauptsächlich im Bereich großer Abstände von Ion und Bindungsstelle zu sehen, die im Bereich des statistischen Fehlers lagen. Die Bindungsenergien waren identisch im Rahmen des Fehlers. Dieses Ergebnis unterstützt die Annahme von unabhängigen Bindungsplätzen.

Der Einsatz vergrößerter Modelle ist nach allen bisherigen Ergebnissen sinnvoll. Eine Basis für deren Entwicklung wurde mit den PMF-Daten hergestellt, die für die Parametrisierung eines Modells herangezogen werden können.

Um einen Eindruck vom Verhalten des Telopeptids zu bekommen, wenn es in Wechselwirkung mit mehreren Ionen steht, wurden mehrere In-silico-Experimente durchgeführt. Es wurden einige Startkonfigurationen vorbereitet, indem bei Schnappschüssen vom Telopeptid aus der Langzeitsimulation mehrere Ionen gleichzeitig direkt an Adsorptionsstellen platziert wurden. Es wurden Schnappschüsse ausgewählt, bei denen einige geladenen Gruppen schon relativ nah beieinander waren. In einem der Experimente bildete sich ein Komplex des Telopeptids mit fünf Ionen, der über 2 ns stabil blieb. Der Komplex war durch viele ionische Wechselwirkungen charakterisiert, wobei die Ionen in ihrer Anordnung auch eine deutliche gegenseitige Stabilisierung zeigten. Die prinzipielle Möglichkeit des Telopeptids, größere stabile Komplexe mit Ionen zu formen, wurde an diesem Beispiel erfolgreich demonstriert.

Die Ergebnisse dieser Arbeit ermutigen zu weiterführenden Untersuchungen. Biokomposite wie das studierte Apatit/Kollagen-System sind sehr komplex und ihre Bildung hängt von vielen Faktoren ab. Dies macht eine systematische Untersuchung sehr schwierig. Die hier durchgeführten In-silico-Experimente können keinen Anspruch auf Vollständigkeit erheben. Sie gehören jedoch zu den ersten Modellstudien, die sich speziell dem komplizierten Thema Biomineralisation zuwenden, und liefern wertvolle Informationen hierzu. Die Basis für zukünftige Arbeiten auf der Basis von effektiveren mesoskopischen Modellen wurde in

dieser Arbeit gelegt. Die Entwicklung eines Modells selbst übersteigt den Rahmen dieser Arbeit, ist aber nach allen hier gefundenen Ergebnissen eine vielversprechende Anwendung für die weitere Forschung auf diesem Gebiet.

Ausblick

Die Struktur des N-terminalen Kollagen-Telopeptids wurde in dieser Arbeit untersucht. Eine Strukturaufklärung des C-Telopeptids muss noch folgen, um die Struktur des Kollagens zu vervollständigen. Dies wird aufwendig sein, da das C-Telopeptid größer ist und eine Faltungs-Analyse notwendig macht. Zudem existieren hierfür bisher noch weniger Daten als für das N-Telopeptid.

Die Entwicklung und Durchführung von mesoskopischen Simulationen stellt die nächstliegende Fortführung der Forschung auf diesem Gebiet dar. Damit wird man auf größeren Zeit- und Raumskalen simulieren können und hoffentlich Erkenntnisse über die Mechanismen der Biomineralisation gewinnen. Dabei könnte ein reduziertes Fragment-basiertes Modell zum Einsatz kommen, bei dessen Parametrisierung die hier präsentierten PMFs verwendet werden können. Eine explizite Simulation des Solvens könnte vermieden werden, da dies schon in den PMFs enthalten ist.

Schließlich können mit den zukünftig steigenden Rechenressourcen Gitter-basierte Analysen von Ionenverteilungen durchgeführt werden. Alternativ könnten Fortschritte in der Integraltheorie die Analyse der Ionenverteilung um ein dynamisches Protein ermöglichen.

7 Literatur

- ¹ A. Boskey; *Connect Tissue Res.* **1996**, *35*, 357.
- ² P. Calvert, P. Rieke; *Chem. Mater.* **1996**, *8*, 1715.
- ³ G. Wegner; *Acta mater.* **2000**, *48*, 253.
- ⁴ R. Kniep, S. Busch; *Angew. Chem. Int. Edit.* **1996**, *35*, 2624.
- ⁵ S. Busch, H. Dolhaine, A. DuChesne, et al.; *Eur. J. Inorg. Chem.* **1999**, *10*, 1643.
- ⁶ S. Busch, U. Schwarz, R. Kniep; *Chem. Mater.* **2001**, *13*, 3260.
- ⁷ S. Busch, U. Schwarz, R. Kniep; *Adv. Funct. Mater.* **2003**, *13*, 189.
- ⁸ R. A. D. Williams, J. C. Elliott; *Basic and Applied Dental Biochemistry*, Churchill Livingstone, **1989**.
- ⁹ H. McDowell, T. M. Gregory, W. E. Brown; *J. Res. Natl. Bur. Stand.* **1977**, *81*, 273.
- ¹⁰ E. C. Moreno, M. Kresak, R. T. Zahradnik; *Nature* **1974**, *247*, 1.
- ¹¹ G. Ciccotti, R. Kapral et al.; *Chem. Phys.*, **1989**, *129*, 241.
- ¹² *NIST/SEMATECH e-Handbook of Statistical Methods*, <http://www.itl.nist.gov/div898/handbook/>, **2006**.
- ¹³ G. Zelniker, F. J. Taylor; *Advanced digital signal processing: theory and applications*; Marcel Dekker, **1994**.
- ¹⁴ S. J. Orfanidis; *Introduction to signal processing*; Prentice Hall, **1996**.
- ¹⁵ D. C. von Grünigen; *Digitale Signalverarbeitung*; Carl Hanser, **2001**.
- ¹⁶ D.R. Ferro, J. Hermans; *Acta Cryst.*, **1977**, *33*, 345.
- ¹⁷ M. Müller, *private communication*.
- ¹⁸ D. J. Prockop, K. L. Kivirikko; *Annu. Rev. Biochem.*, **1995**, *64*, 403.
- ¹⁹ H. Lodish, A. Berk, L. Zipursky, P. Matsudaira, D. Baltimore, J. Darnell; *Molecular Cell Biology*, W.H. Freeman & Company, **1999**.
- ²⁰ S. Busch, *Dissertation*, Darmstadt **1998**.
- ²¹ K. E. Kadler, D. F. Holmes, J. A. Trotter, J. A. Chapman; *Biochem. J.*, **1996**, *316*, 1.
- ²² B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J. D. Watson; *Molecular Biology of the Cell*, Garland Publishing, **1994**.
- ²³ R. Z. Kramer, J. Bella, B. Brodsky, H. M. Berman; *J. Mol. Biol.*, **2001**, *311*, 131.
- ²⁴ J. Stetefeld, S. Frank, J. Engel et al.; *Structure*, **2003**, *11*, 339.
- ²⁵ R. Berisio, L. Vitagliano, L. Mazzarella, A. Zagari; *Protein Science*, **2002**, *11*, 262.
- ²⁶ J. K. Rainey, M. C. Goh; *Protein Science*, **2002**, *11*, 2748.
- ²⁷ A. Bhattacharjee, M. Bansal; *IUBMB Life*, **2005**, *57*, 161.
- ²⁸ A. Otter, G. Kotovych, P. G. Scott; *Biochemistry*, **1989**, *28*, 8003.
- ²⁹ X. Liu, P. G. Scott, A. Otter, G. Kotovych; *J. Biomol. Struct. Dyn.*, **1990**, *8*, 63.
- ³⁰ A. Otter, P. G. Scott, G. Kotovych; *Biopolymers*, **1993**, *33*, 1443.
- ³¹ J. P. Orgel, T. J. Wess, A. Miller; *Structure*, **2000**, *8*, 137.
- ³² L. Vitagliano, G. Némethy, A. Zagari, H. Scheraga; *J. Mol. Biol.*, **1995**, *247*, 69.
- ³³ J. P. Malone, A. George, A. Veis; *Proteins*, **2004**, *54*, 206.
- ³⁴ A. D. MacKerell Jr., M. Karplus et al.; *J. Phys. Chem. B*, **1998**, *102*, 3586.

- ³⁵ W. Smith, T. R. Forester; *DL_POLY molecular simulation package*, The Council for the Central Laboratory of the Research Councils, Daresbury Laboratory at Daresbury, Nr. Warrington, **1994-1996**.
- ³⁶ L. Kale, K. Schulten et al.; *J. Comp. Phys.*, **1999**, *151*, 283.
- ³⁷ D. Zahn, *private communication*, **2003**.
- ³⁸ W. Humphrey, A. Dalke, K. Schulten; "VMD - Visual Molecular Dynamics", *J. Mol. Graphics*, **1996**, *14*, 33.
- ³⁹ J. Brickmann, T. Goetze, W. Heiden, G. Moeckel, S. Reiling, H. Vollhardt, C.-D. Zachmann; "Interactive visualization of molecular scenarios with MOLCAD/SYBYL", in *Data visualization in molecular science*, J.E. Bowie Ed., Addison-Wesley Publishing Company Inc., Reading, Mass., **1995**, 83-97.
- ⁴⁰ Sybyl 6.9.1, Tripos Inc., 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA.
- ⁴¹ C. H. Wu, L. S. L. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Z. Hu, R. S. Ledley, P. Kourtesis, B. E. Suzek, C. R. Vinayaka, J. Zhang, W. C. Barker; The Protein Information Resource. *Nucleic Acids Research*, **2003**, *31*: 345.
- ⁴² C. Liebecq, *Biochemical Nomenclature and Related Documents*, 2nd edition, Portland Press, **1992**.
- ⁴³ N. Roveria, B. Parma et al.; *Mat. Sci. Eng. C*, **2003**, *23*, 441.
- ⁴⁴ L. Stryer, *Biochemistry*, 3rd Edition, W.H. Freeman & Company, **1988**.
- ⁴⁵ P. Ren, J. W. Ponder; *J. Phys. Chem. B*, **2003**, *107*, 5933.
- ⁴⁶ T. E. Klein, C. C. Huang; *Biopolymers*, **1999**, *49*, 167.
- ⁴⁷ J. K. Rainey, M. C. Goh; *Bioinformatics*, **2004**, *20*, 2458.
- ⁴⁸ R. Berisio, L. Vitagliano, L. Mazzarella, A. Zagari; *Protein Science*, **2002**, *11*, 262.
- ⁴⁹ R. Z. Kramer, M. G. Venugopal, J. Bella, P. Mayville, B. Brodsky, H. M. Berman; *J. Mol. Biol.*, **2000**, *301*, 1191.
- ⁵⁰ D. Zahn, *unpublished results*.
- ⁵¹ A. Rich, F. H. C. Crick; *J. Mol. Biol.*, **1961**, *3*, 483.
- ⁵² J. Bella, H. M. Berman; *J. Mol. Biol.*, **1996**, *264*, 734.
- ⁵³ M. Keil, *Dissertation*, Darmstadt, **2002**.
- ⁵⁴ S. Hauptmann, H. Dufner, J. Brickmann, S. M. Kast, R. S. Berry; *Phys. Chem. Chem. Phys.*, **2003**, *5*, 635.
- ⁵⁵ D. Zahn, O. Hochrein; *Phys. Chem. Chem. Phys.*, **2003**, *5*, 4004.
- ⁵⁶ C. F. Schwenk, B. M. Rode; *Pure Appl. Chem.*, **2004**, *76*, 37.
- ⁵⁷ T. Megyes, T. Grósz, T. Radnai, I. Bakó, G. Pálkás; *J. Phys. Chem. A*, **2004**, *108*, 7261.
- ⁵⁸ C.-G. Zhan, D. A. Dixon; *J. Phys. Chem. A*, **2004**, *108*, 2020.
- ⁵⁹ A. Öhrn, G. Karlström; *J. Phys. Chem. B*, **2004**, *108*, 8452.
- ⁶⁰ K. Maksimiak, S. Rodziewicz-Motowidło, C. Czaplewski, A. Liwo, H. A. Scheraga; *J. Phys. Chem. B* **2003**, *107*, 13496.
- ⁶¹ A. Masunov, T. Lazaridis; *J. Am. Chem. Soc.* **2003**, *125*, 1722.
- ⁶² S. A. Hassan, *J. Phys. Chem. B*, **2004**, *108*, 19501.
- ⁶³ R. A. Friedman, M. Mezei; *J. Chem. Phys.*, **1995**, *102*, 419.
- ⁶⁴ X. Rozanska, C. Chipot; *J. Chem. Phys.*, **2000**, *112*, 9691.
- ⁶⁵ A. Liwo, C. Czaplewski, J. Pillardy, H. A. Scheraga; *J. Chem. Phys.*, **2001**, *115*, 2323.
- ⁶⁶ J. Lee, K. Park, J. Lee; *J. Phys. Chem. B*, **2002**, *106*, 11647.

- ⁶⁷ K. Maksimiak, S. Rodziewicz-Motowidlo, C. Czaplewski, A. Liwo, H. A. Scheraga; *J. Phys. Chem. B*, **2003**, *107*, 13496.
- ⁶⁸ A. Liwo, M. Khalili, H. A. Scheraga; *P. Natl. Acad. Sci. USA*, **2005**, *102*, 2362.
- ⁶⁹ M. Khalili, J. A. Saunders, A. Liwo, S. Oldziej, H. A. Scheraga; *Protein Science*, **2004**, *13*, 2725.
- ⁷⁰ H. Grubmüller, R. A. Böckmann; *Angewandte Chemie Int. Ed.*, **2004**, *43*, 1021.
- ⁷¹ G. Pearson; *Inorg. Chem.*, **1988**, *27*, 734.

8 Anhang

Lebenslauf

Persönliche Daten

Name: Thorsten Schepers
Anschrift: Flachsbachweg 17
64285 Darmstadt
Geburtsdatum: 24.9.1976
Geburtsort: Lengerich
Staatsangehörigkeit: deutsch
Familienstand: ledig

Bildungsgang

08/83 - 07/87 Grundschule Intrup, Lengerich

08/87 - 06/96 Hannah-Arendt-Gymnasium Lengerich, Abitur

10/96 - 04/02 Westfälische-Wilhelms-Universität Münster, Diplom Chemie.
Diplomarbeit bei Prof. J. Michl in Boulder, Colorado, USA.
Thema: "Simple models for sigma delocalization in linear oligosilanes".

07/02 - 03/06 Technische Universität Darmstadt, Promotion.
Dissertation bei Prof. J. Brickmann, Thema: "Toward an Efficient Simulation of Biomineralization: A Computational Study of the Apatite/Collagen System".

Verwendete Hilfsmittel

Für die Simulationen in dieser Arbeit kamen die Programme NAMD 2.5, DLPOLY 2.14, DLPROTEIN 2.12 und CHARMM 24g2 zum Einsatz. Die Rechnungen wurden auf den IBM Power 4 Servern des HHLR durchgeführt. Die Analysen und Visualisierungen wurden mit den Programmen Sybyl 6.9.1, VMD 1.8.4, Molcad, sowie in FORTRAN 77, TCL 8.4 und tcsh 6.14.0 selbst geschriebenen Programmen und Skripten durchgeführt. Die Analysen wurden zum Teil auf dem HHLR, zum Teil auf einem Dell Precision 670 (Dual Intel Xeon 3 GHz, 4 GB RAM) mit SuSe Linux 10 und einem HP I2000 (Dual Itanium IA-64 800 MHz, 2 GB RAM) mit Debian durchgeführt. Das Programm histool von B. Schilling wurde zur Konvertierung von Trajektorien benutzt. Der vorliegende Text wurde mit Microsoft Word 2000 geschrieben, Grafiken wurden mit den genannten Programmen und OriginLab Origin 7 und Adobe Illustrator 10 angefertigt.

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, dass ich meine Dissertation selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe.

Darmstadt, den _____

(Thorsten Schepers)

Erklärung

Ich erkläre hiermit, noch keinen Promotionsversuch unternommen zu haben.

Darmstadt, den _____

(Thorsten Schepers)