# Geometry-based Conformational Sampling of Proteins

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaflichen Fakultäten
der Georg-August-Universität zu Göttingen

vorgelegt von
**Daniel Seeliger**
aus Kirchheim/Teck

Göttingen 2007

Referent: Prof. Dr. Bernd Abel
Korreferent: Prof. Dr. Helmut Grubmüller
Tag der mündlichen Prüfung:

# Vorveröffentlichungen der Dissertation

Teilergebnisse dieser Arbeit wurden in folgenden Beiträgen vorab veröffentlicht.

## Publikationen

D. Seeliger and B. L. de Groot: Prediction of Protein Flexibility from Geometrical Constraints. *Biotech International*, 2006, 18, 20-22.

D. Seeliger and B. L. de Groot: Atomic Contacts in Protein Strutcures: A Detailed Analysis of Atomic Radii, Packing and Overlaps. *PROTEINS*, 2007, 68, 565-601.

D. Seeliger, J. Haas and B. L. de Groot: Geometry-Based Sampling of Conformational Transitions in Proteins. *Structure*, 2007, 15, 1482-1492.

U. Zachariae, R. Schneider, P. Velisetty, A. Lange, D. Seeliger, S. Wacker, Y. Karimi-Nejad, G. Vriend, O. Pongs, M. Baldus and B. L. de Groot: An atomistic view links toxin-induced conformational changes to C-type activation in a potassium channel. *submitted*, 2007.

# Contents

**1 Introduction** **9**

**2 Theory and Concepts** **15**

- 2.1 Protein Structure . . . . . . . . . . . . . . . . . . . . . . . . . . 15
- 2.2 Experimental Structure Determination . . . . . . . . . . . . . . . 18
  - 2.2.1 X-ray Crystallography . . . . . . . . . . . . . . . . . . . . 19
  - 2.2.2 Nuclear Magnetic Resonance Spectroscopy . . . . . . . . 22
- 2.3 Protein Motion . . . . . . . . . . . . . . . . . . . . . . . . . . 25
- 2.4 Simulation Methods . . . . . . . . . . . . . . . . . . . . . . . . 26
  - 2.4.1 Molecular Dynamics . . . . . . . . . . . . . . . . . . . . 27
  - 2.4.2 Replica-Exchange Molecular Dynamics . . . . . . . . . . 29
  - 2.4.3 Normal Mode Analysis . . . . . . . . . . . . . . . . . . . 30
  - 2.4.4 Elastic Network Models . . . . . . . . . . . . . . . . . . 31
- 2.5 Geometry-Based Molecular Simulation . . . . . . . . . . . . . . 32
  - 2.5.1 Geometrical Constraints in Protein Structures . . . . . . . 33
  - 2.5.2 Structure Generation: The CONCOORD-algorithm . . . . 34
- 2.6 Principal Components Analysis . . . . . . . . . . . . . . . . . . 36

**3 Parametrization** **39**

- 3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . 39
- 3.2 Experimental Data . . . . . . . . . . . . . . . . . . . . . . . . . 40
- 3.3 The PDBBrowser: A Tool for Flexible Database Queriing . . . . . 41
  - 3.3.1 Program Structure . . . . . . . . . . . . . . . . . . . . . 41
  - 3.3.2 Database Queries . . . . . . . . . . . . . . . . . . . . . . 42
- 3.4 Optimization of Atomic Radii . . . . . . . . . . . . . . . . . . . 44

3.5 Atomic Packing in Protein Structures . . . 47
  3.5.1 Quantitative Assessment of Packing Quality . . . 49
  3.5.2 Packing quality in protein structures . . . 52
3.6 Summary . . . 57

**4 Constraint Definition 59**
4.1 Introduction . . . 59
4.2 Evaluation of Hydrogen Bond Stability . . . 60
4.3 Hydrophobic Clusters . . . 63
4.4 Residue Networks . . . 64
4.5 Manual Constraint Definition . . . 66
4.6 Summary . . . 68

**5 Conformational Sampling 71**
5.1 Introduction . . . 71
5.2 Adenylate Kinase . . . 72
5.3 Calmodulin . . . 74
5.4 Aldose Reductase . . . 75
5.5 T4-Lysozyme . . . 78
5.6 Rigid and flexible regions in proteins . . . 81
5.7 Summary . . . 84

**6 Modeling of Protein Parts 85**
6.1 Introduction . . . 85
6.2 Loop Modeling . . . 86
6.3 Protein Core Repacking . . . 89
6.4 Summary . . . 90

**7 Modeling of Complexes 91**
7.1 Introduction . . . 91
7.2 Conformational Sampling of Small Molecules . . . 93
7.3 Protein-Ligand Complexes . . . 96
7.4 Protein-Protein Complexes . . . 99
7.5 Summary . . . 102

**8 Conclusions** **103**

**9 Acknowledgments** **107**

**10 Appendix** **111**

10.1 Protein Structures . . . . . . . . . . . . . . . . . . . . . . . . . . . . 111
10.2 Protein Structures Used for Hydrogen Bond Statistics . . . . . . . . 114
10.3 Ubiquitin Structures . . . . . . . . . . . . . . . . . . . . . . . . . . . 114

**Bibliography** **115**

# Chapter 1

# Introduction

*Die Natur ist das einzige Buch, das auf allen Blättern großen Gehalt bietet.*

*- Johann Wolfgang von Goethe*

Proteins are macromolecules that are found in every living organism, in every cell and every subunit of the cell. They have structural and mechanical functions, catalyze chemical reactions, pump ions, recognize signals and trigger immune responses. Actually, there is no cellular function in which proteins are not involved. Hence, understanding protein function virtually means understanding life.

The first step to understand the molecular basis of function is structure. The human genome project yielded a huge amount of protein sequence data and the challenge is to turn this data into information about the 3-dimensional structure of proteins. So far $\approx 46000$ protein structures have been resolved and serve to understand the machinery of life on the atomic level.

However, structure is only the first step, as almost always dynamics is essential for function. Regardless of whether a protein functions as enzyme, molecular motor, transport protein or receptor, its function is often coupled to motion. These motions range from side-chain fluctuations to reorientations of entire domains and partial unfolding and refolding. Understanding protein function is thus strongly coupled to insight into dynamics and flexibility. X-ray crystallography, which is still the major source of structural information of proteins, provides mainly static pictures of one conformation, even though a number of proteins has been resolved in different conformations providing insights into protein flexibility directly from experimental data [1]. Structures resolved by NMR-spectroscopy are usually published as an ensemble of conformations that fulfil the experimentally determined restraints and provide more information about protein flexibility. However, the method is still restricted to proteins of limited size.

A particularly important research area is the computational design of novel drugs. Knowledge about protein structures in different conformational substates, either from experimental data or simulation, has been proven to enhance protein-protein docking [2–4] and Structure-Based Drug Design(SBDD) [5–9].

Due to the difficulties associated with derivation of information about protein flexibility from experiments, many computational approaches have been developed and successfully applied. The most widely used methodology to tackle protein flexibility is Molecular Dynamics (MD) simulation. However, despite the enormous increase in computer power and advances in algorithm techniques and parallelisation, MD simulations are computationally expensive and moreover, high energy barriers are often not overcome within accessible time. In order to alleviate the resulting sampling problem, several advanced simulation methods based on MD have been developed and successfully applied to numerous problems within the field of protein research, among them Replica-Exchange Molecular Dynamics (REMD) [10], Conformational Flooding [11, 12] and Targeted Molecular Dynamics (TMD) [13, 14]. However, even these methods are not routinely applicable for the efficient sampling of conformational transitions. Computationally more efficient, but less accurate methods, are based on gaussian network models [15, 16], normal mode analysis [17–20] or graph theoretical approaches [21].

A different approach is the CONCOORD-method [22], which is based on geometrical considerations to predict protein flexibility. A given input structure is analyzed and translated into a geometric description of the protein. Based on this description, the structure is rebuilt, commonly several hundreds of times, leading to an ensemble that can be analyzed and essential degrees of freedom [23], often representing the biological relevant motions in proteins, may be extracted.

Induced fit motions, that proteins often undergo upon binding a ligand, are one of the most challenging problems in structure-based drug design. A commonly accepted theory of the induced fit describes this phenomenon as a consequence of a change in the free energy landscape due to the presence of the ligand with the effect that the conformation with the lowest free energy in the unbound state is not identical with the lowest free energy conformation of the protein/ligand-complex. This problem is not exclusively restricted to structural differences of bound and unbound protein conformations. Different ligands also may cause the protein to adopt different conformations.

This means that even resolved protein structures that have been co-crystallized together with a ligand are not necessarily ideal targets for molecular docking or the derivation of pharmacophore models. Larger ligands with high affinities to the target might not fit into a binding site of a smaller ligand. Consequently interactions of smaller ligands would in such cases be underestimated in the binding sites of large ligands.

Conformational flexibility of the binding site worsens this problem. As the results of molecular docking studies are very sensitive to even minor side-chain movements, the predictive power of these methods, when applied to binding sites with flexible loops, rapidly drops to the level of crystal ball gazing. This is particularly concerning as a considerable number of todays most promising drug targets are channel proteins with flexible binding sites. Thus, incorporation of protein flexibility is crucial to move forward and to enhance the predictive power and reliability of in silico approaches in the field of structure-based drug design [24–26]. The usage of structure ensembles has been shown to improve these efforts. In some rare cases such an ensemble can be compiled directly from experimental data and used for molecular docking, which has been shown to be superior to docking to a single receptor structure [8]. Also snapshots taken from Molecular Dynamics trajectories have been employed [6, 7] and shown to lead to better results in some cases. However, obtaining representative structure ensembles from experimental data, covering the *relevant* conformational space, will also in the future be restricted to a very limited number of proteins. Structure ensembles derived from NMR experiments provide a better estimate of protein flexibility, though this method is still restricted to proteins of limited size. Moreover, the lower re-

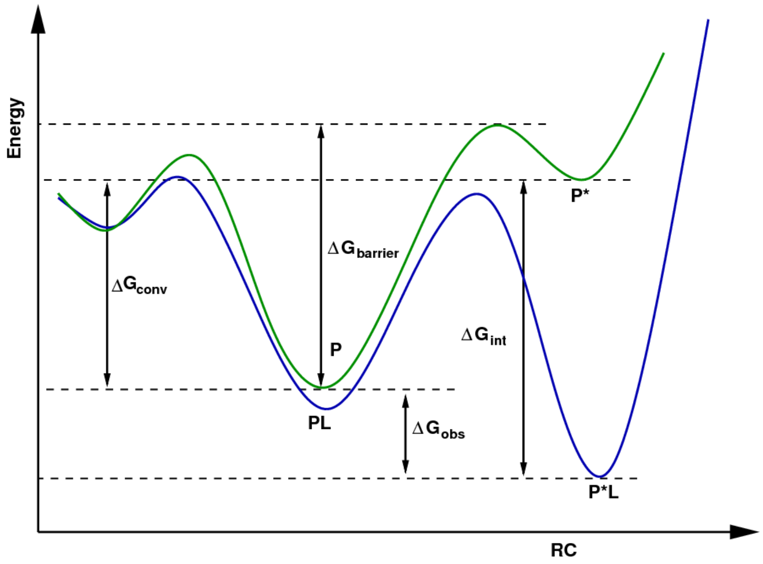


Figure 1.1. **Protein flexibility and ligand binding.** A protein exists in two conformations (P and P*) with energy difference $\Delta G_{conv}$. The ligand (L) can bind the protein (P) to give the a complex (PL), or bind to P* to give a complex (P*L). Although P* has a higher free energy, it might offer greater scope for interactions with L, thereby giving rise to a large, favourable interaction $\Delta G_{int}$. The resulting complex (P*L) has a lower energy than that of the complex PL. The observed affinity of L for the protein conformational ensemble is is governed by $\Delta G_{obs}$. Slow binding kinetics might well be observed, as P* is a higher-energy conformer than P and an energy barrier $\Delta G_{barrier}$ must be surmounted before optimal binding to L can take place. This is also the bottleneck for force-field based simulation methods, as such barriers might not be overcome within accessible time. (Figure adapted from Simon Teague [9]).

solution of protein models derived from NMR data compared to X-ray structures hampers structure-based drug design based on such structures.

Hence, static pictures of protein structures, derived from X-ray crystallography or even from homology modelling are and will be used as starting points for structure based drug design in the future and flexibility properties will have to be derived from in silico methods.

Commonly, Molecular Dynamics simulations are employed to study protein dynamics and thus, are the method of choice for generating protein structures in different conformational states from a given 3-dimensional structure. A hypothetic infinite trajectory contains all possible conformations of the protein together with the corresponding free energy obtained from the phase space density, and thus, all conformations in which ligands can bind to the protein. Here lies the weakness of force field based simulation methods for obtaining structure ensembles to be used for structure-based drug design. As an induced fit upon ligand binding is a consequence of a change in the free energy landscape, the free energy of the corresponding protein conformation with removed ligand is higher, in some cases significantly higher than the lowest free energy conformation. This means, that within accessible time, conformations in which greater scope for interactions with a ligand is possible might not be sampled within the limited timeframe of typical MD simulation. Because of this sampling problem MD-simulations suffer from, it is necessary to augment the effort of finding alternative ways to efficiently generate structure ensembles representing the *relevant* conformational space.

This work focuses on the development of geometry-based molecular simulation techniques and their application to biologically relevant questions. Based on the original CONCOORD method [22], which has been developed to predict conformational ensembles *around* a known structure, a major extension, termed tCONCOORD, was developed that expands the scope of geometry-based molecular modeling to several fields of protein science.

In the following chapter the fundamentals of protein structure and protein structure determination are recapitulated. Furthermore, established computational methods are reviewed.

In the third chapter we present how the wealth of experimental data can be turned into parameter sets for biomolecular simulations. For instance, a novel set of atomic radii has been derived from high-resolution X-ray structures. Using these parameters, we could show that the distance distribution of atomic contacts in protein structures is highly conserved and exclusively resolution dependent [27].

In chapter four we describe how structures are analyzed in tCONCOORD and how geomtrical constraints are defined. Special attention is payed to a novel method to estimate the stability of hydrogen bonds in proteins based on the solvation probabilities of surrounding atoms [28]. Applications of tCONCOORD to biologically relevant questions are the objective of the subsequent chapters.

In chapter five we show how tCONCOORD can be used to predict protein conformational flexibility. Applications to proteins as diverse as the globular protein ubiquitin and the multi-domain protein calmodulin reveal that experimentally observed protein flexibility and conformationl transitions are faithfully reproduced.

Chapter six focuses on predicting conformational flexibility of protein parts. We show how geometry-based molecular modeling has been successfully applied to loop modeling and modeling of a modified protein core for subsequent use in molecular dynamics simulations.

In chapter seven we show how tCONCOORD can be useful in the field of structure-based drug design and in modeling macromolecular assemblies.

# Chapter 2

# Theory and Concepts

> *Es gibt Leute, die glauben, alles wäre vernünftig, was man mit ernsthaftem Gesicht tut.*
>
> *- Georg Christoph Lichtenberg*

## 2.1 Protein Structure

Proteins are polymers comprising 20 (21 if we incorporate seleno-cystein) chemically and structurally different building blocks (amino acids) that fold into highly specific three-dimensional structures.

Naturally occuring proteins and peptides exclusively contain L-$\alpha$-amino acids. The single amino acids in a peptide chain are connected via peptide bonds, forming a dihedral angle of ~180° between H-N-$C_\alpha$-O with the exception of the rare occurence of cis-proline. The backbone of a peptide chain consists of repeating units of the three atoms N, $C_\alpha$ and C. While rotation around the C-N bond ($\Omega$-angle) is limited to a small range around 180°, rotation around the N-$C_\alpha$ bond ($\Phi$-angle) and the $C_\alpha$-C bond ($\Psi$-angle) is possible. Hence, rotation around the backbone $\Phi$- and $\Psi$-angles are the major degrees of freedom underlying protein flexibility.

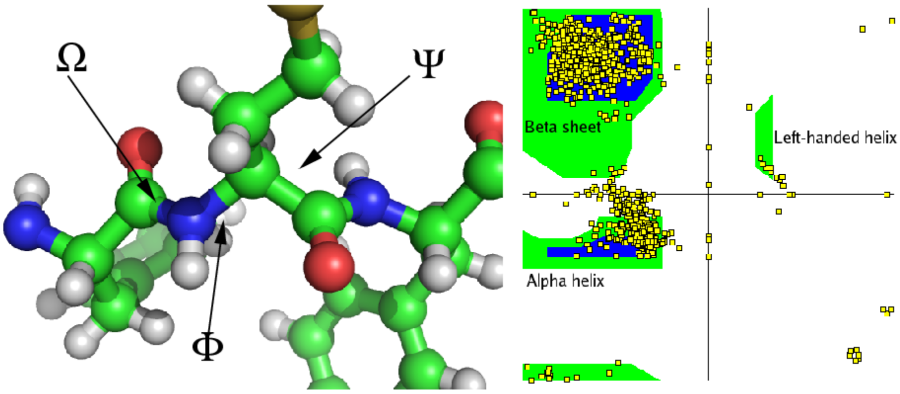


Figure 2.1. Left panel: peptide chain and backbone dihedral angles. Right panel: Ramachandran plot

Due to sterical restrictions the $\Phi$- and $\Psi$-angles of peptide chains in naturally folded proteins only adopt a limited and well-defined part of the dihedral-angle space (see fig. 2.1 right). These $\Phi$-$\Psi$-plots, named Ramachandran-plots after the discoverer G. N. Ramachandran [29], are an important quality criterion for protein structures.

The structural description of proteins is seperated into four levels. Besides the sequence, which is determined by the gene and referred to as the primary structure, the secondary, tertiary and quaternary structure of a protein are distinguished. The secondary structure describes the local fold and is heavily connected to the dihedral angles of the backbone. The DSSP (Dictionary of Protein Secondary Structure) code [30] uses hydrogen bond patterns to classify the secondary structure.

G = 3-turn helix ($3_{10}$ helix). Min length $3$ residues.
H = 4-turn helix ($\alpha$ helix). Min length $4$ residues.
I = 5-turn helix ($\pi$ helix). Min length $5$ residues.
T = hydrogen bonded turn ($3$, $4$ or $5$ turn)
E = beta sheet in parallel and/or anti-parallel sheet conformation (extended strand). Min length 2 residues.
B = residue in isolated beta-bridge (single pair beta-sheet hydrogen bond formation)
S = bend (the only non-hydrogen-bond based assignment)

In DSSP, residues which are not in any of the above conformations is designated as ' ' (space), which sometimes gets designated with C (coil) or L (loop). The helices (G, H and I) and sheet conformations are all required to have a minimal length. This means that 2 adjacent residues in the primary structure must form the same hydrogen bonding pattern. If the helix or sheet hydrogen bonding pattern is too short they are designated as T or B, respectively. Other protein secondary structure assignment categories exist (sharp turns, Omega loops etc.), but they are less frequently used.

The term "tertiary structure" is used to describe the way how the different secondary structure elements are arranged and build the overall fold of the chain. Quaternary structures describe how different peptide chains are arranged to build the protein.

The SCOP database [31–33] (Structural Classification Of Proteins) currently distinguishes approx. 1000 different folds, 1600 super families and 3000 families. It is remarkable that despite the exponential growth of resolved protein structures in the PDB, the last new fold has been determined in 2005. The conformation of the native fold of a protein corresponds to the global minimum on the free energy surface. In globular proteins, tertiary interactions are frequently stabilized by burying hydrophobic amino acid residues in the protein core, from which water is excluded, and by the consequent enrichment of charged or hydrophilic residues on the protein's water-exposed surface.

The prediction of protein structure has been a long-standing problem and is adressed with bioinformatics based methods like homology modeling and physics-based methods like simulations. An overview of the recent progress is given in [34].

## 2.2 Experimental Structure Determination

The three-dimensional structure of proteins is essential for understanding their function and a prerequisite for numerous computational approaches in modern protein research. Thus, great efforts are invested to determine structures at atomic resolution. Once a protein sequence of interest is identified, the protein is either isolated directly from the source cell or tissue, or molecular biology methods are employed to express the protein of interest in a host such as *Escherichia coli*. The latter represents the most common route, where DNA encoding the sequence of the protein is inserted into vectors, facilitating the expression in E. coli.

After expression of the protein, various ways are employed for purification. Centrifugation seperates particles with different mass, but also depends on molecular shape, temperature and solution density. Another common way is “Salting in and salting out”, which makes use of differential solubility of proteins at various ionic strength. The solubility of most proteins increases with growing ionic strength up to a maximum due to increased polarity of the solution. At higher ionic strengths the solubility decreases as ions compete for water molecules against the protein.

Chromatographic methods form the core of most purification protocols. Different proteins can be seperated using various gradients, among them ion exchange chromatography that seperates proteins on the basis of overall charge, size exclusion chromatography that seperates according to the molecular size, hydrophobic interaction chromatography that focuses on differences in surface hydrophobicity and affinity chromatography which is employed if proteins bind a known ligand.

The purified protein is the first step towards structure determination. The Protein Data Bank (PDB) [35] currently contains $44700$ structures of which $\approx 38000$ have been resolved by X-ray diffraction and $\approx 6400$ by NMR-spectroscopy. A small fraction has been resolved by electron microscopy. However, this method does not provide data at atomic resolution.

### 2.2.1 X-ray Crystallography

X-ray crystallography is the pre-eminent technique for the determination of protein structure. X-rays, discovered by Röntgen, were shown to be diffracted by crystals in 1912 by Max von Laue. Bragg interpreted the spots obtained on photographic plates and formulated the relationship between the diffraction pattern and the crystal structure which is known as Bragg's law

$$n\lambda = 2d \sin\Theta, \tag{2.1}$$

where $\lambda$ denotes the wavelength, d the lattice constant and $\Theta$ the angle of the incident radiation. This formula is equivalent to

$$n\lambda = d\cos\Theta_i - d\cos\Theta_r \tag{2.2}$$

with $\Theta_i$ the angle of the incident radiation and $\Theta_r$ angle of the reflected radiation (see fig. 2.2). Extended to three dimensions we obtain the Laue set of equations, where a, b and c refer to the spacing for each of the three dimensions.

$$\begin{aligned} a(\cos\alpha_i - \cos\alpha_r) &= h\lambda \quad \text{(where h=1,2,3,...)} \\ b(\cos\beta_i - \cos\beta_r) &= k\lambda \quad \text{(where k=1,2,3,...)} \\ c(\cos\gamma_i - \cos\gamma_r) &= l\lambda \quad \text{(where l=1,2,3,...)} \end{aligned} \tag{2.3}$$

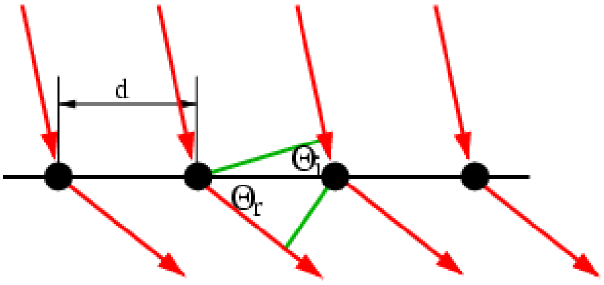


Figure 2.2. **The Laue equations.** The direction of the radiation is represented by red arrows. $\Theta_i$ denotes the angle of the incident radiation, $\Theta_r$ the angle of the reflected radiation. d is the lattice constant.

The Laue equations must be satisfied to assure constructive interference and thus for diffraction to occur.

The unit cell, the basic building block of a crystal, is repeated in three dimensions but is characterized by three vectors (a, b, c) that form a parallelepiped and the three corresponding angles $(\alpha, \beta, \gamma)$. In biological systems, the unit cell may posess internal symmetry containing more than one protein molecule related to others via axes or planes of symmetry. Scattering depends on the properties of the crystal lattice and is the result of interactions between the incedent X-rays and the electrons of atoms within the crystal. Heavy atoms, such as metals or sulphur are very effective at scattering X-rays whereas smaller atoms such as the proton are ineffective. The result of an X-ray diffraction experiment is not a picture of atoms, but a diffraction pattern composed by the reflections of all atoms within a unit cell. As a wave consists of an amplitude $f$ and a phase angle $\psi$, it can be described as a vector

$$\mathrm{f} = f\cos\psi + \mathrm{i}f\sin\psi = f\mathrm{e}^{\mathrm{i}\psi} \tag{2.4}$$

Since all atoms contribute to the observed diffraction pattern, these vectors are summed together and are described by the vector $\mathbf{F}_{hkl}$ known as the structure factor

$$\mathrm{F}_{hkl} = \sum f\cos\psi + \sum \mathrm{i}f\sin\psi \tag{2.5}$$

leading to

$$\mathrm{F}_{hkl} = F_{hkl}(\cos\varphi_{hkl} + \mathrm{i}\sin\varphi_{hkl}) = F_{hkl}\mathrm{e}^{\mathrm{i}\varphi_{hkl}}. \tag{2.6}$$

$F_{hkl}$ is the square root of the intensity of the observed diffraction spot often called $I_{hkl}$, whereas the $\varphi_{hkl}$ term represents the summation of all phase terms constributing to this spot. The structure factor $\mathbf{F}_{hkl}$is the Fourier transform of the electron density. The value of the electron density at a real-space lattice point (x,y,z) denoted by $\rho$(x,y,z) is equivalent to

$$\rho(x,y,z) = \frac{1}{V}\sum_{hkl=-\infty}^{+\infty} F_{hkl}\mathrm{e}^{\mathrm{i}\varphi_{hkl}}\mathrm{e}^{-2\pi\mathrm{i}(hx+ky+lz)} \tag{2.7}$$

where $\rho$ is the value of the electron density at the real-space lattice point (x,y,z), and V is the total volume of the unit cell, and $\varphi$ is the phase information.

To calculate the electron density map from the measured intensities, the determination of $\varphi_{hkl}$ is required, which is known as the *phase problem*. For small molecules it is possible to make guesses about the conformation and to calculate diffraction patterns of the 'guess' and compare the result with the experimentally determined diffraction pattern. For proteins this is not possible. This problem may be overcome by irradiating crystals that have been soaked in the presence of heavy metal ions. From the diffraction patterns of this metal labelled crystals, structure factors can be calculated that enable the derivation of the electron density map of the protein. Other methods to resolve the phase problem include MAD and molecular replacement. The MAD (multiwavelength anomalous diffraction) method analyzes the phase shift that is caused by replacement of methionine with Seleno-methionine. The positions of the methionine residues provide initial phases. Molecular replacement is employed if a structure of a related crystal structure exists, which serves as a search model to determine the orientation and position of the molecules within the unit cell.

The initial electron density map does not resolve individual atoms. Structures before refinement are often at resolutions $> 4.5$ Å where only $\alpha$ helices are observed and the identification of side chains is unlikely. Computer programs fit electron density maps and the process is assisted by assuming standard bond length and angles. Refining models in an iterative fashion progressively improves the agreement with experimental data. A structure is judged by the crystallographic R-factor, defined as the average fractional error in the sum of the differences between calculated structure factors ($F_{cal}$) and observed structure factors ($F_{obs}$) divided by the sum of the observed structure factors

$$\mathrm{R} = \sum \frac{|F_{obs} - F_{cal}|}{\sum F_{obs}}. \tag{2.8}$$

A cross-validated quality criterion is $R_{free}$, which is calculated from a subset ($\approx 10\%$) of reflections that were not included in the structure refinement. A value of $0.20$ is often represented as an R-factor of $20$ percent and 'good' structures have $R_{free}$-factors ranging from $15 - 25$ percent or approximately $1/10^{th}$ of the resolution of the data. The results of protein structure determination are files containing coordinates for all resolved atoms together with their B-factor (Debye-Waller factor), that reflects spreading or blurring of electron density and represents the mean square displacement of atoms in units of $\mathring{A}^2$. High B-factors can either be due to experimental noise or indicate increased mobility and disorder of atoms. Residues on the protein surface and particularly atoms corresponding to long side-chains as in arginine or lysine usually display high B-factors. The occupancy, which denotes the probability of finding the atom in a certain position, is also stored. High resolution structures often provide alternate positions for atoms, e.g. if side chains adopt different conformations.

The major bottleneck of structure determination using X-ray crystallography is the production of protein crystals. Crystallization requires the ordered formation of large (dimensions larger than $0.1$ mm along each axis), stable crystals with sufficently long-range order to diffract X-rays. Structures produced by X-ray diffraction are only as good as the crystals from which they are derived. Finding optimal conditions for crystallizing a protein is difficult as various parameters can be changed. Different reagents to reduce protein solubility, their concentration, pH value and protein concentration are usually varied in screening approaches, nowadays often carried out by robots. However, this approach only works for soluble proteins. Membrane proteins, which are of special interest from pharmaceutical points of view since many of them are potential drug targets, are very difficult to crystallize.

### 2.2.2 Nuclear Magnetic Resonance Spectroscopy

The second important technique for determining protein structure is nuclear magnetic resonance spectroscopy (NMR). In contrast to X-ray crystallography, NMR does not require protein crystals, but the proteins are studied in solution. Underlying the NMR phenomenon is a property of all atomic nuclei called 'spin'. Spin describes the nature of a magnetic field surrounding a nucleus and is characterized by a spin number, I, which is either zero or a multiple of $1/2$. Nuclei whose spin number equals zero have no magnetic field and from NMR standpoint are uninteresting. This occurs when the number of neutrons and the number of protons are even. Spin $1/2$ nuclei represent the simplest situation and arise when the number of protons plus neutrons is an odd number. The most important spin $1/2$ nucleus is the proton with a high natural abundance (~$100\%$) and its occurance in all biomolecules. For nuclei such as $^{12}$C the most common isotope is NMR 'silent' and the active spin $1/2$ nucleus ($^{13}$C) has a low natural abundance of ~$1.1\%$. For spin $1/2$ nuclei application of a magnetic field removes degeneracy and the energetic levels split into parallel and anti-parallel orientations. Spins aligned parallel with external magnetic fields are of slightly lower energy than those aligned in an anti-parallel orientation, hence the population is different and given by the Boltzmann distribution.

$$n_{upper}/n_{lower} = e^{-(\Delta E/k_B T)} \tag{2.9}$$

At thermal equilibrium the number of nuclei in the lower energy level slightly exceeds those in the higher energy level. As a result of this small inequality it is possible to elicit transitions between states by the application of short, intense, radio frequency pulses.

The use of NMR spectroscopy as a tool to determine protein structure is based around several related parameters that influence the observation of signals. These parameters include the chemical shift ($\delta$), spin-spin coupling constant ($J$), the spin-lattice relaxation time ($T1$), the spin-spin relaxation time ($T2$), the peak intensity, the nuclear Overhauser effect (NOE) and Residual Dipolar Couplings (RDC).

For protein structure determination particularly the chemical shift and the nuclear Overhauser effect are important. The chemical shift reflects the chemical nature of groups and mainly depends on the electron density at the proton. As a reference, Trimethylsilane (TMS) is used, which has higher electron densities at the hydrogen atoms than most hydrogen atoms occuring in organic molecules. Its signal is set to zero and other chemical shifts are defined relative to the TMS signal in parts per million (ppm). Low electron densities at the proton, for instance in polar groups, lead to higher chemical shifts. Due to the partial double bond nature of the amide bond, the amide proton of a polypeptide backbone has a chemical shift between $8.0$ and $9.0$ ppm, whereas protons in methyl groups have chemical shifts between $0$ and $2.0$ ppm. The nuclear Overhauser effect is the fractional change in intensity of one resonance as a result of irradiation of another resonance. As a result of dipolar 'through space' interactions the irradiation of one resonance perturbs intensities of neighbouring resonances. The NOE is expressed as

$$\eta = (I - I_0)/I_0 \tag{2.10}$$

where $I_0$ is the intensity without irradiation and $I$ is the intensity with irradiation. The NOE effect is rapidly attenuated by distance and declines with the inverse sixth power of the distance between two nuclei.

$$\eta \propto r^{-6} \tag{2.11}$$

Thus, the NOE provides information about nuclei which are closed in space. Such distance restraints are used to determine the three-dimensional structure of proteins. Usually simulated annealing, restrained molecular dynamics simulations are employed to derive structure models from NMR data and an ensemble of typically $10 - 30$ models, those with the lowest energies, is deposited in the PDB.

The major drawback in NMR structure determination is the so-called assignment problem. Before distance and angle restraints from NOE's can be determined, each resonance from the spectra has to be assigned to a pariticular proton of the protein. As a protein consisting of ~$100$ residues contains about $700$ protons, spectral overlaps usually preclude complete assignment of all protons. Therefore, using NMR spectroscopy for structure determination is still limited to small proteins.

## 2.3 Protein Motion

Changes in protein conformations play a vital role in biochemical processes, from bioploymer synthesis to membrane transport. Depending on the particular function of the protein, these motions range from side-chain movements to re-orientation of complete domains. Table 2.1 shows a classification of protein motions according to their frequency.

| Time Scale [s] | Amplitude [Å] | Description |
|---|---|---|
| $10^{-15} - 10^{-12}$ | $0.001 - 0.1$ | bond stretching, angle bending |
| $10^{-12} - 10^{-9}$ | $0.1 - 10$ | side-chain motion, loop motion |
| $10^{-9} - 10^{-6}$ | $1 - 100$ | domain motion, small peptide folding |
| $10^{-6} - 10^{-1}$ | $10 - 100$ | protein folding |

Table 2.1. Classification of protein motions

Proteins move on a highly complex and rugged free energy landscape with several regions of low free energy that can be seperated by high barriers. Many of these conformations are important for function, e.g. one conformation may allow entrance of a ligand or binding to another protein. Ligands often cause dramatic conformational changes as they alter the free energy landscape.

Such ligand triggered conformational changes are of tremendous importance in signal cascades as they may stabilize a protein in an active conformation enabling the protein to bind another protein, a ligand, or to a specific region of RNA/DNA. Receptor proteins for instance, bind proteins or ligands on the extracellular side, causing a conformational change on the intracellular side that again influences action within the cell. Such allosteric mechanisms denote an elaborate way of information flow.

Despite the complexity of protein structures and the huge number of degrees of freedom, the functionally relevant protein motions are usually not coupled with extensive disturbance of local order. Moreover, many proteins can be described as rigid domains connected by flexible linkers. The domains keep their internal structure, mainly driven by the hydrophobic effect.

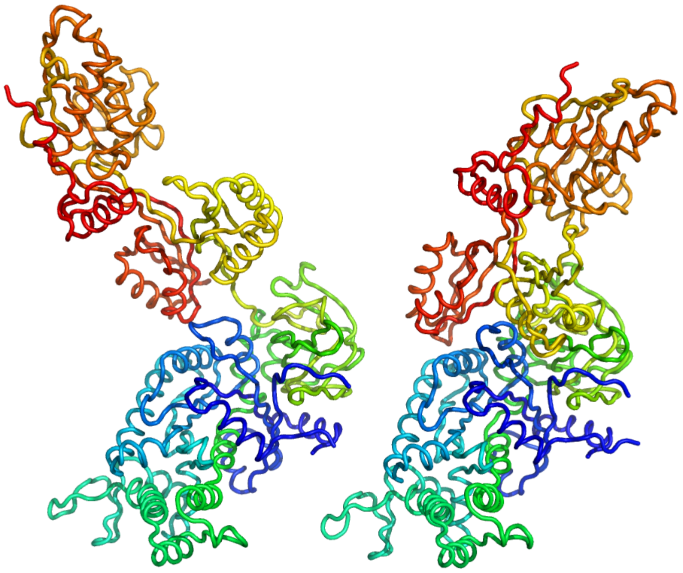


Figure 2.3. **Domain motion in ribosomal translocase EF-2.** Right panel: apo state. Left panel: Sordarin bound state.

## 2.4 Simulation Methods

As the conformational flexibility of proteins is often not directly accessible with experimental methods, this field of research is intensively addressed by computational methods. Predominantly molecular dynamics simulations are employed to obtain dynamic properties of proteins. However, despite the enormous increase in computer power and advances in algorithm techniques and parallelisation, MD simulations are computationally expensive and limited to the nanosecond or microsecond timescale for most systems. Thus, conformational transitions that include crossing of high energy barriers can often not be observed within the accessible time. In order to alleviate this sampling problem, a number of molecular dynamics based simulation methods have been developed and successfully applied to numerous problems within the field of protein research, among them Replica-Exchange Molecular Dynamics

(REMD) [10], Conformational Flooding [11, 12] and Targeted Molecular Dynamics(TMD) [13, 14]. Computationally more efficient, however less accurate methods, are based on gaussian network models [15, 16], normal mode analysis [17–20] or graph theoretical approaches [21].

### 2.4.1 Molecular Dynamics

Molecular Dynamics (MD) simulations describe the evolution of a molecular system in time. In conventional MD simulations atoms are treated as particles, which obey Newton's equations of motion. Therefore, three assumptions are made. (i) nuclear and electronic motions are decoupled (Born-Oppenheimer approximation), (ii) nuclei behave as classical particles, and (iii) the interactions between the particles are described using an empirical force field.
The general idea of the Born-Oppenheimer approximation is the separation of slow and fast degrees of freedom. The wavefunction $\psi$ in the time-dependent Schrödinger equation

$$\widehat{H}\psi = i\hbar\frac{\partial\psi}{\partial t} \tag{2.12}$$

is a function of the cooridnates and momenta of both, nuclei and electrons. Since nuclei are much heavier than electrons, it is a good approximation to regard the nuclear and electronic motion as decoupled. Thus, the electronic wavefunction $\psi_e$ only depends parametrically on the nuclear coordinates and the total wavefunction $\psi_{tot}$ can be seperated into an electronic and a nuclear part.

$$\psi_{\text{tot}}(\mathrm{r}, \mathrm{R}) = \psi_n(\mathrm{R})\psi_e(\mathrm{r}; \mathrm{R}) \tag{2.13}$$

where **R** = ($\mathbf{R}_1$,$\mathbf{R}_2$,...,$\mathbf{R}_N$) denotes the coordinates and momenta of the *N* nuclei and **r** = ($\mathbf{r}_1$,$\mathbf{r}_2$,...,$\mathbf{r}_M$) the coordinates and momenta of the *M* electrons, respectively. The resulting time-independent Schrödinger equation for the electrons

$$\widehat{H_e}\psi_e(R, r) = E_e(R)\psi_e(R, r) \tag{2.14}$$

can then be solved for fixed nuclei positions. Thus, the nuclei now move in an effective potential, given by the ground state energy $E_e(R)$ which describes the influence of the electron dynamics on the nuclei motion. This approximation usually holds very well.
For a typical macromolecular simulation system with thousands of atoms, the solution of the time-dependent Schrödinger equation for the nuclear motion is prohibitively expensive. Therefore, in classical MD it is assumed that particles obey Newton's equations of motion (Newton's second law)

$$-\nabla_i V(\mathrm{R}) = m_i \frac{d^2\mathbf{R}_i(t)}{dt^2}, \text{or} \tag{2.15}$$

$$\mathbf{F}_i = m_i \mathbf{a}_i, \tag{2.16}$$

where $V(\mathrm{R})$ is the potential energy, and $\mathbf{R}_i$ and $m_i$ are the coordinates and mass of atom $i$, respectively. The force $\mathbf{F}_i$ acting on this atom determines its acceleration $\mathbf{a}_i$ which, within a descrete time step $\Delta t$, leads to a change of the velocity and position of the atom. The time step $\Delta t$ has to chosen small enough to capture the fastest motions in the system. Under normal conditions, Newton's second law is a good approximation for macromolecular systems. However, quantum effects such as the behaviour at low temperatures or the tunneling of hydrogen atoms can not be described.

The third approximation is necessary since the evaluation of the potential $V(\mathrm{R})$ by solving the electronic Schrödinger equation is currently still too expensive, rendering extensive simulations of biomolecules in water unfeasible. Therefore, the potential energy is expressed as a sum of simple and easy-to-compute analytical functions, which, in combination with a correspnding set of empirical parameters, make up the so-called molecular mechanical (MM) force field, e.g.,

$$\mathrm{V} = \sum_{\text{bonds } i} \mathrm{V}_B^i + \sum_{\text{bond angles } j} \mathrm{V}_\alpha^j + \sum_{\text{impropers } k} \mathrm{V}_{imp}^k + \sum_{\text{dihedrals } l} \mathrm{V}_D^l + \sum_{\text{pairs } m,n} (\mathrm{V}_{Coul}^{m,n} + \mathrm{V}_{LJ}^{m,n}) \, . \tag{2.17}$$

The number of energy terms, their exact function, and the individual parameters vary between different force fields. Popular force fields are OPLS [36, 37], AM-

BER [38, 39], CHARMM [40] and GROMOS [41, 42]. In all these force fields, atoms are represented as point charges and electrostatic interaction between them is described by the Coulomb law

$$\mathrm{V}_{Coul}(\mathbf{R}, q) = \sum_{\text{pairs } m,n} \frac{q_m q_n}{4\pi\epsilon_0\epsilon_1 \mathrm{R}_{m,n}}. \tag{2.18}$$

Pauli repulsion and Van-der-Waals attraction is typically cast in the form of the Lennard-Jones term,

$$\mathrm{V}_{LJ}(\mathbf{R}) = \sum_{\text{pairs } m,n} \left[ \frac{C_{12}(m,n)}{R_{m,n}^{12}} - \frac{C_6(m,n)}{R_{m,n}^{6}} \right] \tag{2.19}$$

where the parameters $C_{12}$ and $C_6$ are the repulsion and attraction coefficents. Since bonds are approximated by a harmonic potential, bond breaking cannot take place in a molecular mechanics force field. Moreover, since bond vibrations represent the fastest motion in the system and limit the time step, they are often treated as constraints by employing the SHAKE [43] or LINCS [44] algorithm, which allows the time step to be increased to 2 fs. Molecular dynamics simulations of biomolecules are usually carried out in explicit solvent. Frequently used water model are SPC [45], SPC/E [46], TIP3P and TIP4P [47]. A detailed review over different water models is given in [48]. An extensive study on the accuracy of water model/force field combinations is given in [49].

### 2.4.2 Replica-Exchange Molecular Dynamics

Within the short nanosecond time scale accessible to conventional MD simulations, the system often cannot overcome larger energy barriers to regions of the configurational space that are sampled at physiological conditions. In this case, the obtained conformational ensemble often does not cover all functionally relevant conformations. Replica-Exchange Molecular Dynamics (REMD) is an MD-based simulation method which enables enhanced conformational sampling by making use of increased temperatures. In REMD simulations, a number of copies (replicas) of the system are simulated simultaneously at differerent temperatures with conventional MD. Pairwise exchange of replicas is attempted repeatedly after a number of MD steps. This allows the system to overcome energy barriers that would not be surmounted by conventional MD within accessible time. The exchange probability is calculated using the metroplis criterion

$$P = min(1, e^{-\beta(E(i+1)-E(i))}), \tag{2.20}$$

where P is the acceptance probability of an attempted step and $\beta$ denotes the inverse temperature, $\beta = \frac{1}{k_B T}$, with k$_B$ the Boltzmann constant. Although dynamic information gets lost in REMD simulations, the single replicas still represent Boltzmann-ensembles of the system at the respective temperatures.

### 2.4.3 Normal Mode Analysis

Normal mode analysis is one of the major simulation techniques used to probe the large-scale, shape-changing motions in biological molecules [50–52]. These motions are often coupled to function and a consequence of binding other molecules like substrates, drugs or other proteins. In NMA studies it is implicitely assumed that the normal modes with the largest fluctuation (lowest frequency modes) are the ones that are functionally relevant, because, like function they exist by evolutionary “design” rather than by chance.
Normal mode analysis is a harmonic analysis. The underlying assumption is that the conformational energy surface can be approximated by a parabola, which is known to be not correct since functional modes at physiological temperatures are highly anharmonic [51, 53]. To perform a normal mode analysis one needs a set of coordinates, a force field describing the interactions between constituent atoms, and software to perform the required calculations. The performance of a normal mode analysis in Cartesian coordiante space requires three main calculation steps.
1) Minimization of the conformational potential energy as a function of the atomic coordinates.

2) The calculation of the so-called “Hessian” matrix

$$H(f) = \frac{\partial^2 f}{\partial x_i \partial x_j} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix} \tag{2.21}$$

which is the matrix of second derivatives of the potential energy with respect to the mass-weighted atomic coordinates.
3) The diagonalization of the Hessian matrix. This final steps yields eigenvalues and eigenvectors (the “normal modes”).

Energy minimization can require quite a lot of CPU time. Furthermore, as the Hessian matrix is a 3N×3N matrix, where N is the number of atoms, the last step can be computationally demanding.

### 2.4.4 Elastic Network Models

Elastic network models [54] are basically a simplification of normal mode analysis. Usually, instead of an all atom representation, only C$_\alpha$ atoms are taken into account. This means a ten-fold reduction of the number of particles which decreases the computational effort dramatically. Moreover, as the input coordinates are taken as ground state, no energy minimization is required. The potential energy is calculated according to

$$V = \frac{\gamma}{2} \sum_{|r_{ij}^0| < R_C} (r_{ij} - r_{ij}^0)^2 \tag{2.22}$$

where $\gamma$ denotes the spring constant and R$_C$ the cut-off distance. Regarding the drastic assumptions inherent in the normal mode analysis, these simplifications do not mean a severe loss of quality. This together with the relatively low computational cost explain the current popularity of elastic network models.

## 2.5 Geometry-Based Molecular Simulation

Molecular structures are represented by coordinates of atoms. If the topology of the molecule is constant, which means that no chemical changes occur, the flexibility of the molecule is restricted to conformational changes. Conformational isomers are generated by rotating a bond of a molecule. Hence, if we regard bond lenghts and angles as constant, the internal degrees of freedom of a molecule are determined by the number of torsion angles and the number of conformations $\mathcal{C}$ can be calculated according to

$$\mathcal{C} = (\frac{360}{\Delta\varphi})^N, \tag{2.23}$$

with N the number of torsions and $\Delta\varphi$ the torsion angle range used for discretization. Even if we take a large bin size $\Delta\varphi$ of 30 degrees per conformation we obtain $1728$ different conformations for butane with $3$ rotatable bonds, $20736$ conformations for pentane and $248832$ for hexane. This is still manageable on a computer, however it examplifies that due to the power N dependency this approach is limited to molecules of limited size.

Usually only those conformers are of interest which belong to minima on the free energy landscape as these are the conformations most likely to be observed according to the Boltzmann distribution

$$\frac{N_i}{N_j} = \frac{g_i}{g_j} exp \frac{-(E_i - E_j)}{RT} \tag{2.24}$$

where $N_i$ and $N_j$ denote the number of molecules in state $i$ and $j$, respectively. $E_i$ and $E_j$ are the corresponding energies, $R$ the gas constant and $T$ the temperature. $g_i$ and $g_j$ are the degeneracies, or the number of states having the energy $E_i$ or $E_j$, respectively. This holds for small molecules as well as for macromolecules. In this case, only a very small part of the conformational space can be explored. Thus, the aim is to find a way to reduce the conformational space such that it still contains the most *relevant* conformations, more precisely those with low free energies. This can be achieved by introducing additional conditions, namely constraints. In the example of linear alkanes, such an additional condition could be to only regard *staggered* conformations, which reduces the number of conformations per rotatable bond to 3, leading to 27 conformations for butane, and 81/243 for pentane/hexane. Hence, an intelligent choice of constraints is essential to reduce the search space to computationally accessible dimensions.

### 2.5.1 Geometrical Constraints in Protein Structures

For macromolecules like proteins such an intelligent choice of constraints is a difficult task. If we assume physiological conditions the 3-dimensional structure of a protein is determined by its sequence, the solvent and in some cases small molecules that interact with the protein. The amino acid chain arranges such that the free energy is minimal. This is achieved by optimal intramolecular interactions, interactions between protein and solvent and a most favourable entropic contribution achieved by burying hydrophobic residues in the core of the protein. Although protein function is a dynamic process and significant conformational changes have been determined experimentally, most of the protein's local structural properties are conserved. However, only few unfavourable interactions can lead to a dramatic increase of the available conformational space. The discrimination between favourable and unfavourable interactions and thus, the determination of the geometrical constraints of a protein is therefore of major importance to reduce the overall conformational space to the *functionally relevant* one. Consequently, the thourough analysis of protein structures and the interactions determining structure and function capture a significant part of this work.

### 2.5.2 Structure Generation: The CONCOORD-algorithm

As every geometric formation, a molecular structure can be described using *external* or *internal* coordinates. The latter define particular atom positions relative to others. The geometry of three atoms $i$, $j$ and $k$ can be described by the squared distances $\mathrm{d}_{ij}^2$, $\mathrm{d}_{ik}^2$, and $\mathrm{d}_{jk}^2$.

$$\begin{aligned}
\mathrm{d}_{ij}^2 &= (x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2 \\
\mathrm{d}_{ik}^2 &= (x_i - x_k)^2 + (y_i - y_k)^2 + (z_i - z_k)^2 \\
\mathrm{d}_{jk}^2 &= (x_j - x_k)^2 + (y_j - y_k)^2 + (z_j - z_k)^2
\end{aligned} \tag{2.25}$$

Mathematically this is a system of quadratic equations which can be solved using basic linear algebra and yield the *external* coordinates of the atoms of the system. In order to obtain information about the flexibility of a structure, the equalities in 2.25 which serve as internal coordinates of the system are replaced by constraints since relative atom positions are not fixed but allowed to adopt a certain range of values. Constraints can be expressed as inequalities, more precisely quadratic inequalities when applied to distances.

$$\begin{aligned}
\mathrm{d}_{ij}^2(\mathrm{min}) \leq (x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2 \leq \mathrm{d}_{ij}^2(\mathrm{max}) \\
\mathrm{d}_{ik}^2(\mathrm{min}) \leq (x_i - x_k)^2 + (y_i - y_k)^2 + (z_i - z_k)^2 \leq \mathrm{d}_{ik}^2(\mathrm{max}) \\
\mathrm{d}_{jk}^2(\mathrm{min}) \leq (x_j - x_k)^2 + (y_j - y_k)^2 + (z_j - z_k)^2 \leq \mathrm{d}_{jk}^2(\mathrm{max})
\end{aligned} \tag{2.26}$$

In many fields of sciences optimization problems with inequalities as side conditions are addressed. In most cases however, the focus lies on optimizing a certain function (e.g. production costs in economic sciences) with inequalities as side conditions (machine A cannot produce more than x parts per day). When generating protein structures one could think of the free or potential energy as a value to optimize. However, the lowest energy configuration of a protein structure is not the only interesting one, since under physiological conditions the thermal energy causes proteins to adopt different conformations many of which are relevant for function. Therefore, usual optimization techniques are not appropriate.

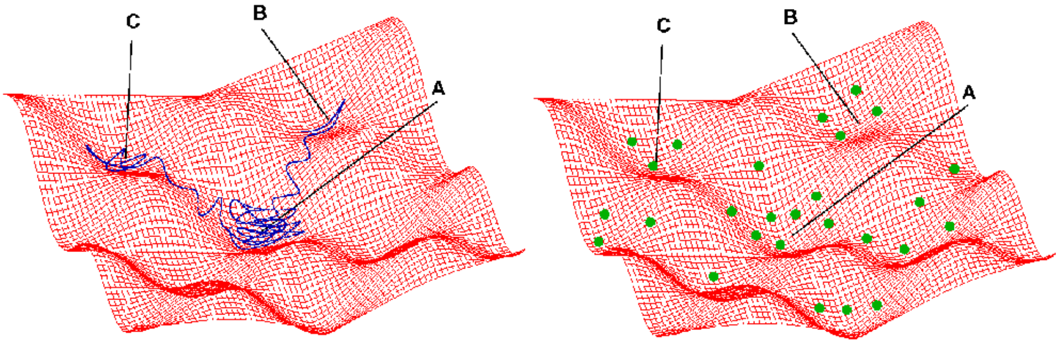


Figure 2.4. **Conformational sampling.** The left panel shows a schematized sampling of a MD-trajectory (blue line). A simulation starting from conformation A has to overcome energy barriers to sample the conformations B and C. Depending on the barrier height, these conformations are not sampled within accessible time. The right panel shows a CONCOORD-sampling. The green dots represent structures which are predicted from geometrical considerations. Energy barriers do not affect the sampling, however no information about the path between conformations is obtained. The choice of geometrical constraints determines the size of the sampled conformational space and the energy of the predicted structures.

Predicting protein conformations with feasible free energies based on geometric considerations is the objective of the CONCOORD-algorithm [22]. Starting from random coordinates, atom positions are adjusted iteratively until all predefined constraints are fulfilled. Repeating this procedure several times leads to an ensemble of structures as a representation of the conformational space which is accessible within the defined constraints. As the initial condition for each run is a random configuration, every generated structure is independent from the previous one. On the one hand this implies that no information is obtained about the path along which two conformations are connected and possible energy barriers between them. On the other hand, this approach enables crossing of even high energy barriers and finding other possible conformations. Hence, the CONCOORD approach does not suffer from a sampling problem like other simulation approaches like MD. Figure 2.4 shows the sampling properties of an MD-simulation (left) and a CONCOORD-ensemble (right) on an idealized energy landscape. In an MD-simulation every configuration is determined by the previous one, the energy landscape is basically explored by a *walk* and the sampling is limited by energy barriers. In a CONCOORD simulation, all configurations are independent. Instead of a *walk jumps* are performed on the energy landscape which enables extensive conformational sampling within few hours of CPU time. Moreover, the generated ensembles also include conformational substates that are seperated by energy barriers which can not be surmounted by MD-simulations within reasonable time. At this point the importance of the constraint selection becomes evident since they implicitly determine the ensemble properties of proposed configurations.

## 2.6 Principal Components Analysis

Protein structure ensembles, either from simulation or experimental data, are often analyzed by a Principal Components Analysis (PCA) to extract the essential degrees of freedom. PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

PCA can be used for dimensionality reduction in a data set by retaining those characteristics of the data set that contribute most to its variance, by keeping lower-order principal components and ignoring higher-order ones. Such low-order components often contain the "most important" aspects of the data.

In protein research, these data can be molecular dynamics trajectories or structure ensembles. The functionally relevant motions of proteins are often the low-frequency motions that correspond to the eigenvectors of the covariance matrix with the largest eigenvalues.

After superposition to a common reference structure, a variance-covariance matrix of positional fluctuations is constructed:

$$\mathbf{C} =< (\mathbf{x}(t)- < \mathbf{x} >)(\mathbf{x}(t)- < \mathbf{x} >)^{\mathrm{T}} > \qquad (2.27)$$

where $<>$ denotes an ensemble average. The coordinates **x** are denoted as a function of time for clarity, but may be provided in any order and do not need to be time dependent. **C** is a symmetric matrix that can be diagonalized by an orthogonal transformation **T**:

$$\mathbf{C} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}^{\mathrm{T}} \tag{2.28}$$

with $\Lambda$ the diagonal (eigenvalue) matrix and **T** containing as comlumns the eigenvectors of **C**. The eigenvalues $\lambda$ correspond to the mean square eigenvector coordinate fluctuation, and therefore contain the contribution of each principal component to the total fluctuation. The eigenvectors are usually sorted such that the eigenvalues are decreasing eigenvalue. For a system of N atoms, **C** is a $3\mathrm{N}\times 3\mathrm{N}$ matrix. If at least $3\mathrm{N}$ configurations are used to construct **C**, then $3\mathrm{N}$-$6$ eigenvectors with non-zero eigenvalues will be obtained. Six eigenvalues should be exactly zero, of which the corresponding eigenvectors describe the overall rotation and translation (that is eliminated by the superposition). If only M configurations are available (with M<3N) then at most M-1 non-zero eigenvalues with corresponding eigenvectors will result. If $\mu_i$ is the $i$th eigenvector of **C** (the $i$th column of **T**), then the original configurations can be projected onto each of the principal components to yield the principal coordinates $\mathrm{p}_i$(t) as follows:

$$\mathrm{p}_i(t) = \mu_i \cdot (\mathbf{x}(t) - <\mathbf{x}>) \tag{2.29}$$

The variance $<\mathrm{p}_i^2>$ equals the eigenvalue $\lambda_{\mathrm{i}}$. These projections can be easily transformed back to cartesian coordinates for visualization purposes as follows:

$$\mathbf{x}'_i(t) = \mathrm{p}_i(t) \cdot \mu_i + <\mathbf{x}> \tag{2.30}$$

Two sets of eigenvectors $\mu$ and $\nu$ can be compared to each other by taking inner products:

$$I_{ij} = \mu_i \cdot \nu_j \tag{2.31}$$

Subspace overlaps are often calculated as summed squared inner products:

$$O_n^m = \sum_{i=1}^{n} \sum_{j=1}^{m} (\mu_i \cdot \nu_j)^2 \tag{2.32}$$

expressing how much of the n-dimensional subspace of set $\mu$ is contained within the m-dimensional subspace of set $\nu$.

# Chapter 3

# Parametrization from Experimental Data

*Durch Heftigkeit ersetzt der Irrende, was ihm an Wahrheit und Kräften fehlt.*

*- Johann Wolfgang von Goethe*

## 3.1 Introduction

The 3-dimensional structure of proteins is determined by covalent bonds, non-covalent interactions like electrostatics and Pauli-repulsion, and entropic contributions. The sum of these interactions leads to a well-defined geometry with restricted conformational flexibility. Although proteins are found to be highly divers in their overall structure, their local geometry is highly conserved. Besides well-defined length distributions of covalent bonds, also a characteristic distribution of backbone dihedral angles and hydrogen bond geometries are observed in all protein structures, regardless of their sequence and function. The satisfaction of such local geometrical constraints is therefore an important quality check for protein structures. Commonly employed structure validation programs like WHATIF [55] and PROCHECK [56] assess the quality of a structure by comparison of local geomtries in the given structure with the distributions of the corresponding geometries in a database of protein structures.

Since tCONCOORD[1] ($t$ stands for transition) builds protein structures, extensive parametrization of simulation parameters is mandatory to generate structures that satisfy the same quality criterions as for experimentally determined structures. In this chapter we describe how simulation parameters are derived from experimental data using a newly developed program termed PDBBrowser. Since interatomic distances are crucial for structure quality we derived a complete set of atomic radii from high-resolution X-ray structures and show furthermore how these radii can be used to describe packing properties in protein structures, thereby revealing that atomic packing is strongly resolution dependent.

## 3.2 Experimental Data

The Protein Data Bank [35] contains data regarding the 3-dimensional structure of proteins. The predominant contingent of this data has been derived from X-ray diffraction on protein crystals, however also data from NMR-experiments and electron microscopy is available. tCONCOORD requires a lot of parameters for constraint definition and structure generation. The quality of generated structures as well as ensemble properties heavily depend on the chosen parameter set. Therefore, the data set used for the derivation of simulation parameters should represent the most reliable data currently available. Since electron microscopy usually does not provide data at atomic resolution, its use is not eligible. Also the reliability of structure models derived from NMR data is not sufficient for this purpose. For the different parametrization processes in this work we exclusively used X-ray structures that have been resolved to high resolution, depending on the particular purpose either $< 1.2\,\text{Å}$ or $< 1.6\,\text{Å}$.

Instead of taking the hydrogen positions that are available for a number of high-resolution X-ray structures, hydrogen positions were generated using the HB2NET module of WHAT IF [57]. We chose this strategy as only few data sets are complete, and because the bond lengths for C-H, N-H and O-H are systematically underestimated in X-ray diffraction [58]. A further advantage of the

---

[1] *http://www.mpibpc.mpg.de/groups/de_groot/dseelig/tconcoord.html*

employed hydrogen placement algorithm is that it evaluates different protonation states and optimizes the hydrogen bond network within the structure, including side-chain flips of histidine, glutamine and asparagine residues, when appropriate.

## 3.3 The PDBBrowser: A Tool for Flexible Database Queriing

Building protein structures with low free energies without actually using explicit energy functions requires extensive knowledge about the underlying structural determinants in atomic detail. Hence, thorough parametrization from experimental data is mandatory to predict reliable protein structures. As these data cannot be obtained by simply using the predefined query features of the Protein Data Bank [35], protein structure data has to be transformed into a queriable storage format, enabling the derivation of any kind of distribution which can be obtained from protein structure files.

To this end, a database query solution has been developed that allows quick, flexible and detailed queriing of properties from structure data, e. g. answering questions of the kind "*What is the distribution of $C_\alpha$-$C_\alpha$ distances if the two corresponding residues form a backbone-sidechain hydrogen bond and at least one residue is a Valine or Leucine and has a Tryptophane residue in its neighborhood?*" This program, termed *PDBBrowser*, has been developed and used to derive all parameters that are used in the newly developed tCONCOORD program.

### 3.3.1 Program Structure

The PDBBrowser consists of a C-library, an interface from the C-library to the object-oriented scripting language Python, a Python module and a Python executable. The kernel, written in C, carries out all computationally demanding operations like neighbor-searching and calculation of distances and angles. Furthermore it assigns atom types, atomic radii and other properties to each atom.

A Python-interface convertes the C-structure into a “Python-readable” structure (PyObject *) that can be accessed from the interpreter level. At the Python-level, the data is converted into comfortable classes (Fig. 3.1) which provide the possibility to select the particularly interesting atoms or residues. A protein structure is stored as an object of the class *Model*, which contains a list of *Chain*-objects, a list of *Molecule*-objects and *Atom*-objects.

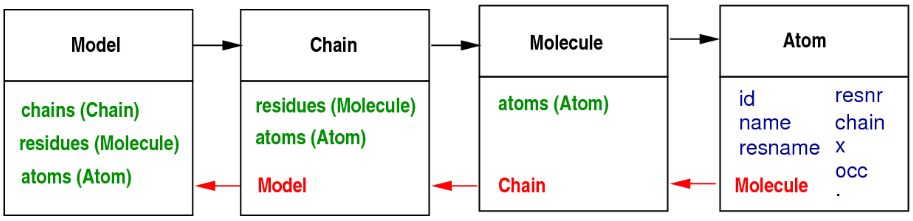


Figure 3.1. Schematised representation of the Python-classes in the PDBBrowser. Each box represents a class. Different data types are indicated by different colors. Black represents a class, green a list, blue an attribut and red refers to the superior class.

```
for atom in model.atoms:     # loop over all atoms
    if atom.name == 'CA':    # select CA-atoms
        print atom.x         # print coordinates

mol = model.residues[0]           # the first residue
mol = model.chains[1].residues[0] # the first residue
                                  # of the second chain
ch  = model.chains[-1]            # the last chain
```

Listing 3.1. Examples for selections in the PDBBrowser

The structure of the Python-classes allows both, an easy way to select atoms, residues or chains of interest and to obtain statistics of particular observables. The latter can be done by built-in statistics functions or by incorporation of Python-modules like scipy or numarray, which provide a broad range of optimized mathematical routines and statistics modules.

### 3.3.2 Database Queries

The PDBBrowser can be used to carry out any kind of database query. As input information, it requires a “job file” which must be written in Python-syntax with additional key words (Listing 3.2). The flexibility of the program becomes evident when non-standard distributions like distances or angles are inquired. The user can define arbitrary functions using the Python-language and optionally other Python-modules, e. g. optimized linear algebra routines from the scipy package [59].

```
%DATABASE = '/storage/structs/'
%MODE     = 'mult'

%newjob

 def ca_n_bond_length():
     output = 'ca_n_blen.dat'         # name of the output file
     histo_size = [1.2,1.6,0.02]      # size of the histogram
     for atom in model.atoms:         # loop over all atoms
         if atom.name == 'CA':        # select CA-atoms
             for at in atom.bonds:    # loop over bound atoms
                 if at.name == 'N':   # select atom
                     dist = atom-at   # calculate distance
                     to_file(dist)    # store this value

%endjob
```

Listing 3.2. Example of job file for the PDBBrowser: Functions, written between the key-words %newjob and %endjob are interpreted by the PDBBrowser as database functions and run over the database defined by the key-word %DATABASE. The %MODE-key-word defines whether the database will be loaded once ('mult') or file by file ('single')

The potency of this ansatz of organizing data is furthermore, that an arbitrary number of queries can be carried out by loading the database only once. Merely the size of the database, or actually the number of structures that can be loaded, is limited by the memory of the computer. For the parametrizations carried out for this work, databases of 200-300 structures were used.

Besides bond and angle parameters, non-covalent atomic contacts are of special interest since they heavily influence structure quality and conformational freedom.

## 3.4 Optimization of Atomic Radii from High-Resolution X-ray Structures

Protein structures are stabilized by many different interactions. The lower limit of distances between atoms can either be dominated by electrostatic repulsion or Pauli repulsion. In our approach we do not make any assumptions about the underlying potential, except that there is a minimum distance d$_{ij}$ for each pair of atoms below which strong repulsion takes place. Accordingly, a distance range $\mathrm{d_{ij}} + \Delta\mathrm{r}$ is defined for atoms to be contacting.

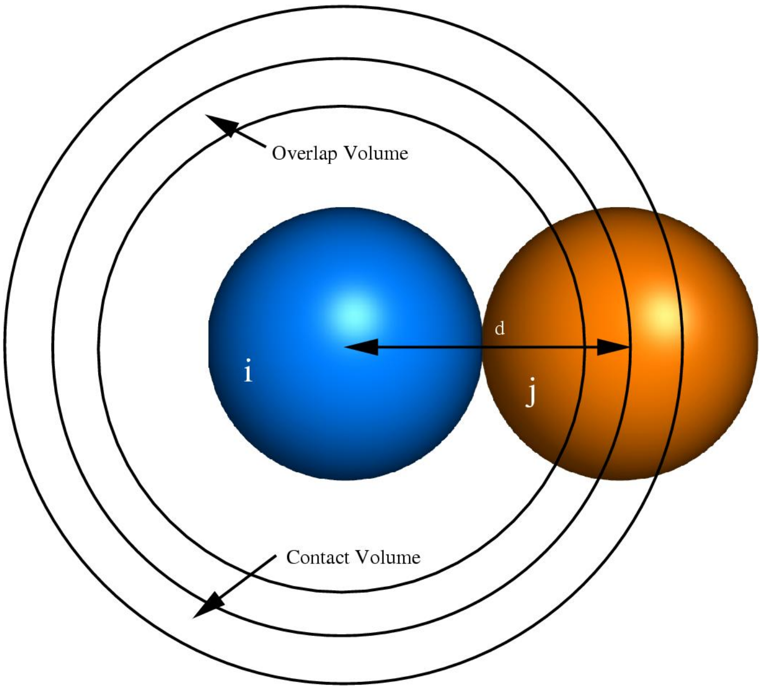


Figure 3.2. **Contact volume and overlap volume.** If an atom j is found within the sphere shell labeled as contact volume it is counted as a favourable contact. If the distance between atom i and j is smaller than the VdW sum it is counted as an overlap and weighted with the overlap volume.

If *universal* atomic radii exist, many atom pairs should be found at their contact distance. To test this hypothesis, atomic radii were derived that maximize the number of contacts while minimizing the number of overlaps in a set of $106$ high resolution X-ray structures (resolution $< 1.2\,\text{Å}$) from the Protein Data Bank (PDB). In total, about $5.7$ million atom pairs were used in the optimization procedure. Atom pairs with fewer than four bonds in between were excluded.

To reflect the chemical nature of atoms in protein residues we defined $35$ atom types adapted from the OPLS-AA force field [36]. VdW distances were taken as the sum of VdW radii $\mathcal{S}_{ij}$ except for a set of polar and charged atom types for which specific combinations were defined to realistically account for hydrogen bonding and electrostatic repulsion. All structures were protonated with the WHATIF [57] software package. Since hydrogens atoms were added computationally, the obtained radii may depend on the chosen method of hydrogen placement. The position of hydrogen atoms is in most cases well defined by the local geometry of the surrounding heavy atoms. Only the hydrogen position in OH, $CH_3$, $NH_3$ and the protonation state of histidines is sometimes ambiguous. However, the number of undefined hydrogen positions is small compared to well defined ones, and are therefore not expected to influence the optimization significantly.
A contact $C_{ij}$ between two atoms i and j is defined as

$$\mathrm{C_{ij}} = \begin{cases} 0 & \text{for } \mathrm{d_{ij}} < \mathcal{S}_{\mathrm{ij}} \\ 1 & \text{for } \mathcal{S}_{ij} \leq \mathrm{d_{ij}} \leq \mathcal{S}_{\mathrm{ij}} + \Delta\mathrm{r} \\ 0 & \text{for } \mathrm{d_{ij}} > \mathcal{S}_{\mathrm{ij}} + \Delta\mathrm{r}. \end{cases} \tag{3.1}$$

Likewise an overlap $O_{ij}$ is counted if the distance is smaller than the sum of the VdW radii.

$$\mathrm{O_{ij}} = \begin{cases} 1 & \text{for } \mathrm{d_{ij}} < \mathcal{S}_{\mathrm{ij}} \\ 0 & \text{for } \mathrm{d_{ij}} > \mathcal{S}_{\mathrm{ij}}. \end{cases} \tag{3.2}$$

Both values $C_{ij}$ and $O_{ij}$ are weighted statistically by their according volume (figure 3.2). The contact volume $V_C$(i,j) is calculated with

$$\mathrm{V_C(i,j)} = \frac{4}{3}\pi\left[(\mathcal{S}_{\mathrm{ij}} + \Delta\mathrm{r})^3 - \mathcal{S}_{\mathrm{ij}}^3\right]. \tag{3.3}$$

For the overlap volume not the whole volume of the sphere is counted, as small overlaps occur more frequently than large ones. Therefore, like for the contacts we used a sphered shell as illustrated in figure 3.2.

$$\mathrm{V_O(i,j)} = \frac{4}{3}\pi\left[\mathcal{S}_{\mathrm{ij}}^3 - (\mathcal{S}_{\mathrm{ij}} - \Delta\mathrm{r})^3\right] \tag{3.4}$$

Now we defined a contact number density $\mathcal{N}_C^k$ for each atom type k.

$$\mathcal{N}_C^k = \sum_i \sum_j \frac{\mathrm{C_{ij}}}{\mathrm{V_C(i,j)}} \quad ; i \in k. \tag{3.5}$$

Likewise, the overlap number density $\mathcal{N}_O^k$ is defined as:

$$\mathcal{N}_O^k = \sum_i \sum_j \frac{\mathrm{O_{ij}}}{\mathrm{V_O(i,j)}} \quad ; i \in k. \tag{3.6}$$

The quantity to be optimised, the effective contact density, therefore is defined as:

$$\mathcal{N}^k = \mathcal{N}_C^k - \mathcal{N}_O^k \tag{3.7}$$

which is maximised through iteration for each atom type k. This way a set of *contact* atomic radii and combinations was derived.

A closer look at the derived contact radii listed in table 3.1 reveals that most carbon, nitrogen and oxygen radii are smaller compared to those from previous work [60–65]. This is mostly caused by the use of explicit hydrogen atoms. In comparison to Lennard-Jones parameters from force-fields, our atomic radii are also generally smaller. This is due to the fact that, in force-fields, the local geometry of atoms is determined also by other interactions, particularly electrostatic interactions. Our approach aims at a geometrical description that reflects the optimal contact distance between atom pairs as a combined effect of all interactions. A number of systematic deviations became evident during the optimisation that are found to be caused by the original classification of the atom types. Hence, a number of additional atom types were introduced. For instance an additional atom type was introduced for $\mathrm{C}_\alpha$ atoms (atom type CA) as we found that $\mathrm{C}_\alpha$'s form much closer contacts than other aliphatic carbon atoms. Likewise hydrogen atoms connected to $\mathrm{C}_\alpha$ atoms (atom type HA) form closer contacts than other un-polar hydrogens making them more similar in size to polar hydrogens (atom type H). This observation may indicate weak $C_\alpha$-H...O bonds that have recently been discussed [66, 67]. This example of a systematic protein-specific deviation shows that atomic radii derived from small molecule crystals [58,62] are not readily transferable to macromolecular structures like proteins and indicates the significance of a protein-specific set of atomic radii derived from atomic-resolution protein structures.

Additionally, a set of specific combinations of atom types was defined to realistically account for electrostatics like small hydrogen-bond distances. The combinations are listed in table 3.2. The very small radius for charged hydrogens (atom type HC) is remarkable but may in part be due to the small number of contacts that these atoms form. Hence, the statistics for this atom type is limited.

The distances calculated from the derived atomic radii and combinations represent the most likely observed distance for specific atom pairs in natively folded protein structures as a result of all interactions. Therefore, they are well suited to serve as parameter set in tCONCOORD or other simulation protocols.
The distribution of favourable contacts and overlaps in protein structures can be interpreted as packing property. Since this property should be very sensitive to small changes of atomic coordinates a resolution dependence might be expected. To test this hypothesis we developed a method to quantify this resolution dependence based on the derived atomic radii.

## 3.5 Atomic Packing in Protein Structures

Billions of years of evolution optimized proteins to fulfill their functions efficiently. Regardless whether the protein functions as enzyme, molecular motor, transport protein or receptor, a prerequisite for optimal function is a fine-tuned structural and dynamical framework, either directly or indirectly provided by the native structure of the protein. An important, but as yet unresolved question is which functional constraints exactly are imposed on a protein structure. Sequence and structure conservation patterns provide valuable hints in this respect, like the conservation of the structure in the catalytic site of an enzyme. However, such information is typically local and restricted to a specific class of proteins. The same holds for other localized structural constraints like disulphide bridges or

| Atom Type | Radius[Å] | Description | Atom Type | Radius[Å] | Description |
|---|---|---|---|---|---|
| H0 | 1.19 | unpolar hydrogens | CH2P | 1.47 | $C_{\beta,\gamma,\delta}$ in P |
| HAR | 1.14 | aromatic hydrogen | CY | 1.87 | $C_\gamma$ in Y |
| HA | 1.03 | $H_\alpha$ | CY2 | 1.63 | $C_\eta$ in Y |
| H | 1.05 | polar hydrogen | CF | 1.83 | $C_\gamma$ in F |
| HC | 0.58 | hydrogen in charged groups (R,K) | CDR | 1.69 | $C_\delta$ in R |
| HDR | 0.67 | $H_\delta$ in arginine | CR1H | 1.75 | $C_{\delta 2}$ in H |
| C | 1.43 | carbon in C=O | CRHH | 1.63 | $C_{\epsilon 1}$ in H |
| CA | 1.48 | $C_\alpha$ | O | 1.41 | oxygen in C=O |
| CH1E | 1.92 | aliphatic carbon with 1 hydrogen | OC | 1.33 | oxygen in $COO^-$ |
| CH2E | 1.89 | aliphatic carbon with 2 hydrogens | OH1 | 1.31 | oxygen in C-O-H |
| CH3E | 1.81 | aliphatic carbon with 3 hydrogens | NH1 | 1.37 | nitrogen with 1 hydrogen |
| CR1E | 1.81 | aromatic carbon | NH2 | 1.45 | nitrogen with 2 hydrogens |
| CR1W | 1.76 | $C_{\zeta 2}$,$C_{\eta 2}$ in W | NH3 | 1.35 | nitrogen with 3 hydrogens |
| C5 | 1.76 | $C_\gamma$ in H | NC2 | 1.45 | $N_\zeta$ in R |
| C5W | 1.86 | $C_\gamma$ in W | NHS | 1.40 | unprotonated N in H |
| CW | 1.74 | $C_\epsilon$ in W | SM | 1.79 | S in M |
| CH2G | 1.76 | $C_\alpha$ in G | S | 1.83 | S in C |

Table 3.1. Atomic radii derived from a set of 106 high-resolution X-ray structures.

specific salt bridges. Hence, the role of global structural determinants underlying or supporting function remains to be determined.

Protein design and engineering studies suggest a crucial role for packing in protein stability and function [68–71], including exact complementarity of neighboring side chains [72–74]. Even conservative mutations of single amino-acids can lead to destabilizations [75, 76]. Additionally, the inclusion of an explicit packing term in protein design algorithms has significantly improved the accuracy of designed predictions [73], indicating that optimal packing is a crucial factor in protein structures. Packing densities in protein cores have been described as high and comparable to solid crystals [60, 64]. However, beyond

| Atom | Types | D[Å] | Atom | Types | D[Å] |
|---|---|---|---|---|---|
| O | O | 3.3 | HC | NHS | 1.84 |
| O | H | 1.86 | H | NHB | 2.00 |
| O | OC | 2.84 | HC | NHB | 1.95 |
| OH1 | O | 2.64 | O | NC2 | 2.82 |
| O | HC | 1.70 | O | NH2 | 2.84 |
| OH1 | H | 1.62 | O | NH3 | 2.60 |
| OC | HC | 1.74 | O | NHS | 2.66 |
| OH1 | HC | 1.70 | NH1 | NHS | 2.88 |
| OC | H | 1.60 | NH2 | NHS | 3.00 |
| O | NH1 | 2.82 | NH3 | NHS | 2.84 |
| H | NHS | 2.0 | O | CA | 3.18 |
| HA | O | 2.30 | | | |

Table 3.2. Lower bounds for distances of specific atom type combinations

average densities and free volume considerations [77], the exact packing extent, in terms of atomic contacts, remains unknown.

Here, we have developed an approach to quantitatively determine the packing efficiency of a large set of protein structures at different levels of resolution. A "packing score" is introduced that allows a robust assessment of the degree of packing efficiency, resting on our derived set of atomic contact radii derived from a set of high resolution protein structures. We show that the distribution of close contacts and overlaps in protein structures is invariant and highly conserved in high-resolution X-ray structures, regardless of function, size and secondary structure.

The implications for protein structure validation, protein dynamics, structure prediction and design are discussed.

### 3.5.1 Quantitative Assessment of Packing Quality

Optimal packing in molecular systems is characterized by a maximum number of interatomic contacts. In proteins, the maximally attainable packing efficiency is primarily limited by the distribution of unequally sized atoms (C,H,N,O,S), by topological restraints imposed by the connectivity between atoms, and by secondary/tertiary structure restraints. In order to assess the relative degree of packing in native protein structures, we therefore quantified the packing efficiency, evaluated this packing score for a large number of proteins, and compared the results to a synthetic reference. The reference was constructed from a set of 1000 freely rotatable amino acids in solution, distributed according the frequency as observed in natural proteins (see tab. 3.3).

| Amino Acid | Number | Amino Acid | Number |
|---|---|---|---|
| ALA | 96 | LEU | 75 |
| ARG | 35 | LYS | 79 |
| ASN | 46 | MET | 15 |
| ASP | 70 | PHE | 43 |
| CYS | 18 | PRO | 46 |
| GLN | 28 | SER | 58 |
| GLU | 55 | THR | 67 |
| GLY | 78 | TRP | 14 |
| HIS | 26 | TYR | 39 |
| ILE | 42 | VAL | 70 |
| SOL | 180 | | |

Table 3.3. Synthetic Reference System.

This system was subjected to 20 ns of molecular dynamics simulation. Snapshots from this simulation were cooled down to 100K with simulated annealing. As this reference shares the restrictions of native protein structures of unequally sized atoms and connectivities, but has no restrictions due to secondary and tertiary structure and also has no surface or active site which may display poorer packing properties, this reference can be considered as upper limit of the packing efficiency for natively folded proteins, and hence may serve to estimate the relative packing efficiency of protein structures.

In contrast to previous studies [60–65], we do not consider packing in terms of occupied volume fractions. Rather, we focus on the thermodynamically determined distribution of favorable atomic contacts and unfavorable overlaps.

Contacts were counted for closely interacting ($d_{ij} \leq r_{ij} \leq d_{ij} * 1.3$), but non-overlapping atoms. The requirement of maximizing the number of contacts while minimizing the number of overlaps ($r_{ij} < d_{ij}$) ensures counting of true contacts in favor of any secondary maxima. A full set of protein atomic radii was obtained by iteratively adapting the atom radii for the different atom types (see chap. 3.4).

These contact radii were used to evaluate a packing score for a large set of protein structures at different levels of resolution. Non-protein residues like water and ions were neglected. The packing score was defined as the average volume-weighted number of contacts per atom minus the average volume-weighted number of overlaps.

Given a set of optimized atomic radii, a packing score $\mathcal{P}$ can be calculated from interatomic distances as:

$$\mathcal{P} = \frac{1}{N}\left[\sum_{i}\sum_{j>i}\frac{\mathrm{C_{ij}}}{\mathrm{V_C(i,j)}} - \sum_{i}\sum_{j>i}\frac{\mathrm{O_{ij}}}{\mathrm{V_O(i,j)}}\right]$$

$\mathcal{P}$ is high if the number of favourable contacts per atom is high in contrast to the number of overlaps per atom. Note that values do not exclusively depend on the quality of the protein structure, but also on the structure itself. Surface residues are mostly surrounded by solvent molecules which were not taken into account by this method. Therefore the number of contacts and also the number of overlaps with other protein atoms is rather small compared to residues located in the core of a protein. Thus $\mathcal{P}$ reaches the highest values for proteins that are almost spherically shaped.

The packing efficiency can be further illustrated by using a reduced radial distribution function $\mathcal{R}$. For this, for every atom i the distance $\mathrm{r_{ij}}$ to the neighboring atoms j is calculated and related to the optimal distance $\mathrm{d_{opt}} = \mathrm{r_i} + \mathrm{r_j}$ for this combination of atom types. The shape of the resulting reduced radial distribution function (fig. 3.3)

$$\mathcal{R}(\mathrm{r}) = \frac{\mathbf{r}_{\mathrm{ij}}}{\mathrm{d_{opt}}}$$

is found to be highly conserved in all high-resolution structures and can be considered as structural constraint on protein architecture. Values lower than 1 represent overlaps, whereas positive values close to 1 represent favourable contacts.

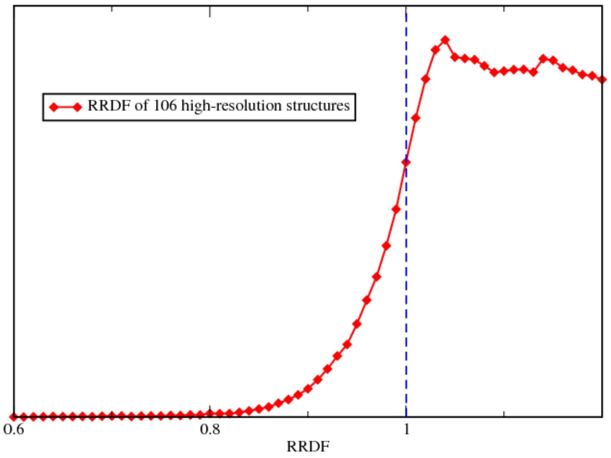


Figure 3.3. **RRDF.** Reduced radial distribution function of an ensemble of 106 high-resolution X-ray structures.

### 3.5.2 Packing quality in protein structures

For the synthetic reference ensemble of structures, built from the final configurations of the simulated annealing simulations, the same procedure for optimizing atomic radii was employed as described in chap. 3.4. Using these radii (data not shown), packing scores were calculated for the synthetic reference ensemble. The average value of these scores was scaled to 1.0 and serves as reference for the packing scores calculated from the experimental structures. The statistical error as estimated from the standard deviation in the ensemble is about 0.01, represented in fig. 3.4 by the thickness of the red line.

Packing scores were calculated for sets of protein structures determined by X-ray crystallography and NMR. X-ray structures were evaluated at different levels of resolution *(see Appendix 10.1)*. NMR structures were compared to refined ensembles from the DRESS [78] database (always the first model was taken from an NMR-ensemble, for this usually represents the lowest energy configuration).

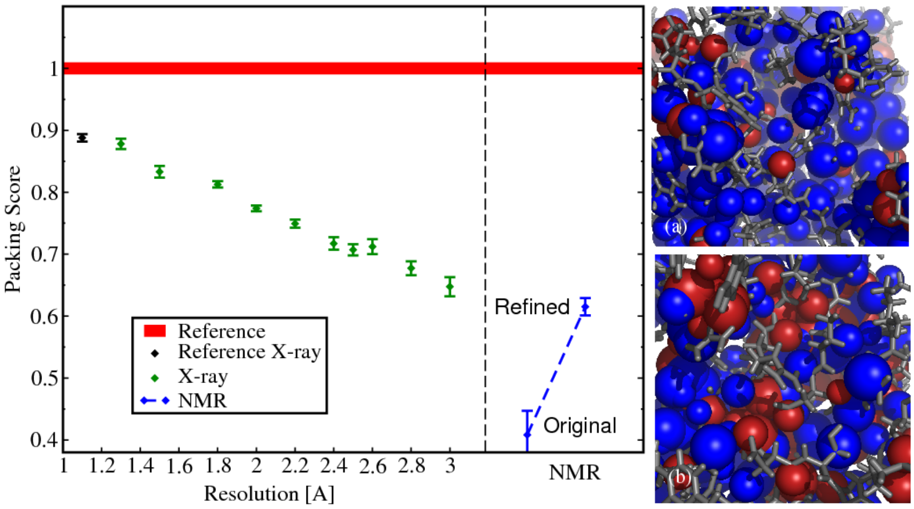


Figure 3.4. **Atomic Packing.** Left Panel: Packing scores. red line (reference): line thickness represents the standard deviation. black: the ensemble of high resolution structures that were used to derive the atomic radii; green: X-ray structures at different levels of resolution; blue: Ensemble of NMR-structures original from the PDB and the same structures from the DRESS database. Right Panel: Two structures of staphylococcal nuclease. (**a**) PDB 1EY4, resolved by X-ray crystallography (Resolution 1.6 Å). (**b**) PDB 1JOR, resolved by NMR. The blue colored atoms are well packed and embedded in their local environment. Red colored atoms cause overlaps with their neighbors.

The results relative to the synthetic reference are shown in fig. 3.4. Remarkably, packing scores of up to 88% of the synthetic reference (in red) were observed, indicating a high packing density for natively folded protein structures resolved at high-resolution. With decreasing resolution the packing efficiency is observed to decrease. While the packing scores for X-ray structures are located in a rather narrow range, values for NMR-structures (blue marks) show much more spread. This behavior is further exemplified for two structures of staphylococcal nuclease, of which one (PDB 1EY4) has been resolved by X-ray crystallography (resolution: 1.6 Å) and the other one by NMR (PDB 1JOR). The right panel of fig. 3.4 shows the difference in atomic packing for fragments of the two structures. In the X-ray structure, apart from surface exposed groups, all atoms are well-packed by nearly ideal contacts (overall packing score: 0.76). In the NMR structure of the same protein, the packing is found to be less ideal due to more overlaps and fewer contacts (overall packing score: 0.45).

The distribution of atomic contacts can be illustrated by a reduced radial distribu-

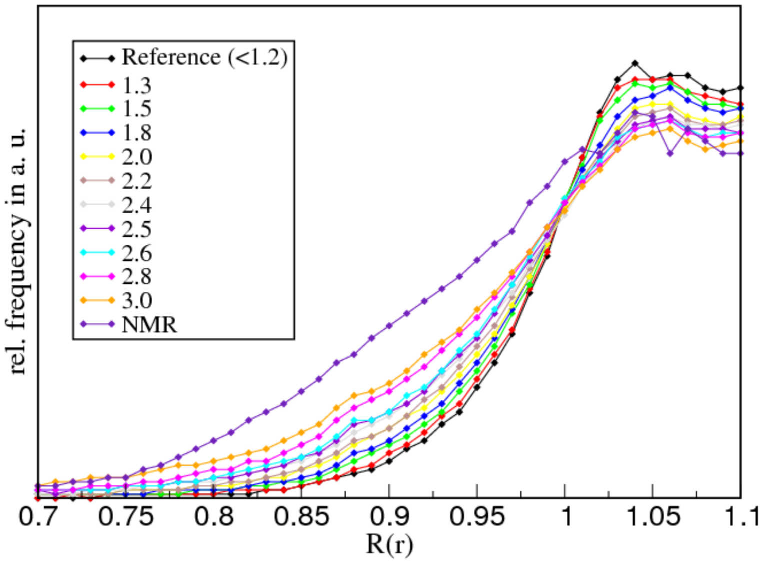


Figure 3.5. **RRDF.** Reduced radial distribution functions of sets of protein structures at different levels of resoultion.

tion function (RRDF), which is a standard radial distribution function normalized to the ideal contacts distance. This function displays all close contacts within a protein structure or an ensemble of structures. Values lower than $1.0$ represent energetically unfavorable overlaps that should occur infrequently according to the Boltzmann distribution. Fig. 3.5 shows the RRDF's of the same structure ensembles as used in fig. 3.4.

At high resolution the curves are found to be steeper, representing a favorable ratio of contacts and overlaps. Furthermore, the plot shows that the distribution of atomic contacts in NMR-structures differs significantly from those in X-ray structures. While the curves for the different levels of resolution basically differ in steepness, the curve corresponding to the NMR-structures shows a systematic deviation. The amount of overlaps is much higher, which can be interpreted as systematic overpacking, in agreement with previous findings [79–84].
The question arises whether the observed resolution dependence reflects protein flexibility or, rather, a resolution-imposed coordinate uncertainty. In other words, could inherent flexibility or disorder that results in limited resolution cause a non-optimal packing ("packing limits resolution") or does limited resolution prevent building of an accurate well-packed model ("resolution limits packing")?

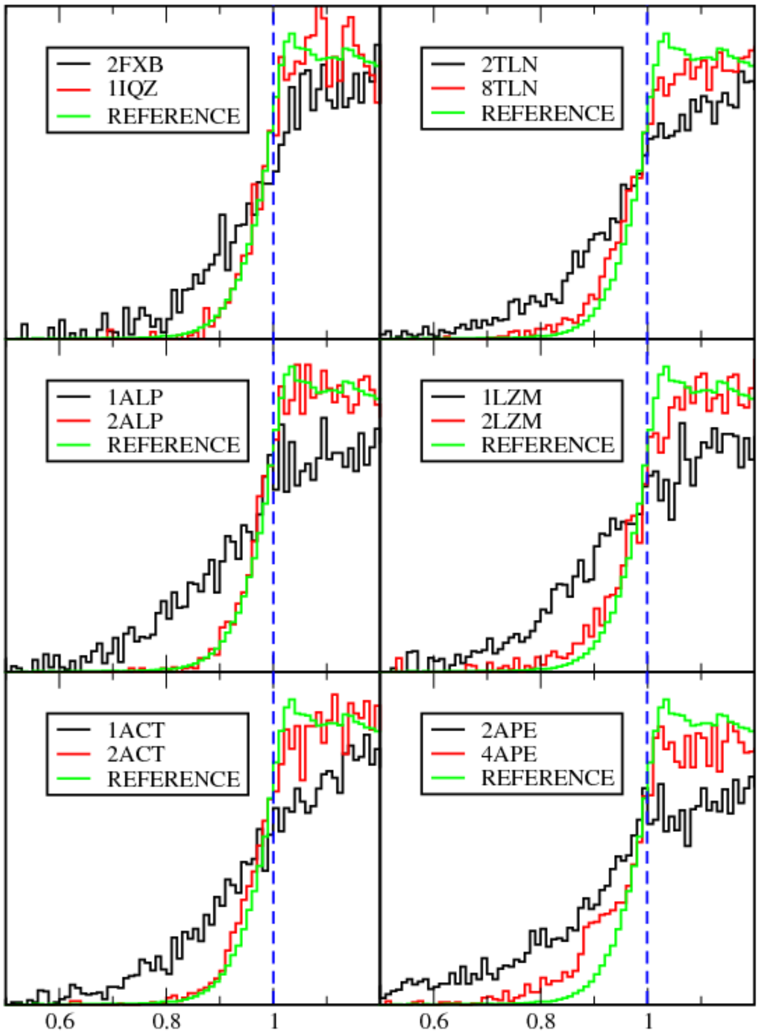


Figure 3.6. Comparison of identical protein structures at different levels of resolution. The black curves represent the reduced radial distribution functions of obsolete protein structures. The red curves represent the same function of the current PDB entries of these proteins. The green curve shows the RRDF of the reference set of $106$ high-resolution X-ray structures.

In order to address this question, we investigated several cases of the same protein structure solved at different levels of resolution. Comparison of packing scores of these identical protein structures shows that packing scores significantly increase at higher resolution (Tab. 3.4). The distribution of overlaps, represented by the left branch of the reduced radial distribution function $(\mathcal{R}(r) < 1)$, is a structural invariant for all protein structures. Fig. 3.6 shows RRDF's of identical

| PDB ID | Resolution[Å] | Packing Score | PDB ID | Resolution[Å] | Packing Score |
|---|---|---|---|---|---|
| 1ACT | 2.8 | 0.46 | 2APE | 2.5 | 0.29 |
| 2ACT | 1.7 | 0.89 | 4APE | 2.1 | 0.73 |
| 1LZM | 2.4 | 0.37 | 2TLN | 2.3 | 0.55 |
| 2LZM | 1.7 | 0.82 | 8TLN | 1.6 | 0.84 |
| 1ALP | 2.8 | 0.39 | 2FXB | 2.3 | 0.59 |
| 2ALP | 1.7 | 0.90 | 1IQZ | 0.92 | 0.90 |

Table 3.4. Comparison of packing scores for identical proteins. Upper line: obsolete structure; lower line: current PDB entry

proteins at different levels of resolution. The curves of the higher resolution versions of these protein structures are remarkably close to the reference curve, strongly supporting the "resolution limits packing" scenario and not the "packing limits resolution" scenario. Hence, overlap distributions and packing considerations could be used as quality check for protein structures. Additionally, these results suggest that a rigorous packing term may aid structure refinement.

Our results show that high resolution natively folded protein structures display a packing efficiency close to that of a condensed phase of free amino acids, regardless of the protein's size and structural and functional origin. Efficient packing therefore represents a universal feature of protein structure. Additionally, efficient packing likely facilitates the restriction of protein dynamics to a limited number of modes essential for function. The calculated packing scores suggest that atomic packing is a structural constraint on protein architecture, offering novel opportunities for the interpretation of sequence alignments and genome data. The fact that packing efficiency shows a marked resolution dependence indicates that rigorous inclusion of an accurate packing term can be expected to enhance structure refinement at low and intermediate resolution levels. Furthermore, it underscores the significance of packing considerations for protein structure prediction, design and docking.

## 3.6 Summary

Generating protein structures with tCONCOORD requires accurate parameters to ensure optimal geometry. In this chapter we described how arbitrary statistical observables can be derived from experimental data using a newly developed program termed PDBBrowser. A complete set of atomic contact radii was derived from high-resolution X-ray structures and used to evaluate packing properties in protein structures. We showed that packing quality and the distribution of favourable contacts and unfavourable overlaps are exclusively resolution dependent. The shape of the introduced reduced radial distribution function (RRDF) is highly conserved in all protein structures and can therefore be regarded as a structural constraint on protein architecture.

# Chapter 4

# Constraint Definition in tCONCOORD

*These results came directly out of a computer and are not to be doubted or disbelieved.*

*- Unknown*

## 4.1 Introduction

The process of protein flexibility prediction in tCONCOORD can be subdivided in two steps. In a first step, a given 3-dimensional structure of a protein is analyzed and translated into geometrical constraints that can be compared to a construction plan of the protein. This task is carried out by the program *tdist*. In a second step, protein structures are built based on the predefined constraints, commonly several hundred times, by the program *tdisco*, which leads to an ensemble of independent structures. Such an ensemble covers the conformational space that is available within the boundaries of the geometrical constraints. For both steps, the constraint definition and the structure generation, detailed knowledge about the geometry of protein structures at the atomic level is mandatory to ensure generation of realistic structures. In the previous chapter we showed how simulation parameters were derived from experimental data. Now we describe how protein structures are analyzed in *tdist* and translated into geometrical constraints. Hydrogen bonds play a crucial role in protein structures. Opening of only one or few hydrogen bonds can lead to a dramtatic increase of the available conformational space. Therefore, we have developed and implemented a method to estimate the opening probability of hydrogen bonds based on the local environment. Also hydrophobic clusters are discussed and how observed structural motifs are translated into constraints.

## 4.2 Evaluation of Hydrogen Bond Stability

During the analysis of experimentally known conformational transitions, it was found that they routinely involve opening of one or more hydrogen bonds. tCONCOORD therefore attempts to predict unstable hydrogen bonds by estimating the solvation probability. This approach is based on the work of Fernandez et al. [85–88] who showed that keeping a hydrogen bond “dry” is a prerequisite for its stability and that protein folding is associated with a systematic desolvation of hydrogen bonds by surrounding hydrophobic groups. Thus, analyzing the neighborhood of a particular hydrogen bond should provide hints for the probability of a water molecule to attack it, which is directly correlated to the opening probability.

To this end, we have analyzed $35$ large-scale molecular dynamics trajectories from different proteins and calculated for each protein atom type $i$ (a total of $167$, hydrogen atoms were not taken into account) the radial distribution function (RDF) with water-oxygen ($\mathrm{O}_{wat}$). Integrating the weighted RDFs according to $P_i = \int_0^d R_{i-Owat}(r)dr$ (with d = $6$ Å) yields a value that may serve as solvation parameter and allows to estimate the probability of finding a water molecule within a certain distance to the particular atom. Since these values were obtained by analyzing a very limited number of trajectories, an accurate statistical error estimation is difficult. Additionally, there is a systematic error, resulting from the low number of different folds and sequences taken into acount for this work. However, previous studies on hydrophobic protection showed that even more simple approaches, such as counting hydrophobic residues around a hydrogen bond, provide valuable hints towards predicting unstable hydrogen bonds [85–88].

| PDB code | simulation time[ns] | PDB code | simulation time[ns] |
|---|---|---|---|
| 1TUX | 110 | 1RAT | 110 |
| 1PGS | 110 | 1UBI | 110 |
| 1CNV | 110 | 1UNE | 110 |
| 135L | 110 | 1VCC | 110 |
| 153L | 110 | 1WBA | 110 |
| 1A3D | 110 | 1A3H | 110 |
| 1AST | 110 | 4ICB | 110 |
| 1BJ7 | 110 | 1CLM | 110 |
| 1BM8 | 110 | 1CSP | 198 |
| 1CPN | 110 | 1EXR | 77 |
| 1DSL | 110 | 1EZM | 110 |
| 1GBG | 110 | 2CHE | 113 |
| 1HYP | 174 | 1MLA | 110 |
| 2APR | 110 | 4AKE | 110 |
| 1CHD | 110 | 1HKA | 110 |
| 1AAJ | 110 | 1KOE | 110 |
| 1ELT | 110 | 1OSA | 110 |
| 1GBS | 110 | | |

Table 4.1. Molecular dynamics trajectories that were used for the derivation of solvation parameters. All simulations were carried out using the Gromacs suite and the OPLS-AA force field with TIP4P water.

The obtained solvation parameters are used to evaluate the surroundings of a particular hydrogen bond. As nearest neighbors of a hydrogen bond we consider all atoms within two intersecting spheres with a radius of 6 Å centered at the hydrogen and the acceptor atom. Bonded and 1–3-neighbors were excluded. Using the solvation parameters from these nearest neighbors, we calculate a solvation score $\mathcal{S}$ according to

$$\mathcal{S} = \frac{1}{N} \sum_{i}^{N} P_i \quad ; \quad \text{N: Number of neighbors.}$$

This score is high if either the number of neighbors is low, which is usually the case at the surface of a protein, or if the neighborhood mostly consists of hydrophilic groups.

In order to incorporate this evaluation method into the constraint definition in tCONCOORD, we calculated the distribution of the introduced solvation score for all hydrogen bonds in $290$ protein structures (see Appendix 10.2) from the Protein Data Bank (PDB) [35] with a resolution higher than $1.6$ Å. (fig. 4.1). For the constraint defintion in tCONCOORD we use thresholds between $2.1$ and $2.2$. A threshold of $2.2$ means that hydrogen bonds with a score higher than $2.2$, and thus exceeding that of $97\%$ of the hydrogen bonds in the analyzed subset of the PDB, are considered to be unstable. Hence, they are disregarded and not translated into constraints.

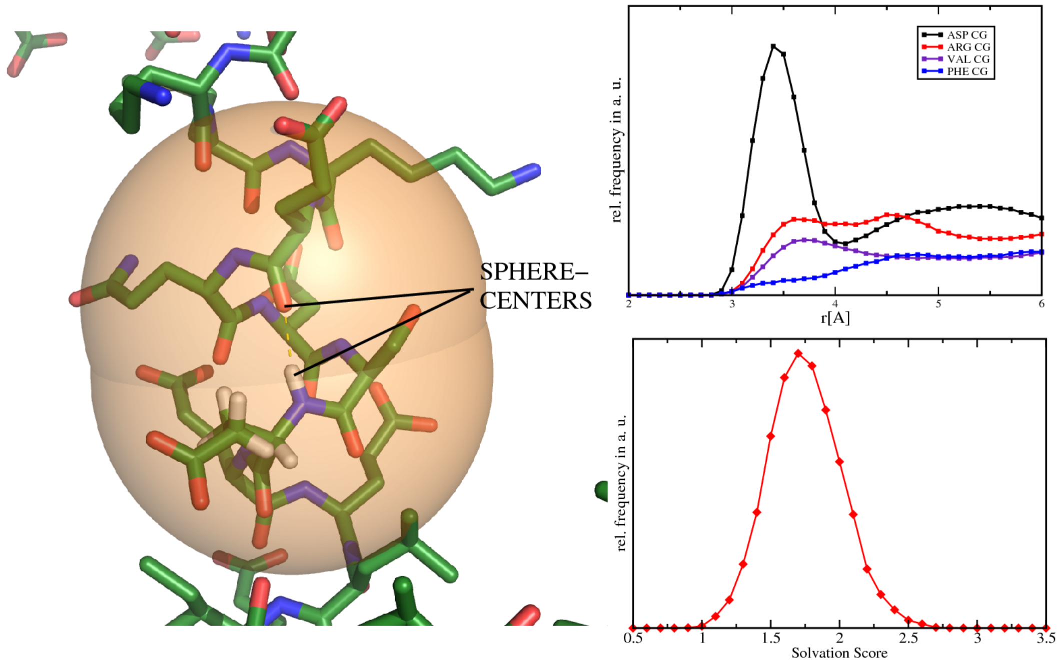


Figure 4.1. Left panel: Hydrogen bond and its neighborhood, which determines the solvation probability. Right upper panel: RDFs for $C_\gamma$-atoms in different amino acids obtained from large scale MD-simulations. Right lower panel: Distribution of solvation scores in a subset of $290$ protein structures from the Protein Data Bank.

## 4.3 Hydrophobic Clusters

The structure of globular proteins is significantly determined by entropy, namely the hydrophobic effect. Exposing hydrophobic residues to the solvent leads to a descreased entropic contribution of the surrounding water to the free energy. Therefore, these residues are usually found in the core of the protein, shielded by hydrophilic residues that interact more favorably with water. Although there is no conventional force in terms of the gradient of a potential energy term keeping hydrophobic residues together, this structural property is usually conserved during conformational transitions. Also in simulations with implicit solvent, the introduction of a hydrophobic potential of mean force (HPMF) has been shown to lead to better free energy estimations. [89]

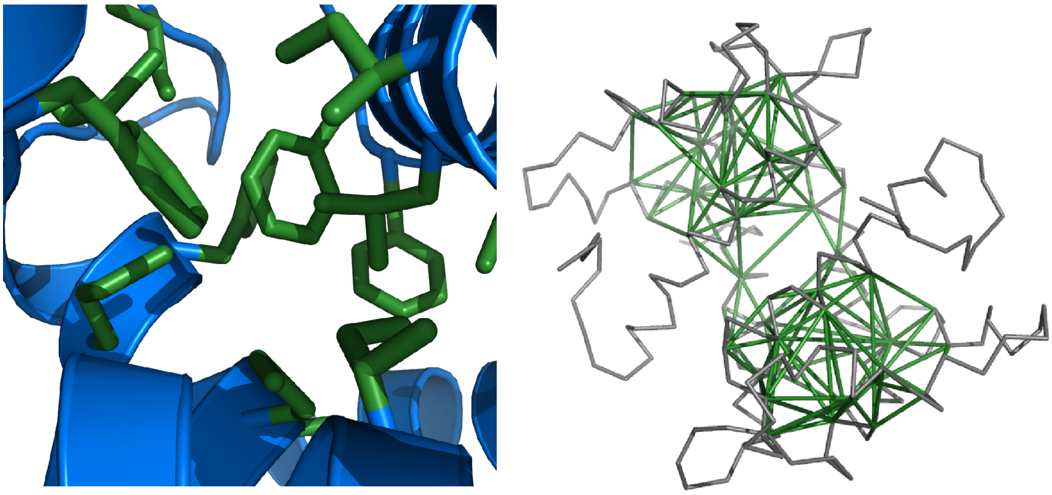


Figure 4.2. Left panel: Hydrophobic residues in a protein core. Right panel: Hydrophobic cluster definition in tCONCOORD. Green sticks represent hydrophibic "interactions"

In tCONCOORD hydrophobic clusters are defined as three-body correlations of hydrophobic residues. The side-chain carbon atoms of the residues ILE, VAL, LEU, MET, PHE, and TRP are considered as hydrophobic atoms. If three hydrophobic atoms from three different residues are found within short distance (in all simulations of this work 6 Å is used), these "interactions" are defined as constraints. The left panel in fig. 4.2 shows such a hydrophobic cluster in a protein core. The right panel shows a schematized representation of all hydrophobic constraints defined in a protein structure. The grey tubes connect $C_\alpha$-atoms according to the proteins sequence. Green tubes connect $C_\alpha$-atoms that are constraint by hydrophobic clusters. As can be seen, these clusters are exclusively found in the core of the protein.

## 4.4 Residue Networks

Since the conformational space of polypeptide chains is enormously large, it is mandatory to reduce this space as much as possible by geometrical constraints in order to faithfully predict protein flexibility computationally in a feasible manner. Therefore, in an approach like tCONCOORD that yields protein structures based on geometrical considerations, it is beneficial to define as many as possible *indirect* constraints in addition to the inclusion of direct interactions from connectivity or hydrogen bonds. Such indirect constraints have to be considered and defined as accuratly as possible. For instance, the $C_\alpha$-atoms of a residue *i* and a residue *i+4* can adopt all distances between their van-der-Waals-distance and three times the $C_\alpha$-$C_\alpha$ distance of $3.8\,\text{Å}$ , roughly $11\,\text{Å}$. However, if residue *i* and a residue *i+4* form a backbone-backbone hydrogen bond as in $\alpha$-helices the range of accessible distances for this pair is significantly reduced. Furthermore, the accessible distances for all atoms and residues connected to residue *i* and a residue *i+4* is reduced, too. Regarding such effects allows assignment of well-defined distance constraints for atoms and residues that are far away in sequence and do not have direct interactions.

tCONCOORD uses a residue network analysis in order to group residues based on the interaction framework. Thereby residue “interaction rings” consisting of four or five residues are identified. Different types of interaction, covalent, hydrogen bonds or hydrophobic interaction, are treated equally at this stage. Secondary structure motifs of proteins can be described as fused “interaction rings”. In $\alpha$-helices for instance, the residues *i* and *i+4* are connected by a backbone-backbone hydrogen bond, hence the residues *i* to *i+4* form an “interaction ring” consisting of five nodes. Consequently the residues *i+1* and *i+5* form a hydrogen bond, too.

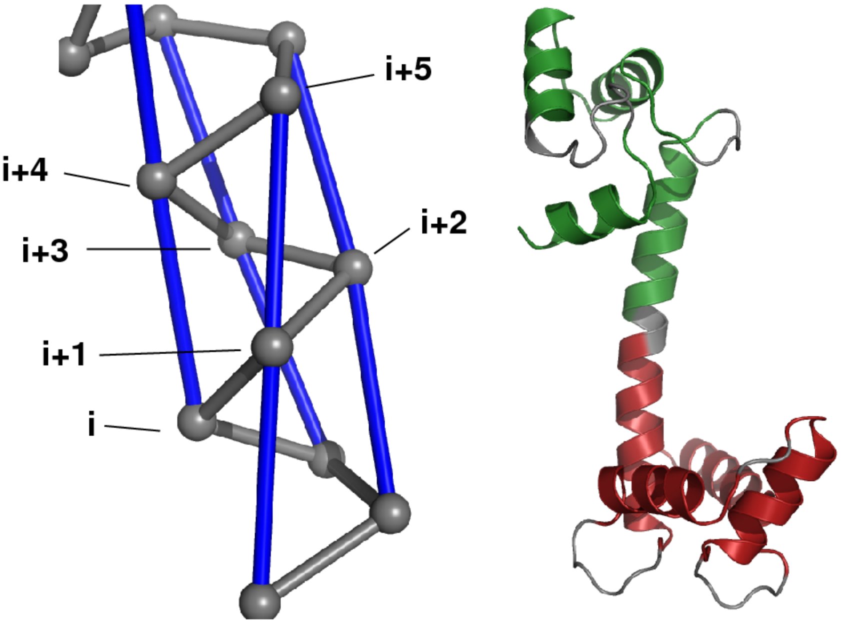


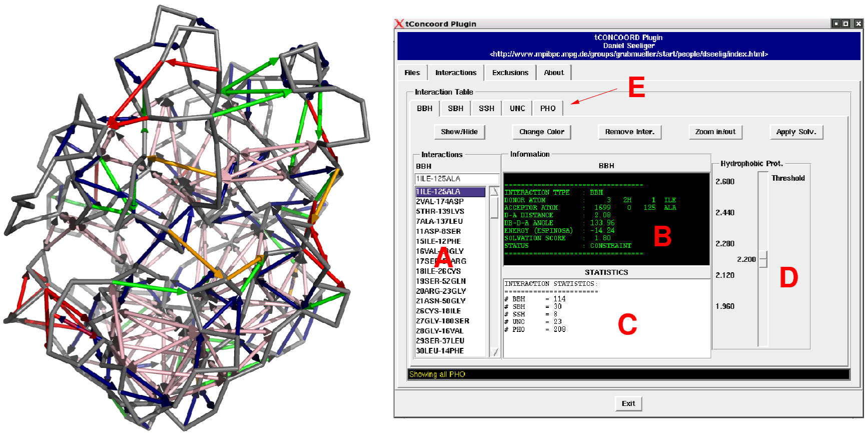
Figure 4.3. Left panel: “Interaction rings” in an $\alpha$-helix. Blue sticks represent hydrogen bonds. The helix can be described as a system of fused rings of residues, connected by covalent interactions and backbone-backbone hydrogen bonds. Right panel: “Interaction groups” as defined in tCONCOORD. The red and green colored regions represent groups with low internal flexibility. Grey colored regions are not grouped and represent flexible regions.

This leads to an “interaction ring” using the nodes $i$, $i+1$,$i+4$ and $i+5$ with two covalent edges and two backbone hydrogen bond edges.

Figure 4.3 illustrates how each residue can be regarded as a node that is part of divers “interaction rings” (left panel), thus disturbance of its coordinates would affect multiple geometric formations. As several “interaction rings” share nodes, their motions are coupled and can therefore be fused into one “interaction group”. Such groups have limited internal flexibility which allows assignment of well-defined constraints within the group. The grouping-algorithm reduces the number of “interaction groups” by merging those groups with a certain number of common members (fig. 4.3, right panel). Depending on the size and structure of the protein, this number ranges from one to a few tens. Some residues however, are not put into groups since they do not interact with other residues and also their neighbors have few interactions. These residues usually represent the most flexible part of the structure, mainly loops located on the protein’s surface.

## 4.5 Manual Constraint Definition

Generating structure ensembles with tCONCOORD is a two step process. In a first step, a given input structure is analyzed and turned into geometrical constraints. This is done by the program *tdist*. In a second step, structures are rebuild based on the predefined constraints using the program *tdisco*. The properties of the resulting ensemble is thus mainly depending on the first part, the constraint definition. tCONCOORD allows to influence the constraint definition manually. Intercations can be switched off or defined interactively enabling the user to generate ensembles covering only parts of the conformational space or to study the influence of mutations to conformational flexibility. For instance, if potential induced fit structures should be generated from an open conformation, the program can be forced to produce only closed conformations by imposing appropriate constraints on certain residues. The program *tdist* therefore writes information about all interactions found in the input structure. These files can be visualized with PyMOL (*www.pymol.org*) using the tCONCOORD plugin and specific changes can be applied with visual support. Figure 4.4 shows the graphical user interface (GUI) of the tCONCOORD plugin. On the left, a protein structure is shown us-



```
==================================
INTERACTION TYPE    : BBH
DONOR ATOM          :     3   2H     1  ILE
ACCEPTOR ATOM       :  1699    O   125  ALA
D-A DISTANCE        :  2.08
DB-D-A ANGLE        : 133.96
ENERGY (ESPINOSA)   : -14.24
SOLVATION SCORE     :  1.80
STATUS              : CONSTRAINT
==================================
```

```
INTERACTION STATISTICS:
======================
# BBH     = 114
# SBH     = 30
# SSH     = 8
# UNC     = 23
# PHO     = 208
```

```
Showing all PHO
```

Figure 4.4. **tCONCOORD Plugin.** The left panel shows a $C_\alpha$-representation of a protein and interactions defined by *tdist*. The right panel shows the tCONCOORD plugin for PyMOL. Every interaction is listed in the listbox A. Diplay B gives detailed information about the interaction and C lists interaction statistics. The slider labeled with D can be used to define solvation score thresholds. To switch between different interaction types, the table E is used.

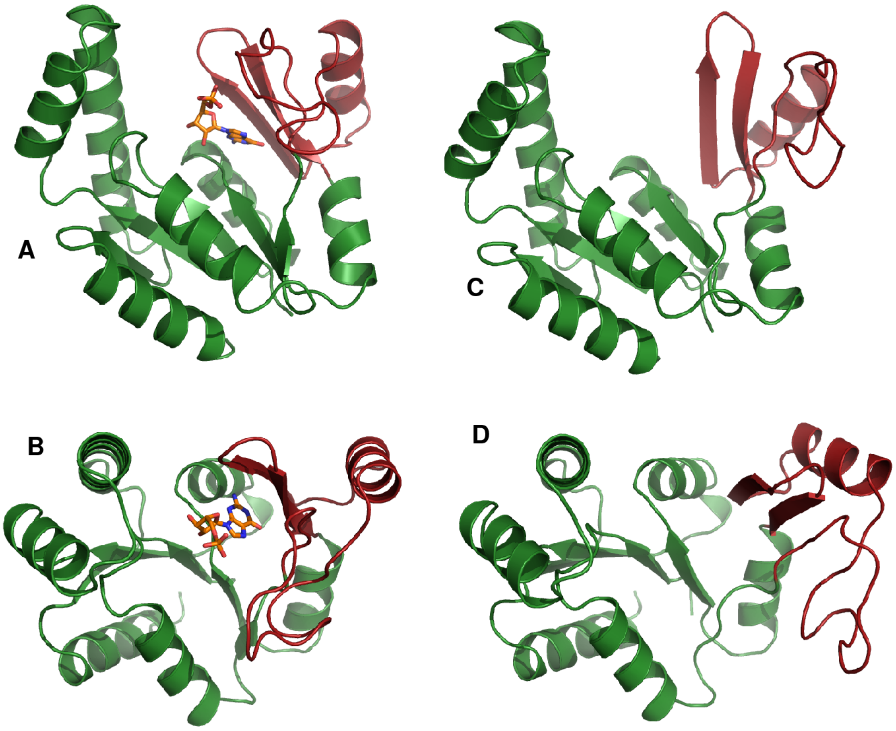


Figure 4.5. **Guanylate Kinase.** A and B show the structure of guanylate kinase bound to the ligand guanosine-5'-monophosphate (PDB 1EX7). C and D show the structure of the apo conformation (PDB 1EX6). Upon binding the ligand, the red colored domain closes over the ligand. The RMSD of this domain between bound and unbound state is $\approx 8$ Å

ing grey tubes for $C_\alpha$-atoms. Interactions are represented with colored arrows. Details about every interaction are listed in the GUI that furthermore allows an interactive definition of solvation parameters and the definition of exclusions. An application is shown in fig. 4.5. Guanylate kinase undergoes a large conformational change upon binding a ligand. The binding process is associated with the closure of a lid (colored red in fig. 4.5) over the ligand. Although a tCONCOORD simulation started from the apo structure (PDB 1EX6, fig. 4.5C+D) samples both, closed and open conformations, most of the generated structures are uninteresting if the focus lies on ligand bound states. Imposing additional constraints that only allow generation of closed conformations, leads to a much better sampling of the conformational space that is *relevant* for binding ligands. The left panel of fig. 4.6 shows additional imposed constraints between two protein domains.

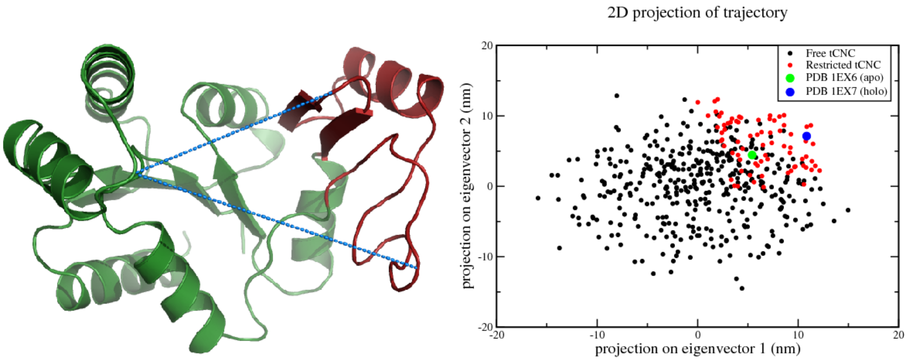


Figure 4.6. **Conformational Sampling.** Left panel: Additional constraints imposed on the open conformation of guanylate kinase (PDB 1EX6). Right panel: Principal component analysis. Green dot: unbound conformation, blue dot: ligand-bound conformation, black dots: free tCONCOORD sampling, red: tCONCOORD sampling with additional constraints.

The distance between residue $42$ and $137$, and $74$ and $137$ are forced to become shorter than observed in the starting structure. The different sampling properties are illustrated in the right panel of fig. 4.6. The ligand-free conformation (PDB 1EX6) is shown as green dot, the ligand-bound state (PDB 1EX7) as blue dot. The free tCONCOORD simulation (black) samples a large conformational space, thereby producing many conformations that are not relevant for ligand binding. The restricted tCONCOORD simulation (red) samples a much smaller area of the conformational space, namely closed conformations which are supposed to be relevent for ligand binding. Hence, such considerations may aid a specific application, like in this case drug design.

## 4.6 Summary

The $3$-dimensional structure of proteins is determined by many interactions. In order to predict protein flexibility and conformational transitions, it is mandatory to distiguish between conserved und non-conserved interactions. In this chapter we showed how interactions in proteins are analyzed and translated into geometrical constraints. The opening probability of hydrogen bonds is estimated by a thorough analysis of the environment and the estimation of the solvation probability. The hydrophobic effect is taken into account by defining hydrophobic clusters. Since the performance of the CONCOORD algorithm increases if long-range constraints can be defined, a network analysis is used to determine protein parts with reduced flexibility. Finally we introduced a PyMOL plugin that, firstly, allows for visual control of the constraint definition process, and secondly, allows to influence constraint definition which might be appropriate to address specific questions. In the next chapter we show applications of tCONCOORD to selected proteins that examplify the scope of geometry-based molecular simulation and its usefulness in protein research.

# Chapter 5

# Geometry-based Sampling of Conformational Transitions in Proteins

> *Mensch: ein vernunftbegabtes Wesen, das immer dann die Ruhe verliert, wenn von ihm verlangt wird, daß es nach Vernunftgesetzen handeln soll.*
>
> *- Oscar Wilde*

## 5.1 Introduction

The fast and accurate prediction of protein flexibility is one of the major challenges in protein science. In this chapter we show applications of tCONCOORD to study the conformational flexibility of proteins with biological relevance. To allow comparison with experimental data, systems have been chosen of which experimental data provides insights into flexibility and functionally relevant protein motions. As first example we chose adenylate kinase as a representative of protein kinases which play important roles in signal transduction and enzyme activation by transferring phosphate groups. Many kinases are involved in cancer and therefore interesting drug targets, however, inherent protein flexibility hampers computational drug design with existing methods.

Calmodulin, the second example, is a ubiquitous, calcium-binding protein that can bind to and regulate a multitude of different protein targets, thereby affecting many different cellular functions. It mediates processes such as inflammation, metabolism, apoptosis, muscle contraction, intracellular movement, short-term and long-term memory, nerve growth and the immune response. Upon binding to proteins or inhibitors calmodulin undergoes large conformational changes associated with partial unfolding.

The third example, aldose reductase, is an enzyme in carbohydrate metabolism that converts aldose to a sugar alcohol, using NADPH as the reducing agent. Its role in diabetes is intensively discussed and several inhibitors have been discovered and co-crystallized, revealing a flexible binding site consisting of several loops.

T4-Lysozyme has been chosen since it has been crystallized in many different conformations, allowing direct interpretation of protein flexibility from experimental data. Moreover, the protein has been extensively studied with MD-simulations, shedding light on the dynamics of conformational transitions.

Staphylococcal nuclease and ubiquitin finally are proteins with completely different flexibility properties. Experimental data of ubiquitin did not reveal any extensive collective conformational flexibility, whereas staphylococcal nuclease has flexible loops. We show that tCONCOORD correctly predicts the flexibility of these loops just as well as it predicts the limited flexibility of ubiquitin.

## 5.2 Adenylate Kinase

Adenylate kinase displays a distinct induced fit motion upon binding to its substrate (ATP/AMP) or an inhibitor (see fig. 5.1B). Structures in different conformations have been resolved [90–93] contributing significantly to the understanding of the catalytic mechanism of this class of enzymes. We carried out two tCONCOORD simulations using the closed conformation of adenylate kinase (PDB 1AKE, see fig. 5.1A) as input. In one simulation the ligand ($AP_5A$) was removed. Additionally, one simulation starting from an open X-ray conformation (4AKE) was carried out. Fig. 5.1C and D show the result of a principal components analysis (PCA) applied to the experimental structures. The first eigenvector (*x*-axis) cor-

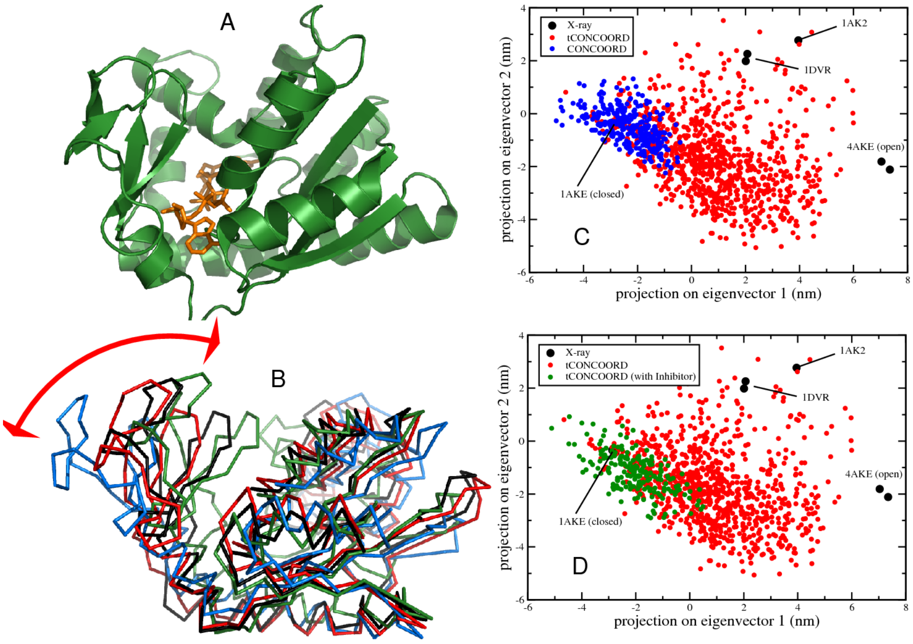


Figure 5.1. **Adenylate Kinase.** A: Crystal structure (PDB 1AKE) of adenylate kinase (green) with bound inhibitor AP$_5$A (orange). B: Superimposition of several X-ray structures in different conformations, indicating the induced fit motion. C and D: Principal components analysis. Experimental structures (black circles) and three simulation ensembles (blue, red and green circles) are projected onto the first two eigenvectors. The blue ensemble was generated with CONCOORD, the red one with tCONCOORD. tCONCOORD correctly predicts the induced fit motion and samples open conformations when started from the closed conformation with removed ligand. If the ligand remains in the input structure, the conformational space is restricted to conformations around the closed state (green).

responds to the induced fit motion indicated by the red arrow in fig. 5.1B. Every dot in the plots represents a single structure. The RMSD from the closed conformation (PDB 1AKE) to the open conformations is 4.0, 5.4, and 7.4 Å for 1DVR, 1AK2, and 4AKE, respectively. The red dots represent the ensemble that has been generated with tCONCOORD using the closed conformation of adenylate kinase without ligand as input. The blue dots in fig. 5.1C represent an ensemble that has been generated using CONCOORD (version 1.2), using the same input. As can be seen, the CONCOORD ensemble (blue) basically samples the conformational space around the input structure, not sampling open conformations. The tCONCOORD ensemble (red) behaves differently. It almost completetly samples the conformational space that is covered by the experimental structures, thereby also visiting open conformations (high $x$-values). The experimental structures were reached with a deviation of $2.4$, $2.6$, and $3.1$ Å $C_\alpha$-RMSD for 1DVR, 1AK2, and 4AKE, respectively. For comparison, for the CONCOORD cluster these RMSD values are much higher with $3.4$, $4.4$, and $5.9$ Å . In structure-based drug design often the reverse problem, predicting induced-fit structures from an open conformation, needs to be addressed. A tCONCOORD starting from an open conformation (PDB 4AKE) as input produces closed conformations with comparable accuracy as the open conformations are sampled when starting from a closed structure. The experimentally determined structures are reached with RMSD's of $2.5$, $2.9$, and $3.3$ Å for 1DVR, 1AK2, and 1AKE, respectively.

The conformational flexibility changes significantly if the ligand remains in the input structure. Fig. 5.1D shows a comparison of an ensemble with the ligand present in the input structure (green dots) with the previously discussed ensemble, generated without ligand (red dots). As can be seen, the presence of the ligand leads to a reduction of the conformational space that is sampled by the protein and open conformations are not sampled anymore.

## 5.3 Calmodulin

The structure and dynamics of calmodulin has been studied extensively by X-ray crystallography and NMR. In its activated ($Ca^{2+}$-bound) conformation [94], calmodulin exposes hydrophobic residues to the solvent enabling binding to a target, either a protein or an inhibitior. The binding process itself requires a large conformational change involving the unfolding of the central helix in order to allow rotation of the C-terminal domain to form the binding site [95](fig. 5.2A and B).

A tCONCOORD simulation of this particularly challenging case has been carried out. The instability of a number of hydrogen bonds in the central helix of the activated form (PDB 1CLL) was correctly identified (see fig. 5.2C) and incorporated accordingly into the constraint definition.

The resulting ensemble (fig. 5.2E, left) can be described as two freely rotating domains connected by a linker. These results are in good agreement with

NMR-studies of calmodulin [96] (fig. 5.2E, right). In fig. 5.2F the projections of the tCONCOORD ensemble (green cloud), the NMR-ensemble (red dots), the X-ray structures of the activated form (orange dot), and the ligand bound conformation (blue dot) onto the first three eigenvectors of a PCA are shown. The tCONCOORD-ensemble represents an extended sampling of the conformational space, comprising all experimentally determined structures. The RMSD between the activated conformation of calmodulin and the bound conformation is 14.6 Å. The closest match of a structure from the ensemble, generated with tCONCOORD, to the experimentally known ligand bound conformation is as low as 2.8 Å (fig. 5.2D). This is an example of a case where a ligand bound conformation of the protein is predicted using only the structurally completely different unbound state as input. The possibility of such predictions is of obvious relevance for applications in the field of structure based drug design.

## 5.4 Aldose Reductase

Aldose reductase (AR) is believed to play an important role in diabetes and therefore is a potential drug target [97, 98]. It adopts a TIM-barrel fold and uses NADPH as cofactor to reduce various aldehydes. AR has been crystallized with different inhibitors. A remarkable fact concerning these inhibitors is that they have very different structures, sizes and molecular weights [98]. AR is able to bind these structurally different inhibitors because of a very flexible binding site. Figure 5.3 shows the structure of AR (PDB 2FZD) with bound cofactor (red) and the inhibitor Tolrestat (orange). The regions that are responsible for the formation of a hydrophic sub-pocket are labeled with A and C. The B-loop is responsible for binding the cofactor. In order to study the influence of both, the ligand and the cofactor, on the conformational flexibility of aldose reductase, tCONCOORD simulations were carried out for the entire complex (AR+NADP+Tolrestat), the complex with removed inhibitor (AR+NADP) and free AR. To compare the flexibility of the different systems, a principal component analysis was applied to the combined ensembles of all three runs. Subsequently the ensembles for each system were projected onto the eigenvectors with the largest eigenvalues. The

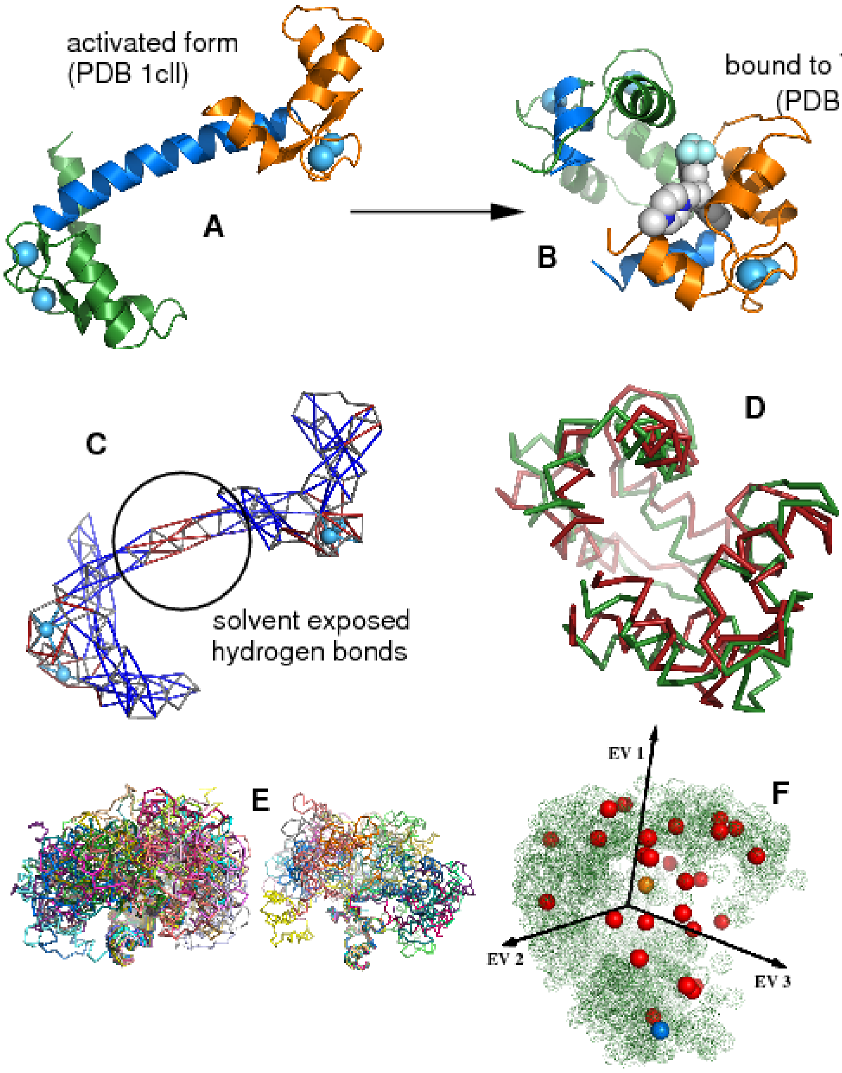


Figure 5.2. **Calmodulin.** A shows the activated form of calmodulin (PDB 1CLL) used as input for tCONCOORD. B shows the structure of calmodulin bound to Trifluoroperazine (TFP). The RMSD between these two structures is $14.6$ Å. C shows the result of the hydrogen bond analysis of tCONCOORD. Red sticks represent hydrogen bonds with high solvation probabilities and are not regarded as constraints in the tCONCOORD simulation. D shows the superimposition of the ligand bound conformation (green) and the closest match of a structure from the tCONCOORD ensemble (red) with an RMSD of $2.8$ Å. E shows a tCONCOORD ensemble and a NMR- ensemble (PDB 1CFF) fitted on the C-terminal domain. F shows the projection onto the $3$ eigenvectors with the largest eigenvalues of a PCA. The tCONCOORD-ensemble is shown as green cloud, the NMR-ensemble as red dots. The orange dot represents the X-ray structure of the open (activated) conformation, the blue dot represents the closed (ligand bound) state.

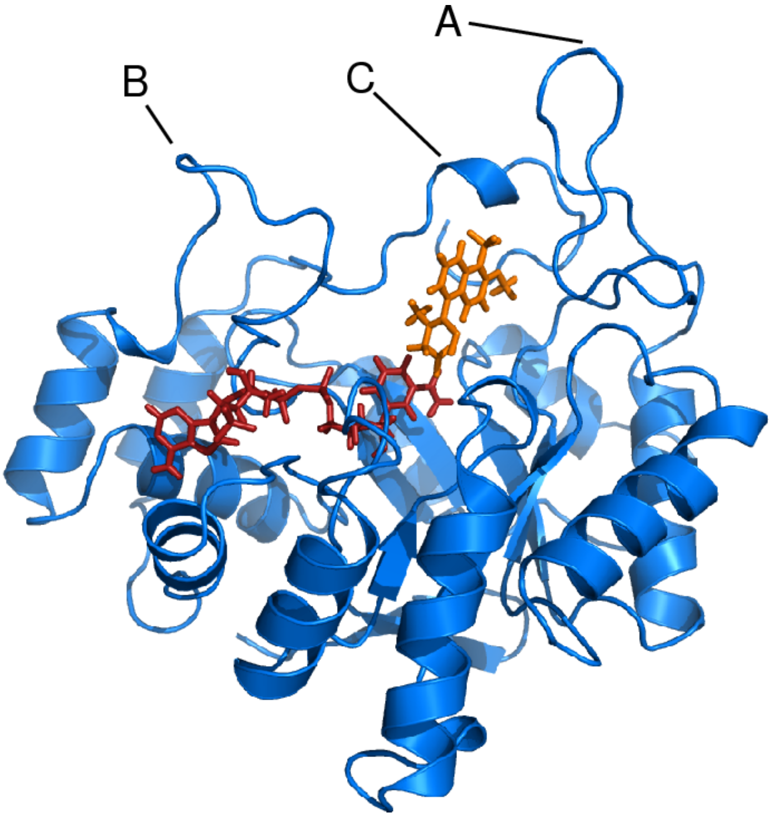


Figure 5.3. **Aldose Reductase.** Structure of human aldose reductase (PDB 2FZD) with bound cofactor (NADP, red) and inhibitor (Tolrestat, orange). The loops labeled A and C form parts of the Tolrestat binding site. Loop B interacts with the cofactor.

eigenvectors $1$ and $2$ mainly correspond to movements of the A-loop in AR, as indicated in fig. 5.4, right panel. The projection of the ensembles onto these eigenvectors (fig. 5.4, left) reveals the same flexibility along these eigenvectors for the free AR ($\sigma_1^{free} = 5.15\,\mathrm{nm}$, $\sigma_2^{free} = 4.34\,\mathrm{nm}$) and the AR with bound cofactor ($\sigma_1^{holo} = 5.07\,\mathrm{nm}$, $\sigma_2^{holo} = 4.25\,\mathrm{nm}$). In the third system, where also Tolrestat is bound, the flexibility is reduced significantly due to interaction of the ligand with the A and C-loop ($\sigma_1^{tol} = 3.13\,\mathrm{nm}$, $\sigma_2^{tol} = 3.28\,\mathrm{nm}$). Figure 5.5 compares the motions along the eigenvectors $3$ and $4$. The motions corresponding to eigenvector $3$ predominantly represent a movement of loop B, which is involved in binding the cofactor. Here we observe high flexibility for the free AR ($\sigma_3^{free} = 3.41\,\mathrm{nm}$), whereas the fluctuation for the holo-form ($\sigma_3^{holo} = 2.76\,\mathrm{nm}$) and the entire complex ($\sigma_3^{tol} = 2.96\,\mathrm{nm}$) along this mode is comparable.

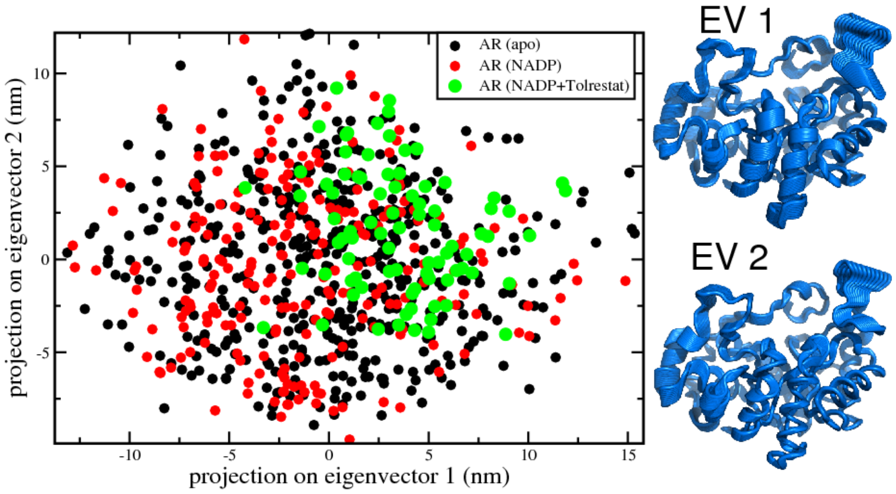


Figure 5.4. Projection of tCONCOORD ensembles of aldose reductase onto eigenvector 1 and 2 of a principal compnents analysis. The structures on the right represent the predominant motions along these vectors. On the left, the 2-dimensional projection of 3 different ensembles is shown. The green dots represent the ensemble of the entire complex, the red dots represent the holo form, and the black dots the apo form. The projection shows the reduced flexibility of the binding site in the presence of Tolrestat. Binding of NADP, however has no effect on these modes

Eigenvector $4$ again reveals a clear difference between the holo form and the complete complex systems. As the main component of this mode is a movement of the C-loop, flexibility of this region is dramatically reduced by Tolrestat ($\sigma_4^{tol} = 1.30\,\text{nm}$), whilst free AR and holo AR display comparable and somewhat higher flexibility along this eigenvector ($\sigma_4^{free} = 2.01\,\text{nm}$, $\sigma_4^{holo} = 2.10\,\text{nm}$).

## 5.5 T4-Lysozyme

Bacteriophage T4-Lysozyme (T4L) is one of the rare cases where conformational flexibility can be directly estimated from X-ray structures [99]. It has been crystallized in many different conformational states shedding light on the dynamical behavior. The main collective motion is a hinge-bending mode that is presumably necessary to allow entrance and release of the substrate. This mode is described by the first eigenvector of a principal components analysis, carried out on the experimental data. In order to predict open conformations when using the closed

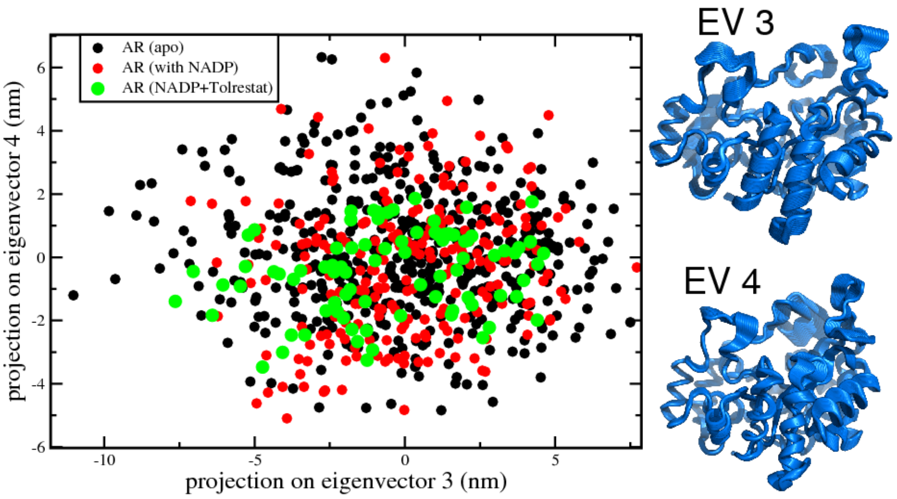


Figure 5.5. Projection of tCONCOORD ensembles of aldose reductase onto eigenvector 3 and 4 of a principal compnents analysis. The structure figures on the right represent the predominant motions along these vectors. On the left, the 2-dimensional projection of 3 different ensembles is shown. The green dots represent the ensemble of the entire complex, the red dots represent the holo form and the black dots the apo form. The projection shows increased flexibility along eigenvector 3 if NADP is removed, because loop B is predominantly involved in this motion. Eigenvector 4 mainly represents a movement of loop C which leads to decreased flexibility for the ensemble with Tolrestat bound.

conformation as input for tCONCOORD, a correct detection of unstable hydrogen bonds is mandatory. As can be seen in fig. 5.6, a hydrogen bond that is formed between Glu22 and Arg137 in the closed conformation (PDB 2LZM, left structure) is not present in the open conformation (PDB 149L, right structure) and the distance from the C$\delta$ of Glu22 to C$\zeta$ of Arg137 changes from 3.8Å to more than 18Å. The hydrogen bond analysis method of tCONCOORD correctly predicts the instability of this hydrogen bond as indicated in the picture in the central upper panel of fig. 5.6. The blue sticks represent stable hydrogen bonds, whereas red sticks mark those that display high probabilities of water attack. The latter are not defined as constraints. Figure 5.6 also shows the projection of the experimental structures, a tCONCOORD ensemble and 3 MD trajectories, which have been started from different conformational states, onto the first two eigenvectors obtained from a PCA carried out using the X-ray structures. It can be seen that the tCONCOORD ensemble, started from a closed state (PDB 2LZM), also samples open conformations. A closer look at the MD trajectories reveals that the longest

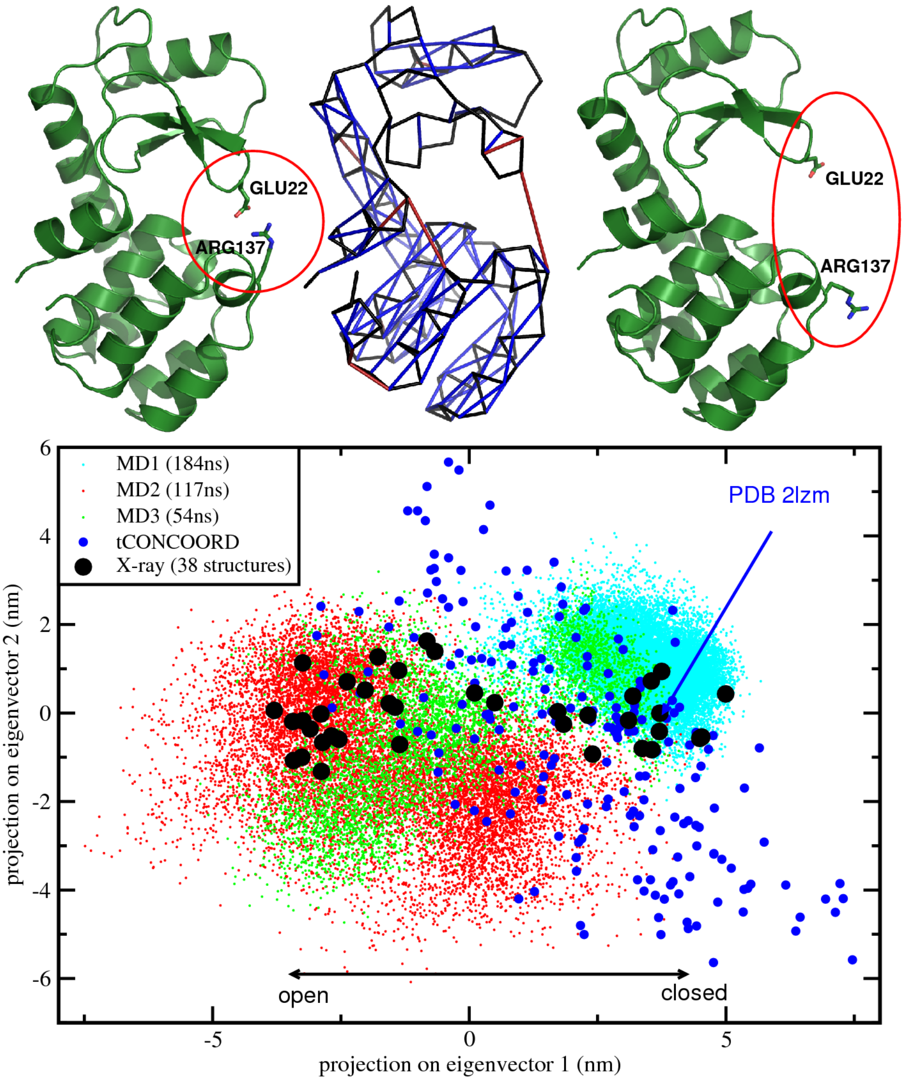


Figure 5.6. **T4-Lysozyme.** The upper left panel shows the structure of the closed conformation of T4L (PDB 2LZM). This structure has been used as input for tCONCOORD. The picture in the middle shows the hydrogen bond stability analysis carried out by tCONCOORD. Red marked hydrogen bonds, like the bond between GLU22 and ARG137, are predicted to be unstable. The right picture shows the structure of an open conformation of T4L (PDB 149L). Indeed, in this conformation this hydrogen bond is not present anymore. The lower panel shows the result of a principal component analysis applied to the experimental structures. The experimental structures (black), the tCONCOORD ensemble (blue) and three MD trajectores (cyan, red and green) are projected on the first two eigenvectors.

trajectory (cyan, 184ns) does not sample open conformations at all, whereas the shorter simulations (red and green) cover more of the conformational space. The phase space density produced by the MD simulations indicates an energy barrier between the closed and the open conformation, which is not overcome in the simulation represented by the cyan circles. The tCONCOORD sampling however, is not affected by energy barriers and samples most of the space covered by the MD trajectories. Although the tCONCOORD ensemble samples both open and closed conformations it does not completely sample the conformational space sampled by the MD simulations that started from open conformations. This is due to the fact that tCONCOORD defines constraints from a single input structure, in this particular case a closed conformation. If unstable interactions are not entirely detected in the constraint definition process this can lead to an exclusion of regions of the conformational space. The tCONCOORD ensemble furthermore samples regions of the conformational space that are not visited by the MD simulations and the experimental structures. This could be either due to an energy barrier that is too high to be overcome by MD simulations within the accessible time-scale, or the energy of this region of the conformational space is too high to be part of the *relevant* conformational space.

## 5.6 Rigid and flexible regions in proteins

Functional studies on protein structures benefit significantly from information about the flexibility and rigidity of protein parts. The calculation of root mean square fluctuations (RMSF) from tCONCOORD ensembles can provide valuable hints regarding these properties. To test the reliability of flexibility predictions we chose two applications with completely different structure and different flexibility properties, which have been experimentally determined. As a first testcase we chose ubiquitin, a small 70 residue protein of which 46 X-ray structures are available in the PDB (see Appendix 10.3). The RMSF determined from the X-ray structures (fig. 5.7 red curve) shows that the protein is relatively rigid, with the only noteworthy flexibility at the C-terminus and a loop around residue 8. The RMSF profile calculated from the tCONCOORD ensemble generated using PDB 1UBI [100] as input (fig. 5.7 black curve) represents the same flexibility properties as the experimental data. Although the flexibility level of the tCONCOORD ensemble is constantly above the X-ray ensemble (which may be due to the fact that most X-ray structures were solved at low temperatures), the overall picture of a rigid protein with flexible C-terminus is reproduced (correlation coefficient $0.95$). For comparison the RMSF of an ensemble generated with an elastic network model [101, 102] is shown (fig. 5.7 green curve). This fast and efficient method is routinely employed to predict protein flexibility and reproduces the experimental fluctuations only slightly worse than tCONCOORD (correlation coefficient $0.9$). However, the structures from the tCONCOORD ensemble all have reasonable geometry (bond lengths, angles, dihedrals, interatomic distances), which is not always the case for single structures derived from elastic network models.

As second application we chose staphylococcal nuclease of which an NMR-ensemble [103] (PDB 1JOR) is available and provides information on the flexibility of the protein. The RMSF calculated from the NMR ensemble (fig. 5.8, red curve) shows that mainly one loop around residue $42$ is very flexible. Furthermore, the loops around residues $80$ and $110$ show increased flexibility. The RMSF

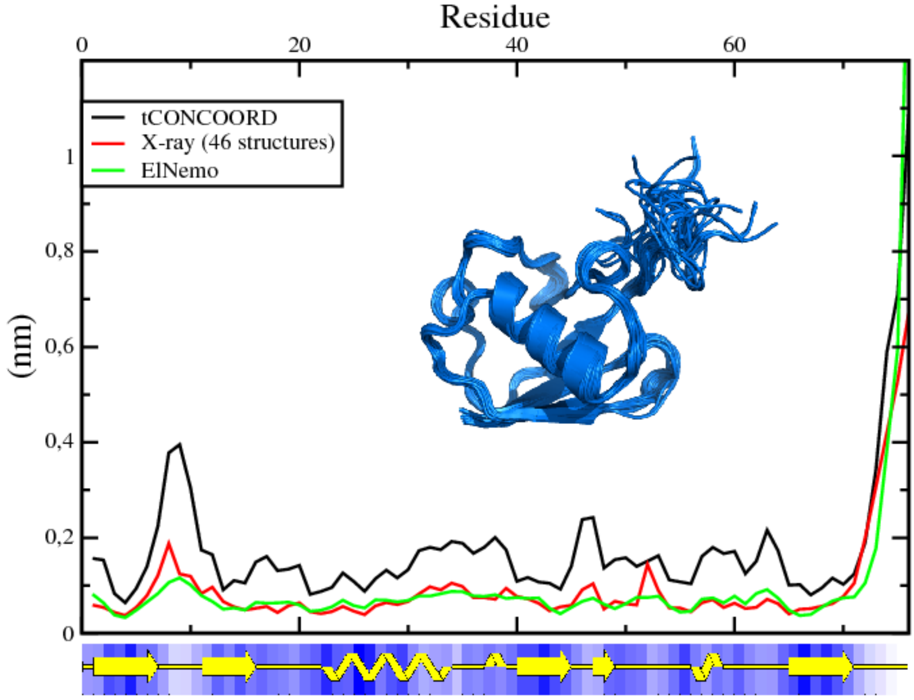


Figure 5.7. **Ubiquitin.** Root mean square fluctuation in ubiquitin.

profile calculated from a tCONCOORD ensemble (fig. 5.8, black curve), using an X-ray structure (PDB 1EY4) [104] as input, yields qualitatively the same picture. The most flexible regions detected by the tCONCOORD ensemble are again in good agreement with the experimental data (correlation coefficient $0.8$) and again slightly better than the elastic network model (green curve, correlation coefficient $0.78$). The tCONCOORD ensemble predicts higher flexibility for some parts of the protein than observed in the NMR-ensemble. This might be either due to interactions tCONCOORD underestimates, or due to an overtight representation of the NMR data which is sometimes caused by imposing time- and ensemble-averaged experimental properties onto single structures during refinement [105–107].

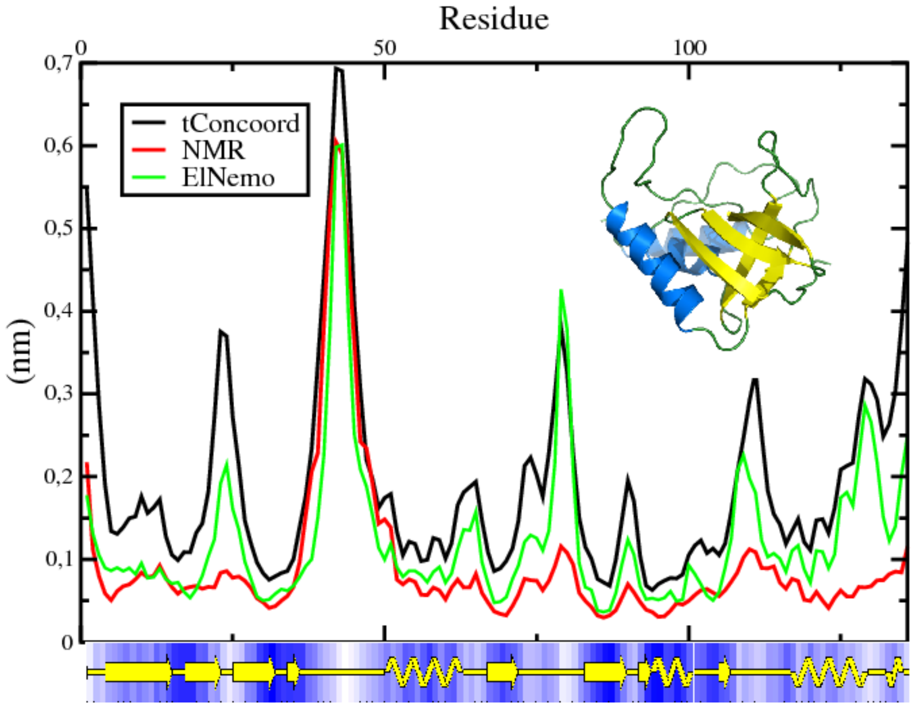


Figure 5.8. **Staphylococcal Nuclease.** Root mean square fluctuation in staphylococcal nuclease.

## 5.7 Summary

In this chapter we reported a novel approach to accurately predict large conformational transitions in proteins and its application to selected systems with biological relevance. Information about conformational transitions is often a prerequisite to understand protein function. With tCONCOORD we provide an efficient simulation approach to predict protein conformational transitions. The resulting ensemble can be either used to study the essential degrees of freedom of a protein, to identify flexible and rigid parts in a structure or to obtain different starting points for other simulation protocols. Furthermore, incorporation of protein flexibility by tCONCOORD ensembles, e.g in docking protocols, is expected to enhance the efforts of structure based drug design.

# Chapter 6

# Molecular Modeling of Protein Parts

*Homology modeling. This is the dark side of folding.*

*- Unknown*

## 6.1 Introduction

In the previous chapter we have shown how structure ensembles generated by tCONCOORD can be useful to gain insights into protein function, thereby focusing on conformational flexibility of the complete protein. Many questions related to protein function only concern restricted areas of protein structure, binding sites for instance represent such a case. In tCONCOORD such a restricted sampling can be carried out either by imposing additional constraints or keeping parts of the protein fixed. In any case, a template structure is required for the constraint definition.

In this chapter we show how geometry-based structure modeling of protein fragments can be carried out without having a complete template. The first part focuses on loop modeling, which denotes the prediction of a protein fragment on the surface. In the second part we show that structure modeling also can be applied to reconstruct areas in the protein core which is often required in the field of protein design.

## 6.2 Loop Modeling

Loops often determine the functional specificity of a given protein framework and contribute to active and binding sites and loop modeling is therefore an intensively studied field of research and has been recently reviewed by Rossi et. al. [108]. Experimental data however, e.g. X-ray crystallography does not always provide structure models with all parts of the protein resolved at atomic resolution. Flexible parts of proteins, particularly loops, are often not resolved and hence, need to be modeled computationally. Basically two different approaches are employed.
i) Physics-based methods. The modeling process is regarded as a mini folding problem and loop conformation are produced by employing distance restraints to force the loop to the anchor positions. Subsequent minimization, heating and again minimization yields loop conformations with low energies.
ii) Knowledge-based methods. The missing loop is built using homology modeling, thereby searching databases for fragments with the same sequence as the missing loop.
tCONCOORD's loop modeling approach can be regarded as a combination of both. Geometrical constraints are derived from experimental data similar to knowledge-based methods. However, a database of protein fragments with known geometry is not required. Molecular geometry is a result of all forces acting on atoms, thus incorporation of geometrical constraints yield energetically meaningful configurations. tCONCOORD's loop modeling approach is applicable to loops of arbitrary size and sequence and can be used to generate ensembles of loop configurations which can be used in subsequent refinement or simulation protocols. tCONCOORD can build loops using geometrcial constraints, thereby keeping the resolved part of the protein fixed. A script ('do_loop.py') has been developed to prepare the structure and the loop to be inserted. The anchor positions, the last resolved protein residues, are attached to the loop and their coordinates stored in a second input file. These coordinates serve as target positions in the actual loop modeling process. Figure 6.1 shows an application to a 12 residue loop. The anchor residues in the right are drawn as ball and stick. The coordinates of these residues are used as target coordinates.

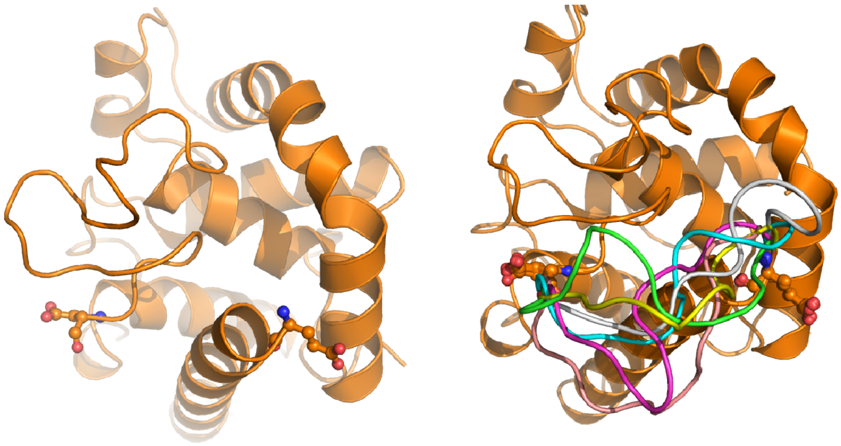


Figure 6.1. **Loop Modeling.** Left: Structure of PDB 153L with removed residues 98-109. The anchor residues 97 and 110 are shown as ball and stick representation. Right: Loop conformations generated with tCONCOORD.

The right picture shows loops generated with tCONCOORD. The advantage of this approach becomes evident if very long loops have to be modeled. Other methods often use one anchor as starting position and try to find the other anchor by exploring the torsional degrees of freedom of the backbone, which denotes an exponential increase of computer time. tCONCOORD has been tested with fragments up to $40$ residues, thereby producing geometrical acceptable structures. Loops generated by tCONCOORD can be used for subsequent optimizations, e. g. for multiple copy simulations, simulated annealing or other simulation protocols to find the best loop conformations. Moreover, knowlegde-based methods can be combined with tCONCOORD, e. g. if a part of the loop should adopt a certain type of secondary structure, this can be easily incorporated into the constraint defintion. Figure 6.2 shows an application to a complex case. The Bacteriophage $\Phi 29$ connector consists of $12$ identical chains, forming a large pore through which DNA is released. The loops which interact with the DNA are not resolved in the X-ray structure, however, they are required for setting up simulations to study the mechanism of this system. The missing fragments are $16$ residues long and have to adopt conformations that, i) do not form knots whith each other, and ii) leave enough space for the DNA strand to be placed in the center of the pore. Loops have been modeled using the monomeric chain and

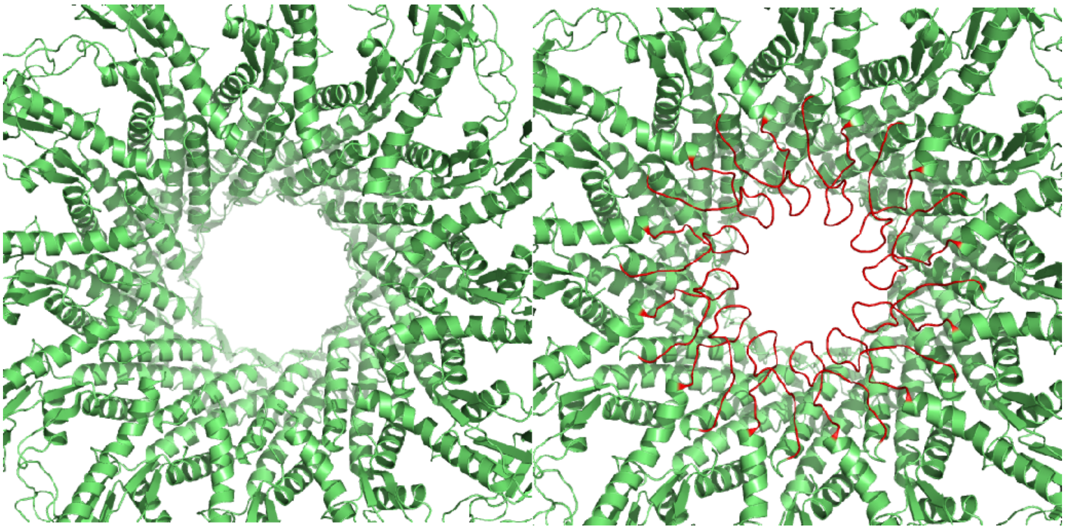


Figure 6.2. **Complex loop modeling.** Left: X-ray resolved structure of the Bacteriophage $\Phi29$ connector (PDB 1H5W). Right: The red loops were modeled with tCONCOORD.

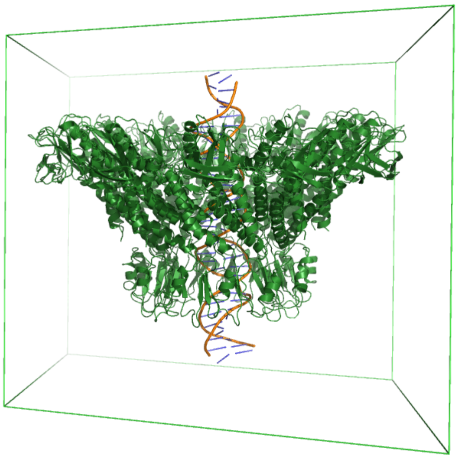


Figure 6.3. **Simulation system.** Complete system of the Bacteriophage $\Phi29$ connector with modeled loops and DNA strand.

a DNA strand as input, thus forcing the loop to leave space for the DNA. From the loop ensemble, configurations have been selected for subsequent simulations. The complete system is shown in fig. 6.3.

## 6.3 Protein Core Repacking

Manipulations of protein cores are of interest for industrial biotechnology since they may stabilize or destabilize the fold, and thus affect the function of proteins. Mutations in the densely packed protein core can cause reorientation of sidechains, usually referred to as repacking. tCONCOORD can be used to generate such repacked structures for further use in other simulation protocols. For tCONCOORD, generating structures with a mutated amino acid works exactly as the loop modeling procedure, since it is basically the insertion of a short loop, not at the surface but in the protein core. Merely the atoms around the inserted amino acid should not be kept fixed to allow for reorientation.

A more complicated example is the insertion of disulfide bridges, which has been carried out at the structure of F1-ATPase as part of a larger project in which experimental findings should be investigated by MD-simulations. F1-ATPase is an intensively studied molecular motor which produces ATP. Figure 6.4A shows the X-ray structure of F1-ATPase. During its functional cycle the central rotor (colored red) rotates - driven by proton flow - inside the F1-part (green), thereby sythesizing ATP.

Experimental studies on F1-ATPase [109] revealed that crosslinking the rotor with the stator by disulfide bridges in central and bottom positions prevents rotation and therefore ATP hydrolysis/synthesis. However, crosslinking at a top position of the stator did not affect function. Molecular Dynamics studies of the system are expected to provide insights into these findings. In order to set up such simulation systems, the disulfide bridges need to be inserted at different positions in the protein. Figure 6.4B shows the region of the system where one of the disulfide bridges should be introduced. The two cysteine residues that form the disulfide bridge are $\approx 7.4\,\text{Å}$ away from each other in the native X-ray structure. If these two cysteines should form a bond this part of the protein core needs to be re-constructed, thereby imposing a distance constraint on the two sulfur atoms. Figure 6.4D shows the same region of the protein core in space fill represenation which examplifies the dense atomic packing in the protein core. The sulfur-sulfur bond has been defined manually in *tdist*. The positions of all atoms residing more than $10\,\text{Å}$ away from the to cysteines where kept fixed. Afterwards, the region has been reconstructed

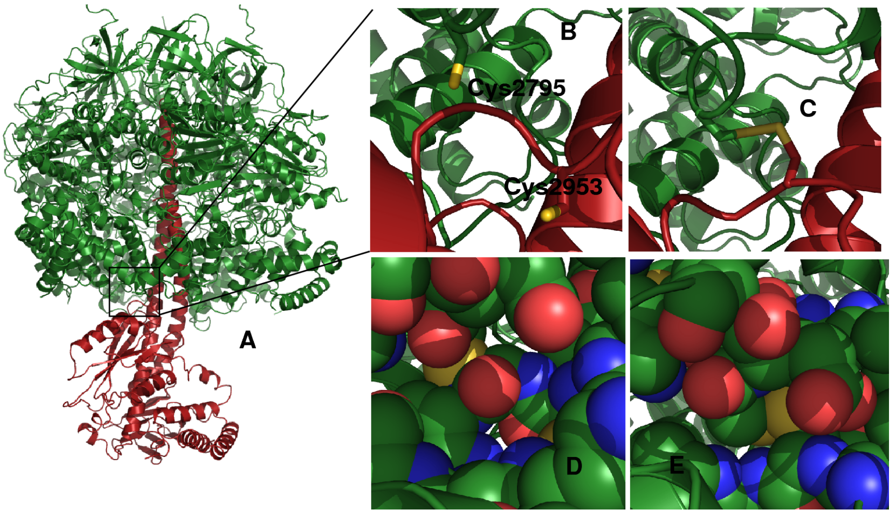


Figure 6.4. **F1-ATPase.** A: Crystal structure of F1-ATPase. B: Region where disulfid bridge should be intoduced. C: Reconstructed region with disulfid bridge. D and E: Space fill models of the region with and without disulfid bridge.

with *tdisco*, thereby generating structures with a disulfide bridge present between the two cysteins (fig. 6.4C+E). The generated *engineered* structures are used to study the structural fundamentals of the experimentally observed behaviour by MD simulations.

## 6.4 Summary

Modeling of protein parts is frequently required in protein science. We have shown that geometry-based methods can be applied to model loops of arbitrary length which is a prerequisite for subsequent simulation or docking protocols. We furthermore showed that tCONCOORD can be used to reconstruct protein parts, thereby allowing for large structural rearrangements. These capabilities are expected to be useful for drug design and protein engineering studies.

# Chapter 7

# Molecular Modeling of Complexes

> *Um ein tadelloses Mitglied einer Schafherde zu sein, muß man vor allem ein Schaf sein.*
>
> *- Albert Einstein*

## 7.1 Introduction

Non-covalent assemblies of molecules are encountered in almost every process in living cells. Communication and control is conducted by small molecules like neurotransmitters or hormones binding to proteins, proteins binding to other proteins, or proteins binding to RNA and DNA. Our understanding of cell processes is thus strongly coupled to the understanding of interactions of participating molecules. Since each drug somehow influences signal cascades or enzymes by specific binding to a protein, our ability to take purposeful influence on metabolic function or malfunction requires knowledge about their structure. The prediction of molecular assemblies is therefore an intensively studied field of research.

Finding or designing small molecules that specifically bind to a target protein is the objective of structure-based drug design (SBDD). Despite experimental milestones like combinatorial chemistry and high-throughput screening (HTS), the number of chemically feasible, drug-like molecules, which has been estimated to be in the order of $10^{60} - 10^{100}$ [110], prohibits exhaustive searching. Computational methods are desired to reduce the number of candidates for experimental testing. Such Virtual Screenings (VS) comprise several methodologies.

Docking small molecules (ligands) in macromolecular structures and estimating the affinity of the resulting complex is a widely used method in structure-based drug design and was pioneered in the early 1980s [111]. The process of molecular docking consists of two steps.

i) The generation of a protein-ligand complex, and
ii) the scoring/estimation of the binding energy

A major shortcoming in current docking protocols is the neglect of protein flexibility. Most docking programs treat the protein as rigid or only allow rotations of selected side-chains, but do not account for backbone mobility.

A second widely used method is the comparison of compound libraries with a pharmacophore model. Such a pharmacophore model is a simple geometric description of a molecule that is assumed complementary to the binding site of a protein. It is either derived from known ligands or the protein structure. Afterwards, the molecules of the compound library are screened towards their ability to adopt a conformation that fits the pharmacophore model.

Geometrical constraints can not only be used to generate structure ensembles, as already shown in chap. 5, but also to generate protein-ligand complexes, thereby allowing both, the protein and the ligand to be fully flexible. In contrast to the prediction of protein flexibility, which requires one set of constraints, this approach requires a geometric description of the protein, a geometric description of the ligand, and constraints that reflect the interaction between both. The first part, the geometric description of the protein has been discussed in previous chapters and applied to different structures. The functionality of generating structure ensembles of arbitrary small molecules was implemented, which requires recognition of hybridization based on the geometry of the input structure. In this chapter, we describe how geometry-based methods can be useful in structure-based drug design and in generating protein-ligand and protein-protein complexes.

## 7.2 Conformational Sampling of Small Molecules

One of the frequently employed methods in receptor-based drug design is the calculation of a pharmacophore model from the receptor structure and its comparison with large libraries of small molecules, thereby checking which ligands are suitable to adopt a conformation complementary to the receptor structure. Thus, a prior knowledge about possible ligand conformations is mandatory. Ususally energy-based methods are employed to sample the conformational space of the ligand. Afterwards a subset of conformations, corresponding to the minima on the energy landscape, is used for comparison with the pharmacophore model. Although a systematic search can be very effective for molecules with limited conformational flexibility, the exponential growth of the search space with the number of rotatable bonds, as well as problems associated with ring closures, limit its utility as a general conformational sampling technique.

Moreover, a recent report based on an examination of $510$ crystal structures concluded that bioactive conformations of ligands often have significantly higher energies than their corresponding energy minima [112]. HIV-1 protease inhibitors

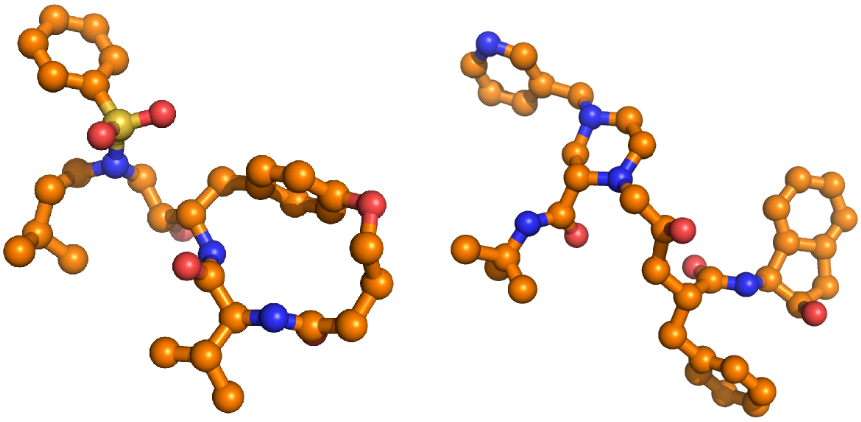


Figure 7.1. **HIV-1 protease inhibitors.** Left picture: The macrocyclic pepidomimetic inhibitor HBB for HIV-1 protease (from PDB 1Z1H). The molecule contains a flexible ring system involving $15$ atoms, $10$ freely rotatable bonds, $2$ peptide bonds with restricted flexibility and a planar aromatic group making generation of conformers with torsion-based methods extremely difficult. The right picture shows a the inhibitor MK1 from PDB 1HSG. In the bound state, the inhibitor adopts a very extended conformation with few intramolecular interactions, hence a energetically unfavoured conformation

are paricularly challenging cases. Figure 7.1 shows two inhibitors of HIV-1 protease. The macrocyclic peptidomimetic inhibitor HBB from the structure with PDB identifier 1Z1H (left) contains a flexible ring system involving 15 atoms, 10 freely rotatable bonds, 2 peptide bonds with restricted flexibility and a planar aromatic group. This renders generation of conformers with torsion-based methods extremely challenging. A second known inhibitor (MK1 from PDB 1HSG) is shown in the right picture of fig. 7.1. In the bound state, the inhibitor adopts a very extended conformation with few intra-molecular interactions, hence a energetically unfavoured conformation. In a Virtual Screening (VS) these inhibitors would probably not have been detected as potential binders due to either having problems to generate conformers or due to not taking in account extended conformations that have higher energies.

tCONCOORD's ability to generate conformational ensembles of arbitrary small molecules may alleviate this obstacle. The generation of several, say 100, conformations of a drug-like molecule, which usually represents a good sampling of the conformational space, takes only a few seconds, hence making it applicable for large compound libraries. The generated ensemble contains geometrically accessible conformations, which not necessarily correspond to a minimum of the potential energy but may represent conformations the molecule adopts upon binding. Such compound libraries may be used for subsequent docking studies or experimental testing. A first step in Virtual Screening is the derivation of a pharmacophore model. Such a model represents a rough geometric description of molecules that are believed to bind to the target protein. In a Virtual Screening approach pharmacophore models are derived either from the receptor structures or from known binders, for instance two hydrogen bond donor positions and one acceptor position that form a triangle with defined geometry. Such models can be derived using tCONCOORD's *tpharm* program, which calculates preferred positions for certain atoms from the receptor structure. Figure 7.2 (left, upper) shows the structure of the serine protease gamma chemotrypsin (PDB 8GCH) with a bound tri-peptide. The pharmacophore model calculated with *tpharm* is visualized with a red mesh for preferred acceptor positions and a blue mesh for preferred donor positions. Figure 7.2 (right, upper) shows that the tri-peptide adopts a conformation such that donor and acceptor atoms reside in the preferred regions.

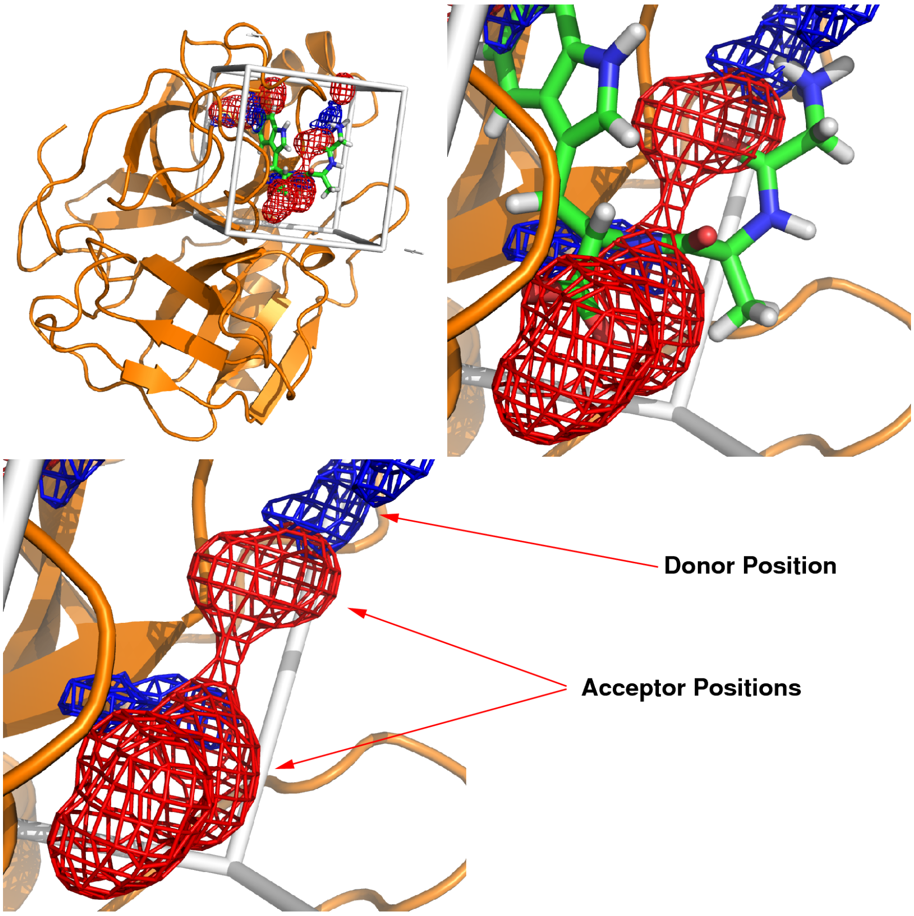


Figure 7.2. **Pharmacophore model.** The upper left picture shows the serine protease gamma chemotrypsin with a bound tri-peptide (PDB 8GCH). In the upper right picture the binding site is shown. The blue mesh represents preferred postions for hydrogen bond donors, whereas the red mesh highlights areas where hydrogen bond acceptors are preferred. The preferred positions for donor and acceptor atoms have been calculated using the tCONCOORD program *tpharm* using the protein structure with removed ligand. In the upper right picture it can be seen that the conformation of the tri-peptide fits nicely to the calculated areas. The calculated preferred areas can be transferred to pharmacophore model using geometrical constraints (lower picture).

A minimum requirement for a potential binder for this protein should therefore be the satisfaction of the pharmacophore model, meaning it must have the ability to adopt a conformation in which two acceptor positions and one donor position satisfy the spatial constraints.

In order two identify such molecules a two step approach is used. First, structure ensembles of a database of drug-like molecules are generated using the *tdist* and *tdisco* programs. Afterwards this data is analyzed using the program *tsearch* which compares the structure ensembles with the pharmacophore model. Those molecules which satisfy the pharmacophore model ($\approx 5\%$ of the screened library) can used for subsequent filtering or for docking studies.

## 7.3 Protein-Ligand Complexes

Obtaining high-resolution structures from protein-ligand complexes is a difficult task and a major bottleneck in structure-based drug design [113, 114]. Once a target protein has been crystallized successfully, protein-ligand complexes are tried to be obtained by soaking ligands into the crystal, which often causes breaking of the crystal. Also co-crystallization of protein and ligand often requires completely different conditions as crystallizing the protein alone. The structure of unbound proteins often differs significantly from the ligand bound conformation, rendering them useless for docking studies and other receptor-based drug design methods [8]. Generating protein conformations of potential ligand bound states is therefore of great interest in the field of structure-based drug design. Geometry-based structure prediction is a helpful instrument to address this question. In tCONCOORD, two different approaches can be employed. The first

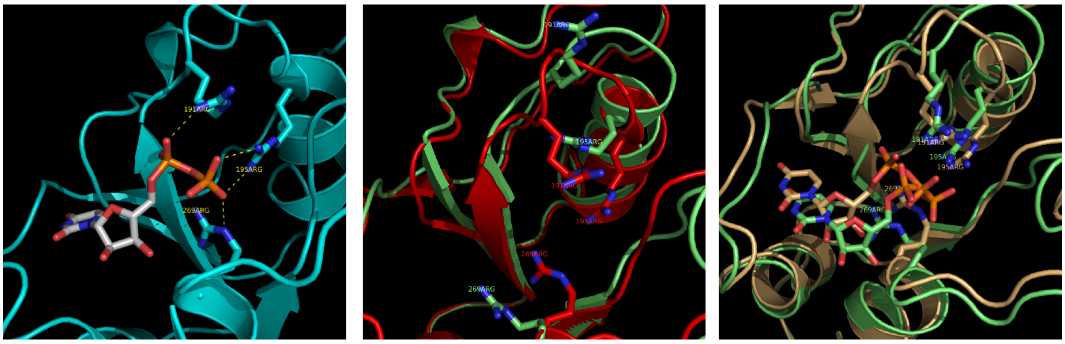


Figure 7.3. **Prediction of Holo-Structures.** Left panel: X-ray structure of the binding site of DNA beta-glucosyltransferase bound to uridine-5 -diphosphate (PDB 1JG6). Middle panel: Holo-structure (red) together with the apo-structure (PDB 1JEJ). Arg191 moves as much as $9$Å upon ligand binding. Right panel: Overlay of the holo-X-ray structure (brown) and a docked pose generated by *tdock* (green). The *tdock* simulation started from the apo X-ray structure (green in the middle panel).

method is expected to be helpful for cases where the apo-structure of the protein is known as well as a binder, the natural substrate for instance. Based on few known interactions between the protein and the ligand, either from mutational studies or NMR-experiments, a newly developed program termed *tdock* generates structures of protein-ligand complexes, thereby allowing both, the ligand and the receptor to be fully flexible. Figure 7.3 shows an application to DNA beta-glucosyltransferase. The left panel shows the X-ray structure of the binding site of DNA beta-glucosyltransferase bound to uridine-5 -diphosphate (PDB 1JG6). The middle panel shows this X-ray structure (in red) together with the apo-X-ray structure (in green, PDB 1JEJ), clearly illustrating the conformational change upon binding. The loop including Arg191 moves as much as 9 Å upon binding the ligand. Using the apo conformation and the ligand together with geometrical constraints between both as input for *tdock* structures are obtained (green) which reproduce the experimentally determined binding mode (brown) shown in the overlay in the right panel of fig. 7.3. The generated protein-ligand complexes can be subjected to molecular dynamics simulation or structure refinement protocols. A second approach is useful for cases where no information about ligand binding is available. Upon ligand binding, many proteins undergo conformational changes, mostly referred to as induced fit. With the example of calmodulin in chapter 5 we already showed that tCONCOORD ensembles started from unbound conformations contain conformations of ligand bound states. Current docking methods fail to rank ligand libraries correctly [115], however, they reproduce experimentally observed binding modes in most cases. Hence, docking a known ligand into an ensemble of protein structures should produce protein-ligand complexes close to the experimentally determined if the protein adopts a conformation close to the ligand bound state. To test this hypothesis, we generated an ensemble from guanlyate kinase, which undergoes a distinct induced fit motion upon binding guanosine-5'-monophosphate. The ensemble, started from the unbound conformation (PDB 1EX6), contains structures close to the experimentally determined ligand bound state. Subsequent docking of the ligand guanosine-5'-monophosphate into such a structure reveals a binding mode close to the experimentally determined structure (fig. 7.5).

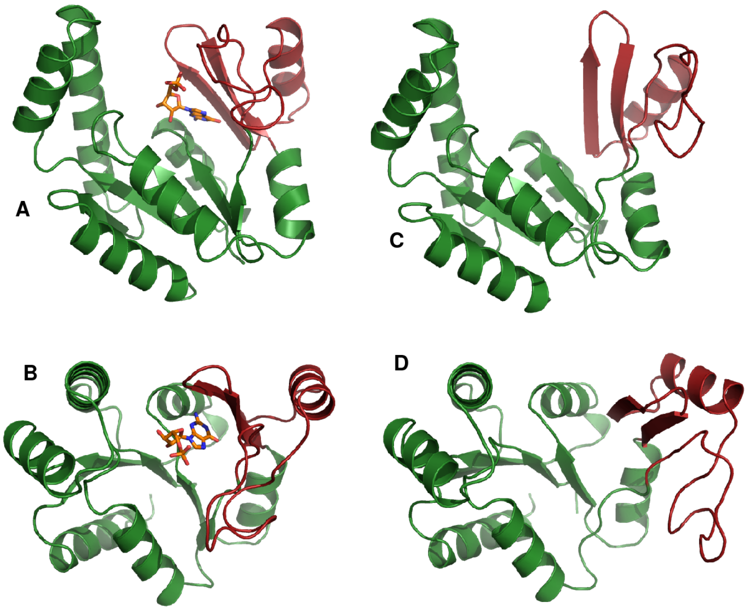


Figure 7.4. **Guanylate Kinase.** A and B show the structure of guanylate kinase bound to the ligand guanosine-5'-monophosphate (PDB 1EX7). C and D show the structure of the apo conformation (PDB 1EX6). Upon binding the ligand, the red colored domain closes over the ligand. The RMSD of this domain between bound and unbound state is $\approx 8$ Å.

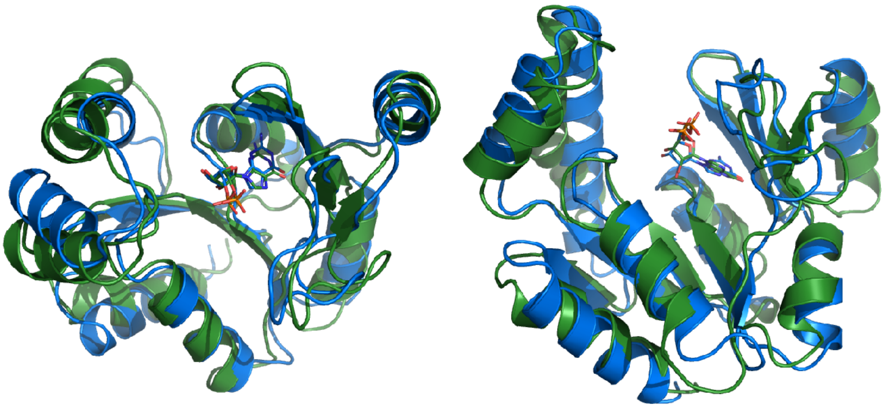


Figure 7.5. **Guanylate Kinase with Ligand.** The experimentally determined structure of guanylate kinase with bound ligand (PDB 1EX7) is shown in blue. The protein-ligand complex, obtained by docking the ligand into a structure from a tCONCOORD ensemble, using the unbound conformation as input, is shown in green.

## 7.4 Protein-Protein Complexes

Conformational sampling of molecular assemblies is not limited to protein-ligand systems. As part of a larger project, which addresses questions related to potassium channel blocking by scorpion toxins, tCONCOORD has been used to study conformational flexibility of a protein-protein assembly. Scorpion toxins, polypeptides of 35-40 amino acids length, bind to the extracellular entrance of potassium channels and efficiently block ion conduction [116–118]. Related peptides constitute major toxic agents in the venoms of spiders, snakes, and sea anemones. The interactions between peptidic toxins and potassium channels range among the strongest of all known protein-protein complexes [118]. Kaliotoxin (KTX), a 38-residue peptide, contains an $\alpha$-helix and two antiparallel $\beta$-strands rigidified by three disulfide bonds, and specifically blocks the voltage-gated $K^+$ channel Kv1.3. KTX binds to a KcsA-Kv1.3 chimera with high specificity and a very high affinity of 30 pM [117, 119]. Although a large number of experimental and theoretical studies have been carried out to address the interaction between toxin peptides and potassium channels, atomic structures of these tight complexes are not available so far. The most detailed information on the structure of the complex comes from computational studies [120], double mutation binding cycles [121], and a recent solid state-NMR (ssNMR) study by Lange et al. [119]. The ssNMR experiment revealed that KTX binding to the KcsA-Kv1.3 chimera changes the conformational states of both KTX and the channel. Based on the assessment of changes in the chemical shift of residues from both kaliotoxin and the K+ channel chimera upon complex formation, it was shown that toxin binding does not simply plug the channel entrance but is also accompanied by a conformational change in the selectivity filter of the channel [119]. The most significant changes of backbone chemical shifts were observed in the region of the extracellular selectivity filter entrance, i.e. at residues Gly77, Gly79, and especially at Tyr78. Among sidechain signals, the most substantial chemical shift changes were seen next to the selectivity filter at Glu71 and Asp80. However, the exact molecular mechanism underlying this set of chemical shift changes remained elusive. Another important question arising from the ssNMR study of the KTX:KcsA-Kv1.3 complex [119] was why the

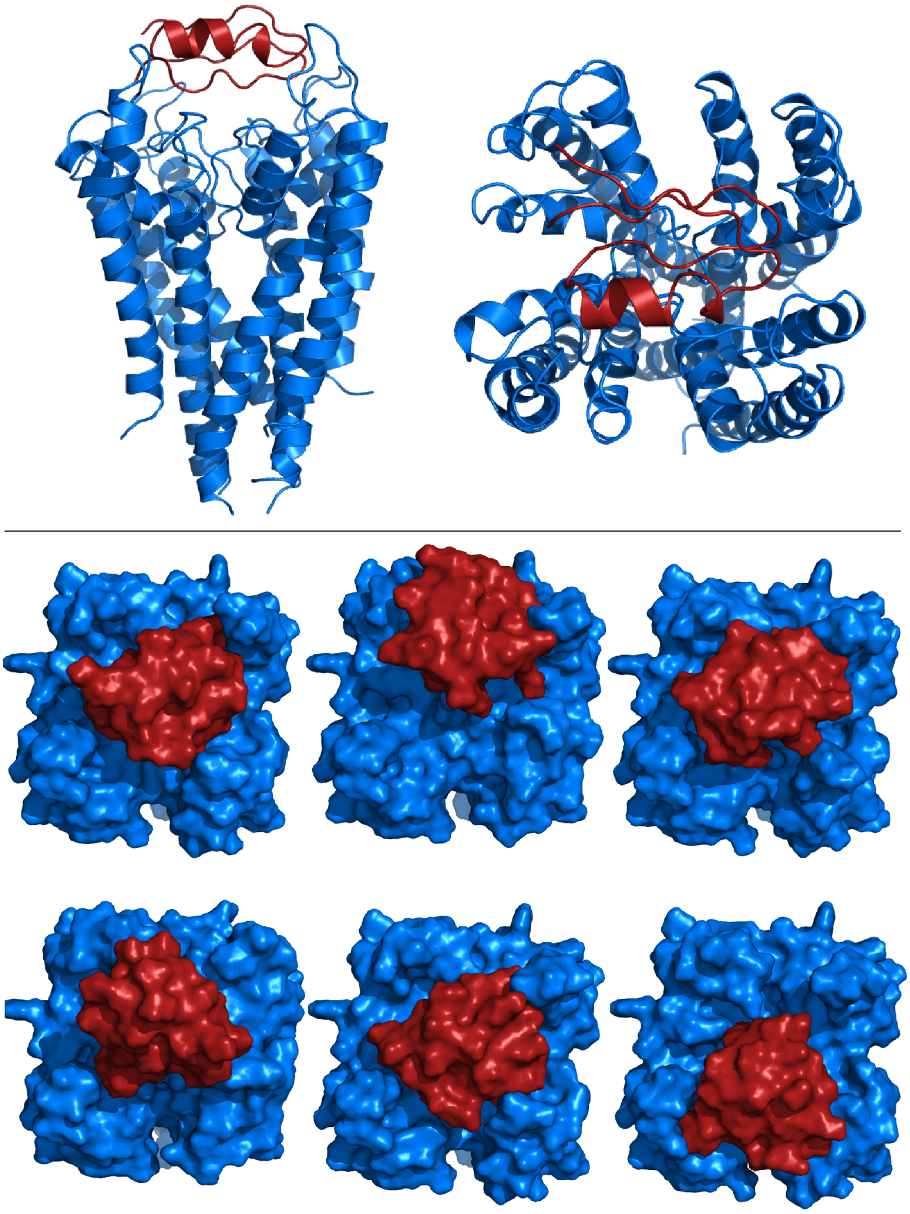


Figure 7.6. **Kaliotoxin binding.** The upper panel shows a side and top view of KTX (red) bound to the potassium channel. This configuration has been derived from MD-simulations. The lower panel shows conformations generated with *tdock*. KTX adopts a variety of conformations that were stable in subsequent MD-simulations.

symmetry between the four channel subunits was kept intact in spite of tight binding of the nonsymmetric toxin peptide. This result cannot be explained by averaging out the effects induced by toxin binding over the four subunits, since a contact between the asymmetric KTX and the single channel, stable during the timescale of the NMR experiment, should be expected to lead to an asymmetric signal.

An intriguing observation made in the ssNMR experiment, therefore, was the retention of the four-fold symmetry of the channel after association of the asymmetric toxin. KTX binding was expected to induce anisotropic chemical shift changes in the channel tetramer due to its non-symmetric shape. A possible explanation is structural heterogeneity in the bound states, i.e. an ensemble of tight structures formed after binding of KTX to the channel. Such an ensemble would average out local breaches of symmetry in the tetramer.

To test this hypothesis, we produced an ensemble of complexes from our MD structural model using tCONCOORD (fig. 7.6). As constraints, we assumed that KTX Lys27 is inserted into the selectivity filter and that the sidechains of Asp80 and Glu71 are charged, as seen in the MD simulations. This resulted in a heterogenous ensemble of bound configurations, all equally geometrically feasible. We tested the stability of ten of these structural models in 10-ns molecular dynamics simulations, which showed a wide variation in the position and orientation KTX adopts in the complex. The large majority of the models remained stably bound in the simulations (fig. 7.6). It is worth noting that spontaneous backbone flips of Tyr78 were observed in these models, i.e. they are consistent with ssNMR. This result indicates that an ensemble of toxin-bound states, rather than a single complexed structure, may in fact be formed by KTX binding to KcsA-Kv1.3. The conformational changes triggered by KTX association at Asp80 and Glu71 and the region between Gly77 and Gly79 of the selectivity filter may be sufficient to allow tight binding of KTX and channel blockade by Lys27. Additionally, binding heterogeneity may increase the affinity of KTX toward KcsA-Kv1.3, by entropic stabilization. Heterogeneity of the complexes might also be an explanation for the fact that crystallization of toxin-channel complexes has not been achieved so far.

## 7.5 Summary

The prediction of molecular assemblies, protein-ligand and protein-protein complexes, is of tremendous importance for understanding function of biological processes. The treatment of protein flexibility in protein-ligand and protein-protein docking, however, is still in its infancy. Fast and efficient conformational sampling with geometry-based methods overcomes current limitations and opens possibilities for the development of new simulation protocols. In this chapter we have shown first steps towards incorporation of geometry-based molecular modeling in structure-based drug design and the prediction of molecular assemblies. Geometry-based conformational sampling can be beneficial for different fields of interest from virtual screening of ligand libraries, prediction of ligand bound conformations from unboud conformations to generation of protein-protein complexes. Especially further development of simulation protocols that enable the prediction of ligand bound conformations from apo structures is expected to alleviate current obstacles in receptor-based drug design.

# Chapter 8

# Conclusions

> *Man kann nicht die Fackel der Wahrheit durch die Menschenmenge tragen,*
> *ohne die Bärte zu versengen.*
>
> *- Georg Christoph Lichtenberg*

The fast and accurate prediction of protein flexibility is one of the major challenges in protein science. Since information about protein flexibility is frequently not experimentally accessible, computational methods are often the only way to bridge the gap between structure, motion and function. In this thesis the tCONCOORD program is developed and applied to diverse fields of protein research.
The methods rests on a the translation of structural data into geometrical constraints on the basis of which structures are reconstructed subsequently. Extensive parametrization was carried out using experimental data, thereby deriving a novel set of atomic radii. These radii were used to study packing properties in protein structures, revealing that the distance distribution of atomic contacts in proteins is exclusively resolution dependent. These findings are expected to enhance protein structure prediction, structure refinement and quality assessment of protein strutcures.
A thorough analysis of interactions in a given protein structure is the basis for defining geometrical constraints. A novel method to estimate the stability of hydrogen bonds was developed and implemented in tCONCOORD. This method allows to predict conformational transitions in proteins which has been demon-

strated at several proteins with diverse folds. The induced fit motion that adenlyate kinase undergoes upon ligand binding was correctly predicted in both directions, when starting from a closed and from an open conformation. Also the large conformational transition of calmodulin, which is associated with partial unfolding and where the activated and ligand bound conformations differ as much as $15\,\text{Å}$ RMSD, was faithfully reproduced. The ligand bound state was reached with $2.8\,\text{Å}$ RMSD when using the activated conformation as input.

The ability to model loops was also implemented in tCONCOORD. Based on geometrical considerations, loop conformations of arbitrary length can be built. The method was applied to model missing loops in the dodecameric bacteriophage $\Phi 29$ connector to allow for subsequent molecular dynamics simulations. Furthermore, tCONCOORD was used to partly reconstruct the core of F1-ATPase, thereby introducing disulfide bridges which required extensive repacking.

In structure-based drug design protein conformational flexibility has been outlined to be a major challenge. The newly implemented capability to handle arbitrary small molecules with tCONCOORD is a first step towards establishing geometry-based sampling in this field. We demonstrated that tCONCOORD can be used for screening compound libraries and to extract those compounds that fit a given pharmacophore model. Even more encouraging is that protein structure ensembles that have been generated with tCONCOORD using an unbound structure as input often sample conformations that correspond to ligand bound states. At the example of guanylate kinase we showed that docking a known ligand into such a structure correctly identifies the binding site and that the proposed binding mode is comparable to the experimentally determined one. A different approach of finding induced fit structures with tCONCOORD can be employed if information about the binding mode is available, e. g. from experimental data. In such a case, the program *tdock* can be used to generate ligand bound protein conformations that were shown to be close to the experimentally observed structure. With the same methodology, protein-protein complexes in different conformations were generated to study scorpion toxin binding to a potassium channel. Many of these conformations remained stable in subsequent molecular dynamics simulations suggesting binding heterogenity that provides a plausible interpretation of experimental findings.

# Outlook

Geometry-based methods, based on the CONCOORD algorithm, are useful tools for protein science. The present work provides an overview on the different fields of protein research where geometry-based methods have been shown to be beneficial so far. Future developments could go into diverse directions. Extensive parametrization of nucleic acids to predict conformational flexibility of DNA, RNA and nucleic acid/protein complexes is an obvious extension. The most promising field of application however, is structure-based drug design where the neglect of protein flexibility in todays established methods is a serious limitation. Both in predicting ligand bound conformations from unbound states and towards development of a fully flexible docking protocol, geometry-based methods are expected to overcome this current limitation. tCONCOORD can be easily extended and combined with other simulation methods. Incorporation of an energy function, e.g. ROSETTA [122,123] or PFF02 [124] should further enhance the quality of generated structure ensembles.

# Chapter 9

# Acknowledgments

Im letzten Kapitel möchte ich all jenen danken, die zum erfolgreichen Abschluss dieser Arbeit auf ganz unterschiedliche Art und Weise beigetragen haben. An erster und herausgehobener Stelle möchte ich meinen Eltern danken, die mich über all die Jahre nicht nur finanziell und moralisch unterstützt haben, sondern auch von Anfang an sämtliche mich anfallenden Begeisterungswellen entweder mitgetragen oder zumindest tapfer ertragen haben. Obwohl nicht katholisch haben sie Pilgerfahrten nach Lourdes und Santiago de Compostela gelobt, sollte ich die Unternehmungen Abitur, Studium und Promotion zu einem erfolgreichen Abschluss bringen. Der richtige Zeitpunkt dafür wäre jetzt! Auch meinen Großeltern danke ich für die jahrelange Unterstützung. Zu wissen, dass ihr und die Familie stolz auf mich seid, war immer Antrieb und Motivation.

Bei Dr. Bert de Groot möchte ich mich ganz besonders herzlich bedanken. Die Art der Betreuung, die ich von ihm während meiner Doktorandenzeit erfahren habe, kann ich nur jedem Doktoranden wünschen. Die stoische Ruhe, mit der er sich neue Ideen, darunter auch minder geniale, gelassen anhört, hat mich nachhaltig beeindruckt. In keinem unserer Gespräche hatte ich den Eindruck mit einem Chef zu sprechen, der auf Autorität qua Amt angewiesen ist, sondern eher mit einem Kollegen, der über viel mehr Wissen, Können und Erfahrung verfügt. Der Titel, unter dem diese Arbeit begonnen wurde, war nie Dogma, sondern eher Orientierung. Mit sämtlichen Freiheiten, eigene Ideen zu verfolgen, zu verwerfen und wiederzubeleben. Prof. Dr. Bernd Abel danke ich,

dass er bereit war, Gutachter für diese Arbeit sein. Ebenso danke ich Prof. Dr. Helmut Grubmüller, dass er sich als Gutachter für diese Arbeit zur Verfügung stellte und darüber hinaus für das einzigartige Arbeitsumfeld und Arbeitsklima in seiner Abteilung.

Eveline Heinemann, in Personalunion Sekretärin, Feldwebel und Wirtschaftskontrolldienst der Abteilungsküche, ist mir unheimlich. Es wird mir immer unbegreiflich bleiben, wie man gleichzeitig telefonieren, email schreiben, ein Gespräch führen und mit einem Blick sehen kann, dass meine Reiskostenabrechnung nicht korrekt ausgefüllt ist. Der gemeine Wissenschaftler, hier schon vom Zusehen überfordert, fühlt sich in solchen Momenten als Gelegenheitsautist.

Einen ebenso unverzichtbaren Anteil am reibungslosen Betrieb der Abteilung haben unsere Systemadministratoren Ansgar Esztermann, Martin Fechner, Oliver Slawik und Ingo Hoffmann, die sich liebevoll um ungefähr 1800 Cluster-CPU's und etwa 40 Arbeitsplatzrechner kümmern. Während mein Verhältnis zu Computern eher nüchtern und pragmatisch ist (Hauptsache er tut was ich will, wie ist mir egal), habe ich bei ihnen häufig dein Eindruck, dass sie eine Art symbiotische Beziehung zu Computern pflegen, deren Wesen mir immer verborgen bleiben wird. In meiner Zeit als Doktorand haben zwei Festplatten, ein Mainboard und ein Bildschirm das Zeitliche gesegnet. In keinem dieser Fälle hat es länger als 24 Stunden gedauert, bis ich mit neuem Gerät und ohne Datenverlust weiterarbeiten konnte. Vielen Dank dafür.

Ich danke ebenso allen Kollegen, die mit Diskussionen, Ratschlägen und Kritik wesentlichen Anteil an dieser Arbeit haben. Besonders herausheben möchte ich Dr. Jürgen Haas, der mir Unmengen von Simulationsdaten zur Verfügung stellte und Dr. Carsten Kutzner für die Hilfe beim Umzingeln besonders hinterhältiger Segmentation faults. Für das Lesen von Manuskripten danke ich Dr. Ira Tremmel, Prof. Dr. Helmut Grubmüller, Prof. Dr. Gert Vriend, Prof. Dr. Christian Griesinger, Prof. Dr. Daan van Aalten und PD Dr. Markus Wahl.

Maik Götte, mit dem ich das Büro teile, verdanke ich eine Vielzahl von Erkenntnissen, die in etablierten Lehrbüchern bis dato keine Würdigung erfahren haben. Das osmotische Perpetuum Mobile, sowie die Deutung geschlechtsspezifischer Verhaltensmuster unter besonderer Berücksichtigung

steinzeitlichen Selektionsdrucks stehen exemplarisch für die schöpferische Kraft (im streng nicht-kreationistischen Sinne), die sich hinter der Tür zu Büro E33 verbirgt. Viel Vergnügen bereitete mir auch immer Maik's Verhältnis zu seinem Rechner, das auch mit viel Milde nicht als friedliche Co-Existenz bezeichnet werden kann. Psychologen bezeichnen eine Haltung wie die seine als intentionalen (absichtsunterstellenden) Standpunkt. Leider sind die Absichten, die Maik's Rechner verfolgt, grundsätzlich bösartiger Natur. Matthias Müller, der auch häufig Zeuge dieser Ereignisse war und mein Staunen teilte, möchte ich ebenso danken. Ganz besonders für das Lesen dieser Arbeit.

Am Schluß möchte ich Tanja danken. Für die Erkenntnis, dass nicht alles mit Wissenschaft, Logik und Rationalität greifbar ist.

# Chapter 10

# Appendix

## 10.1 Protein Structures at Different Levels of Resolution

| $\leq$ 1.2Å | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1a6g | 1a6k | 1a6m | 1a6n | 1amm | 1b9o | 1bkr | 1bz6 | 1bzp | 1bzr | 1d2u | 1d4t |
| 1ds1 | 1ea7 | 1exr | 1f86 | 1f98 | 1f9i | 1gmx | 1gu2 | 1gyo | 1h4a | 1h4x | 1i3h |
| 1i8o | 1ifc | 1ixg | 1ixh | 1j0o | 1j0p | 1j98 | 1jbc | 1jbe | 1jet | 1jf8 | 1jm1 |
| 1jse | 1k5n | 1koi | 1kt6 | 1kwn | 1lf7 | 1lqt | 1ls1 | 1luq | 1m2d | 1mc2 | 1mn8 |
| 1mwq | 1naz | 1nls | 1nwz | 1o7i | 1obo | 1odv | 1oe3 | 1ot6 | 1ot9 | 1ota | 1otb |
| 1p5f | 1pm1 | 1psr | 1q35 | 1r2q | 1rg8 | 1rqw | 1rw1 | 1rwy | 1rxj | 1ryo | 1sau |
| 1sf3 | 1sf5 | 1sfd | 1sfh | 1soa | 1swu | 1sxw | 1sxx | 1sxy | 1sy0 | 1sy1 | 1sy2 |
| 1sy3 | 1t1e | 1t1g | 1tu9 | 1ug6 | 1ugu | 1uzv | 1v8h | 1vk1 | 1vyr | 1x8n | 1x8o |
| 1.3Å | | | | | | | | | | | |
| 1amm | 1atg | 1bxa | 1bz6 | 1ctq | 1e29 | 1f41 | 1f4p | 1fcy | 1flm | 1g61 | 1gnl |
| 1gnt | 1gyo | 1gzt | 1hb2 | 1ikj | 1ird | 1j2r | 1jbc | 1jet | 1jeu | 1jev | 1jhg |
| 1jr0 | 1jw8 | 1kmt | 1kt7 | 1kwn | 1kyf | 1lq9 | 1lxz | 1lzl | 1m2b | 1m70 | 1mf7 |
| 1mjn | 1ml7 | 1obn | 1obo | 1ooh | 1oqv | 1otd | 1oxc | 1qau | 1qks | 1r29 | 1rro |
| 1rtt | 1ryo | 1s2p | 1t1i | 1tjy | 1tu9 | 1ugu | 1usc | 1usf | 1v70 | 1v8h | 1xub |
| 1.5Å | | | | | | | | | | | |
| 1abs | 1bab | 1bvc | 1bz0 | 1c1l | 1ccr | 1d7p | 1do1 | 1e30 | 1e87 | 1elk | 1f46 |
| 1fl0 | 1flp | 1g7n | 1gd0 | 1ggz | 1gmu | 1hbg | 1hd2 | 1i0r | 1i54 | 1icm | 1ike |
| 1j3w | 1jr8 | 1jzf | 1jzl | 1kr7 | 1l7l | 1lfm | 1lmi | 1ln4 | 1m2a | 1mbc | 1mg4 |
| 1n0r | 1na5 | 1noa | 1np4 | 1ntv | 1o3y | 1o7u | 1o85 | 1oaq | 1ocy | 1pee | 1pmy |
| 1pvm | 1q1f | 1qto | 1rat | 1rhb | 1rnc | 1roc | 1sh8 | 1shu | 1st9 | 1szh | 1thb |
| 1tp6 | 1tua | 1uxa | 1vl7 | 1whi | 1x91 | 1xb3 | 1xrk | 1y2t | 2arc | 2bfq | 2hbg |
| 2mbw | 2mcm | 2rat | 2sns | 3ezm | 3hbi | 3rat | 4cpv | 4rat | 5cyt | 5rat | 6rat |

| 1.8Å | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1a01 | 1a3n | 1a3o | 1a7d | 1a7e | 1aaj | 1ag9 | 1aiz | 1akt | 1atz | 1azc | 1azl |
| 1b0w | 1b1i | 1bbh | 1bd8 | 1beb | 1bj7 | 1bje | 1bwh | 1bwi | 1bze | 1c02 | 1c44 |
| 1c7b | 1c7c | 1c7d | 1c8w | 1c9x | 1cbm | 1cbs | 1cgo | 1ch2 | 1ch9 | 1cj6 | 1cj7 |
| 1cj8 | 1cj9 | 1cjw | 1ckc | 1ckd | 1ckf | 1clm | 1cmb | 1cmc | 1co8 | 1dly | 1dqe |
| 1dt1 | 1duz | 1dxt | 1dxu | 1dxv | 1e4h | 1e5a | 1eef | 1ekg | 1enj | 1eo6 | 1euj |
| 1evh | 1f1m | 1f63 | 1fao | 1fd7 | 1fhj | 1fld | 1fnl | 1g8e | 1gbu | 1gdi | 1gdk |
| 1gdl | 1gn0 | 1gqa | 1h8u | 1hn2 | 1huq | 1hxl | 1hxz | 1i4y | 1i53 | 1i7u | 1i8k |
| 1ibe | 1ijt | 1ilk | 1iq4 | 1iu1 | 1j22 | 1jah | 1jai | 1jie | 1jyh | 1jzj | 1k2e |
| 1k6k | 1kdi | 1keb | 1kgi | 1kl5 | 1kn3 | 1kpe | 1kzb | 1kze | 1lav | 1law | 1lb6 |
| 1lds | 1len | 1lfa | 1lhu | 1m6m | 1mfi | 1mlk | 1mlm | 1mlo | 1mtk | 1mwd | 1my5 |
| 1mz4 | 1n6o | 1n71 | 1n8u | 1n9f | 1n9h | 1nbc | 1nco | 1ncx | 1ncz | 1no5 | 1o1l |
| 1o1o | 1o1p | 1ofj | 1ofk | 1oqc | 1ow3 | 1oxj | 1p90 | 1pc5 | 1pgv | 1pxd | 1py9 |
| 1pza | 1pzb | 1q5z | 1qah | 1qi8 | 1qpw | 1qy0 | 1r1y | 1r9h | 1rbr | 1rbs | 1rbt |
| 1rbu | 1rbv | 1rdi | 1rdj | 1rdk | 1rdn | 1rds | 1rg0 | 1rha | 1rlh | 1rtm | 1rtx |
| 1rzy | 1s7i | 1sdk | 1sdl | 1swm | 1tf1 | 1tr0 | 1tt6 | 1tyr | 1u29 | 1uhi | 1ulk |
| 1uy1 | 1v2z | 1vfa | 1vfb | 1wad | 1wba | 1wej | 1wou | 1wsb | 1xau | 1xt6 | 1xxo |
| 1ye2 | 1ytt | 2aae | 2aza | 2bm3 | 2cdv | 2che | 2chf | 2cmm | 2eif | 2fax | 2fcr |
| 2flv | 2fox | 2fvx | 2lal | 2mgd | 2mgf | 2mgg | 2myc | 2myd | 2pab | 2rnt | 2spc |

| 2.0Å | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1a00 | 1a0z | 1a1x | 1a2j | 1a4f | 1a6u | 1a78 | 1a7n | 1a7p | 1a7q | 1a7r | 1a86 |
| 1aan | 1acf | 1afa | 1afd | 1aiu | 1akv | 1aly | 1amx | 1av5 | 1azi | 1b1e | 1b1j |
| 1b8c | 1b9a | 1bch | 1bff | 1bft | 1bhd | 1bht | 1bm7 | 1bm9 | 1bmz | 1bre | 1btn |
| 1bv1 | 1byr | 1bys | 1bz8 | 1c3k | 1c3m | 1c7f | 1c9h | 1cdy | 1cgq | 1ch3 | 1chp |
| 1ckh | 1cp0 | 1cs3 | 1czy | 1d00 | 1d01 | 1d0a | 1d2o | 1d2z | 1d9c | 1dck | 1dd3 |
| 1dm9 | 1do6 | 1dpf | 1dqk | 1dqt | 1duo | 1dvo | 1dvq | 1dvs | 1dvx | 1dy2 | 1e3v |
| 1e5y | 1eei | 1enk | 1esl | 1euo | 1evx | 1ezl | 1f7s | 1f9a | 1fcg | 1fhg | 1fil |
| 1flv | 1fn0 | 1fso | 1ftg | 1fzv | 1g17 | 1g1k | 1g73 | 1g8z | 1gbv | 1gcs | 1gcv |
| 1gd7 | 1gmb | 1gob | 1gr3 | 1gxj | 1h52 | 1h53 | 1hby | 1he1 | 1hmd | 1hmo | 1hy2 |
| 1i04 | 1i05 | 1i1b | 1i1o | 1i4m | 1i55 | 1iii | 1iik | 1iob | 1is5 | 1iul | 1iz6 |
| 1jb2 | 1jlm | 1jra | 1jvl | 1jwg | 1jyj | 1k1k | 1k5u | 1kjt | 1knc | 1kpa | 1kpb |
| 1kuj | 1kwv | 1kwx | 1kwy | 1kx0 | 1kxg | 1l2w | 1l5b | 1l5z | 1lgp | 1lh1 | 1lh2 |
| 1lh3 | 1lh5 | 1lh6 | 1lh7 | 1lho | 1lhs | 1lht | 1lin | 1ljt | 1lki | 1lob | 1loc |
| 1lpj | 1m4r | 1m7b | 1mbi | 1mbn | 1md0 | 1mff | 1mlf | 1mlg | 1mlh | 1mlj | 1mln |
| 1mlq | 1mlr | 1moc | 1mod | 1mp9 | 1mq9 | 1msc | 1mx4 | 1mx6 | 1myi | 1n0s | 1n2d |
| 1np1 | 1np8 | 1npl | 1npu | 1nxv | 1o1k | 1ob9 | 1obp | 1obu | 1oc3 | 1ocw | 1ogc |
| 1oqw | 1oux | 1ox3 | 1p1l | 1p27 | 1p4p | 1pbv | 1pi1 | 1pne | 1py0 | 1q2y | 1q5h |
| 1q5u | 1q5x | 1qc5 | 1qhe | 1qoi | 1qsr | 1qy7 | 1r7l | 1rcd | 1rci | 1rei | 1rfj |
| 1rgl | 1ris | 1rj4 | 1rkb | 1rl6 | 1row | 1rsm | 1rte | 1rtp | 1s3p | 1sct | 1sgm |
| 1sko | 1spe | 1sra | 1swk | 1swp | 1swt | 1tha | 1tjj | 1tjl | 1tn3 | 1tou | 1tow |
| 1ts3 | 1tvq | 1tw4 | 1twu | 1u90 | 1urv | 1uxe | 1v4u | 1v74 | 1vc1 | 1vlg | 1vxa |
| 1vxb | 1wdc | 1wdj | 1wmy | 1wpb | 1xb8 | 1xd5 | 1xd6 | 1xiz | 1xtq | 1xz4 | 1y2f |
| 1y45 | 1y4f | 1yf9 | 1yh2 | 1yih | 1ylk | 1yma | 1ymc | 1yqb | 1zib | 1zon | 1zop |
| 21bi | 2afg | 2ang | 2c2c | 2clr | 2cym | 2e2c | 2fam | 2gal | 2hbe | 2i1b | 2ifb |
| 2lh1 | 2lh2 | 2lh3 | 2lh5 | 2lh6 | 2lh7 | 2lig | 2mgb | 2mgh | 2mgi | 2mgj | 2mgk |
| 2mgl | 2mhb | 2mya | 2np1 | 2rox | 2scp | 2tir | 2tn4 | 2try | 31bi | 3cbs | 3mba |

| 2.2Å | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1a0k | 1aa0 | 1aj9 | 1ao3 | 1aqe | 1awi | 1azb | 1b1u | 1b78 | 1b7m | 1bfa | 1bfc |
| 1bfs | 1bin | 1ca4 | 1cbq | 1cdl | 1cf0 | 1ckg | 1cpw | 1cqk | 1cxa | 1cxz | 1d1j |
| 1dqo | 1dy0 | 1dy1 | 1dyn | 1ecw | 1eje | 1eni | 1ep8 | 1ete | 1etp | 1f6r | 1f6s |
| 1fdb | 1few | 1fga | 1ftp | 1fvc | 1fy9 | 1fya | 1g43 | 1gao | 1gjy | 1h6y | 1hbh |
| 1hda | 1hdb | 1hkf | 1hro | 1i1y | 1i7r | 1i8n | 1ihk | 1ils | 1is1 | 1ise | 1iwn |
| 1j7s | 1joc | 1jot | 1jpg | 1juo | 1jv5 | 1jyb | 1jzk | 1jzn | 1k7u | 1ked | 1kj1 |
| 1kpc | 1l8d | 1lcw | 1loa | 1lt6 | 1lta | 1mnj | 1mnk | 1mob | 1n1q | 1nq3 | 1nzr |
| 1odd | 1of2 | 1oxn | 1p4u | 1pbo | 1pcz | 1pug | 1pxu | 1q21 | 1qew | 1qjh | 1qsd |
| 1qsn | 1qua | 1qvc | 1r28 | 1r5p | 1r6y | 1rcg | 1rpf | 1rph | 1rr7 | 1s9w | 1sce |
| 1si4 | 1sl7 | 1smt | 1sql | 1t1n | 1tk6 | 1tkp | 1tp0 | 1tul | 1ufh | 1ulg | 1vi8 |
| 1vmo | 1wrp | 1xdd | 1xwa | 1xwb | 1y0g | 1y5k | 1yca | 1ycs | 1yh9 | 1yn5 | 2dhn |
| 2hbd | 2hbf | 2hhd | 2mga | 2mjp | 2q21 | 2roy | 2tsa | 3cln | 3tmy | 421p | 4cln |

| 2.4Å | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1a2b | 1ap2 | 1azr | 1bai | 1bii | 1bjf | 1bwu | 1c4p | 1czv | 1dd4 | 1e96 | 1en7 |
| 1f99 | 1fb8 | 1fgb | 1fue | 1fyr | 1g1q | 1gx8 | 1gyw | 1h3q | 1hul | 1i1r | 1i5i |
| 1i8i | 1i9h | 1id1 | 1j4t | 1j9g | 1jhl | 1job | 1joe | 1jrk | 1k7t | 1l9b | 1lb5 |
| 1lxd | 1mup | 1n1i | 1nbw | 1nt3 | 1oek | 1pdk | 1qmt | 1rcc | 1rce | 1rd4 | 1s3l |
| 1squ | 1sys | 1u6m | 1u74 | 1ugy | 1umr | 1uoj | 1ury | 1uvy | 1ux9 | 1v5h | 1vyg |

| 2.5Å | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 155c | 1a1r | 1a4r | 1a9e | 1adw | 1alb | 1b0g | 1b0o | 1b4a | 1b86 | 1b88 | 1btg |
| 1c2r | 1c3a | 1ca7 | 1cdj | 1ch4 | 1cje | 1d0j | 1d8l | 1dp8 | 1dtp | 1e4j | 1ewa |
| 1ewj | 1f4o | 1fdl | 1fl9 | 1frx | 1fx3 | 1g96 | 1gke | 1gli | 1gnq | 1gp9 | 1hhi |
| 1hhj | 1hhk | 1hlb | 1htl | 1htm | 1hup | 1ie4 | 1ies | 1iiu | 1itb | 1ith | 1iuh |
| 1ixx | 1j0r | 1j42 | 1jaf | 1ji5 | 1jnp | 1jsg | 1jy8 | 1l9g | 1le4 | 1ljm | 1lzw |
| 1mbs | 1mi7 | 1mpu | 1mqa | 1msp | 1nk1 | 1npb | 1nwi | 1om9 | 1onl | 1oqe | 1ouu |
| 1p1g | 1p6p | 1pbx | 1pf5 | 1pl5 | 1pmb | 1prq | 1psp | 1pvh | 1pyb | 1qd0 | 1qhh |
| 1qil | 1qpf | 1r14 | 1r5v | 1rtb | 1rvw | 1s3m | 1s3n | 1s9x | 1sj7 | 1szb | 1toq |
| 1u5o | 1uiz | 1vhi | 1vpf | 1wq1 | 1wwa | 1x8s | 1xbn | 1ye0 | 2a2u | 2acg | 2ans |

| 2.6Å | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1aby | 1ahn | 1azn | 1b1b | 1b6e | 1baj | 1bj3 | 1bql | 1byw | 1cav | 1caw | 1cax |
| 1cd8 | 1cqp | 1div | 1doa | 1dp9 | 1dql | 1dxm | 1eh1 | 1f33 | 1ffp | 1h0x | 1hac |
| 1hhg | 1hik | 1hrs | 1ice | 1jgc | 1jnu | 1jts | 1jtz | 1k2f | 1kac | 1lfq | 1lft |
| 1lkt | 1ltb | 1mif | 1mst | 1n1l | 1nih | 1nob | 1oqd | 1q8m | 1qq2 | 1rin | 1rjz |
| 1s1c | 1s1g | 1stp | 1tbp | 1tdq | 1tnf | 1ulc | 1vgf | 1viv | 1ycr | 1yhr | 2bjy |
| 2.8Å | | | | | | | | | | | |
| 1ar2 | 1asx | 1avo | 1b0v | 1baw | 1bmp | 1bq7 | 1bz9 | 1cid | 1e0r | 1f6l | 1fe3 |
| 1gfw | 1gmv | 1h0y | 1h2p | 1hfv | 1hng | 1htn | 1hv4 | 1i1f | 1i7t | 1im9 | 1j95 |
| 1jvm | 1k8f | 1kd7 | 1knk | 1kx8 | 1lfv | 1m4m | 1mfr | 1n0f | 1n0g | 1n9o | 1nlx |
| 1nwn | 1oxz | 1ozb | 1pkp | 1r3k | 1rdd | 1scm | 1sk3 | 1uh0 | 1uh1 | 1uvh | 1v4l |
| 1xni | 1xts | 1y1l | 1yen | 1yo7 | 2ara | 2cbr | 2dhb | 2mm1 | 2ms2 | 2pgh | 2snv |
| 3.0Å | | | | | | | | | | | |
| 1cjq | 1cmy | 1cry | 1dcm | 1dov | 1efx | 1fbi | 1gff | 1gxk | 1h9v | 1hbs | 1hhh |
| 1hij | 1i8l | 1ict | 1ij9 | 1jh5 | 1kq5 | 1l8i | 1le2 | 1lem | 1mfh | 1mva | 1mvb |
| 1niv | 1ny7 | 1qb3 | 1rfb | 1s0h | 1tp8 | 1uot | 1vcp | 1vf5 | 1wat | 1ypo | 1zoo |
| NMR | | | | | | | | | | | |
| 1a57 | 1ajw | 1ak6 | 1bbn | 1blr | 1bsh | 1c8p | 1cfc | 1ck2 | 1ck9 | 1cn7 | 1cz4 |
| 1dc2 | 1e3y | 1e41 | 1egx | 1eiw | 1eo1 | 1eza | 1ezo | 1f2h | 1f3y | 1fmm | 1fo7 |

## 10.2 Protein Structures Used for Hydrogen Bond Statistics

| Resolution | $\leq 1.6$Å | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1A6G | 1A6K | 1A6M | 1A6N | 1A7S | 1AGY | 1B6G | 1B9O | 1BKR | 1BS9 | 1BXO | 1BYI |
| 1BZP | 1BZR | 1C7K | 1CEX | 1CXQ | 1CZB | 1D2U | 1D4T | 1DS1 | 1DY5 | 1EA7 | 1EB6 |
| 1EXR | 1F86 | 1F98 | 1F9I | 1F9Y | 1FN8 | 1FY4 | 1FY5 | 1G4I | 1G66 | 1GCI | 1GDN |
| 1GDQ | 1GDU | 1GMX | 1GOK | 1GQV | 1GU2 | 1GVK | 1GVT | 1GVU | 1GVV | 1GVW | 1GVX |
| 1GWM | 1H11 | 1H2J | 1H4A | 1H4X | 1H5V | 1H97 | 1HDO | 1HF6 | 1HJ8 | 1HJ9 | 1I1W |
| 1I1X | 1I40 | 1I8O | 1IC6 | 1IEE | 1IFC | 1IXG | 1IXH | 1J0O | 1J0P | 1JBC | 1JBE |
| 1JF8 | 1JK3 | 1JM1 | 1JSE | 1JSF | 1K2A | 1K4I | 1K4O | 1K4P | 1K5C | 1K6A | 1K7C |
| 1KCD | 1KF2 | 1KF3 | 1KF4 | 1KF5 | 1KF7 | 1KF8 | 1KMS | 1KMV | 1KNG | 1KOI | 1KOU |
| 1KQ6 | 1KT6 | 1L3K | 1LJN | 1LKK | 1LKS | 1LQP | 1LS1 | 1LU4 | 1LUG | 1LUQ | 1LWB |
| 1M2D | 1M40 | 1M9Z | 1MC2 | 1MFM | 1MJ5 | 1MOO | 1MWQ | 1N55 | 1N9B | 1NAZ | 1NKI |
| 1NLS | 1NNF | 1NQJ | 1NWZ | 1O8S | 1OCQ | 1OD3 | 1OD8 | 1ODV | 1OE2 | 1OE3 | 1OEW |
| 1OEX | 1OH0 | 1ONG | 1OT9 | 1OTA | 1OTB | 1OXD | 1OXE | 1P5F | 1P6O | 1P7V | 1P7W |
| 1PJX | 1PM1 | 1PMH | 1PQ5 | 1PQ7 | 1PQ8 | 1PSR | 1Q0E | 1QJ4 | 1QNJ | 1QTW | 1QV0 |
| 1QV1 | 1QXY | 1R0R | 1R2Q | 1RG8 | 1RQW | 1RTQ | 1RW1 | 1RWY | 1S0Q | 1S0R | 1SAU |
| 1SEN | 1SF3 | 1SF5 | 1SFD | 1SFH | 1SFS | 1SL9 | 1SSX | 1SWY | 1SWZ | 1SX2 | 1SX7 |
| 1SXW | 1SXX | 1SXY | 1SY0 | 1SY1 | 1SY2 | 1SY3 | 1T1G | 1T2D | 1T3Y | 1TJ9 | 1TJM |
| 1TJX | 1TK4 | 1TKJ | 1TQG | 1TT8 | 1U7R | 1UFY | 1UNQ | 1UOW | 1UOZ | 1US0 | 1UTN |
| 1UTO | 1UTQ | 1UZ3 | 1V0K | 1V0L | 1V0M | 1V0N | 1V7S | 1V7T | 1VL9 | 1VZI | 1W0N |
| 1W3L | 1W66 | 1WKQ | 1WTN | 1X6X | 1X6Z | 1X8N | 1X8O | 1X8P | 1X8Q | 1XJU | 1XMT |
| 1XOD | 1XQO | 1XT5 | 1XVM | 1XVO | 1Y55 | 1Y93 | 1YLJ | 1YLT | 1YS1 | 1YWA | 1YWB |
| 1YWC | 1YWD | 1Z2U | 1Z53 | 1Z70 | 1ZJY | 1ZJZ | 1ZK4 | 1ZLB | 1ZWP | 2A6Z | 2AGE |
| 2AGI | 2AH4 | 2ANV | 2ANX | 2AWK | 2AXW | 2AYW | 2B3H | 2BAX | 2BOE | 2BOG | 2BZV |
| 2BZZ | 2C71 | 2C9V | 2CAL | 2CCW | 2CHH | 2CI1 | 2CWS | 2EUT | 2F01 | 2FHL | 2FHZ |
| 2FJ8 | 2FOS | 2FOU | 2FOV | 2FRG | 2G58 | 2GH7 | 2GZ5 | 2PVB | 3LZT | 3PYP | 4LZT |
| 7A3H | 8A3H | | | | | | | | | | |

## 10.3 Ubiquitin Structures

| Ubiquitin | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1aar_1 | 1aar_2 | 1cmx | 1f9j_1 | 1f9j_2 | 1nbf_1 | 1nbf_2 | 1p3q_1 | 1p3q_2 | 1s1q_1 |
| 1s1q_2 | 1tbe_1 | 1tbe_2 | 1ubi | 1ubq | 1uzx | 1wr6_1 | 1wr6_2 | 1wr6_3 | 1wr6_4 |
| 1wrd | 1xd3_1 | 1xd3_2 | 1yd8_1 | 1yd8_2 | 1yiw_1 | 1yiw_2 | 1yiw_3 | 2ayo | 2c7m |
| 2c7n_1 | 2c7n_2 | 2c7n_3 | 2c7n_4 | 2c7n_5 | 2c7n_6 | 2d3g_1 | 2d3g_2 | 2fcq_1 | 2fcq_2 |
| 2fid | 2fif_1 | 2fif_2 | 2fif_3 | 2g45_1 | 2g45_2 | | | | |

# Bibliography

[1] Gerstein, M. and Krebs, W. A database of macromolecular motions. *Nucleic Acids Res.* **15**(26), 4280–4290 (1998).

[2] Bonvin, A. M. J. J. Flexible protein-protein docking. *Curr. Opin. Struct. Biol.* **16**, 194–200 (2006).

[3] Mustard, D. and Ritchie, D. W. Docking essential dynamics eigenstructures. *Proteins: Struct., Funct., Bioinf.* **60**, 269–274 (2005).

[4] Ehrlich, L. P., Nilges, M., and Wade, R. C. The impact of protein flexibility on protein-protein docking. *Proteins: Struct. Funct. Genet.* **58**, 125–133 (2005).

[5] Knegtel, R. M. A., Kuntz, I. D., and Oshiro, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **266**, 424–440 (1997).

[6] Carlson, H. A. Protein flexibility and drug design: How to hit a moving targe t. *Curr. Opin. Chem. Biol.* **6**, 447–452 (2002).

[7] Meagher, K. L. and Carlson, H. A. Incorporating protein flexibility in structure-based drug desi gn: Using hiv-1 protease as a test case. *J. Am. Chem. Soc.* **126**, 13276–13281 (2004).

[8] McGovern, S. L. and Shoichet, B. K. Information decay in molecular docking screens against holo, apo and modeled conformations of enzymes. *J. Med. Chem.* **46**, 2895–2907 (2003).

[9] Teague, S. J. Implications of protein flexibility for drug discovery. *Nature Rev. Drug Discovery* **2**, 527–541 (2003).

[10] Sugita, Y. and Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).

[11] Grubmueller, H. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Phys. Rev. E* **52**, 2893–2906 (1995).

[12] Lange, O. F., Schaefer, L. V., and Grubmueller, H. Flooding in gromacs: accelerated barrier crossings in molecular dynamics. *J. Comp. Chem.* **27**, 1693–1702 (2006).

[13] Schlitter, M., Engels, M., and Kruger, P. Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. *J. Mol. Graph.* **12**, 84–90 (1994).

[14] van der Vaart, A. and Karplus, M. Simulation of conformational transitions by the restricted perturbation-targeted molecular dynamics method. *JCP* **122**, 114903 (2005).

[15] Bahar, I., Atilgan, A. R., Demirel, M. C., and Erman, B. Vibrational dynamics of folded proteins: Significance of slow and fast motions in relation to function and stability. *Phys. Rev. Lett.* **80**(12), 2733–2736, Mar (1998).

[16] Haliloglu, T., Bahar, I., and Erman, B. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.* **79**(16), 3090–3093, Oct (1997).

[17] Go, N., Noguti, T., and Nishikawa, T. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. USA* **80**(12), 3696–3700 (1983).

[18] Brooks, B. and Karplus, M. Harmonic dynamics in proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. USA* **80**(21), 6571–6575 (1983).

[19] Krebs, W., Alexandrov, V., Wilson, C., Echols, N., Yu, H., and Gerstein, M. Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic, (2002).

[20] Alexandrov, V., Lehnert, U., Echols, N., Milburn, D., Engelman, D., and Gerstein, M. Normal modes for predicting protein motions: A comprehensive database assessment and associated web tool. *Protein Sci.* **14**, 633–643 (2005).

[21] Jacobs, D. J., Rader, A. J., Kuhn, L. A., and Thorpe, M. F. Protein flexibility predictions using graph theory. *Proteins: Struct. Funct. Genet.* **44**, 150–165 (2001).

[22] de Groot, B. L., van Aalten, D. M. F., Scheek, R. M., Amadei, A., Vriend, G., and Berendsen, H. J. C. Prediction of protein conformational freedom from distance constraints. *Proteins: Struct. Funct. Genet.* **29**, 240–251 (1997).

[23] Amadei, A., Linssen, A. B. M., and Berendsen, H. J. C. Essential dynamics of proteins. *Proteins: Struct. Funct. Genet.* **17**, 412–425 (1993).

[24] Carlson, H. A. Protein flexibility is an important component of structure-based drug discovery. *Curr. Pharm. Design* **8**(17), 1571–1578 (2002).

[25] Erickson, J. A., Jalaie, M., Robertson, D. H., Lewis, R. A., and Vieth, M. Lessons in molecular recognition: The effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem.* **47**, 45–55 (2004).

[26] Anderson, A. C. The process of structure-based drug design. *ChemBiol* **10**, 787–797 (2003).

[27] Seeliger, D. and de Groot, B. L. Atomic contacts in protein structures: a detailed analysis of atomic radii, packing and overlaps. *PROTEINS: Struct. Funct. Bioinf.* **68**, 595–601 (2007).

[28] Seeliger, D., Haas, J., and de Groot, B. L. Geometry-based sampling of conformational transitions in proteins. *Structure* **15**, 1482–1492 (2007).

[29] Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99 (1963).

[30] Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

[31] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. SCOP, a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).

[32] Lo Conte, L., Brenner, S. E., Hubbard, T. J. P., Chothia, C., and Murzin, A. SCOP database in 2002: refinements accomodate structural genomics. *Nucl. Acid Res.* **30**(1), 264–267 (2002).

[33] Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T., Chothia, C., and Murzin, A. G. SCOP database in 2004: refinements integrate structure and sequence familiy data. *Nucl. Acid Res.* **32**(D), 226–229 (2004).

[34] Dill, K. A., Banu Ozkan, S., Weikl, T. R., Chodera, J. D., and Voelz, V. A. The protein folding problem: when will it be solved? *Curr. Opin. Struct. Biol.* **17**, 342–346 (2007).

[35] Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. The protein data bank, (2000).

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

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

[38] Ponder, J. W. and Case, D. A. Force fields for protein simulations. *Adv. Prot. Chem.* **66**, 27–85 (2003).

[39] Cheatham III, T. E. and Young, M. A. Molecular simulations of nucleic acids. sucesses, limitations and promise. *Biopolymers* **56**, 232–256 (2001).

[40] MacKerell, A. D., Bashford, D., Bellot, M., Dunbrack, R. L., Evanseck, J. D., Field, M., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenckrich, M., Smith, J. C., Stote, R., Straub, J., Wtanabe, M., Wioekiewicz-Kuczera, J., Yin, D., and Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**(18), 3586–3616 (1998).

[41] Van Gunsteren, W. F. and Berendsen, H. J. C. *Gromos manual.* BIOMOS, Biomolecular Software, Laboratory of Physical Chemistry, University of Groningen, The Netherlands, (1987).

[42] Van Gunsteren, W., Billeter, S., Eising, A., Hünenberger, P., Krüger, P., Mark, A., Scott, W., and Tironi, I. *Biomolecular simulation: the GROMOS96 manual and user guide.* Biomos b.v., Zürich, Groningen, (1996).

[43] Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints; molecular dynamics of n-alkanes. *J. Comp. Phys.* **23**, 327–341 (1977).

[44] Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comp. Chem.* **18**, 1463–1472 (1997).

[45] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., and Hermans, J. Interaction models for water in relation to protein hydration. In Intermolecular Forces, Pullman, B., editor, 331–342. D. Reidel Publishing Company, Dordrecht (1981).

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

[47] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).

[48] Guillot, B. A reappraisal of what we have learnt during three decades of computer simulations on water. *J. Mol. Liquids* **101**, 219–260 (2002).

[49] Hess, B. and van der Vegt, N. D. A. Hydration thermodynamic properties of amino acid analouges: A systematic comparison of biomolecular force fields and water models. *J. Phys. Chem. B* **110**, 17616–17626 (2006).

[50] Gō, N., Noguti, T., and Nishikawa, T. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. USA* **80**, 3696–3700 (1983).

[51] Brooks, B. R. and Karplus, M. Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. USA* **80**, 6571–6575 (1983).

[52] Levitt, M., Sander, C., and Stern, P. S. The normal modes of a protein: Native bovine pancreatic trypsin inhibitor. *Int. J. Quant. Chem.: Quantum Biology Symposium* **10**, 181–199 (1983).

[53] Austin, R. H., Beeson, K. W., Eisenstein, L., Frauenfelder, H., and Gunsalus, I. Dynamics of ligand binding to myoglobin. *Biochemistry* **14**(24), 5355–5373 (1975).

[54] Tirion, M. M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical Review Letters* **77**(9), 186–195 (1996).

[55] Vriend, G. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graphics* **8**, 52–56 (1990).

[56] Laskowski, R. A., MacArthur, M. ., Moss, D. S., and Thornton, J. M. Procheck: a program to check the stereochemical quality of protein structures. *J. Appl Cryst.* **26**, 283–291 (1993).

[57] Hooft, R. W. W., Sander, C., and Vriend, G. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. *Proteins: Struct. Funct. Genet.* **26**, 363–376 (1996).

[58] Rowland, R. S. and Taylor, R. Intermolecular nonbonded contact distances in organic crystal structures: Comparison with distances expected from van der waals radii. *J. Phys. Chem.* **100**, 7384–7391 (1996).

[59] Jones, E., Oliphant, T., Peterson, P., et al. SciPy: Open source scientific tools for Python, (2001–).

[60] Chothia, C. Structural invariants in protein folding. *Nature* **254**, 304–308 (1975).

[61] Li, A.-J. and Nussinov, R. A set of van der waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. *PROTEINS: Structure, Function and Genetics* **32**, 111–127 (1998).

[62] Tsai, J., Taylor, R., Chotia, C., and Gerstein, M. The packing density in proteins: Standard radii and volumes. *J. Mol. Biol.* **290**, 253–266 (1999).

[63] Bondi, A. Van der waals volumes and radii. *J. Phys. Chem.* **68**, 441–451 (1964).

[64] Richards, F. M. Areas, volumes, packing, and protein structure. *Ann. Rev. Biophys. Bioeng.* **6**, 151–176 (1977).

[65] Iijima, H., Jr., J. B. D., and Marshall, G. R. Calibration of effective van der waals atomic contact radii for proteins and peptides. *PROTEINS: Structure, Function and Genetics* **2**, 330–339 (1987).

[66] Derewenda, Z. S., Lee, L., and Derewenda, U. The occurence of c-h...o hydrogen bonds in proteins. *J. Mol. Biol.* **252**, 248–262 (1995).

[67] Kang, B. S., Devedjiev, Y., Derewenda, U., and Derewenda, Z. The pdz2 domain of synthenin at ultra-high resolution: Bridging the gap between macromolecular and small molecule crystallography. *J. Mol. Biol.* **338**, 483–493 (2004).

[68] Dahiyat, B. I. and Mayo, S. L. De novo protein design: Fully automated sequence selection. *Science* **278**, 82–87 (1997).

[69] Ventura, S., Vega, M. C., Lacroix, E., Angrand, I., Spagnolo, L., and Serrano, L. Conformational strain in the hydrophobic core and its implications for protein folding and design. *Nat. Struct. Biol.* **9**, 485–493 (2002).

[70] Kuhlmann, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).

[71] Walsh, S. T. R., Cheng, H., Bryson, J. W., Roder, H., and DeGrado, W. F. Solution structure and dynamics of a de novo designed three-helix bundle protein. *Proc. Natl. Acad. Sci. USA* **96**, 5486–5491 (1999).

[72] Desjarlais, J. R. and Handel, T. M. De novo design of hydrophobic cores of proteins. *Protein Science* **4**, 2006–2018 (1995).

[73] Dahiyat, B. I. and Mayo, S. L. Probing the role of packing specifity in protein design. *Proc. Natl. Acad. Sci. USA* **94**, 10172–10177 (1997).

[74] Kellis, J. T., Nyberg, K., and Fersht, A. R. Energetics of complementary side-chain packing in a protein hydrophobic core. *Biochemistry* **28**, 4914–4922 (1989).

[75] Godoy-Ruiz, R., Perez-Jimenez, R., Ibarra-Molero, B., and Sanchez-Ruiz, J. A stability pattern of protein hydrophobic mutations that reflects evolutionary structural optimization. *Biophysical Journal* **89**, 3320–3331 (2005).

[76] Chen, J. and Stites, W. E. Packing is a key selection factor in the evolution of protein hydrophobic cores. *Biochemistry* **40**, 15280–15289 (2001).

[77] Liang, J. and Dill, K. A. Are proteins well-packed? *Biophysical Journal* **81**, 751–766 (2001).

[78] Nabuurs, S. B., Nederveen, A. J., Vranken, W., Doreleijers, J. F., Bonvin, A. M., Vuister, G. W., Vriend, G., and Spronk, C. A. DRESS: a database of refined solution NMR structures. *PROTEINS: Structure, Function and Bioinformatics* **55**, 483–486 (2004).

[79] Word, J. M., Lovell, S. C., LaBean, T. H., Tayler, H. C., Zalis, M. E., Presley, B. K., Richardson, J. S., and Richardson, D. C. Visualizing and quantifying molecular goodness-of-fit: Small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.* **285**, 1711–1733 (1999).

[80] Ratnaparkhi, G. S., Ramachandran, S., Udgaonkar, J. B., and Varadarajan, R. Discrepancies between NMR and x-ray structures of uncomplexed barstar: Analysis suggests that packing densitites of protein structures determined by NMR are unreliable. *Biochemistry* **37**, 6958–6966 (1998).

[81] Doreleijers, J., Rullmann, J. A. C., and Kaptein, R. Quality assessment of NMR structures: a statistical survey. *J. Mol. Biol.* **281**, 149–164 (1998).

[82] Doreleijers, J. F., Vriend, G., Raves, M. L., and Kaptein, R. Validation of nuclear magnetic resonance structures of proteins and nucleic acids: Hydrogen geometry and nomenclature. *PROTEINS: Structure, Function and Genetics* **37**, 404–416 (1999).

[83] Spronk, C. A. E. M., Ringe, J. P., Hilbers, C. W., and Vuister, G. W. Improving the quality of protein structures derived by NMR spectroscopy. *J. Biol. NMR* **22**, 281–289 (2002).

[84] Garbuzynskiy, S. O., Melnik, B. S., Lobanov, M. Y., Finkelstein, A. V., and Galzitskaya, O. V. Comparison of x-ray and NMR structures: Is there a systematic difference in residue contacts between x-ray- and nmr-resolved protein structures? *PROTEINS: Structure, Function and Genetics* **60**, 139–147 (2005).

[85] Fernandez, A., Colubri, A., and Berry, R. S. Three-body correlations in protein folding: the origin of coopertivity. *Physica A* **307**, 235–259 (2002).

[86] Fernandez, A., Sosnick, T. R., and Colubri, A. Dynamics of hydrogen bond desolvation in protein folding. *J. Mol. Biol.* **321**, 659–675 (2002).

[87] Fernandez, A. and Berry, S. Extend of hydrogen-bond protection in folded protein: A constraint on packing architectures. *Biophysical Journal* **83**, 2475–2481 (2002).

[88] Fernandez, A., Rogale, K., Scott, R., and Scheraga, H. A. Inhibitor design by wrapping packing defects in hiv-1 proteins. *Proc. Natl. Acad. Sci. USA* **101**, 11640–11645 (2004).

[89] Lin, M. S., Fawzi, N. L., and Head-Gordon, T. Hydrophopbic potential of mean force as a solvation function for protein structure prediction. *Structure* **15**, 727–740 (2007).

[90] Müller, C. W. and Schulz, G. E. Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor $\mathrm{Ap_5A}$ refined at 1.9 Å resolution. *J. Mol. Biol.* **224**, 159–177 (1992).

[91] Müller, C. W., Schlauderer, G. J., Reinstein, J., and Schulz, G. E. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* **4**(2), 147–156 (1996).

[92] Schlauderer, G. J. and Schulz, G. E. The structure of bovine mitochondrial adenylate kinase: Comparison with isoenzymes in other compartments. *Protein Sci.* **5**(3), 434–441 (1996).

[93] Schlauderer, G. J., Proba, K., and Schulz, G. E. Structure of a mutant adenylate kinase ligated with an atp-analogue showing domain closure over atp. *J. Mol. Biol.* **256**(2), 223–227 (1996).

[94] Chattopadhyaya, R., Meador, W. E., Means, A. R., and Quiocho, F. A. Calmodulin structure refined at 1.7 angstroms resolution. *J. Mol. Biol.* **228**(4), 1177–1192 (1992).

[95] Cook, W. J., Walter, L. J., and Walter, M. R. Drug binding by calmodulin: crystal structure of a calmodulin-trifluoperazine complex. *Biochemistry* **33**(51), 15259–15265 (1994).

[96] Elshorst, B., Hennig, M., Försterling, H., Diener, A., Maurer, M., Schulte, P., Schwalbe, H., Griesinger, C., Krebs, J., Schmid, H., Vorherr, T., and Carafoli, E. Nmr solution structure of a complex of calmodulin with a binding peptide of the $\mathrm{ca}_{2+}$ pump. *Biochemistry* **38**, 12320–12332 (1999).

[97] Brownlee, M. Biochemistry and molecular cell biology of diabetic complications. *Nature* **414**, 813–820 (2001).

[98] Steuber, H., Zentgraf, M., Gerlach, C., Sotriffer, C. A., Heine, A., and Klebe, G. Expect the unexpected or caveat for drug designers: Multiple structure determinations using aldose reductase crystals treated under varying soaking and co-crystallisation conditions. *J. Mol. Biol.* **363**, 174–187 (2006).

[99] de Groot, B. L., Hayward, S., van Aalten, D. M. F., Amadei, A., and Berendsen, H. J. C. Domain motions in bacteriophage T4 lysozyme; a comparison between molecular dynamics and crystallographic data. *Proteins: Struct. Funct. Genet.* **31**, 116–127 (1998).

[100] Love, S. G., Muir, T. W., Ramage, R., Shaw, K. T., Alexeev, D., Sawyer, L., Kelly, S. M., Price, N. C., Arnold, J. E., Mee, M. P., and Mayer, R. J. Synthetic, structural and biological studies of the ubiquitin system: synthesis and crystal structure of an analogue containing unnatural amino acids. *Biochem J* **323**(3), 727–737 (1997).

[101] Suhre, K. and Sanejouand, Y. H. ElNemo: a normal mode web-server for protein movement analysis and the generation of templates for molecular replacement. *Nucl. Acids Res.* **32**, 610–614 (2004).

[102] Suhre, K. and Sanejouand, Y. H. On the potential of normal mode analysis for solving difficult molecular replacement problems. *Act. Cryst. D* **60**, 796–799 (2004).

[103] Wang, J., Truckses, D. M., Abildgaard, F., Dzakula, Z., Zolnai, Z., and Markley, J. L. Solution structures of staphylococcal nuclease from multidimensional, multinuclear NMR: nuclease-h124l and its ternary complex with $Ca^{2+}$ and thymidine-3',5'-bisphosphate. *J Biomol NMR* **10**(2), 143–164 (1997).

[104] Chen, J., Lu, Z., Sakon, J., and Stites, W. E. Increasing the thermostability of staphylococcal nuclease: implications for the origin of protein thermostability. *J. Mol. Biol.* **303**(2), 125–130 (2000).

[105] Spronk, C. A. E. M., Nabuurs, S. B., Bonvin, A. M. J. J., Krieger, E., Vuister, G. W., and Vriend, G. The precison of NMR structure ensembles revisited. *J. Biomol. NMR* **25**, 225–234 (2003).

[106] Bonvin, A. M. and Brünger, A. T. Conformational variability of solution nuclear magnetic resonance structures. *J. Mol. Biol.* **250**(1), 80–93 (1995).

[107] Cuniasse, P., Raynal, I., Yiotakis, A., and Dive, V. Accounting for conformational variability in nmr structure of cyclopeptides: Ensemble averaging of interproton distance and coupling constant restraints. *J. Am. Chem. Soc.* **119**(22), 5239–5248 (1997).

[108] Karen A. Rossi, C. A. W., Nayeem, A., and Krystek jr., S. R. Loopholes and missing links in protein modeling. *Protein Sci.* **16**, 1999–2012 (2007).

[109] Gumbiowski, K., Cherepanov, D., Müller, M., Pänke, O., Promto, P., Winkler, S., Junge, W., and Engelbrecht, S. F1-ATPase: forced full rotation of the rotor despite covalnt crosslink with the stator. *J. Biol. Chem.* **276**(45), 42287–42292 (2001).

[110] Schneider, G. and Fechner, U. Computer-based de novo design of drug-like molecules. *Nature Rev. Drug Discovery* **4**, 649–663 (2005).

[111] Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161**, 269–288 (1982).

[112] Kirchmair, J., Laggner, C., Wolber, G., and Langer, T. Comparative analysis of protein-bound ligand conformations with respect to catalyst's conformational space subsampling algorithms. *J. Chem. Inf. Model.* **45**, 422–430 (2005).

[113] Danley, D. E. Crystallization to obtain protein-ligand complexes for structure-aided drug design. *Act. Cryst. D* **62**, 569–575 (2006).

[114] McNea, I. W., Kan, D., Kontopidis, G., Patterson, A., Tayler, P., Worrall, L., and Walkinshaw, M. D. Studying protein-ligand interactions using protein crystallography. *Crystallography Reviews* **11**(1), 61–71 (2005).

[115] Warren, G. L., Andrews, C. W., Capelli, A.-M., Clarke, B., LaLonde, J., Lambert, M. H., Lindvall, M., Nevins, N., Semus, S. F., Senger, S., Tedesco, G., Wall, I. D., Woolven, J. M., Peishoff, C. E., and Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **49**, 5912–5931 (2006).

[116] Hidalgo, P. and MacKinnon, R. Revealing the architecture of a $K^+$ channel pore through mutant cycles with a peptide inhibitor. *Science* **268**, 307–310 (1995).

[117] Ayiar, J., Withka, J. M., Rizzi, J. P., Singleton, D. H., Andrews, G. C., Lin, W., Boyd, J., Hanson, D. C., Simon, M., and Dethlefs, B. Topology of the pore-region of a $K^+$ channel revealed by the NMR-derived structures of scorpion toxins. *Neuron* **15**(5), 1169–1181 (1995).

[118] Garcia, M. L., Gao, Y. D., McManus, O. B., and Kaczorowski, G. J. Potassium channels: from scorpion venoms to high resolution structure. *Toxicon* **39**, 739–748 (2001).

[119] Lange, A., Giller, K., Hornig, S., Martin-Eauclaire, M.-F., Pongs, O., Becker, S., and Baldus, M. Toxin-induced conformational changes in a potassium channel revealed by solidstate nmr. *Nature* **440**, 959–962 (2006).

[120] Ranganathan, R., Lewis, J. H., and MacKinnon, R. Spatial localization of the $K^+$ channel selectivity filter by mutant cycle-based structure analysis. *Neuron* **16**, 131–139 (1996).

[121] Eriksson, M. A. L. and Roux, B. Modeling the structure of agitoxin in complex with the shaker $K^+$ channel: A computational approach based on experimental distance restraints extracted from thermodnamic mutant cycles. *Biophys. J.* **83**, 2595–2609 (2002).

[122] Bonneau, R., Strauss, C. E., Rohl, C. A., Chivian, D., Bradley, P., Malmström, L., Robertson, T., and Baker, D. De novo prediction of three-dimensional structures for major protein families. *J. Mol. Biol.* **322**(1), 65–78 (2002).

[123] Bradley, P., Chivian, D., Meiler, J., Misura, K. M., Rohl, C. A., Schief, W. R., Wedemeyer, W. J., Schueler-Furman, O., Murphy, P., Schonbrun, J., Strauss, C. E., and Baker, D. Rosetta predictions in casp5: successes, failures, and prospects for complete automation. *PROTEINS: Struct. Funct. Bioinf.* **53**, 457–68 (2003).

[124] Herges, T. and Wenzel, W. An all-atom forcefield for teriary structure prediction of helical proteins. *Biophys. J.* **87**, 3100 (2004).

# Lebenslauf von Daniel Seeliger

---

## Persönliche Daten:

Name: Daniel Christian Seeliger
Adresse: Spandauer Straße 2a , 37120 Lenglern
Geburtsdatum/-ort: 24. August 1977, Kirchheim/Teck
Nationalität: Deutsch

---

## Ausbildung:

| | |
|---|---|
| 07.1988-06.1995 | Schlossgymnasium Kirchheim/Teck<br>Abschluss: Mittlere Reife |
| 07.1995-06.1998 | Technisches Gymnasium der Max-Eyth-Schule, Kirchheim/Teck, Abschluss: Abitur |
| 10.1998-11.2003 | Chemiestudium, Universität Ulm,<br>Abschluss: Diplom-Chemiker |
| 04.2003-10.2003 | Diplomarbeit am Forschungszentrum Jülich,<br>Institut für Werkstoffe und Verfahren der Energietechnik,<br>Betreuer: Prof. Dr. Eckhard Spohr<br>Thema: Computersimulation von protonierten Nafion/Wasser-Systemen |
| 03.2004- | Max-Planck-Institut für biophysikalische Chemie, Göttingen,<br>Beginn der Doktorarbeit mit dem Arbeitstitel:<br>Geometry-based Conformational Sampling of Proteins,<br>Betreuer: Dr. Bert de Groot |