



---

**Forschungszentrum Karlsruhe**  
in der Helmholtz-Gemeinschaft

---

**Wissenschaftliche Berichte**  
FZKA 7171

# **Free-Energy Simulations using Stochastic Optimization Methods for Protein Structure Prediction**

**A. Schug**  
Institut für Nanotechnologie

September 2005



Forschungszentrum Karlsruhe

in der Helmholtz-Gemeinschaft

Wissenschaftliche Berichte

FZKA 7171

Free-Energy Simulations  
using Stochastic Optimization Methods  
for Protein Structure Prediction

Alexander Schug

Institut für Nanotechnologie

vom Fachbereich Physik der Universität Dortmund  
genehmigte Dissertation

Forschungszentrum Karlsruhe GmbH, Karlsruhe  
2005

**Impressum der Print-Ausgabe:**

**Als Manuskript gedruckt  
Für diesen Bericht behalten wir uns alle Rechte vor**

**Forschungszentrum Karlsruhe GmbH  
Postfach 3640, 76021 Karlsruhe**

**Mitglied der Hermann von Helmholtz-Gemeinschaft  
Deutscher Forschungszentren (HGF)**

**ISSN 0947-8620**

**urn:nbn:de:0005-071717**

## Abstract

The prediction of protein tertiary structure and the understanding of the folding process remains one of the outstanding challenges in biological physics. While theoretical models for protein structure prediction that partially rely on experimental information have shown consistent progress, the development of de-novo strategies that rely on sequence information alone is much more complicated. If one assumes that a protein is in thermodynamic equilibrium with its environment, its native state corresponds to the global minimum of its free-energy landscape. The free-energy of the system is accessible by ensemble averaging of the combined internal energy of protein and solvent, or directly in a free-energy forcefield. Here an implicit solvation model approximates interactions of the system protein and solvent as well as most of the entropic contributions. As a major challenge remains the search for the global minimum for which different stochastic optimization methods were developed and applied. Using these methods it was possible to predict the structure of helical proteins from different protein families ranging in size from 20 to 60 amino acids. Starting with random initial conformations we achieved a high agreement with experimental data.

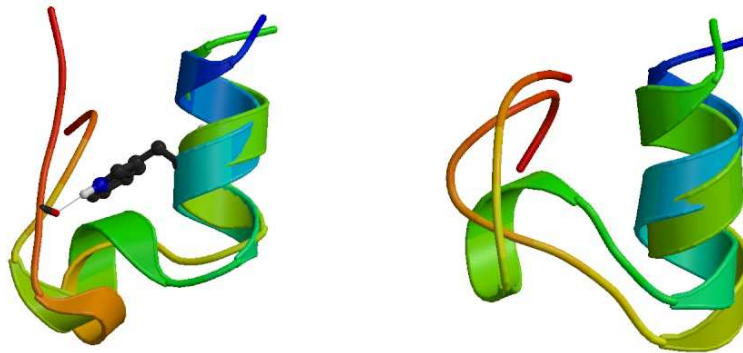
## Freie Energie Simulationen mittels stochastischer Optimierungsmethoden für die Proteinstruktur

### Zusammenfassung

Die Vorhersage der Tertiärstruktur eines Proteins und das Verständnis des zugehörigen Faltungsvorganges stellen eine große Herausforderung in der biologischen Physik dar. Obwohl theoretische Modelle für die Proteinstrukturvorhersage, die auf experimentelle Daten zurückgreifen, ständigen Fortschritt zeigen, ist die Entwicklung von Methoden, die einzig auf die Sequenzinformation zurückgreifen, um ein vielfaches aufwändiger. Unter der Annahme, dass der native Zustand eines Proteins im thermodynamischen Gleichgewicht mit seiner Umgebung ist, entspricht dieser dem globalen Minimum seiner freien Energie. Die freie Energie des Systems ist entweder über eine Ensemble Mittlung der kombinierten inneren Energien von Protein und Solvent zugänglich oder man greift direkt auf ein geeignetes Kraftfeld für die freie Energie zurück. In dieser Arbeit nähert ein implizites Lösungsmittelmodell die Wechselwirkungen des Systems Protein-Solvent sowie die wesentlichen entropischen Beiträge an. Es verbleibt die Herausforderung der globalen Minimierung, für die verschiedene stochastische Optimierungsstrategien entwickelt und angewandt worden sind. Mittels dieser Methoden war es möglich, die native Struktur verschiedener helikaler Proteine der Größe 20 bis 60 Aminosäuren erfolgreich von zufälligen Startstrukturen ausgehend vorherzusagen. Diese Strukturen sind in hoher Übereinstimmung mit experimentellen Daten.



Free-Energy Simulations using Stochastic Optimization  
Methods for Protein Structure Prediction



Dissertation

University of Dortmund  
Department of Physics

Alexander Schug

March 2005

*to my father, Dr. Heinrich Schug*

The picture on the front page shows an overlay of measured and best estimation of the global minimum in PFF01 using the Stochastic Tunneling method. The theoretical structure in the left image shows a correctly folded protein and agrees well with the NMR result. The image on the right is slightly misfolded but may represent an important intermediate step in the folding process. Title Page of Phys. Rev. Letters 10th October 2003[105].



# Inhaltsverzeichnis

<b>1</b>	<b>Preamble</b>	<b>5</b>
<b>2</b>	<b>Basics of Proteins</b>	<b>7</b>
2.1	Living Systems . . . . .	7
2.2	Amino Acids . . . . .	8
2.2.1	Peptide Binding and Peptides . . . . .	12
2.2.2	Dihedral Angles . . . . .	12
2.3	Proteins . . . . .	12
2.3.1	Primary Structure . . . . .	16
2.3.2	Secondary Structure . . . . .	16
2.3.3	Disulfide Bridges . . . . .	18
2.4	Protein Stability and Thermodynamic Hypothesis . . . . .	18
<b>3</b>	<b>Protein Structure Prediction</b>	<b>21</b>
3.1	Protein Structure and Evolution . . . . .	21
3.2	Homology Searching and Multiple Sequence Alignment . . . . .	22
3.3	Prediction of Secondary Structure . . . . .	22
3.4	Prediction of Tertiary Structure . . . . .	23
<b>4</b>	<b>Forcefields for Protein Folding</b>	<b>25</b>
4.1	Thermodynamics . . . . .	25
4.1.1	Interaction of chemically bonded atoms . . . . .	26
4.1.2	Interactions of non-bonded atoms . . . . .	28
4.1.3	Molecular Dynamics . . . . .	31
4.1.4	Established forcefields . . . . .	32
4.1.5	Discussion on the Transferability of individual Forcefield Terms . . . . .	34
<b>5</b>	<b>Protein Force Field 01, PFF01</b>	<b>35</b>
5.1	Interactions of the Forcefield . . . . .	36
5.1.1	Potential Types . . . . .	36

<b>6</b>	<b>Stochastic Minimization</b>	<b>45</b>
6.1	Stochastic Minimization . . . . .	45
6.1.1	Introduction . . . . .	45
6.1.2	The basic Idea: Monte-Carlo . . . . .	46
6.1.3	Simulated Annealing . . . . .	50
6.1.4	Freezing Problem . . . . .	51
6.1.5	Stochastic Tunneling . . . . .	51
6.2	Basin Hopping . . . . .	52
6.3	Genetic Algorithms . . . . .	55
6.4	Distributed Computing: Server-Client Model using Screen Savers . . . . .	55
6.5	Energy Landscape Paving . . . . .	56
6.6	Adaptive Parallel Tempering . . . . .	56
<b>7</b>	<b>Prediction of tertiary structures with PFF01</b>	<b>59</b>
7.1	Test on theoretically modeled $Ala_{10} - Gly_5 - Ala_{10}$ . . . . .	60
7.2	Trp-Cage Protein . . . . .	62
7.2.1	Stochastic Tunneling . . . . .	62
7.2.2	Adaptive Parallel Tempering . . . . .	67
7.2.3	Energy Landscape Paving . . . . .	69
7.3	Protein A . . . . .	77
7.4	HIV-Accessory Protein . . . . .	79
7.5	Villin Headpiece . . . . .	80
7.6	Bacterial Ribosomal Protein L20 . . . . .	81
7.7	Discussion on the Efficiency of Optimization Techniques . . . . .	83
<b>8</b>	<b>Summary</b>	<b>87</b>
<b>A</b>	<b>Used programs and definitions</b>	<b>91</b>

# Kapitel 1

## Preamble

Proteins are the most versatile macromolecules in biological systems[11]. They fulfill a multitude of different tasks from storing other molecules like oxygen in human blood to catalytic function. They are responsible for the stability of macro-complexes, such as hair, and facilitate membrane transport or the transmission of electric signals in the human brain[124]. In spite of this importance their properties and function are still far from being fully understood by scientists. Their basic composition is that of linear polymers composed of sequences of amino acids. This amino acid sequence is encoded by DNA. It already contains all necessary information about the unique three-dimensional structure of a protein which is directly correlated to its function. The formation of this three-dimensional structure in their physiological environment, mostly water, out of its linear amino acid sequence takes place in a complex process called *protein folding*. Regardless of the starting point or de-folding it by changing the environmental conditions many proteins will finally always assume the same structure.<sup>1</sup> This unique three-dimensional structure is called the *native state* of a protein which can be obtained to atomic resolution for many proteins by X-ray scattering or NMR. For lack of suitable experimental techniques, time resolved information on the the actual folding process is currently not available with similar detail.

In order to *theoretically understand* the process of protein folding scientists in different fields, such as biology, information science, mathematics, biochemistry and physics, have pursued various often interdisciplinary approaches. Knowledge based approaches use databases of experimentally determined structural information for proteins to predict structures for other proteins[118, 36]. In recent years these methods have made steady progress towards *de-novo protein structure prediction*, although they require substantial sequence similarity to yield usable results. Unfortunately they give only indirect evidence regarding the mechanisms by which proteins assume their unique three-dimensional structure.

The challenge for more sophisticated models motivated by physical ideas is the size and complexity of the system. Proteins consist of numerous different amino acids with hundreds of atoms but have no exploitable higher symmetries. In order to gain more insight into the

---

<sup>1</sup>Very few proteins show an alternative native state under minimal changes in environment like prions.

folding process simple, but tractable *lattice models* representing protein structure have been developed[58, 81]. The necessary simplifications however leave a wide gap between these models and actual protein structure. The simulation of the protein folding process by molecular dynamics using existing biomolecular forcefields like AMBER or CHARMM has yielded increasingly valuable insights into the folding mechanisms. Such simulations permit the understanding of the folding process by an interpretation of the gained data, but are presently limited to very small proteins by the extremely high computational demands[30, 112, 101]. One major source of complexity arises from the strong influence the environment has on the folding process and protein structure. Thus the appropriate inclusion of solvent effects has led to controversial debates[120, 22, 91]. It is also presently unclear which families of proteins are adequately folded using established forcefields with molecular dynamical simulations. In some simulations the forcefields fail to stabilize the native state, which raises questions about their applicability.

In this thesis the validity of an alternative, atomically resolved approach to the folding of proteins based on models of the underlying physical interactions is investigated. The thermodynamic hypothesis[3] postulates that most proteins are in thermodynamic equilibrium with their environment. Therefore it should be possible to represent this unique native state as the global minimum of an appropriate free-energy model[40, 42]. In this thesis we will use an *all-atom free-energy-forcefield* based on physically motivated interactions called PFF01[53, 48] to represent the underlying interactions governing protein structure formation. Starting from random initial conditions we attempt to predict protein structure de-novo, by finding the global optimum of the forcefield. The computational demands of this approach are significantly less than molecular dynamical simulations while still allowing insight into the forces responsible for stabilizing the native state or driving the folding process.

We will concretely validate the forcefield against experimental data, i.e. the experimentally determined native state should correspond to the global minimum. Given the forcefield the remaining challenge is finding this global minimum in the rough energy landscape representing the high-dimensional conformational space of a given protein.

Accordingly this thesis will address two central questions:

- Protein folding is ultimately governed by complicated quantum-mechanical effects, such as the formation of hydrogen bonds, Fermi-repulsion of electronic clouds and interaction of the protein surface with a complex environment: *Can the folding of a protein be understood and represented by a classical free-energy-forcefield and, if yes, how can it be done in a computationally treatable way?*
- Proteins have many degrees of freedom and no exploitable symmetries. It is known that global minimization of rough and high-dimensional energy landscapes, like those in spin-glass theory, is very difficult. Therefore: *Due to the complexity of such a forcefield, are there optimization methods allowing to find the global minimum and what about the efficiency of these methods?*

# Kapitel 2

## Basics of Proteins

Proteins occur in very many different activities in biological systems. Some exist in solvent (water) only while others are fully or partially embedded in membranes. They fulfill a multitude of different tasks from controlling the flow of ions and molecules in and out of cells over giving stability to macroscopic structures to enzymatic functions[11, 124].

This chapter gives a short introduction to the chemical and physical properties of proteins. It is shortly explained how they are built by peptide bonding and what their basic composition is. Their biological function and importance is only touched.

### 2.1 Living Systems

The dry matter of biological systems consists to over 98% of a very limited set of elements: C, N, O, H, Ca, P, K and S. <sup>1</sup> The mechanism by which biological life has developed out of these elements is still not completely understood. Different, competing scenarios exist. One of these scenarios investigates the development and chemical interactions of the atmosphere on early earth[11, 124].

About 4.5 billion years ago the atmosphere on early earth consisted of H<sub>2</sub>O, N<sub>2</sub>, CO<sub>2</sub>, CH<sub>4</sub>, N<sub>3</sub>, SO<sub>2</sub> and H<sub>2</sub>. The ultraviolet radiation of the sun and/or electric discharge facilitated the construction of more complex organic molecules like simple organic acids or amino acids[80]. The interaction and competition of these first molecules to develop forward to more complexity or to break up again in their components lead to kind of critical moment in the development of early life: the transition from randomly built molecules to small systems of molecules which were capable of self-replication. During this process the first proteins have developed which is also reflected in their name which derives from the Greek *proteios*, the first, primary. These systems, including proteins, must have been able to deal with changing environments. Out of these small systems finally cells formed which had the evolutionary advantage of protecting

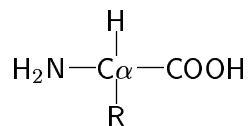
---

<sup>1</sup>Other elements like Fe may play vital roles in some biochemical interactions, like Fe for the oxygen transport in blood, but are very rare.

their interior against the environment. After the emergence of cells, evolution of complex life-forms on earth proceeded.

## 2.2 Amino Acids

For sake of simplicity when we speak in the following of amino acids (AA) we refer to the 20 *standard* amino acids commonly found in nature[124, 11].<sup>2</sup> They belong to the group of  $\alpha$ -amino acids. The  $\alpha$ -amino acids consist of a central carbon  $C_\alpha$ , surrounded by a primary amino-group (-NH<sub>2</sub>), a carboxyl-group (-COOH), a hydrogen (-H) and a specific side-chain (R).<sup>3</sup> In addition it is remarkable that all amino acids in proteins are of the L-form, with the D-form absent. The reasons for this preference of the chiral L-form are unknown[124, 11, 24]. The general structure of an amino acid is shown in schema 2-1. The difference between the



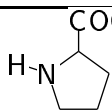
Schema 2-1: The chemical structure of amino acids. The side-chains ('R' for residue) differ between the different amino acids and give them their unique chemical and physical properties.

20 amino acids lies in their side-chains, which show a different composition and therefore give each amino acid unique chemical and physical properties. One classification divides these 20 amino acids into three groups by the polarity and charge of the side-chains[124]. A listing of these amino acids is as follows:

---

<sup>2</sup>Also less common amino acids exist in addition to the 20 standard amino acids.

<sup>3</sup>Proline is an exception from this rule since it consists of a secondary not a primary amino-group.

Name	Three-letter-code	One-letter-code	structural formula
Amino acids with apolar side-chains			
Alanine	Ala	A	$\begin{array}{c} \text{H} \\   \\ \text{H}_2\text{N}-\text{C}-\text{COOH} \\   \\ \text{CH}_3 \end{array}$
Glycine	Gly	G	$\begin{array}{c} \text{H} \\   \\ \text{H}_2\text{N}-\text{C}-\text{COOH} \\   \\ \text{H} \end{array}$
Isoleucine	Ile	I	$\begin{array}{c} \text{H} \\   \\ \text{H}_2\text{N}-\text{C}-\text{COOH} \\   \\ \text{H}_3\text{C}-\text{CH}-\text{CH}_2 \\   \\ \text{CH}_3 \end{array}$
Leucine	Leu	L	$\begin{array}{c} \text{H} \\   \\ \text{H}_2\text{N}-\text{C}-\text{COOH} \\   \\ \text{CH}_2 \\   \\ \text{H}_3\text{C}-\text{CH}-\text{CH}_3 \end{array}$
Methionine	Met	M	$\begin{array}{c} \text{H} \\   \\ \text{H}_2\text{N}-\text{C}-\text{COOH} \\   \\ \text{CH}_2 \\   \\ \text{CH}_2 \\   \\ \text{S} \\   \\ \text{CH}_3 \end{array}$
Phenylalanine	Phe	F	$\begin{array}{c} \text{H} \\   \\ \text{H}_2\text{N}-\text{C}-\text{COOH} \\   \\ \text{CH}_2 \\   \\ \text{C}_6\text{H}_5 \end{array}$
Proline	Pro	P	
Tryptophan	Trp	W	$\begin{array}{c} \text{H} \\   \\ \text{H}_2\text{N}-\text{C}-\text{COOH} \\   \\ \text{CH}_2 \\   \\ \text{C}_8\text{H}_6\text{N}_2 \end{array}$
Valine	Val	V	$\begin{array}{c} \text{H} \\   \\ \text{H}_2\text{N}-\text{C}-\text{COOH} \\   \\ \text{H}_3\text{C}-\text{CH}-\text{CH}_3 \end{array}$

Name	Three-letter-code	One-letter-code	structural formula
Amino acids with uncharged apolar side-chains			
Asparagine	Asn	N	$\begin{array}{c} \text{H} \\   \\ \text{H}_2\text{N}-\text{C}-\text{COOH} \\   \\ \text{CH}_2 \\   \\ \text{CONH}_2 \end{array}$
Cysteine	Cys	C	$\begin{array}{c} \text{H} \\   \\ \text{H}_2\text{N}-\text{C}-\text{COOH} \\   \\ \text{CH}_2\text{SH} \end{array}$
Glutamine	Gln	Q	$\begin{array}{c} \text{H} \\   \\ \text{H}_2\text{N}-\text{C}-\text{COOH} \\   \\ \text{CH}_2 \\   \\ \text{CH}_2 \\   \\ \text{CONH}_2 \end{array}$
Serine	Ser	S	$\begin{array}{c} \text{H} \\   \\ \text{H}_2\text{N}-\text{C}-\text{COOH} \\   \\ \text{CH}_2\text{OH} \end{array}$
Threonine	Thr	T	$\begin{array}{c} \text{H} \\   \\ \text{H}_2\text{N}-\text{C}-\text{COOH} \\   \\ \text{CHOH} \\   \\ \text{CH}_3 \end{array}$
Tyrosine	Tyr	Y	$\begin{array}{c} \text{H} \\   \\ \text{H}_2\text{N}-\text{C}-\text{COOH} \\   \\ \text{CH}_2 \\   \\ \text{C}_6\text{H}_4 \\   \\ \text{OH} \end{array}$

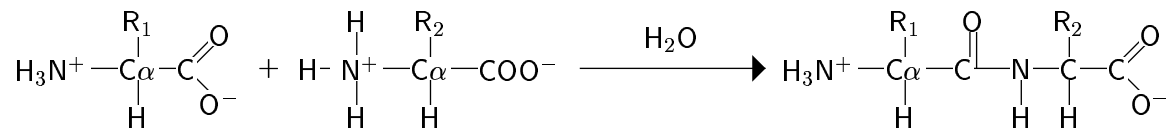


Name	Three-letter-code	One-letter-code	structural formula
Amino acids with charged polar side-chains			
Arginine	Arg	R	$  \begin{array}{c}  \text{H} \\    \\  \text{H}_2\text{N}-\text{C}-\text{COOH} \\    \\  \text{CH}_2 \\    \\  \text{CH}_2 \\    \\  \text{CH}_2 \\    \\  \text{NH} \\    \\  \text{H}_2\text{N}-\text{C}-\text{NH}  \end{array}  $
Aspartate	Asp	D	$  \begin{array}{c}  \text{H} \\    \\  \text{H}_2\text{N}-\text{C}-\text{COOH} \\    \\  \text{CH}_2 \\    \\  \text{COOH}  \end{array}  $
Glutamate	Glu	E	$  \begin{array}{c}  \text{H} \\    \\  \text{H}_2\text{N}-\text{C}-\text{COOH} \\    \\  \text{CH}_2 \\    \\  \text{CH}_2 \\    \\  \text{COOH}  \end{array}  $
Histidine	His	H	$  \begin{array}{c}  \text{H} \\    \\  \text{H}_2\text{N}-\text{C}-\text{COOH} \\    \\  \text{CH}_2 \\    \\  \text{N} \\  \diagup \quad \diagdown \\  \text{N} \quad \text{N}-\text{H}  \end{array}  $
Lysine	Lys	K	$  \begin{array}{c}  \text{H} \\    \\  \text{H}_2\text{N}-\text{C}-\text{COOH} \\    \\  \text{CH}_2 \\    \\  \text{CH}_2 \\    \\  \text{CH}_2 \\    \\  \text{CH}_2\text{NH}_2  \end{array}  $

Table 2.1: Overview of the different amino acids and a possible classification by charge and polarity of the side-chains[124]. Each of the 20 standard amino acids is presented with its one- and three-letter codes. In addition its chemical structural formula is shown.

## 2.2.1 Peptide Binding and Peptides

Under physiological conditions the amino-group protonizes and the carboxyl-group deprotonizes. This can lead to peptide-bonding between two amino acids.:



One very important aspect of peptide bonding is its  $\pi$ -bond-character. This reduces the degrees of freedom for the peptide since the  $\pi$ -bond-character constrains rotations around the peptide bond to very small angles. In practice the peptide plane can be seen as rigid[24]. Longer chains of amino acids connected by this chemical reaction are simply called peptides. One classifies them into di-,tri-,oligo- or poly-peptides when two, three, some (3-10) or many amino acids are linked via this peptide-bond. These peptides are *linear* chains because every amino acid can be linked only to two others. No covalent bonds between side-chains belonging to different amino acids occur except in the special case of disulfide bridges<sup>4</sup>. The counting of the amino acids starts with the *N-terminus* (or amino terminus) and stops at the amino acid with the free carboxyl group, the *C-terminus*[124, 11].

## 2.2.2 Dihedral Angles

Due to the  $\pi$ -character of the peptide plane only the *dihedral angles* remain as degrees of freedom for the main chain of a protein. In this case both bonds and the peptide plane are seen as rigid. The dihedral angles are the two angles connecting the two planes which sit next to the  $C_\alpha$  of each amino acid. They are, starting from the N-terminus, called  $\Phi$  and  $\Psi$ [24]. The normal arrangement of a protein is in trans form, as illustrated in figure 2.1. One can plot the angles  $\Phi$  and  $\Psi$  against each other in form of a diagram, where each spot indicates a sterically allowed conformation. Such a diagram is called *Ramachandran-plot*[99]. One example of such a plot is given in figure 2.2.

## 2.3 Proteins

Proteins are molecules consisting of one or more poly-peptides with a well defined three dimensional structure. Different kind of proteins exist. *Globular proteins* have their three-dimensional structure with all atoms closely packed. They take their structures depending on the solvent, which is normally water, and form a hydrophobic core. In appearance look like spheres. Other proteins are called *transmembrane proteins* because they are partly or fully embedded inside a membrane. Despite these differentiations all proteins have several common structural elements. Since globular proteins are most common we will speak in the following

<sup>4</sup>A disulfide bridge is a covalent linking between the sulfurs of two cysteines.

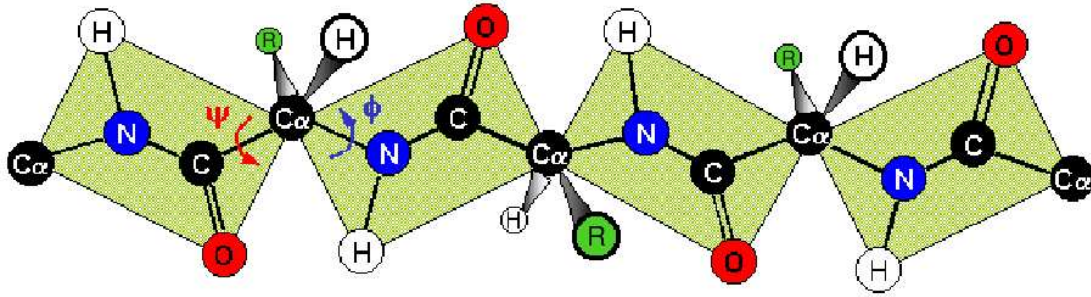


Abbildung 2.1: This picture shows an maximally extended peptide chain with  $\Phi$  and  $\Psi = 180^\circ$  each. The indicated planes can be regarded as rigid due to the nature of peptide bonding.

of them when not specifically indicating other kind of proteins are meant.

A protein takes a unique three-dimensional structure during the *folding* process. This unique structure is called its *native state*. It is possible to change this structure by changing the environmental conditions like temperature or PH-value. However given the same environmental conditions afterwards again, a protein will always regain the same native state as before as long as the amino acid sequence stays unchanged.

The figures 2.3, 2.4 and 2.5 illustrate the difficulty in properly displaying this three dimensional structure as measured in experiment.<sup>5</sup> In picture 2.3 already a simplification has taken place since the atoms are displayed as hard spheres and not by the wavefunction of the electrons. However this picture still lacks the ability to properly emphasize the important structural attributes of a protein. In figure 2.4 only the *backbone* including the peptide-bonding is displayed which allows a good overview of the basic structure of this protein. Most commonly a presentational form as in figure 2.5 displays the backbone according to its *secondary structure*.

The knowledge about structure of a specific protein can be put in a hierarchy. Different common substructures arise in protein composition[124, 11]. This allows to characterize the structural information depending on the level of details included. This data is normally given for the native state of a protein.

- *Primary structure* is the information about the amino acid sequence of the protein alone. No structural information about the three-dimensional conformation of the protein is included. When numbering the amino acids one starts with the N-terminus of the protein and ends with the C-terminus.
- *Secondary structure* includes information about basic properties of the three-dimensional conformation. Each amino acid is classified as belonging to a specific secondary structure element, such as *helix* or  *$\beta$ -sheet*.

<sup>5</sup>All figures show different presentations of the experimentally determined native conformation of the Bacterial Ribosomal Protein L20 (pdb[12]-code: 1GYZ) which was ab-initio folded in-silico[108].

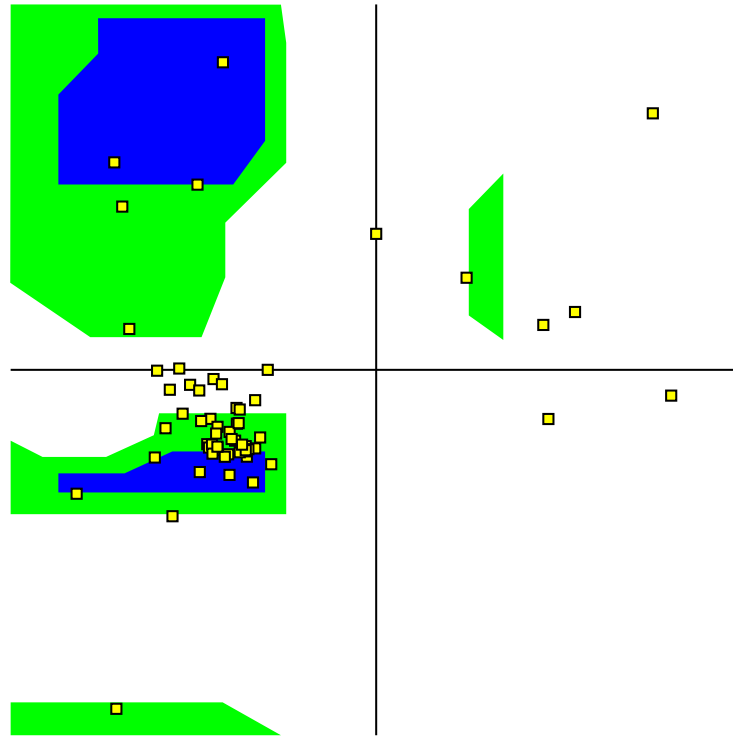


Abbildung 2.2: This picture shows a typical Ramachandran plot. The plotted dihedral angles ( $\Phi$  and  $\Psi$  on horizontal and vertical axis respectively ranging each from  $-\pi$  (left bottom corner) to  $\pi$  (upper right corner)) are taken from the native state of the Bacterial Ribosomal Protein L20 shown in figures 2.3, 2.4 and 2.5 and plotted as yellow dots. In the diagram the white areas correspond to conformations where atoms may clash according to calculations on polypeptide chains with hard spheres. These regions are therefore sterically problematic. The blue regions correspond to conformations where no steric clashes are possible and the green areas show the regions computed by using slightly shorter van-der-Waals radii in the calculations, i.e. the atoms are allowed to come a little closer to each other. The upper left region is associated with  $\beta$ -sheets, the other two regions with helices. The lower left region is related to right-handed  $\alpha$ -helices, the upper right region to left-handed ones.

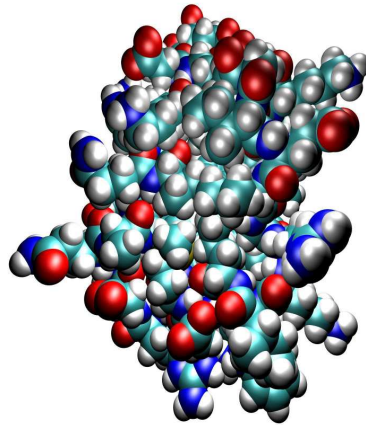


Abbildung 2.3: This picture shows the experimentally determined structure of the *Bacterial Ribosomal Protein L20* 1GYZ. The different atoms are colored according to their types (C light blue, N dark blue, H white, S yellow, O red) and have sizes according to their van-der-Waals-radii.

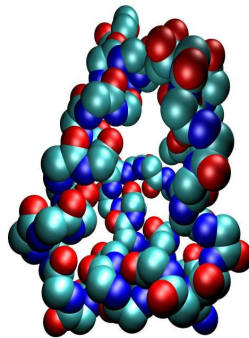


Abbildung 2.4: This picture shows the experimentally determined structure of the *Bacterial Ribosomal Protein L20* 1GYZ. The different atoms are colored according to their types (C light blue, N dark blue, H white, S yellow, O red) and have sizes according to their van-der-Waals-radii. Displayed are only the backbone atoms.

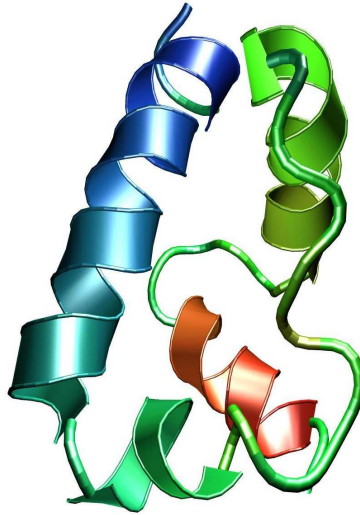


Abbildung 2.5: This picture shows the experimentally determined structure of the *Bacterial Ribosomal Protein L20* 1GYZ. The coloring goes from the N-terminus (blue) to the C-terminus (red). The protein is displayed according to its *secondary structure*. Helices are displayed as areas with broad ribbons, coil-regions with strands. This method is called *cartoon representation* and is the most common form of presentation for proteins.

- *Tertiary structure* includes complete spatial information about the full three-dimensional structure of one amino acid sequence.
- *Quarternary structure* becomes important when proteins consist of multiple amino acid sequences and describes how these different polypeptide units are organized three-dimensionally with regard to each other.

### 2.3.1 Primary Structure

Primary structure describes the sequence of amino acids starting from the N-terminus and ending with the C-terminus. The amino acids are normally given in one- or three-letter codes as presented in table 2.1. Since undamaged proteins can unfold by the change of environmental conditions and re-fold given the original conditions again all information needed for taking a unique native state is already encoded in this primary sequence. This means that in principle it *should be possible to predict higher order structure information from the primary structure, i.e. the amino acid sequence, alone*[3]. This prediction of tertiary structure is the main topic of this thesis.

### 2.3.2 Secondary Structure

Secondary structure describes the most often occurring elements of protein structure in a convenient way. Each amino acid is classified as belonging to one specific secondary structure

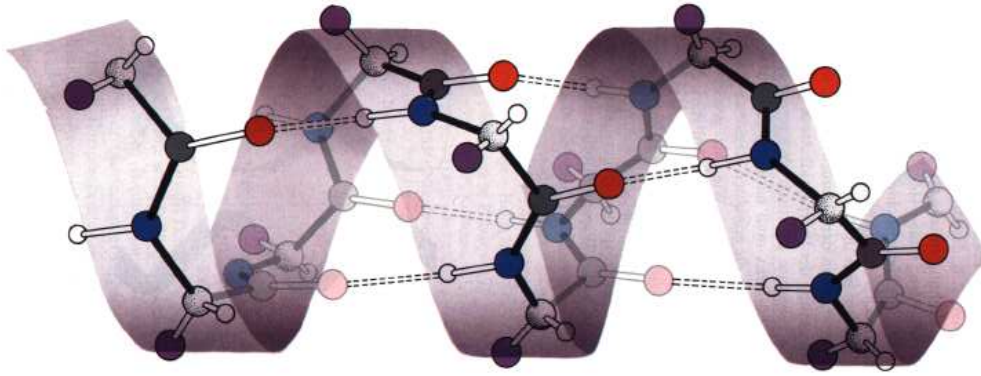


Abbildung 2.6: This picture shows the most common helix, the  $\alpha$ -helix. 3.6 amino acids are included per turn of the helix and one can see the hydrogen-bonding between the  $CO$  and the  $NH$  groups. [124]. As an overlay the cartoon presentation of a helix is indicated. The different atoms are colored according to their types (C black, N dark blue, H white, O red).

element. There are different possibilities describing it depending on how much one looks into detail. The program DSSP[61] assigns each amino as belonging to one of eight different states, including different kind of helices or bends. For us it is sufficient to describe secondary structure in a one letter code as belonging to either *Helix*, *Sheet* (also *E* for *Extended strand* is used) or *Coil*. This *three state description* will be used consistently in this thesis. Therefore we coarsen the DSSP classification to these three states (see appendix A).

### Helix

The helix is a very often occurring element of protein structure. It is stabilized by hydrogen bonding between the  $CO$ -group of amino acid  $A$  and the  $NH$ -group of amino acid  $A + N$  (typically is  $N=4$ ). Different kind of helices exist. They are defined by the number of amino acids per turn of the helix  $n$  and the number of atoms  $m$  integrated into the hydrogen bonding as  $n_m$ -helix. The most commonly occurring helical structure is the  $\alpha(3.6_{13})$ -helix (see figure 2.6). Others are the  $3_{10}$  or the  $\pi(4.4_{16})$ -helix (see figure 2.7).

### $\beta$ -Sheet

The  $\beta$ -sheet is another commonly occurring secondary structure element. It consists of multiple  $\beta$ -strands which are stabilized by hydrogen-bonding between the individual strands. There are two variants of  $\beta$ -sheets, the parallel and the antiparallel one whose strands run in according directions. The name *sheet* derives from its planar form (see figure 2.8). In proteins this sheet is often slightly drilled in itself.

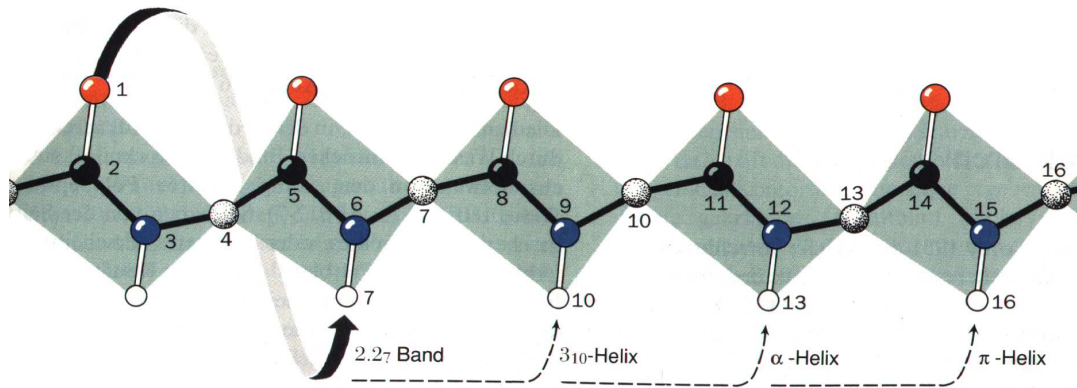


Abbildung 2.7: This picture shows the difference between the different types of helices. [124]. The arrow indicates possible hydrogen bonds. The different atoms are colored according to their types (C black, N dark blue, H white, O red).

One important difference is that helices are stabilized by local hydrogen bonding between amino acids close in sequence.  $\beta$ -sheets however are stabilized by long ranged hydrogen bonding between amino acids which can be far apart in sequence.

## Coil

Everything other than helix or  $\beta$ -sheet is referred to as coil. It is most often only poorly defined and more loose in three dimensional protein structure than helix or sheet.

### 2.3.3 Disulfide Bridges

When two side-chains of cysteine become spatially close to each other disulfide bridging can occur. This is bonding or bridging a single covalent bond between the oxidized sulfurs of the cysteine side-chains. Thereby the stability of the protein increases but a strong topological constraint is put on the protein at the same time. This is the only covalent bond that can be formed or broken during protein folding.

## 2.4 Protein Stability and Thermodynamic Hypothesis

Under physiological conditions most proteins assume their unique native state. Developing this structure under these conditions is called the folding of the protein. Precisely the native state is not a single state but a macro-state consisting of an ensemble of micro-states very



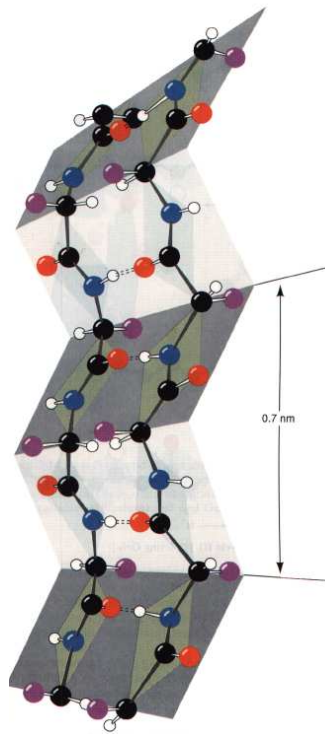


Abbildung 2.8:  $\beta$ -sheets are stabilized by hydrogen bonding as indicated in this schematics. Please note their common zip-zap appearance[124]. The different atoms are colored according to their types (C black, N dark blue, H white, O red, Side-chain violet).

similar to each other. However a protein is stable in its structure only under very narrowly defined circumstances[94, 95]. Slight changes in temperature or the PH-value of the solvent can force the protein to adapt a new structure under these changed conditions[7]. By returning to the original conditions the protein develops the same structure as previously again, including reforming possibly broken disulfide bridges, i.e. the protein re-folds.

Following the ideas of Anfinsen [3] the native state occupies the lowest free-energy for the system protein and solvent. Therefore *the folding of a protein is independent of the starting point. The environment has a strong effect on protein structure. The native state is in thermodynamic equilibrium for the system protein and environment. The amino acid sequence alone determines the native three-dimensional structure. This native configuration can be described as the lowest energy of its free-energy surface.*<sup>6</sup>

Protein stability can be understood as a gap in the inner energy between the native structure and other competing compact structures[71, 102, 18]. This gap must be big enough to overcome entropy since the native ensemble occupies a much smaller region in conformational space than the unfolded ensemble. The folding of a protein is believed to be the result of an equilibrium between two dominant forces, the hydrophobic interaction and loss of conformational entropy[7, 25, 26, 47]. It is believed that the *contact order* of a protein vastly determines the *folding speed*[63] of a protein. It is the mean sequence-distance between residues forming contact in the native structure. This can be rationalized by reminding that a low contact-order allows a protein to form native-like contacts earlier since the average distance between these residues is low. However proteins with a high contact-order need most of its native structure already be formed in order to have native contacts which slows folding down. However it is still subject of discussion whether the relative or absolute contact order (see appendix A, *Absolute contact order*) is more important[59, 63, 19].

These conclusions are also criticized because the correlation between the energy gap and folding ability is quite weak. Additionally this criterion concentrates on only a very small part of the energy surface, whereas the folding process could depend on complex characteristics of the entire energy landscape[17].

In an alternative approach to understand protein folding the energy landscape takes the form of a *folding funnel*[27, 47, 89, 90, 109]. This is a set of convergent pathways leading to the global minimum, i.e. the whole energy landscape is biased towards the native ensemble. These different pathways make the global minimum kinetically accessible from the ensemble of misfolded structures[67]. Both quantitative models and experimental data support this *new sight* on protein folding. The probing of hidden parts of the free energy surface could allow a microscopic theory of folding[40].

---

<sup>6</sup>This is the so called thermodynamic hypothesis.

# Kapitel 3

## Protein Structure Prediction

This chapter gives a short overview of methods for protein structure prediction, for which several approaches have been developed in the last years. First of all one has to decide what level of protein structure should be predicted. If it is sufficient to know the secondary structure computationally inexpensive methods can be used. To predict ab-initio the tertiary structure from the primary sequence much more computationally demanding methods are needed[118]. Further one has to distinguish between physical methods and knowledge based methods[36]. The first use physical approaches by modeling intra-molecular and inter-molecular interactions of the protein with its environment. The latter look for sequence similarities in existing databases and derive structure by comparing the composition of similar proteins and estimating from them a sensible prediction. Since the aim of this thesis is ab-initio structure prediction with physical forcefields, other methods are only peripherally covered.

### 3.1 Protein Structure and Evolution

Protein structure is directly correlated to their function in organisms. Especially the tertiary structure determines the surface of a protein by which it interacts with its environment and performs its function. During natural evolution mutations can change the amino acid sequence of proteins.<sup>1</sup> This can result in changes in tertiary structure which is related to the function of a protein. It seems reasonable that changes in structure can render an organism unfit in evolution since important tasks by mutated proteins can be done less (or sometimes more) effectively. Therefore two sequentially similar proteins often share a similar evolution and fulfill also similar tasks in an organism. Defining sequential similarity between proteins should also allow an estimation of their structural and functional similarity.

However the relation protein sequence to protein structure and function is not easy. Most of the time only a very limited set of amino acids determines the task of a protein. There are proteins which have almost identical tertiary structure but fulfill very different tasks in an organism. One example would be the Lysozyme/Lactalbumin superfamily of proteins. They

---

<sup>1</sup>The amino acid sequence of proteins built by living cells is encoded by the DNA.

have a very similar structure with an RMSD-B between Lysozyme (pdb-code[12]8LYZ) and  $\alpha$ -Lactalbumin (pdb-code 1ALC) of under 2 Å. Their sequence has a 35% sequence identity (determined with GAP[56]). But in spite of these similarities they fulfill different tasks in an organism. Thus though the simplifying saying *structure determines function* is true in most cases, there are exceptions like this one. *Backbone structural similarity does not necessarily imply functional similarity*. There are also rare cases in which sequentially and structurally dissimilar proteins fulfill similar tasks in an organism.

## 3.2 Homology Searching and Multiple Sequence Alignment

Searching for *homologues* is often a first step looking for similar folded proteins. It may help later structure predictions[118]. For doing so the primary sequences are compared against other sequences from databases with known tertiary structure. Different schemes exist for this task. Most work with a 20x20 matrix whose entries stand for the similarities between two given amino acids, i.e. amino acids with similar character give higher scores, those dissimilar lower ones. One problem that arises now is the dealing with gaps or insertions in the sequence. How familiar are two amino acid sequences of different length? For giving an answer to this question gaps and insertion must be included in the comparison. This is done by comparing all possible segments of sequence A with all possible segments of sequence B and taking the highest score. Thus the comparison of two sequences gives a score for their similarity.

*Multiple sequence alignment* means considering not only two but many structures in the alignment. This multiple sequence alignment results in a set of sequences which are similar to the given starting sequence. Due to the connection between sequence and structure of a protein this set with known structures serves as a good starting point for following structure predictions.

## 3.3 Prediction of Secondary Structure

Secondary structure prediction should be computationally inexpensive since each amino acid is dealt with only in a general and vague way. This prediction basically tries to assign each amino acid of a sequence a probable element of according secondary structure, like helix or  $\beta$ -sheet, without detailed knowledge of the exact location of the atoms or amino acid groups. Most methods use first homology searching and multiple sequence alignment to get a good starting point. The methods for the prediction of secondary structure can be divided into two broad fields of methods[118, 24]:

- Statistical methods
- Neural networks

Both methods require the amino acid sequence as an input. Each amino acid, a window of surrounding amino acids and other parameters like total number of amino acid in protein

and length of the amino acid sequence are variables to calculate probabilities for this amino acid being part of a specific secondary structure element. Overall neural network methods are more successful in accurate prediction but do not allow much insight into understanding the forces driving the folding of a protein. Statistical methods allow more insight but seem to be less accurate[82, 13, 65]. However, common to all these methods is, that helices are predicted with far better accuracy than  $\beta$ -sheets. This can be easily understood by knowing that helices are stabilized by local interactions (mostly hydrogen bonding) between amino acids close in sequences (*local* interaction). However  $\beta$ -sheets are stabilized by hydrogen bonding between amino acids often far apart in sequence (*non-local* interactions). These *non-local* interactions are much more difficult to predict.

### 3.4 Prediction of Tertiary Structure

The prediction of protein tertiary structure is much more complicated than the prediction of secondary structure since a position for each atom of a protein has to be calculated. Due to the dense packing and the lack of higher symmetry this poses a big challenge. In principle two different approaches can be made. On the one hand there is the *homology modeling* approach in which, like in secondary structure prediction, by homology searching existing databases of structures are investigated for likely similar structures. Afterwards these similar structures are weighted and allow by combination the prediction of a possible structure.

On the other hand there are the physical methods which try to use forcefields for either *molecular dynamics* or direct *structure prediction by minimization*. In the following chapters this last approach will be investigated in detail by presenting the forcefield *PPF01* and different successful minimization methods.



# Kapitel 4

## Forcefields for Protein Folding

This chapter deals with methods for protein modeling. Because a protein is normally not an isolated molecule but in interaction with a solvent these methods have not only to describe the protein itself but also its surrounding environment. During protein folding no new covalent bonds are formed (with the notable exception of disulfide bridges). The energy is determined by interactions between the atoms. An exact quantum mechanical calculation is not possible due to the size of the system. However a description by a *classical* forcefield seems viable.<sup>1</sup> Each point in conformational space is given an according energy or force. The solvent can now be treated *explicitly* or *implicitly*. In the case of explicit treatment of the solvent not only the positions of the atoms of the protein are taken into consideration but also many atoms of the solvent, which results in high computational demands for calculations. A simplification is the usage of an *implicit* solvent model which contains a term for the description of the solvent according to the positions of the *protein's atoms*. This approach will be discussed in detail in the next chapter where the forcefield PFF01 will be described.

### 4.1 Thermodynamics

The thermodynamic hypothesis from Anfinsen [3] postulates that the native structure of a protein corresponds to the global minimum of the free-energy. Therefore protein structure can be predicted by describing the free-energy under physiological conditions and identifying its global minimum.

Different forcefields use various approaches to achieve this aim. However certain common features can be found. Most describe all atoms as points in space and characterize them with classical attributes like mass, polarization and charge. It is important to note that *two forcefields might share the same functional form but use a different parameterization thus resulting in different energies*. So these two fragments of a forcefield, functional form and

---

<sup>1</sup>The term *forcefield* is applied both for functions describing potential energies and actual forces. This derives from the fact that originally forcefields were used to calculate driving forces in molecular dynamics simulations. Later the term was also applied to free energy potentials.

parameterization are of equal importance[38]. The comparison between different forcefields are empirical and speaking of applicable parameterization or functional form depends purely on the results.

#### 4.1.1 Interaction of chemically bonded atoms

Interactions between chemical bond atoms are described according to the number of subsequent covalent bonds involved. They are named 1-2, 1-3 or 1-4 interactions counting the number of the interacting atoms involved. Other interactions are taken as interactions of non-bonded atoms, even though they might be atoms of the same molecule like atoms participating in an inner-molecular hydrogen bond[38].

Figure 4.2 illustrates these chemically bonded interactions.

##### 1-2 interaction or bond stretch

The 1-2 interactions or bond stretch interactions are vibrations of the chemical bond. A good description can be reached by the *Morse function*

$$V_{1-2} = k_1(1 - e^{-A\Delta x})^2$$

$$A = \sqrt{\frac{k_2}{2D_e}}$$

This function requires the parameters  $D_e$  which is the depth of the potential energy minimum, the force constants  $k_n$  and the equilibrium bond length  $x_0$  (included in  $\Delta x = x - x_0$  above). This potential allows the dissociation of the bond but since this is computationally expensive and most forcefield do not allow this dissociation very often not the Morse function but a simple harmonic potential is used:

$$V_{1-2} = k\Delta x^2$$

(force constant  $k$ ,  $\Delta x$  indicates difference to equilibrium bond length). This reduces the needed parameters to two and lowers computational demands. Another possibility is to add higher order terms to the harmonic potential for an increase in accuracy:

$$V_{1-2} = \frac{k_1}{2}\Delta x^2(1 + k_2\Delta x + k_3\Delta x^2 + k_4\Delta x^3 + \dots)$$

The Morse potential and the harmonic potential are plotted in figure 4.1

##### 1-3 interaction or angle bend

Interactions of the 1-3 type or angle bends describe changes around the angle of bending  $\theta$ . Usually classical forcefields describe this also as harmonic, i.e.

$$V_{1-3} = k\Delta\theta^2$$



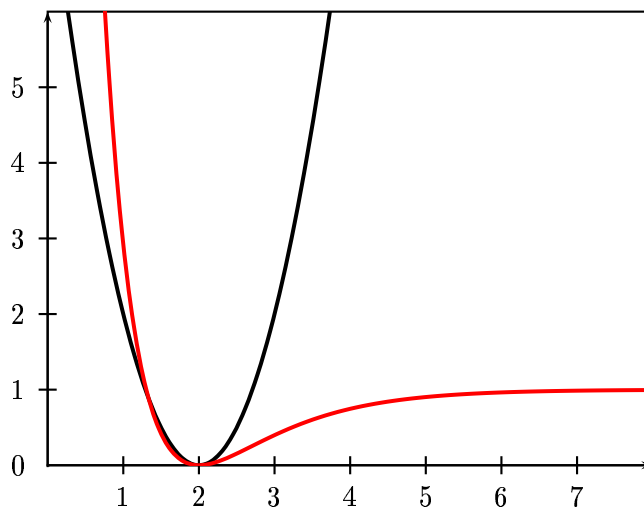


Abbildung 4.1: Comparison of the bond-stretch potentials. Plotted in black is a simple harmonic potential  $\sim \Delta x^2$  while in red the more sophisticated Morse potential  $\sim (1 - e^{-A\Delta x})^2$ , which also allows dissociation of bonds, is plotted. The units are arbitrary.

( $k$  force constant,  $\Delta\theta$  indicates difference to equilibrium angle of bending). Increasing accuracy can be done by including higher order contributions:

$$V_{1-3} = k_1\Delta\theta^2(1 + k_2\Delta\theta + k_3\Delta\theta^2 + k_4\Delta\theta^3 + \dots)$$

( $k_n$  force constants,  $\Delta\theta$  indicates difference to equilibrium angle of bending).

#### 1-4 interaction or torsional term

The third kind of chemical bonded interactions is the 1-4 type or torsional term. It describes rotation around the angle of the plane spanned by the atoms one to four. This potential is mostly described in the form:

$$V_{1-4} = \sum_n V_n (1 + \cos(n\phi + \gamma))$$

( $n$  gives the number of minima of the function during a full 360° rotation with the phase  $\gamma$  describing the exact point of the minimum,  $\phi$  is the angle of torsional bend and the  $V_n$  are often called barriers of rotation). A simple example to understand this potential would be rotations around the central C-C bond in X-C-C-X configurations like ethane  $H_3C - CH_3$ . The hydrogens can block each other (non-bonded interaction of hydrogens belonging to the two carbons) when the central bond is rotated. AMBER for example parameterizes this with  $n = 3$  and  $\gamma = 0$ .

It is worth noting that the torsional term has a close relationship with non-bonded 1-4 like interactions of the involved atoms. AMBER for example mixes the 1-4 bonded and non-bonded interactions by applying both a torsional term and scaling down the non-bonded Coulomb term.

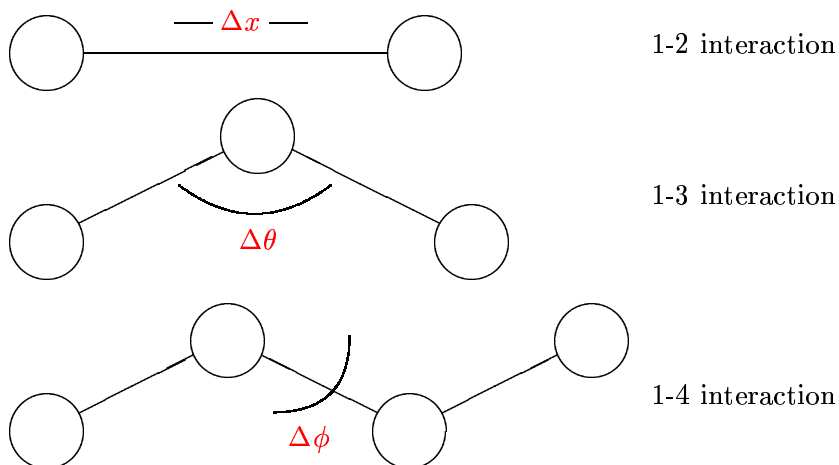


Abbildung 4.2: Illustration of the different types of bonded interactions. Please note that the length and angles  $\Delta x$ ,  $\Delta \theta$  and  $\Delta \phi$  all designate changes from the original reference configuration, i.e. equilibrium distance.

### Cross-Terms

To allow combination of two or more bonded interactions in a forcefield cross-terms can be introduced. One (simple) example would be the water molecule: reducing the angle  $\theta$  between the H-O-H bonds results in extension of the O-H bond-length and therefore a combined movement of 1-2 and 1-3 interactions. The above example could be described as

$$V_{CT} = k(\Delta x_1 + \Delta x_2)\Delta \theta$$

( $k$  force constant,  $\Delta x_n$  distance between  $O$  and  $H_n$ ,  $\theta$  angle H-O-H) or as a harmonic movement of the atoms 1 against 3

$$V = k\Delta r_{13}^2$$

( $k$  force constant,  $r_{13}$  is the spatial distance of the two atoms 1 and 3). Other combinations of bonded interactions in a forcefield are -of course- also possible.

#### 4.1.2 Interactions of non-bonded atoms

For sake of simplicity interactions of non-bonded atoms are usually understood as point-like interactions. Though the atoms involved in these interactions may be of the same molecule they are understood as non-linked and separated.

### Van-der-Waals-Interaction, Lennard-Jones-Potential

Fluctuations in the charge distribution of atoms can result in temporary dipoles. This is a deviation from the ideal behavior and leads to dipole-dipole or dipole-induced-dipole interactions. These forces are called Van-der-Waals forces and can be formulated as a Lennard-Jones-6-12-potential:

$$V_{ij}^{LJ-6-12} = V_0 \left[ \left( \frac{\tau}{r_{ij}} \right)^{12} - 2 \left( \frac{\tau}{r_{ij}} \right)^6 \right]$$

( $V_{ij}$  give the strength of the interaction between atoms  $i$  and  $j$ ,  $V_0$  describes the depth of the potential well,  $\tau$  is the equilibrium distance and  $r_{ij}$  gives the spatial distance of atoms  $i$  and  $j$ ). The energy function is a combination of an attractive term resulting from the interaction of induced dipoles by the power of 6 and a repulsive term resulting from the repulsion of the electric hills according to Pauli's principle by a power of 12. This formulation is only qualitatively correct. The power of 12 can be quickly calculated by squaring the first term. It is too large compared to experimental results. Often this error is corrected by a changed parameterization of other terms.

In order to increase accuracy some forcefields use the Buckingham-potential with a replaced repulsive term:

$$V_{ij}^{Buckingham} = V_0 \left[ \frac{6}{\alpha - 6} e^{-\alpha \left( \frac{r_{ij}}{\tau} - 1 \right)} - \frac{\alpha}{\alpha - 6} \left( \frac{\tau}{r_{ij}} \right)^6 \right]$$

( $V_{ij}$  gives the strength of the interaction between atoms  $i$  and  $j$ ,  $V_0$  describes the depth of the potential well,  $\tau$  is the equilibrium distance and  $r_{ij}$  gives the spatial distance of atoms  $i$  and  $j$ ). The new parameter  $\alpha$  can be adjusted and gives a Lennard-Jones kind potential for values between 14 and 15. Although this potential is more accurate than Lennard-Jones it has the disadvantage of becoming attractive for very small values of  $r_{ij}$  which can be overcome by correction terms.

### Coulomb interaction

The interaction  $V_{ij}$  of two ions or charges  $q_i$  and  $q_j$  which are separated by a distance of  $r_{ij}$  in a medium with the dielectric constant  $\epsilon_r$  is according to Coulomb's law

$$V_{ij}^{Coulomb} = \frac{q_i q_j}{4\pi \epsilon_0 \epsilon_r r_{ij}}$$

( $\epsilon_0$  is the dielectric constant of vacuum). One very difficult parameter is  $\epsilon_r$  which describes the effect of the medium. In the case of proteins the medium is often *not* homogeneous. For example some water molecules might be trapped inside a protein which differ strongly in their dielectric constant ( $\approx 80$ ) from the rest of the protein ( $\approx 2 - 4$ ). A distance depending dielectric constant might be used as well as taking approximate values for the interior of proteins[48].

The above Coulomb-energy has to be calculated for each pair of atoms which means the computational costs are  $O(N^2)$ . Another approach would be using the multi-pole expansion,

based on electric moments of multi-poles like charges, dipoles, quadrupoles etc. This technique, sometimes called *fast multi-pole expansion* avoids some computationally demanding calculations. Recent methods use fast Fourier transformations on multi-poles (FFTM) which have a computational complexity of  $O(N^a)$ , where  $a$  ranges from 1.0 to 1.3 [88].

### Hydrogen bonding

A hydrogen bond appears when a donor (hydrogen) is bonded to a strong electronegative partner like oxygen in water or nitrogen in the backbone of an amino acid. This positive polarized hydrogen can interact with a negative polarized partner like oxygen. Most often this is described as dipole-dipole interaction. This interaction is very important for protein structure since it stabilizes secondary structure like  $\beta$ -sheet or  $\alpha$ -helix. The inclusion of hydrogen bonding has been done in very different forms. Three examples, HB-1, HB-2 and HB-3, are given below but a good functional form to include hydrogen bonding is still subject of scientific discussions[41, 117].

$$V_{ij}^{HB-1} = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r}$$

$$V_{ij}^{HB-2} = \frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{10}}$$

$$V_{ij}^{HB-3} = \cos(\theta) \left( \frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^6} \right) + (1 - \cos(\theta)) \left( \frac{C}{r_{ij}^{12}} + \frac{D}{r_{ij}^6} \right)$$

( $V_{ij}^{HB-x}$  gives the strength of the interaction between atoms  $i$  and  $j$ ,  $q_i$  gives the charge of atom  $i$ ,  $\epsilon_0$  and  $\epsilon_r$  are the dielectric constants of vacuum and medium,  $A$ ,  $B$ ,  $C$  and  $D$  are parameters determining the strength of hydrogen bonding,  $r_{ij}$  is the spatial distance between atoms  $i$  and  $j$  and  $\theta$  the angle of the hydrogen bond).

### Solvent term

The inclusion of an appropriate solvent term is very difficult[22, 91]. One has to make sure that the electrostatic both inner- and inter-molecular interactions are treated in a proper way, which generates problems with the dielectric constant  $\epsilon_r$ . A solution is including dielectric functions or average values for  $\epsilon_r$ .

Also one has to chose between including the solvent in an explicit or implicit way. In the first case the computational demand rises strongly since not only the protein but also the water molecules must be simulated. Depending on the actual implementation this can raise the treated number of atoms to a tenfold compared to the protein alone. Less computationally expensive but also less accurate are implicit solvent models. Following the work of Eisenberg and McLachlan [31] one assumes that the contribution of an atom to the solvent energy is proportional to its solvent-exposed surface:

$$V^{ImplicitSolvent} = \sum_{Atoms} \sigma_{T_i} A_i$$

( $\sigma_i$  are parameters in  $(\text{kcal/mol}) \cdot \text{\AA}^{-2}$  which give the energy contributions per solvent accessible surface of each atom with respect to the type  $T_i$  the atom belongs to (please note that the type of an atom can differ if belonging to different amino acids),  $A_i$  is solvent accessible surface area for each atom  $i$ ). This approximation can be considered as standard for implicit solvent interactions. Comparing the accuracy of this approach with ones from explicit solvent interactions [115] shows that applying implicit solvent interaction offer an excellent tradeoff. Computational demands are strongly lowered while accuracy stays almost the same.

### 4.1.3 Molecular Dynamics

Molecular dynamical simulations were introduced by Alder and Wainwright [1]. Assuming that all occurring forces are treated in an approximate way the forcefield allows the propagation of a protein structure and its movement in time under classical theory. As starting point a configuration of the protein is used (often the experimentally determined structure geometry-optimized in the forcefield) and the velocities are randomized assuming a Maxwell distribution. In all following equations  $x$  represents the atomic coordinates of the system.

The simplest way to do Molecular Dynamics is solving Newton's equation of motion  $F = ma = m\ddot{x}$  for this system

$$\frac{\partial V(x_i)}{\partial x_i} = F_i = m_i a_i = m_i \frac{\partial^2 x_i}{\partial t^2}$$

( $V$  is the potential of the according forcefield,  $m_i$  the atomic mass of atom  $i$ ,  $a_i$  the acceleration of atom  $i$  and  $t$  the time). This approach is used for simulations with explicit water treatment. Another approach is using the *Langevin equation*, a stochastic differential equation, for all atomic coordinates[87].

$$M\ddot{x} + C\dot{x} + \nabla V(x) = D\dot{W}$$

( $M$  matrix of the system with atomic masses on the diagonal,  $C$  damping matrix of the system,  $V$  potential of the according forcefield,  $D$  random force of the system,  $\dot{W}$  uniform random noise (commonly white noise)). The diffusional matrix  $D$  and the damping matrix of the system  $C$  have the following relation:

$$DD^T = 2k_B TC$$

( $T$  the Temperature of the system,  $k_B$  the Boltzmann constant). For low temperatures and no damping this equation simplifies into Newton's equation above. Typically one uses a scalar damping constant  $\gamma$  and  $C = \gamma M$ .

By now typical molecular dynamics simulations are limited to small proteins of 20-40 amino acids for a short period of time. They have difficulties in stabilizing the native structure of the protein even when using high amounts of computational time[30] when not applying parameters specifically adjusted to one protein.

#### 4.1.4 Established forcefields

This short paragraph tries to describe the basic properties of different forcefields. Commercial available packages usually include them. Since our focus is not on the actual implementation details of these packages will not be presented. A recent good overview with a detailed look is given in [72].

Many forcefields use the following composition of the energy function following the ideas given in the above sections. These forcefields are called *class I* force fields.

$$\begin{aligned}
 V(\vec{r}) = & \sum_{bonds} K_b(x - x_0)^2 + \sum_{angles} K_a(\theta - \theta_0)^2 + \sum_{torsions} V_t(1 + \cos(n\phi + \gamma)) \\
 & + \sum_{cross-term} K_{ct}(Y - Y_0)^2 + \sum_{atoms} V_{ij} \left[ \left( \frac{\tau}{r_{ij}} \right)^{12} - 2 \left( \frac{\tau}{r_{ij}} \right)^6 \right] \\
 & + \sum_{atoms} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}}
 \end{aligned}$$

( $\vec{r}$  is a configuration of the protein,  $K_x$  give different parametric constants,  $x$  is the bond length,  $\theta$  the angle valence angle,  $\phi$  the torsion angle,  $\gamma$  phase angle,  $Y$  are cross-terms,  $V_{ij}$  the constants for the Lennard-Jones potential,  $\tau_{ij}$  the minimum interaction radii, the  $r_{ij}$  the spatial distance between atoms  $i$  and  $j$ ,  $q_i$  the charge of atom  $i$ ,  $\epsilon_0$  the dielectric constant in vacuum,  $\epsilon_r$  the dielectric constant of the medium). This basic functional composition is common to many force fields including AMBER[21], CHARMM [73], GROMOS [122] or OPLS[60]. Forcefields which also use higher order functions than in the above equation like the Morse potential are referred to as *class II* force fields. These terms allow an increase in accuracy while having higher computational demands.

The following list gives a quick overview over the most commonly used forcefields. The given functional forms may change slightly between different versions. For example earlier versions of AMBER included hydrogen bonding in form of a 10-12-potential while later ones included recalculated parameters for the forcefield which allowed inclusion of hydrogen bonding in the Lennard-Jones potential.

The newly developed forcefield PFF01 [50, 53, 48] which was used for the simulations in this thesis will be subject of the next chapter and discussed in detail there.

- AMBER (Assisted Model Building with Energy Refinement)[92, 93]

<http://amber.scripps.edu>

This forcefield was developed in the group of Peter Kollmann at UCSF. Different sets for parameters exist. Its functional form is equivalent to the above formulation (Form. 4.1. Older version use a 10-12 potential for the hydrogen bonds:

$$V_{AMBER-Hydrogen} = \sum_{H-bonds} \left[ \frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right]$$

( $C$  and  $D$  are parameters,  $r_{ij}$  the spatial distance between atoms  $i$  and  $j$ ), while newer versions included hydrogen bonding in Lennard-Jones and electrostatics.

- CHARMM (Chemistry at HARvard Molecular Mechanics)[73]

<http://yuri.harvard.edu/>

CHARMM was developed, as the name suggests, in Harvard University and exists as both academic (CHARMM) and commercial (CHARMm) version. All kinds of different parameter sets exist. The development was started and is still run by Martin Karplus. CHARMM uses the same basic composition as above (Form. 4.1). A Urey-Bradley (cross-term)

$$V_{UB} = K_{UB}(r - r_0)^2$$

describes movements of the atoms 1 and 3 against each other ( $r$  spatial distance,  $r_0$  equilibrium distance).

- CVFF/CFF (consistent *valence* forcefield)[74, 75]

<http://struktur.kemi.dtu.dk/cff/cffhome.html>

CFF was developed in the 1960's and is one of the oldest forcefields. The basic functional form of CFF is the same as above. Something special about this forcefield is the fact that it is *not empirical*. The authors refer to it as a *quantum mechanical forcefield* since observables calculated by in-initio calculations determine the energy hypersurface[74]. These values are slightly scaled afterwards since force constants derived from Hartree-Fock calculations have the drawback of being too large while bond lengths are underestimated.

CFF is especially known for a good description of spectroscopic properties. It also includes non-harmonic contributions up to the quadric term to increase accuracy at the cost of higher computational demands.

- GROMOS (Groningen Molecular Simulation)[121]

<http://www.igc.ethz.ch/gromos>

GROMOS is a package for molecular dynamics simulations. It was developed in Groningen in the Netherlands by Wilfred van Gunsteren and Hermann Berendsen. The recent version is the Gromos 43A1 force field[110]. The form of the potential is equivalent to the above general formulation. A special emphasis is put on the cross-terms. This version has been especially optimized for proteins.

- ECEPP (Empirical Conformational Energy Program for Peptides)[86, 85, 32]

This force field differs from other force fields by lowering the degrees of freedom of a protein. It fixes bond length and some angles inside a protein. Doing so reduces the computational effort for calculating energies with the forcefield and minimizing its energy. It is similar in functional form but dissimilar in parameterization from PFF01 with an additional torsional term:

$$E_{\text{tors}} = \sum_i U_i (1 \pm \cos(k_n \phi_n))$$

( $U_i$  torsional barriers of rotation,  $k_n$  multiplicity of the torsion angle  $\phi_n$ ).

### 4.1.5 Discussion on the Transferability of individual Forcefield Terms

The transferability of individual forcefield terms is very difficult. Especially in free-energy forcefields directly transferring single energy terms is not possible: *It is in general not possible to simply add or exchange a energy term from another force field to improve accuracy.* To understand this behavior one starts with the idea of calculating the inner energy and adds a new Hamiltonian to the existing one, as done in [48]. The results and major steps are discussed below:

$$U = \langle H \rangle = \langle H_1 \rangle + \langle H_2 \rangle$$

( $U$  inner energy,  $H$  Hamiltonians for inner energy,  $\langle X \rangle$  thermal expectation values for property  $X$  at a specific temperature). The free energy can be calculated:

$$F = \frac{\ln \langle e^{+\beta H} \rangle}{\beta}$$

( $F$  free energy,  $\beta = \frac{1}{k_B T}$  with  $T$  the temperature and  $k_B$  the Boltzmann constant) By including also the entropy in the form

$$S = -\frac{\partial F}{\partial T} = \frac{\langle H \rangle}{T} + \beta^{-1} \int \int d\vec{p}d\vec{q} e^{-\beta H}$$

( $S$  is the entropy) one can expand the free energy by approximating the exponential and logarithmic function. The free energy then becomes:

$$F = \langle H \rangle + \frac{1}{2}\beta(\langle H^2 \rangle - \langle H \rangle^2) + O(\beta^2) = U - TS.$$

Thus

$$\begin{aligned} F &= F_1 + F_2 + \beta(\langle H_1 H_2 \rangle - \langle H_1 \rangle \langle H_2 \rangle) + O(\beta^2) \\ &= F_1 + F_2 - TS_{12} = U_1 + U_2 - TS, S = S_1 + S_2 + S_{12} \\ &\neq F_1 + F_2. \end{aligned}$$

One can see that the total free energy  $F$  consists of contributions from the individual free energies  $F_1$  and  $F_2$  and another contribution from the entropy of the system. However separation of the entropic contribution is difficult and depending on the possibility of decoupling the system, for example the distinction between bond and non-bond interactions[72].

A simple exchange of parameters between two force fields is therefore in general not possible. The interaction with the solvent has contributions from Lennard-Jones, electrostatics and hydrophobic effects. However it is, for example, possible to exchange the functional form of terms by re-parameterization like done between different versions of the AMBER force field. The hydrogen bonding can be described by both a 10-12 potential or, using a re-parameterized set of parameters, by a combination of the 6-12 potential and the electrostatic interaction.



# Kapitel 5

## Protein Force Field 01, PFF01

The forcefield PFF01 (Protein Force Field 01) is a refinement from prior works by the group of Moulton at CARB. It has been published in [50, 53, 48] where the formulas and parameters of this paragraph have been taken from. It is an all-atom free-energy force field and <sup>1</sup> designed to make predictions of protein structure by finding the lowest free-energy for a given protein following the thermodynamic hypothesis [3]. This global optimization neglects the actual folding process by using efficient optimization methods. These methods are discussed in detail in the following chapters.

The force-field is modeling physical interactions in a protein and between a protein and water at a fixed temperature of 300K. The contributions to the forcefield are

- Electrostatics
- Hydrogen bonding
- Lennard-Jones-6-12
- Surface-depending implicit solvent model

No vibrational terms are included. The bond-length and the peptide planes are kept fixed. The only degrees of freedom are the dihedral angles of both backbone and side-chains of the protein. Therefore the spatial coordinates  $\{\vec{r}\}$  can be directly translated into dihedral angles  $\{\vec{\theta}\}$  and vice versa. Both are equivalent descriptions of a specific conformation of a protein. The parameters for these force field terms derive from a family of proteins. This family of proteins was selected to represent a wide spread of different protein structures. However it has proven necessary to further optimize these parameters. This has been done on the Villin headpiece, a protein intensively investigated by different groups using AMBER [30] and ECEPP/2 [45]. Using these optimized parameters other non-homologous helical proteins could be successfully folded without further optimization of the parameters, meaning that

---

<sup>1</sup>Apolar groups of the type  $CH_N$  are modeled as big super atoms. Modeling these hydrogens explicitly would increase computational demands without increasing accuracy of the forcefield. All other atoms are modeled explicitly.

from random initial conditions the structure corresponding to the lowest found energy is equivalent to the one found in experiment. By now this is the first force-field based on physical interactions able to stabilize proteins from different sizes (up to 60 amino acids[108]) and families of proteins (no homologue sequence identity) in their native states.

## 5.1 Interactions of the Forcefield

As said above PFF01 has contributions from different interactions. These interactions must meet a balance between computational demands and accuracy in modeling the free-energy surface. One important simplification was the exclusion of explicitly modeling the apolar hydrogens in  $CH_N$  groups. The degrees of freedom for the forcefield are the dihedral angles of the backbone and the side-chain. In the following  $\vec{r}$  means a configuration of the given protein.

### 5.1.1 Potential Types

The atoms in the force field are classified according to their chemical characteristics. These potential types are used for the values of the different force field parameters as in table 5.1.

#### Lennard-Jones

The Lennard-Jones interaction is included as a 6-12 potential.

$$V_{LJ}(\vec{r}) = V_0 \sum_{ij} \left[ \left( \frac{R_{ij}}{r_{ij}} \right)^{12} - \left( \frac{2R_{ij}}{r_{ij}} \right)^6 \right]$$

(here  $i, j$  represent the atoms included in the force field,  $r_{ij}$  is the distance between these atoms,  $R_{ij}$  are the Lennard-Jones radii ( $R_{ij} = \sqrt{R_{ii}R_{jj}}$ ). The parameters for the Lennard-Jones potential derive from a potential of mean application to experimental data. By fitting short-range (2Å - 5 Å) radial distributions of a set of 138 different proteins<sup>2</sup> we got as result the radii given in table 5.2.

In our simulations the attractive part of the Lennard-Jones potential plays a very minor role. Much more important is the repulsive part which prohibits clashes of atoms according to the Pauli-principle.

---

<sup>2</sup>These proteins are believed to represent a wide span of different folds[6].

Amino acid	Potential type
ALA	CME
ILE	4xCME
LEU	4xCME
MET	CME CME S CME
PHE	CME 6x CR
PRO	3x CME
TRP	CME 3xCR N <sub>1</sub> H 5xCR
VAL	3xCME
ASN	CME CP O <sub>2</sub> N <sub>2</sub> H H
CYS	CME S
GLN	CME 3xCR N <sub>1</sub> H 5xCR
SER	CP O <sub>1</sub> H
THR	CP CME O <sub>1</sub> H
TYR	CME 6xCR O <sub>1</sub> H
ASP	CME CP O <sub>2</sub> O <sub>2</sub>
ARG	3xCME N <sub>1</sub> H CP 2x(N <sub>1</sub> H H)
GLU	CME CME CP O <sub>2</sub> O <sub>2</sub>
HIS	CME CR N <sub>1</sub> H <sub>2</sub> CR CR N <sub>1</sub> H
LYS	3xCME CP N <sub>3</sub> 3xH
Main Chain	N <sub>1</sub> HM CME CP O <sub>1</sub>
N-terminus	N <sub>3</sub> H H H CME CP O <sub>1</sub>
C-terminus	N <sub>1</sub> HM CME CP O <sub>1</sub> O <sub>2</sub>

Tabelle 5.1: List of the different potential types according to the amino acids. The list is starting from the  $C_{\beta}$  atom outwards.

Potential type	$R_{ii}$	$\sigma_i$
CME	4.10	84
CP	4.10	-6
CR	3.28	93
N <sub>1</sub>	3.55	-30
N <sub>2</sub>	3.55	15
N <sub>3</sub>	3.55	-45
O <sub>1</sub>	3.10	-30
O <sub>2</sub>	3.10	15
S	3.80	84
H	1.95	according to bound partner
HM	2.25	according to bound partner

Tabelle 5.2: This table gives the values for the Lennard-Jones Radii in Å and the solvation enthalpies in kcal/(mol Å<sup>2</sup>).

g	1	2	3	4	5	6
1	0.375731	0.375731	0	0.143396	0.143396	0.043222
2		0.375731	0.161852	0.143396	0.143396	0.031012
3			0	0	0.161852	0.045452
4				0.143396	0.143396	0.043222
5					0.143396	0.031012
6						0.013097

Tabelle 5.3: Parameters for the inverse group-specific di-electrical constants  $\epsilon_{g(i)g(j)}^{-1} = \epsilon_{g(j)g(i)}^{-1}$ .

## Electrostatics

The electrostatics can be divided into contributions from the main- and the side-chain.

$$V_{ele}(\vec{r}) = V_{main}(\vec{r}) + V_{side}(\vec{r}) = \sum_{ij} \frac{q_i q_j}{\epsilon_0 \epsilon_{g(i)g(j)} r_{ij}}$$

They are included in a standard way with group-specific dielectric constants (here  $i, j$  represent the atoms included in the force field,  $q_i$  and  $q_j$  are the according partial charges,  $r_{ij}$  is the distance between these atoms,  $\epsilon_0$  is the dielectric constant,  $\epsilon_{g(i)g(j)}$  are group-specific dielectric constants). The group specific dielectric constants are given according to different types of electrostatic interaction. This represents the characteristics of the atoms as being part of different amino acids and takes their specific partial charges, orientation or accessibility to the solvent into account. This is a crude approximation to the real situation, as only the interacting amino acids and not the complete environment is taken into consideration. The parameters for  $g(i)$  and  $g(j)$  are given in table 5.4, the parameters for  $\epsilon_{g(i)g(j)} = \epsilon_{g(j)g(i)}$  are given in table 5.3. This parameterization excludes some parts or even complete side-chains (like PHE, GLY, MET, PRO) from contributions to the electrostatics.

The parameters for  $g(i) = 1, 2$  are used to describe the hydrogen bonding for the main chain as dipole-dipole interaction. They are the biggest contribution from electrostatics.  $g(i) = 3, 4, 5$  describe interactions of the partially charged  $OH$ ,  $CO$  and  $NH_2$  groups of the (ASN, GLN, SER, THR, TRP)-side-chains, which are smaller in their contributions. The interaction of the charged  $COO^-$  and  $NH_x^{(+)}$  of (ASP, GLU, ARG, LYS, HIS, TRP) are the smallest contributions to the electrostatic interaction.

The electrostatics of the side-chains contribute only in minor quantities to the total free energy of the protein.

Amino acid	atoms	potential	g
Main chain	N	n1	1
	HN	hn	1
	C	co	2
	CO	o1	2
ASN	CG	cp	5
ASN	OD1	o2	5
ASN	ND2	n2	4
ASN	HNA, HNB	h	4
ASP	CB	cme	6
ASP	CG	cp	6
ASP	OD1, OD2	o2	6
GLN	CD	cp	5
GLN	OE1	o2	5
GLN	NE2	n2	4
GLN	HNA, HNB	h	4
GLU	C,CDG	cme	6
GLU	OE1, OE2	o2	6
SER	CB	cme	3
SER	OG	o1	3
SER	HOG	h	3
THR	CB	cme	3
THR	OG1	o1	3
THR	HOG	h	3
TYR	CZ	cr	3
TYR	OH	o1	3
TYR	HOH	h	3
ARG	CD	cme	6
ARG	NE	n1	6
ARG	HNE, HHA, HHB, HHC, HHD	h	6
ARG	CZ	cp	6
ARG	NH1, NH2	n1	6
LYS	CD	cme	6
LYS	CE	cp	6
LYS	NZ	n3	6
LYS	HZA, HZB, HZC	h	6
HIS	CB	cme	6
HIS	CG, CD2, CE1	cr	6
HIS	ND1, NE2	n1	6
HIS	HD1, HE2	h	6
TRP	NE1	n1	6
TRP	HNE	h	6

Table 5.4: The parameters for  $g$  according to the atoms of the different amino acids. Please note that for all other atoms not listed above  $g = 0$ .

## Hydrogen Bonding

Though experimental measurements of the effects of hydrogen bonding on protein folding vary with between  $-2.8\text{kcal/mol}$  to  $+1.9\text{kcal/mol}$  strongly [5, 77] it is generally considered a vital contribution to protein folding[11, 124]. It is especially important for the formation of secondary structure in proteins. Hydrogen bonding can be modeled as part electrostatics and part Lennard-Jones as done in some versions of CHARMM or AMBER. However in PFF01 hydrogen bonding and solvent interaction are considered the two major contributions to protein folding. Therefore special emphasis is placed on also including some quantum-mechanical effects not modeled by pure electrostatics in this classical force field.

When taking only the dipole-dipole interaction of the amino- and carboxyl-groups of the main-chain, long-range interaction are overemphasized due to cooperative effects

$$V_{hydrogen-ij-dipole} = \frac{0.38 \cdot 0.28e^2}{4\pi\epsilon\epsilon_0} \left( \frac{1}{r_{C_iH_j}} - \frac{1}{r_{C_iN_j}} - \frac{1}{r_{O_iH_j}} + \frac{1}{r_{O_iN_j}} \right)$$

( $i, j$  counts the amino acids with  $i$  possessing the carboxyl- and  $j$  the amino-group,  $e$  equals one elementary charge,  $r_{X_iY_j}$  gives the distance of the atoms  $X$  from amino acid  $i$  and  $Y$  from amino acid  $j$ ). Since this cooperative effect gets stronger for longer helices, difficulties when including this equation alone exist. Therefore an additional short-ranged corrective term for hydrogen bonding was included. It takes the alignment of the hydrogen bond with respect to the donor and acceptor groups into account[114]:

$$V_{hb} = \lambda V_{hydrogen-ij-dipole} + (1 - \lambda) V_{corr}$$

( $\lambda$  gives the strength of correction between [0..1] with  $\lambda = 1$  meaning that the hydrogen bonding is modeled by pure dipole-dipole interaction. PFF01 uses as optimal value  $\lambda = 0.75$ ).

$$V_{corr} = V_0 \sum_{ij} R(r_{H_iO_j}) \Lambda(\alpha_{ij}, \beta_{ij})$$

( $V_0 = -2.12 \text{ kcal}/(\text{mol } \text{Å})$ ),  $\alpha$  is the NHO angle,  $\beta$  the angle between the CO and NH-dipoles,  $R(r)$  gives the radial and  $\Lambda(\alpha)$  the angular depending part of the correction potential).

$$R(r) = s_{2.4,0.075}(r)$$

$$s_{A,B}(x) = \frac{1}{2} \left[ 1 - \tanh \left( \frac{x - A}{B} \right) \right]$$

$$\Lambda(\alpha, \beta) = s_{45,5}(\alpha) s_{40,5}(\beta) s_{1.5,0.05} \left( \sqrt{\frac{\alpha^2}{30} + \frac{\beta^2}{24}} \right)^2$$

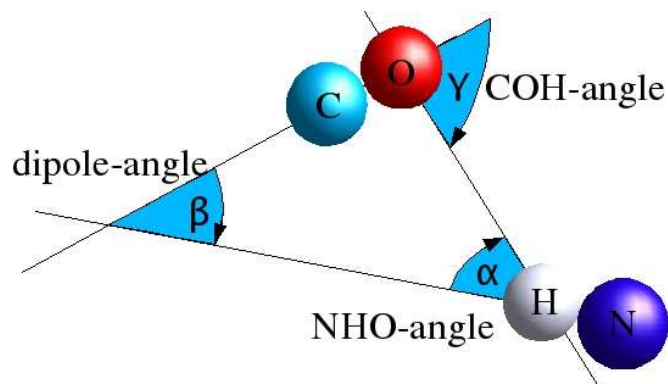


Abbildung 5.1: Definition of the angles  $\alpha$ ,  $\beta$ ,  $\gamma$  occurring in hydrogen bonding.

### Solvation effect

Since PFF01 is a forcefield describing the free energy of a given protein the inclusion of the effect of the solvent occurs in form of an *implicit solvent model*. Implicit solvent model means that the movement of the individual water molecules is not simulated. Instead the movement of the water molecules and the resulting hydrogen bonding between water molecules and protein is included in an *averaged* way. Thus the contribution of the entropy of the protein and of the dynamic system water/protein is slightly less accurate than in other force fields including explicit solvent. However this simplification saves large amounts of computational time in the simulation of the protein.

In order to estimate the contribution of solving the protein in water we use the idea that each atom has different physical/chemical properties and interacts with water mainly by its surface. On the surface two different kind of interactions are important:

- hydrophobicity, entropy of the water molecules
- entropic contributions from configurational entropy of the protein, esp. from the side-chains

The first part is easily understood as contributions from the solvent. The later however demands some thought. On the surface of the protein side-chains are less limited in movement than in the inner part with its dense packing. Therefore moving a side-chain from the surface to the inner parts of a protein gives a loss in configurational entropy. Since the main-chain is much more limited in movement its contribution to configurational entropy is minor.

In experiments the transfer-energy for bringing a peptide from the apolar octanole to water has been measured. Octanole represents the inner of a protein. Thus the transfer energy from octanole to water should represent bringing an inner amino acid to the surface of the protein. Following the thoughts of Eisenberg and McLachlan which are widely used in biophysics [31] two main ideas arise:

- transfer-energy of each atom is proportional to its surface exposed to water

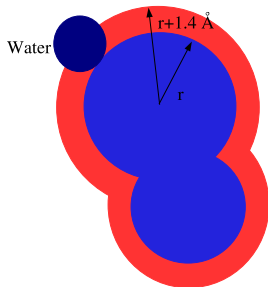


Abbildung 5.2: Schematics representing the calculations of the SASA-surface of a protein by rolling a water-sphere of 1.4 Å radius over two atoms. Each point on the surface belongs to the closest atom and contributes to its SASA-surface.

- transfer-energy of an amino acid is the sum of the transfer energies of the individual atoms

In PFF01 we first calculate the *Solvent Accessible Surface Area* (SASA) of the protein[66]. Water is treated as a sphere with the radius 1.4 Å. This sphere is rolled over the protein surface which is defined by the Lennard-Jones-radii. The area is defined by the the position of the middle of the water-spheres. Each point of the surface belongs to the nearest atom of the protein. The whole process is illustrated in Fig. 5.2.

The SASA's of the individual atoms allow a summation of the transfer-energies:

$$\Delta F = \sum_i \sigma_{PT(i)} A(i)$$

( $i$  counts all atoms,  $PT(i)$  is the potential type of atom  $i$ ,  $\sigma_{PT(i)}$  gives the Atomic Solvent Parameter (ASP) according to the potential types,  $A$  gives the SASA of the atom  $i$ ). The parameters  $\sigma$  are calculated using the above equation using data from experiment. These data are the transfer energies from tripeptides in the form Gly-X-Gly from water to n-octanole[33]. In [33] the above thoughts about dividing the solvent contribution into hydrophobic effect and configurational effect are elaborated. N-octanole is large than water and limits the movements of the side-chains much stronger. Therefore a correction has to be made since the configurational contribution in n-octanole can be estimated to be effectively zero. Later works include further corrections for volume of the different solvents and hydrophobicity of different proteins



[111, 16]. During the development of PFF01 the Lennard-Jones parameters have been further optimized, it was necessary to recalculate new solvation parameters[48]. These are given in table 5.2. Since these parameters are measured at 300K in experiment *our implicit solvent model is fixed at the physical temperature of 300K.*



# Kapitel 6

## Stochastic Minimization

The determination of an appropriate free-energy function as in the previous chapter is only part of the task for protein structure prediction. Equally important is the actual search for the global minimum as demanded by the hypothesis from Anfinsen. In this chapter various approaches for this task, called *global optimization*, will be presented. We search for the global minimum by *stochastic minimization*. Different methods are presented.

In general stochastic minimization has many different applications, may it be the famous traveling salesman problem or wiring electronic circuits[62]. Many of these problems belong to the class of *NP-complete* (nondeterministic polynomial time complete) for which no method for their exact solution is available for big system sizes[37]. One typical problem is the high frustration of spin-glasses[79, 4, 14]. Another one are Lennard-Jones clusters which also have a high number of frustrated minima[35, 28]. In these problems a solution via *divide-and-conquer* like approaches is impossible due to the frustration of the systems. However in this chapter we want to concentrate on the application of stochastic minimization to the protein folding problem[113, 34, 116, 44, 105, 46].

### 6.1 Stochastic Minimization

#### 6.1.1 Introduction

Different major challenges appear in the search for the global minimum in protein-forcefields:

- frustrated and rough energy landscape with many local minima that occur far apart in configurational space
- high-dimensional energy space, which grows exponentially with system size <sup>1</sup>
- no higher symmetries applicable which may help simplifying the problem

---

<sup>1</sup>Even very small proteins like the trp-cage protein (pdb-code 1L2Y, 20 amino acids) have more than 50 degrees of freedom.

Since the native structure of a protein is found by the actual folding process in the huge conformational space one can expect that the topology of the space simplifies the search. This idea leads to the so-called *funnel* hypothesis[15, 14, 27, 47, 89] which means that the topology of the free-energy space has a bias towards the global minimum. A molecular dynamics simulation tries to follow the folding process from a starting configuration. However today's computational resources seem not be sufficient to allow this approach for any but the smallest and simplest proteins for simulations on a timescale similar to the folding times of proteins[30, 112]. To predict the structure of bigger proteins further simplifications need to be made. One idea is to disregard the actual folding process altogether and just look for the ending point, i.e. the global minimum of the free-energy landscape. The analytical solution of such a problem is obviously impossible for above reasons, i.e. system size and lack of symmetry. Also a pure random search failed due to the size of the configurational space and its high frustration. Another approach are *stochastic minimization methods*[78] which will be presented now.

### 6.1.2 The basic Idea: Monte-Carlo

Statistical physics allows to calculate properties of a system out of its probability density function

$$\rho(\vec{q}) = \frac{e^{-\beta H(\vec{q})}}{\int d\vec{q} e^{-\beta H(\vec{q})}}$$

( $\vec{q}$  designates an ensemble of protein structures in conformational space,  $H$  is the Hamiltonian,  $\beta = \frac{1}{k_B T}$ ,  $k_B$  Boltzmann constant,  $T$  temperature). This function gives each state a statistic weight. By using the equation

$$\langle X \rangle = \int d\vec{q} X(\vec{q}) \rho(\vec{q})$$

one can calculate expectation values of observables out of this probability density function. In order to find the native state we want to find the global minimum of its free energy:

$$F = U - TS = \beta^{-1} \ln \langle e^{\beta H} \rangle$$

Strictly spoken the native state is not a single configuration but the *ensemble close to the native state* in conformational space. One can neglect vibrational terms in the protein structure by assuming they contribute similarly for all structures. Then the difference in the free-energy  $\Delta F$  between between two structures of a given protein is:

$$\Delta F = \Delta E - T \Delta S_{con}$$

The conformational entropy depends on the amount of low-energy minima next to the investigated structure. For densely packed proteins the core of the protein is strongly constrained by the repulsion of the atoms. Therefore the major contribution between well defined metastable conformations comes from the surface area of the protein in contact with the solvent. Obviously a protein tries to move its hydrophilic side-chains to the outside and the hydrophobic

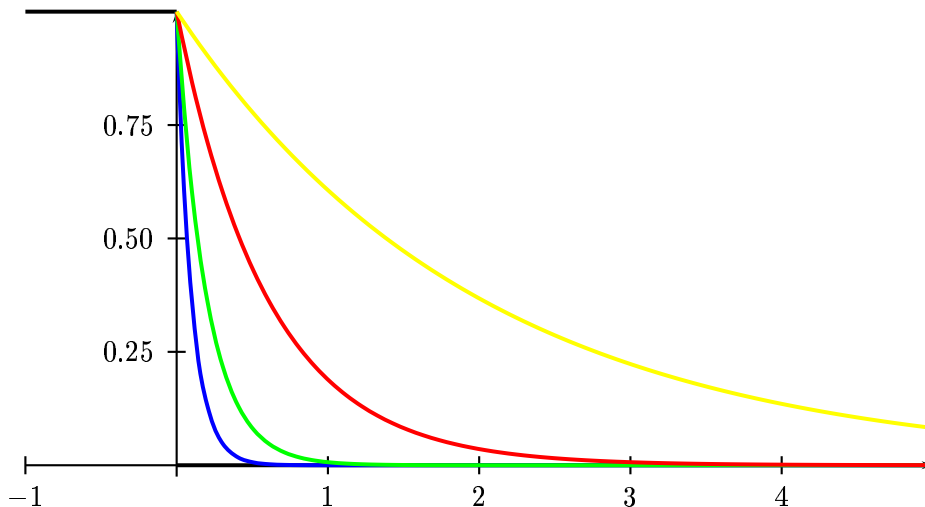


Abbildung 6.1: Plot of the Metropolis criteria (acceptance ratio)  $P_{old,new}$  versus the energy  $E \cdot \frac{kcal}{mol}^{-1}$  for different temperatures  $T = \{0K, 50K, 100K, 300K, 1000K\}$  which are drawn black, blue, green, red, yellow respectively. For all temperatures  $P_{old,new}$  equals one for  $x < 0$  and is drawn in black only.

side-chains away from the solvent to the core of a protein.

Now the Monte-Carlo (MC) algorithm allows to get an approximation of  $\rho(\vec{q})$ . This approximation is done by creating conformations  $\vec{q}_i$  with their weighted probabilities:

$$\frac{\rho(\vec{q}_i)}{\rho(\vec{q}_j)} = e^{-\beta(H(\vec{q}_i) - H(\vec{q}_j))}$$

The theory of Markov Chains and the Metropolis algorithm helps in understanding its implications.

### Markov Chain

The Markov Chain is a general concept often applied in computer science or numerical systems[39]. The sentence *the future depends on the past only through the present* gives its basic idea.

A short elaboration is as follows: Consider a set of states (also called sites)  $\{S_i\}$  and transition probabilities between these states  $p_{ij}$  with  $\sum_j p_{ij} = 1$ . Thus  $p_{ij}$  gives the probability leaving state  $s_i$  (or simply  $i$ ) and entering state  $j$ . When we now generate a sequence of states at discrete times  $k$  the state at the time  $k + 1$  only depends on where we are at the time  $k$  and the transition properties  $p_{ij}$ . The state at the time  $k - 1$  is not important at all for the time  $k + 1$ .

### Metropolis algorithm

We start a fictitious dynamical process in the configurational space, which shall sample configurations according to their thermodynamic relevance. We start at a random configuration  $\vec{q}_i$ ,

$i = 1$  and apply a perturbation, or, in other words, a slight random change to the configuration to generate a structure  $\vec{q}_{new}$ . This structure  $\vec{q}_{new}$  is accepted according to the probability

$$P_{old,new} = \min\left(1, e^{-\beta(E_{new}-E_{old})}\right)$$

(Metropolis criteria, compare figure 6.1). If  $\vec{q}_{new}$  is accepted  $\vec{q}_{i+1}$  is equal to  $\vec{q}_{new}$  else  $\vec{q}_{i+1}$  is the same as  $\vec{q}_i = \vec{q}_{old}$ . In this way we generate a chain of configurations  $\{\vec{q}\}$  and can calculate expectation values as

$$\langle X \rangle = \frac{\sum_{\{\vec{q}\}} X(\vec{q})}{\sum_{\{\vec{q}\}} 1}.$$

In the limit of an infinitely long walk the sampling gives an appropriate description of the important regions of the conformational space. The time average of  $\langle X \rangle$  gives the expectation value for  $X$  provided the free walk is *ergodic*. Therefore we do not have to save all configurations  $\{\vec{q}\}$  but only the  $\{X(\vec{q})\}$ .

The Metropolis criteria fulfills the criteria of *detailed balance*. Both detailed balance and ergodicity have their origin in statistical physics. Ergodicity means that each region of the conformational space must be reachable by the algorithm. The probability for the dynamic process  $M_{ij}^N$  reaching from configuration  $i$  in  $N$  steps the configuration  $j$  is:

$$M_{ij}^N = \sum_{k_1, k_2, \dots, k_N} M_{ik_1} M_{k_1 k_2} \dots M_{k_{N-1} k_N} M_{k_N j}$$

with

$$M_{ij} = \begin{cases} P_{ij} C_{ij} & \text{for } i \neq j \\ 1 - \sum_{i \neq k} P_{ki} & \text{for } i = j \end{cases}$$

( $C_{ij}$  probability for suggesting change from state  $i$  to  $j$ ). To fulfill the demand for ergodicity  $M_{ij}^N > 0$  has to be true for all  $i, j$  with  $N > 0$ .

Detailed balance is a sufficient, but not necessary condition that all states or configurations of the system appear according to the probability density distribution  $\tilde{\rho}(\vec{q})$ :

$$M_{ij} \tilde{\rho}(\vec{q}_i) = M_{ji} \tilde{\rho}(\vec{q}_j)$$

For very long random walks

$$\rho(\vec{q}) = \tilde{\rho}(\vec{q}) = \lim_{n \rightarrow \infty} M_{ij}^n$$

is valid.

The Monte-Carlo method can be therefore shortly described as specific Markov-chain using the Metropolis algorithm. Figure 6.2 shows a sketch of an implementation. MC can be applied in two different ways. One application calculates the expectation values or generates the  $\tilde{\rho}$  at certain temperatures  $T$ . In another application MC samples low-lying minima of the given forcefield.

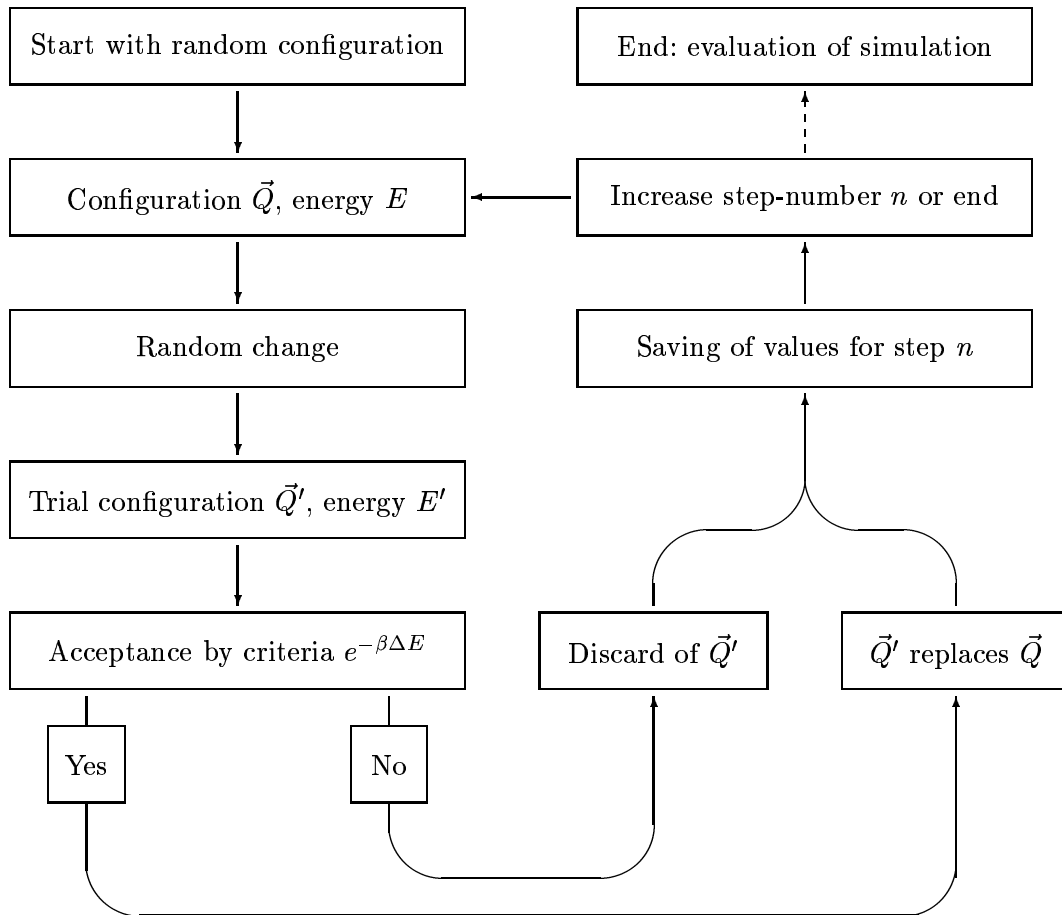


Abbildung 6.2: Simple flow diagram of a Monte-Carlo simulation. Please note that the end of a Monte-Carlo simulation is only defined by the number of steps  $N$  for the simulation. This number has to be sufficiently large to sample the conformational space.

### Remark on Temperature

In standard (equilibrium) thermodynamics the probability for finding a state is proportional to the weight  $e^{-\beta E}$  ( $\beta = 1/(k_B T)$  with  $k_B$  the Boltzmann constant and  $T$  the temperature). By applying MC to calculate expectation values the temperature in the MC-algorithm therefore defines a physical temperature.

However it is important to note that some forcefields, such as the forcefield *PPF01*, include parts which are fixed at a certain temperature such as the implicit solvent model of *PPF01* which is parameterized at a temperature of 300K. *Often there is a non-trivial and unknown dependence of the forcefield parameters on the temperature, which therefore cannot be easily changed during simulation.* In this case we have two different temperatures: the physical temperature, for which the forcefield was parameterized, and the temperature of the optimization algorithm. In this case the temperature of the algorithm loses its physical meaning.

Many Monte-Carlo based optimization algorithms change temperatures during the simulation. If this change is non-adiabatic one must carefully consider the specific meaning of the temperatures. In the course of this chapter we do not apply non-equilibrium thermodynamics but use the temperature just as a tool for finding global minima. Therefore *the temperatures used by stochastic optimization methods have often no physical meaning.*

### 6.1.3 Simulated Annealing

Simulated annealing (sa)[62] generalizes the concept of Monte-Carlo simulations to optimization problems. While during a Monte-Carlo simulation one stays at the same temperature for the duration of the simulation simulated annealing tries to copy the natural relaxation process in which complex structures are annealed, e.g. in the way a liquid or metal freezes. We want to find the global minimum of the potential energy surface. Therefore we replace the temperature  $T$  at which a Monte-Carlo simulation is run with a fictitious temperature  $T_n$  which may change during the simulation at each step  $n$ . This fictitious temperature influences the Boltzmann acceptance of new configurations. We start at a very high temperature which is lowered during the duration of the simulation to adopt a specific final temperature. A high temperature allows to overcome barriers separating local minima while a low temperature confines the search to low-energy conformations. The idea of cooling down like in nature experiences often the problems associated with the high frustration of protein forcefields (like in spin-glasses).

The idea is to cool the system slow enough that the thermodynamic equilibrium is conserved (adiabatic cooling). Using the theory of Markov-chains one can prove that it is possible to generate a cooling schedule which guarantees finding the ground state:

$$\beta_n \leq \epsilon \cdot \ln(n), n \in \mathbb{N}$$

(for  $\epsilon$  small enough). For practical reasons this is far too slow since the necessary evaluation of the forcefield at each step of the simulation adds up in taking a lot of time. For faster temperature reduction one *quenches* the system which may cause entrapment in local minima.



Therefore it is always necessary to run not single but many simulations to overcome these entrapments. A geometric cooling schedule, i.e.

$$\beta_n = \beta_0 \cdot \alpha^n, n \in \{1, 2, \dots, N\},$$

is often used, containing the specific parameters  $\alpha$ ,  $\beta_0$ , and  $N$ . The optimal parameter choices depend on the problem and it is often a matter of trial-and-error to find good parameters as the transferability of the parameters to other optimization problems is very limited.

#### 6.1.4 Freezing Problem

One central problem which arises during stochastic optimization of a system is the *freezing problem*. Standard sa experiences it strongly since quenching the system by quickly decreasing the temperature often entraps the algorithm in local minima without finding the global minimum. To understand this one has to remind oneself that a frustrated potential energy surface is characterized by many local minima which are separated by high barriers.

To sample low-energy conformations and explore a minimum thoroughly *downhill movement* which is connected to *low temperature* is needed. But at the same time this prevents the algorithm from leaving these local minima due to the high barriers surrounding them. Therefore the algorithm experiences *freezing*.

To escape local minima and overcome the high barriers one has to allow *uphill movement* on the potential energy surface which is connected to *high temperatures*. However at the same time high temperatures limit local optimization.

This competition between local optimization and need to overcome high energy barriers is the central problem of stochastic optimization. Sa tries to solve it by starting at high temperatures to overcome possible barriers and then quickly lowering temperature to explore possible minima. However this often leads to local entrapment in case of strong frustration without any chance of leaving a minimum again.

#### 6.1.5 Stochastic Tunneling

Stochastic Tunneling (stun) [126] tries to overcome some of the encountered difficulties with the frustration of the forcefield. The optimized function is subjected to a dynamic nonlinear transformation. For sake of simplicity we speak of an energy to be optimized although stun may also be applied to other problems.

$$E'_n = 1 - e^{-\gamma(E_n - E_0)}$$

( $E_n$  current non transformed energy at step  $n$ ,  $E_0$  equals the best estimation of the global minimum so far). In stun the conformational space is explored by a dynamic process on a transformed potential energy surface. This resembles a normal MC with a Boltzmann acceptance not on  $\Delta E$  but on  $\Delta E'$ . Whenever a new lower energy is encountered  $E_0$  is set to this new value. What this transformation now does is lowering high regions in the energy landscape while at the same time emphasizing deeper lying regions stronger. This allows to overcome

high-energy barriers which are difficult to overcome in simulated annealing simulations or comparable optimization schemes. In total local entrapment is reduced. An illustration is shown in figure 6.3. Since the stun transformation limits the energy-surface to values  $\in [0..1]$  it cannot use the same temperatures as sa simulations as obviously visible by looking at the *Metropolis-criteria*.

Applying this mechanism to protein folding showed the problem that high-energy regions are, once a low-lying minimum has been found, practically equal to 1. They lose any gradient showing in the direction of lower-lying structures. The landscape becomes golf-course like, meaning it is practically flat and one can find the hole in it only by luck. Therefore we adjusted the transformation slightly to have a stronger gradient also in high-energy regions [105]:

$$E_n'' = \ln \left( \gamma(E_n - E_0) + \sqrt{\gamma(E_n - E_0)^2 + 1} \right)$$

The comparison between the two different transformations are shown in figure 6.4, where the second transformation has been slightly modified to ease optical comparison. Both transformations need different values for  $\gamma$ .

## 6.2 Basin Hopping

This version of the basin hopping (bh) is a slight variation from the original idea [125, 29, 103]. Basin hopping simplifies the potential energy surface by mapping local points corresponding to local minima. Our version uses many subsequent sa-simulations with slightly changing parameters between two runs. Each sa-run starts at a high-temperature and cools down to a low temperature within  $N$ -sa-steps. This allows to overcome energetic barriers at high-temperature and good local minimization during the low-temperature part of the simulation. The exact values depend on the forcefield, for PFF01 the starting temperature were in the range of  $600K..1000K$ , with final temperatures being in the range of  $1K..5K$ . After each such sa-simulation using these temperatures bh decides whether to keep the ending configuration by a simple comparison with the energy before the run via a threshold:

$$E_{\text{before sa}} - E_{\text{after sa}} > E_{\text{threshold}}$$

If this criteria is met the ending configuration becomes the starting configuration of a new run else it is discarded. One such sa-run with following threshold-acceptance is a single bh-step. Basin hopping is now a multitude of these bh-steps following each other. It proved efficient to increase the number of simulation steps in sa for later bh-steps. For the threshold values in the range  $3(\text{kcal/mol})..5(\text{kcal/mol})$  proved optimal. When no further gain in energy for many subsequent bh-steps is noticed the lowest energy and corresponding conformation is taken as good approximation of the global minimum.

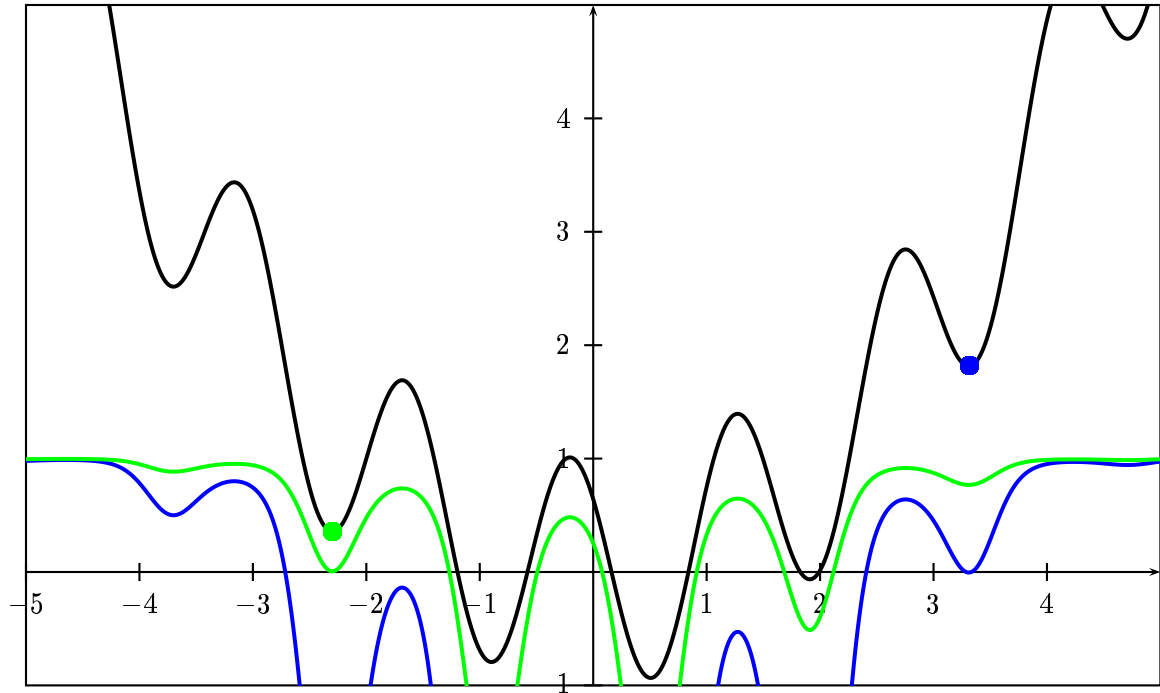


Abbildung 6.3: In *black* a fictitious energy function is shown. This energy function is dynamically transformed during a *stun*-simulation. When *stun* has found no lower energies than the *blue* dot the energy function is transformed to the blue function. When *stun* finds new lower energies the function gets dynamically transformed. When the low-lying energy at the *green* dot has been found the new effective energy function is shown in green. Please note that it becomes easier to overcome higher energy regions as lower energies are found. Here  $\gamma$  equals unity.

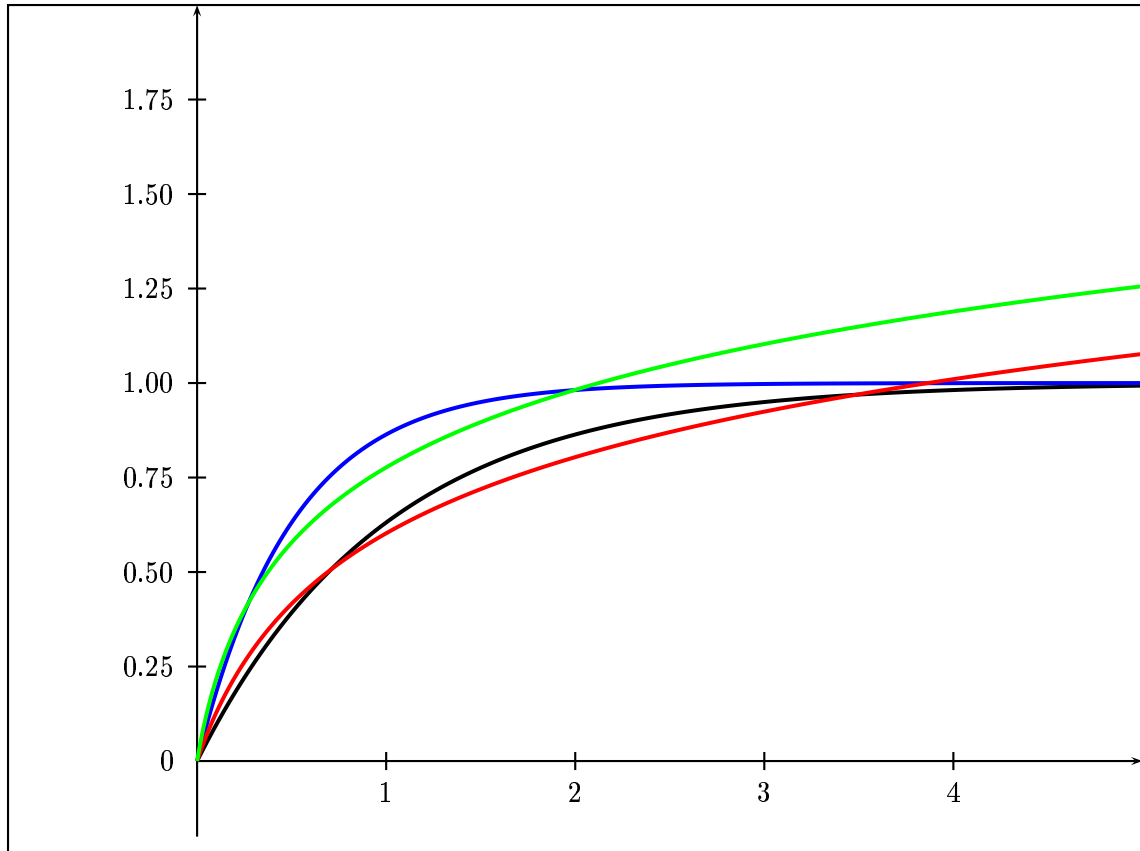


Abbildung 6.4: Comparison of different stun-variants. The x-axis shows  $E_n - E_0$  and the y-axis the transformed energy. In black and blue is the original transformation  $E'_n = 1 - e^{-\gamma(E_n - E_0)}$  with  $\gamma = \{1, 2\}$  respectively applied. In red and green (according again to  $\gamma = \{1, 2\}$ ) is  $E''_n = 0.3 \cdot \ln \left( 5 \cdot \gamma(E_n - E_0) + \sqrt{5 \cdot \gamma(E_n - E_0)^2 + 1} \right)$  shown as transformed energy. The important change is taking away the limitation of the transformation being  $\in 0..1$  therefore allowing a gradient for higher values  $E_n - E_0$  and avoiding golf-course scenarios.

## 6.3 Genetic Algorithms

Genetic algorithms (ga) have been successfully applied to other optimization problems like the Traveling-Salesman problem[64] or Lennard-Jones clusters[127]. There have also been approaches applying it to simplified protein models[23]. Genetic Algorithms try to emulate natural selection by optimization of a target. First we seed a starting population, consisting of individuals. At each step this set of individuals is sorted according to a fitness function. Now this set of individuals is slightly changed. One way of changing is exchanging genes (in a meaning of attributes) between different individuals called *crossing*. Another possible change is a random *mutation* to an individual. Afterwards the fitness function is applied again and the whole procedure continues.

We applied a genetic algorithm to protein folding. As fitness function we took simply the energy in the forcefield PFF01 after short local relaxation. The genes of each individual were its dihedral angles or its secondary structure in form of constraints. However we ran into several problems. Most important was the frequent clash of conformations after crossing. Proteins are very densely packed and in almost all cases the repair of the atomic clashes was impossible rendering the individual unfit. Another problem was the needed long relaxation time even for non-clashing conformation making a quick test of the fitness of a new individual a time-consuming step. In total we did not succeed in protein folding using genetic algorithms even on small proteins like the trp-cage (pdb-code 1L2Y) with its 20 amino acids.

## 6.4 Distributed Computing: Server-Client Model using Screen Savers

### Aims

In protein folding the use of large amounts of computational power is required. Therefore not only single CPU's or dedicated computer clusters could be used but also desktop computers with Internet connection in case of free resources (for example during the night). This approach to protein folding is pursued by the group of Pande (folding@home)[128] who use screen savers to do molecular dynamics simulations on proteins. In this they followed the idea of seti@home[119], one of the first attempts gathering distributed computational resources on a global scale.

We applied a similar implementation as server-client model. In this approach a central server administrates the organization of the simulations run on the clients. Some important questions are:

- What is an efficient scheme for running simulations?
- How to deal with problems of stability of the network? How to deal with interrupted simulations or clients which are removed from our network?

- Most people would allow screen savers to run on their computers as long they do not save data and do not access local files. Which kind of optimization method can work with these restrictions?

### Implementation of the method

First a set of  $N$  starting structures  $\{1..N\}$  is created on the server sorted by total energy. Each time a client connected to the server indicating it has free resources, one structure  $\vec{q}_{selected}$  was selected out of this set and a sa-simulation started on the client. It has shown efficient to increase the number of steps during these sa-simulations. So later sa-simulations get an increased number of steps compared to earlier ones. After the simulation finishes the resulting structure  $\vec{q}'$  is structurally compared against the set of structures. A comparison structure out of the set of starting structures is chosen. First the structure is found with which  $\vec{q}'$  had the lowest RMSD-values. If this value is not below a threshold, the structure highest in energy was instead chosen, so one gets a comparison structure  $\vec{q}_i$  either by closeness according to RMSD or because it is the structure worst (highest in energy) in the set. Now the energies  $E(\vec{q}')$  and  $E(\vec{q}_i)$  are compared and in case  $\vec{q}'$  has the lower total energy it replaces  $\vec{q}_i$ , meaning the losing structure is discarded. This process is repeated for a sufficient time to find low-lying energetic structures. In case a client running a job is not responding for a defined amount of time the corresponding calculation is given to another client again. Due to the nature of stochastic optimization this ending is ill defined. It is assumed that when for a long time, i.e. many subsequent sa-simulations, no further gain in energy was noticed a continuation of further simulations would make no sense. Therefore the simulations are then stopped.

## 6.5 Energy Landscape Paving

Energy landscape paving[45] (elp) is a stochastic optimization method which tries to avoid entrapment in local minima. By performing Monte-Carlo simulations with a modified energy expression the search is subtly steered away from regions in the energy space already visited:

$$w(E') = e^{-\frac{E'}{k_b T}}, E'(\vec{q}) = E(\vec{q}) + f(H(p(\vec{q}), t))$$

$W$  is the weight of a configuration,  $T$  a (sufficient low) Temperature,  $\vec{q}$  a configuration of the protein,  $E'$  the modified energy expression of the energy  $E$  and  $f$  a function of the histogram  $H(p(\vec{q}), t)$  which is depending on a pre-chosen order parameter  $p(\vec{q})$ . We tested several parameters  $q$  like total energy, helicity and radius of gyration. Most successful proved the helicity, as shown in the next chapter[109].

## 6.6 Adaptive Parallel Tempering

The parallel tempering technique was developed to overcome difficulties in the evaluation of thermodynamic observables on rough energy surfaces[76, 70] and on protein folding [43, 10]. It

is related to the replica exchange method and was previously applied to molecular dynamical simulations[69, 100]. The idea of parallel tempering (pt) is to perform several concurrent simulations of different replicas of the same system at different temperatures and to exchange replica (or temperatures, which results in the same effect) between the simulations. Therefore it allows to gain knowledge about thermodynamic expectation values over a wide range of temperatures at the same time, as all simulations are in thermodynamic equilibrium with regard to their specific temperatures.

The probability of exchange between two replicas is

$$P = \min(1, e^{-(\beta_j - \beta_i)}(E_i - E_j))$$

( $\beta_x = 1/(k_b T_x)$ ,  $T_x$  the temperatures and  $E_x$  the energies of the replicas). This exchange mechanism allows to overcome energetic barriers between metastable conformations for low-temperature simulations. It guarantees that all simulations (replicas) remain in thermodynamic equilibrium at their temperatures. The range of temperatures is selected on basis of the system and its configurational space.

We want to use parallel tempering as an optimization technique. For the simulations on proteins we used a span of temperatures between 2 and 600 K which allowed both local optimization and a wide search of the configurational space.

The lowest temperature in general yields the best estimation for the global minimum while the higher-temperature-replicas are needed to find new conformations. The computational effort rises linearly with the number of replicas and the efficiency of pt therefore decreases when more than the minimally needed number of replicas is used. Investigations on protein folding with pt showed that the standard implementation is inferior to basin hopping[52, 51, 83]. The major problem is that a high gap in energy between adjacent replicas results in prohibiting exchanges between these temperature levels. This can happen between several replicas and renders pt much less effective. One can imagine this effect as transforming pt into some kind of independent MC-simulations at fixed temperatures.

Therefore we developed two mechanisms to overcome these difficulties in standard pt. One is an *adaptive temperature control*. We monitored the rate of exchange between adjacent temperatures. If the exchange ratio between temperatures  $i$  and  $i + 1$  was below 0.5% all temperatures above  $T_i$  were lowered by 10% of the difference  $t_{i+1} - t_i$ . If the exchange ratio exceeded 2% the opposite mechanism took place and all these temperatures were increased by the same amount.

The low-temperature states tend to find near native conformations but sometimes the overcome of barriers by exchange between neighboring levels can be slow or cyclic ( $A$  exchanges with  $B$  and later with  $A$  again with no other structures involved). A second mechanism which further increased the efficiency of apt was the introduction of a *replication step*. Every 250,000 simulation steps the lowest-energy information replaced the conformation at the highest temperature and was kept at the highest temperature for 10,000 steps with no possibility of exchange. The latter was necessary to keep it at this temperature for a short time in order to allow some structural change, otherwise it directly and rapidly falls down the temperatures

again. This replication step allowed an alternative possibility to overcome barriers and results in a rapid and large-scale exploration of the configurational space close the presently best conformation.



# Kapitel 7

## Prediction of tertiary structures with PFF01

We investigated whether the forcefield PFF01 [53, 48], presented in chapter 5, was able to predict the native structure for proteins of different sizes starting from random initial conditions. Random initial conditions means that the starting structure has no clashes of the atomic shells but otherwise totally randomized dihedral angles. We applied different stochastic optimization methods finding that the structure with the lowest minimum found in the simulations was equivalent to the native state of each protein. A summarization about the folded proteins is given in table 7.1. These proteins are from different protein families and show little or no sequence identity. They all include high amounts of helical content. In this chapter we focus on proteins for which we have run sufficiently many simulations to observe reproducible folding. Other proteins like 1ENH have also been investigated, but no long enough simulations from random starting conditions have been made so far to predictively find the native state. Given the high computational costs for these simulations as shown in table 7.1 we concentrated on selected proteins spanning a range of families in the possible folds of proteins. Four out of five helical proteins investigated have been successfully folded ab-initio from random starting conditions. In one case, Protein A (1BDD), the energy gap between the folded state and with an RMSD-B of  $\approx 9 \text{ \AA}$  dissimilar unfolded state is too small to consider 1BDD ab-initio folded in PFF01. However also for this protein the best estimation for the global minimum known is in the native folded state starting from initial random conditions. Attempts to fold the protein 1BHI which contains a  $\beta$ -sheet failed by resulting in non-folded conformations better in energy than those similar to the native state. This is presented in figure 7.1. We conclude that PFF01 must be re-parameterized for prediction of  $\beta$ -sheets.

To date PFF01 is the only bio-molecular forcefield which is able to stabilize several proteins in their respective native state. Even more PFF01 is not only stabilizing these proteins but predicting their native state from random initial conditions.

*The simulations reported here validate the forcefield PFF01 against experimental data for se-*

Protein	PDB-Database entry	# Amino acids	RMSD-B	Costs
Trp-Cage Protein	1L2Y	20	2.83 Å	1/2
HIV-Accessory Protein	1F4I	40	2.46 Å	3
Villin Headpiece	1VII	36	3.65 Å	4
Bacterial Ribosomal Protein L20	1GYZ	60	4.64 Å	40
Protein A	1BDD	60	2.69 (*)	N/A

Tabelle 7.1: This table lists the proteins folded from random initial conditions with the forcefield PFF01. The RMSD-B gives the RMSD of the backbone of the structure with the lowest energy found to the native structure. In case of multiple native measurements like in NMR measurements the first structure given in the PDB database given is taken. The last column gives the computational costs in accumulated CPU-years of standard off-the-shelf PC (around 1GHz) for the described simulations. Protein A is a special case, as described in the text. By now the energy gap between the folded state and a dissimilar unfolded state is with  $\approx 1kcal/mol$  too small to say with certainty that we found the native state in ab-initio prediction.

*veral different proteins. Although not suitable for  $\beta$ -sheets this result demonstrates the effectivity and feasibility, given present day computational resources, of protein tertiary structure prediction using an all-atom free-energy forcefield* During the optimization of these proteins different stochastic optimization methods were tested and refined with regard to protein structure prediction. We always noted a competition between good local optimization which usually meant low temperatures for the optimization technique and high temperature for an thorough investigation of the folding space. The previous chapter introduced basic concepts of minimization techniques. We want to explain their application and adoption on the specific problem of protein folding.

## 7.1 Test on theoretically modeled $Ala_{10} - Gly_5 - Ala_{10}$

This polypeptide is an artificial model. It has been modeled in simulation with ECEPP2[2]. Experimental data are not available for this polypeptide chain. However the limited size of 25 amino acids and its very easy sequence incorporating only two different kinds of sidechains suggests it to take a two helix structure. In our simulations it quickly folded from random initial conditions (using standard sa). It took a two helix structure with the bend lying in the Gly region. When simulating without solvent we got a single long helix in contrast to the Simulations of [2], which got the two helical structure also in vacuum. We conclude that our forcefield stresses hydrophobicity more strongly than ECEPP2.

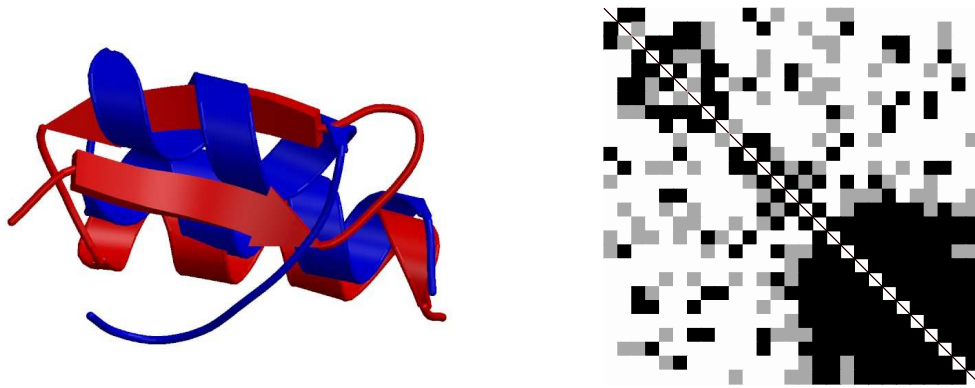


Abbildung 7.1: Shown is on the left an overlay of the native structure (red) and the best structure found during the simulations for the protein 1BHI. The RMSD-B is  $4.75 \text{ \AA}$ , which is already pretty high for a protein with only 38 amino acids. In addition one can easily see on the picture both slightly different topology and a different secondary structure of this protein in the  $\beta$ -sheet part. The competing structure in blue indicates a possible bias of PFF01 towards helical structures and is typical for unfolded structure we gained for 1BHI. On the right the according  $C_\beta$  matrix is shown. One can see the good agreement in the helical part and the almost white part where the  $\beta$ -sheet lies indicating no high agreement there.

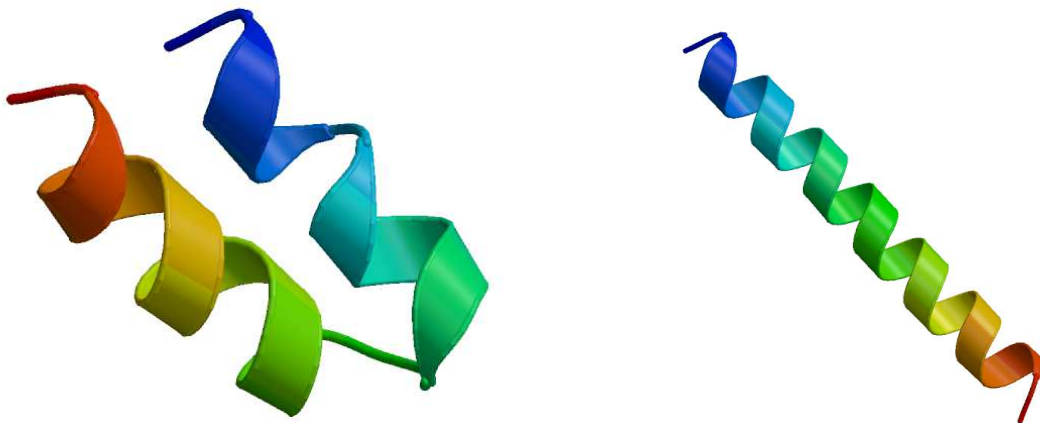


Abbildung 7.2: Structures gained in simulations with PFF01 for the polypeptide  $Ala_{10} - Gly_5 - Ala_{10}$ . The lowest energy structure is shown on the left. The structure on the right is worse in energy using the implicit solvent term of PFF01, but wins in vacuum. Overlay pictures or a  $C_\beta$  matrices are not possible since no experimental data are available.

Sequence & secondary structure	Weight	# Atoms	# Atoms in PFF01	RMSD-B
NLYIQWLKDGGPSSGRPPPS CHHHHHHHCCHHHHCCCCC	2151 <i>D</i>	304	189	2.83 Å

Tabelle 7.2: The basic data about the trp-cage protein 1L2Y. The structure is given in 1-letter-code, the secondary structure in 3-state-code. The  $RMSD - B$  are between the best folded structure and the experimentally measured structure. It consists of two helices which are separated by a small coil region. The three prolines close to the C-terminus prevent helix-formation in this part of the protein.

## 7.2 Trp-Cage Protein

With 20 amino acids the trp-cage protein is very small. The sequence of this protein was artificially modified to generate a fast folder whose folding speed was measured to be  $4\mu s$  only [84, 96]. It has been subject to many investigations and successfully folded in both MD[112] and free-energy-minimization simulations[105, 107, 104].

The basic data about this protein are given in table 7.2.

Due the small size of this protein we tested several folding methods with regard to their efficiency on this protein and refined them further.

### 7.2.1 Stochastic Tunneling

The basic idea of stun is flattening a rough landscape with high barriers between local minima in all regions which lie significantly above the best estimate for the global minimum  $E_0$  in order to search for the global minimum [126]. This can be understood since at finite temperatures the dynamics of the system become diffusive for  $E \gg E_0$  independent of the relative energy differences- which can be called *tunneling* through energy barriers. The applied transformation was:

$$E_{stun1}(E) = 1 - e^{-\gamma(E-E_0)}$$

Applying this original transformation to protein folding proved difficult. As soon as the first atoms clashed this transformation created a golf-course landscape. Further clashes gave no increase in energy due to the strong non-linear transformation. The entire clashing conformational space becomes a featureless plane with no gradient towards non-clashing conformations. Attempts to find applicable values for  $\gamma$  or lowering the temperatures gave no satisfying results. A typical simulation is presented in plot 7.3. Therefore we searched for new transformations. The important characteristic was the inclusion of a gradient also for high energies. This should ensure that there is a mechanism for returning to non-clashing conformations

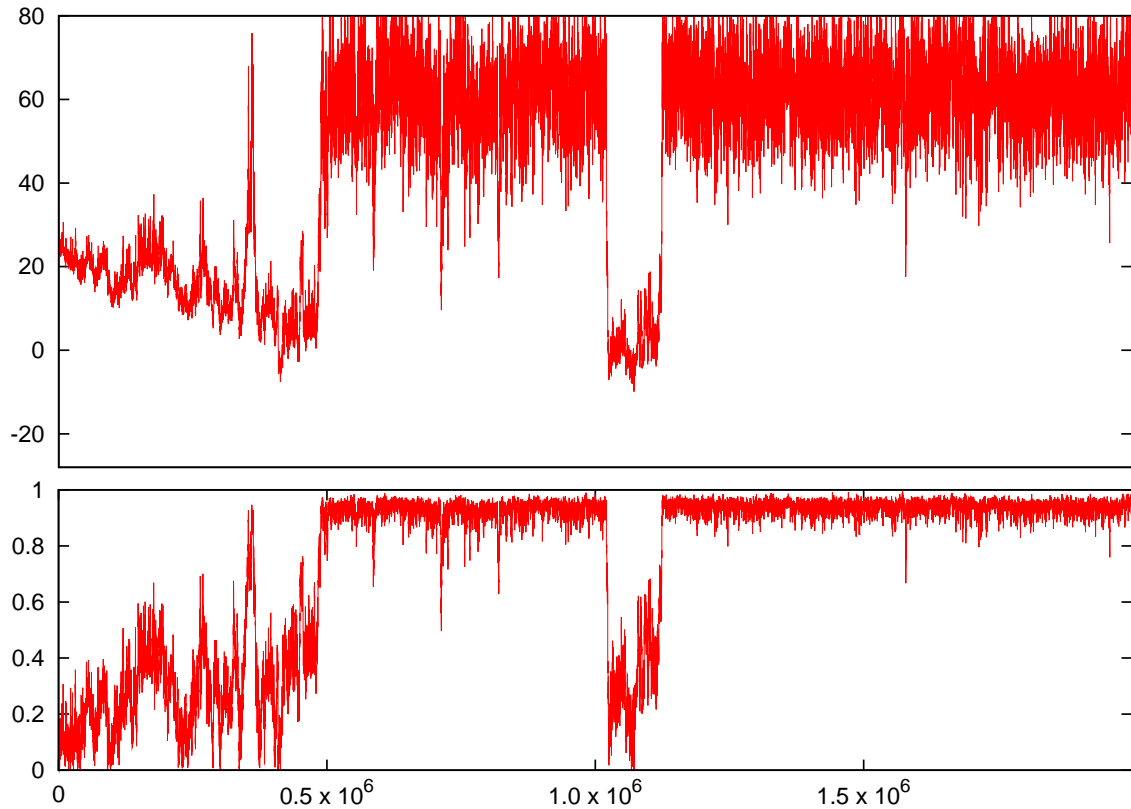


Abbildung 7.3: Typical simulation with the original stun-transformation  $E_{stun1}(E) = 1 - e^{-\gamma(E-E_0)}$ . In the upper plot the total energy in kcal/mol is plotted on the vertical axis. In the lower plot the transformed energy is shown. This transformed energy has no dimension. The horizontal axis gives the number of steps. It can be seen that after some time stun only continues to tunnel and does rarely go back to low energies for local optimization. This problem did not appear in the modified stun-transformation. During this simulation the native state was not found.

easily. The transformation which proved best was

$$E_{stun2}(E) = \ln \left( \gamma(E - E_0) + \sqrt{\gamma(E - E_0)^2 + 1} \right)$$

in connection with changing the temperature for the optimization between low-temperature local search and high-temperature global search/tunneling phases[105]. Finally we ran a simulation with the optimized parameters ( $\gamma = 0.5(kcal/mol)^{-1}$ , T adjusted during the simulations as described later). We ran not a single but in total 25 simulations for one week each on standard CPU's ( $\approx 1Ghz$ ) which corresponds to about 12,500,000 steps. As seen in table 7.3 6 out of 25 simulations resulted in estimates of the global minimum within  $1kcal/mol$  of the best energy found. All of these 6 give a good estimation of the experimentally found NMR-structure (the first two were closer than 3 Å RMSD-B). However even for this small protein the best energy found in the individual simulations still spans a wide range. Therefore it is necessary to run a multitude of simulations to gain a reliable estimation of the global minimum.

As an example for one of these runs the simulation resulting in the best estimate of the global minimum is shown in figure 7.6. Whenever a new lowest energy  $E_0$  is found the dynamic transformation changes the effective energy to zero. This graph also illustrated the alteration between low temperature local searches and high temperature tunneling phases. Whenever a new promising local minimum was found an adaptive temperature algorithm slightly lowered the temperature to facilitate finding the local minimum. Afterwards, when no further gain in energy was expected, a tunneling phase was initiated in which a slight rise in temperature enabled the stun-transformation to overcome high-energy barriers again. After some time the temperature was lowered again to begin a new cycle starting with local optimization via low temperature.

This approach results also in a large scale exploration of the conformational space. The tunneling sometimes initiates a transition to a new and very different local minimum. This exploration is needed to ensure a reliable search for global minimum and presented in figure .

When carefully comparing between predicted structures in the forcefield and experimentally determined native states one sees that the secondary structure is predicted with high accuracy. The first helix from residues 1-10 is almost always predicted correctly and the second helix from 11 to 14 often. The end of protein is quite sloppy and dominated by 3 helix-breaking prolines. This part is most of the time correctly predicted as coil. In almost all simulations it shows that the last 2 residues bend to the back as shown in the overlay picture 7.5. This can be understood since a helix leaves little place for conformational deviations. A coil however is more free and obviously the forcefield has difficulties in arranging it in the native orientation. The NMR-measurement shows also that this coil region of the protein is not fixed in place. When comparing these results to MD simulations [112] also using implicit solvent we gain a comparable degree of accuracy. Explicit solvent simulations reach a higher degree of accuracy but cannot be applied in the optimization approach since the entropic contributions of the solvent must be parameterized. Therefore the application of the implicit solvent model may be the factor limiting the precision of prediction with PFF01.

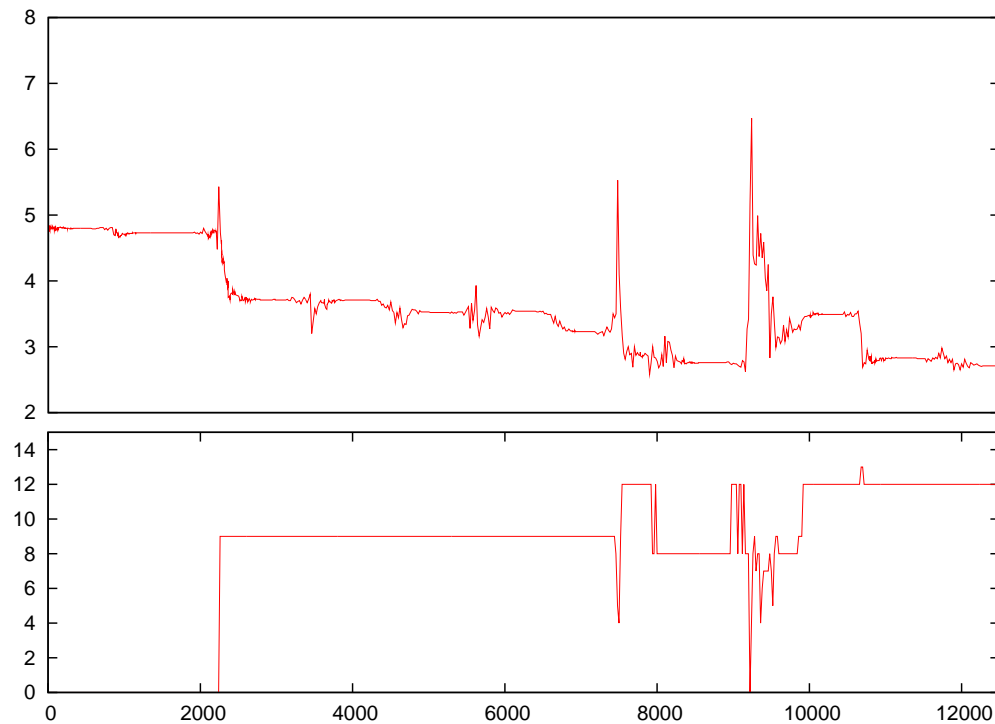


Abbildung 7.4: The vertical axis plots the RMSD-B to the native structure (upper) and the helical content (lower) during the *stun*-simulation resulting in the best estimation of the global minimum. The horizontal axis gives the number of steps in 1000s. It can be seen that tunneling events happen (esp. after 9 Mio steps) resulting in strong changes in structure. Also during the beginning of the simulation first the long helix close to the N-terminus forms while the second helix forms much later after the tunneling event.

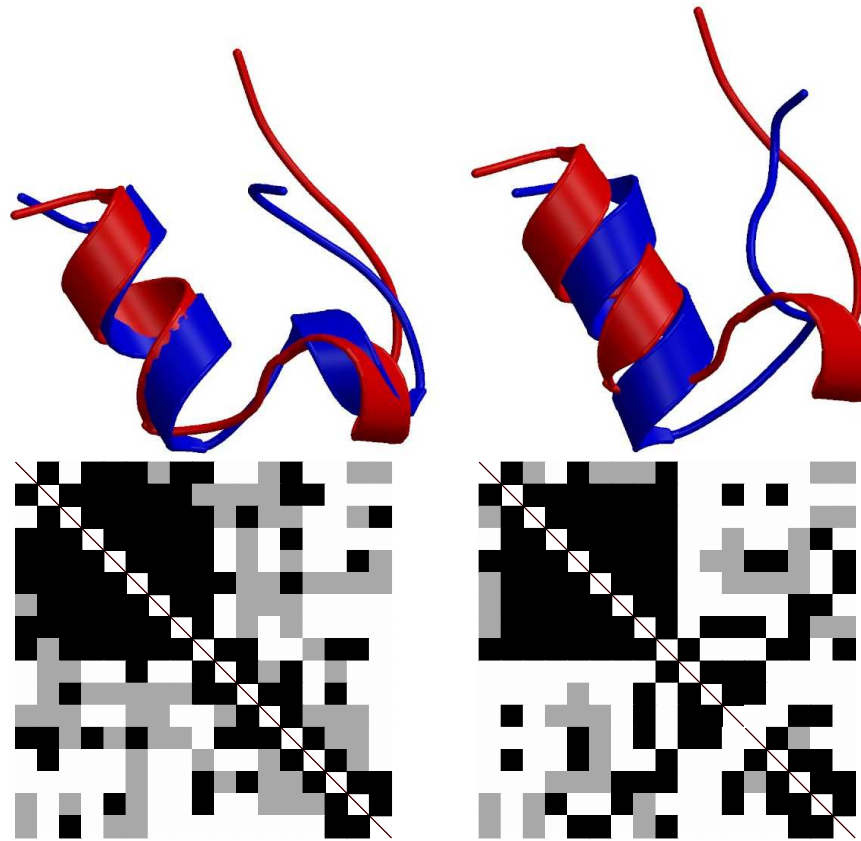


Abbildung 7.5: Upper picture: To the left an overlay of the native structure (red) and the best structure found during the simulations is shown. The right presents an overlay of the native structure (red) and the misfolded structure (\*) (blue) in the table 7.3 missing the second helix. One can see the difficulty in stabilizing the sloppy C-terminus. Lower picture: The according  $C_\beta$  matrices.



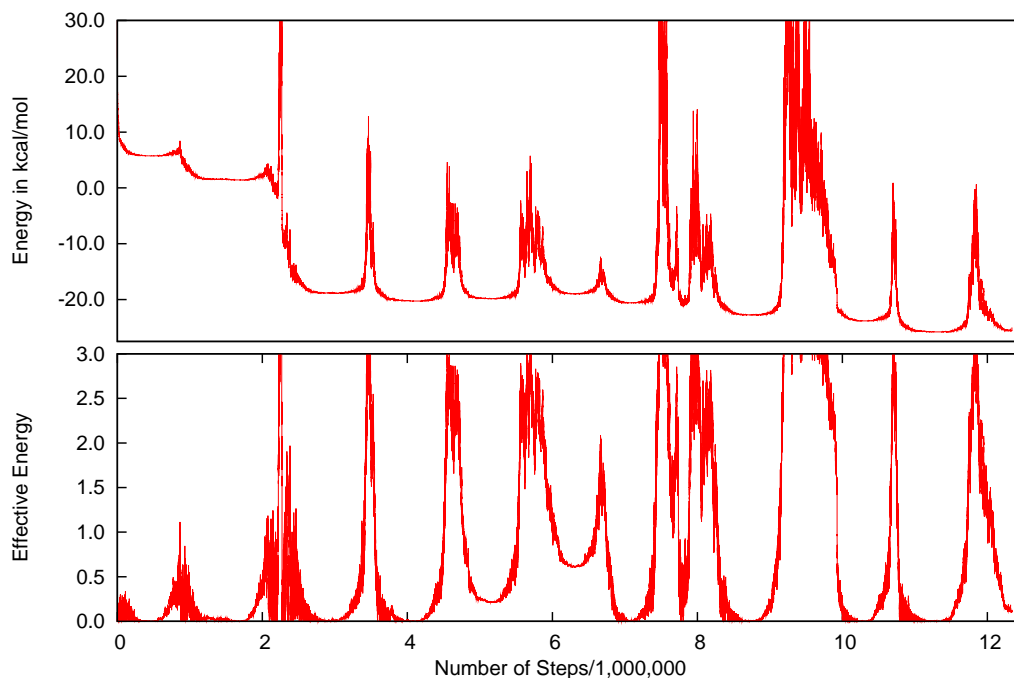


Abbildung 7.6: Total and effective energy in the simulation resulting in the best estimate of the global minimum for the trp-cage protein using stun. The energy is given in kcal/mol, the effective energy has no dimension. The horizontal axis gives the number of MC-steps in millions. Due to the nature of the dynamic transformation whenever a new lowest total energy is found the effective energy becomes zero.

In total stun applied to the force field PFF01 is able to predict the native state of the trp-cage protein. By running a multitude of simulations reliable results are gained since the energy gives a rational criterium for interpretation of individual runs against each another.

### 7.2.2 Adaptive Parallel Tempering

To test the efficiency of adaptive parallel tempering (apt) we performed a series of apt-simulation with 4,8,14 and 30 replicas starting in all replicas with random conformations at high temperatures [107, 104]. We investigated the optimum number of temperature levels (replicas). Starting with high temperatures gives no bias and further demonstrates the feasibility of the adaptive temperature control.

One such simulation is shown in figure 7.8. After an initial short period of relaxation the temperatures converge to stable levels which only slightly fluctuate during the simulations. The touching of the energies from neighboring replicas indicates exchanges between them. During these simulations the structure with the lowest energy found is in excellent agreement with the native structure (see figure 7.7) with a RMSD-B value of 2.01 Å .

Energy	RMSD-B-1	RMSD-B-2	Secondary structure
19.29	2.61	0.00	CHHHHHHHHCCHHHHCCECCC
-25.73	0.00	2.61	CHHHHHHHHCCHHHHCCECCC
-25.79	1.81	2.83	CHHHHHHHHCCHHHHCCECCC
-25.31	2.52	3.05	CHHHHHHHHCCHHHHCCECCC
-25.25	2.55	3.13	CHHHHHHHHCCHHHHCCECCC
-25.25	3.30	4.26	CHHHHHHHHCCHHHHCCECCC
-25.24	2.56	3.13	CHHHHHHHHCCHHHHCCECCC
-25.15	2.57	3.13	CHHHHHHHHHHHHHHCCECCC
(*) -24.15	3.98	4.80	CHHHHHHHHCCECCCEEECC
-24.06	4.43	4.73	CHHHHHHHHCCECCCEEECCC
-23.99	4.50	4.95	CHHHHHHHHCCECCCEEECCC
-23.64	3.50	3.86	CHHHHHCCCHHHHHCCCCC
-23.64	3.70	4.54	CHHHHHHHHHCCCEEECECC
-23.45	2.58	3.18	CHHHHHHHHCCHHHHCCECCC
-23.30	2.96	3.83	CHHHHHHHHCCHHHHCCECCC
-23.27	2.51	2.72	CHHHHHHHHCCHHHHCCECCC
-22.82	4.67	4.73	CHHHHHHHHECCCCCCCCC
-22.53	3.66	4.37	CHHHHHHCCEHHHHHCCECC
-22.49	5.10	4.87	CHHHHHHHHCCEEEECCECC
-22.45	4.01	4.59	CHHHHHHHHHHCCECCCCCCC
-22.23	4.68	5.08	CHHHHHHHHCCEECCEEECCC
-21.27	3.43	2.88	CHHHHHHHHCCHHHHCCECCC
-20.31	5.63	5.77	CHHHHHHHHHHECCCEEECC
-20.20	3.57	4.37	CHHHHHHHHHHCCECCCEEECC
-20.16	3.22	3.31	CHHHHHHHHHHCCHHHHCCECCC
-19.82	3.78	3.89	CHHHHHHHHHHCCECCCCCECC
-18.70	4.64	4.83	CHHHHHHEEHHHHHCCECC

Tabelle 7.3: Overview of the lowest energy conformations found during 25 stun-runs. The first two lines give data for the original/relaxed NMR structure. Energies given are in kcal/mol. The RMSD-B are to the relaxed NMR (1) and the NMR-structure (2) and given in Å . The secondary structure was generated with DSSP and further refined to 3-State structure. Structure (\*) is presented as example for a misfolded structure in figure 7.5.

# Replicas	# simulations	# simulations converged to native conformation
4	5	1
8	4	2
14	1	1
30	1	1

Tabelle 7.4: Results of the adapted pt runs with different number of replicas. Converged to native conformation meant within 2 kcal/mol of the best known estimate of the global minimum and an RMSD-B value smaller than 3 Å . Simulations with low number of replicas ( $\leq 8$ ) are not as reliable as the simulations with higher number of replicas (14 or 30).

Table 7.4 evaluates the simulations with different numbers of replicas. The simulations with a small number of replicas  $\leq 8$  all had difficulties to converge. This can be rationalized, because high and low temperatures are both needed in the simulations to make sure local optimization as well as large-scale searching of the conformational space is done.

The dependence of the error of the simulation (energetic difference to the best known estimation of global minimum) and their dependence on the number of replicas is not surprising. Higher number of replicas yield a lower energy. We find kind of exponential convergence of the method. A small number of replicas means that the simulation depends on rare events. Such simulations get trapped in metastable local minima which are hard to leave without sufficient exchange with other temperature levels. This interpretation is further underlined by table 7.4 since simulations with a low number of replicas tend to be unreliable (getting stuck in these local minima). For larger number of replicas the reliability of the method increases until saturation with 14 replicas. In the 30 replicas-simulation many replicas have temperatures below 1K and optimize only locally. After relaxation the remaining temperatures obey a geometric distribution which is often applied in regular pt simulations.

### 7.2.3 Energy Landscape Paving

Energy landscape paving (elp) is one approach to overcome the multiple minima problem. It does so by performing low-temperature Monte Carlo simulations with a modified energy expression that steers the search away from regions already explored:

$$E_{elp} = E + f(H(q, t))$$

( $f$  is a function weighting the histogram  $H(q, t)$  which depends on pre-chosen order parameters  $q$  and the time  $t$ ). It follows that the probability of leaving a local minimum increases with the time spent there. The histogram is biasing against re-visiting regions already visited before.

We investigated elp on the trp-cage protein [109] for the histogram-function  $f(H(q, t)) = cH(q, t)$  with  $c$  being a free parameter. We chose different order parameters  $q$  and varied

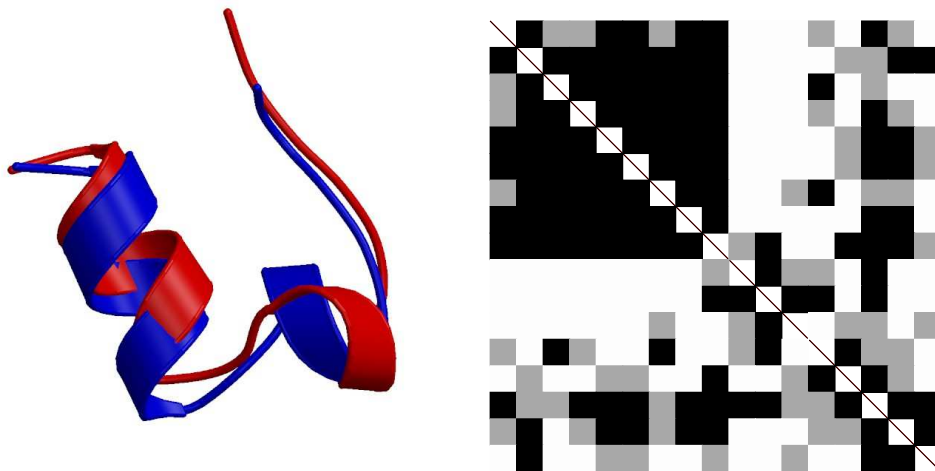


Abbildung 7.7: To the left an overlay of native (red) and best folded structure (blue) during the pt-simulation with 30 replicas of the trp-cage protein. They differ by an RMSD-B value of 2.55 Å. To the right the according  $C_{\beta}$ -matrix.

the strength of  $c$ . The simplest and most obvious choice for  $q$  is the energy  $E$ . However for a given energy different configurations of a protein exist. Therefore we also investigate other parameters and combinations of them, like amount of helicity in the protein, energy and helicity, energy and radius of gyration, energy and end-to-end distance or helicity and end-to-end distance. In addition the temperature  $T$  was varied during the simulations. We quickly learned that the deletion of the whole histogram when entering the high-energy region ( $E > 40 \text{ kcal/mol}$ ) or finding a new local minimum improved the performance of elp. The first re-setting ensures that the system loses all memory of its previous exploration after reaching the high-energy region while the latter increases the time spent on exploring a new local minimum.

In any good global optimization technique the goal is to explore low energy configurations without getting trapped in local minima. Elp accomplishes this task by temporary smoothing locally the energy landscape. The resulting walk into and out of local minima can be seen in figure 7.9a where we show a typical elp run of the trp-cage protein. Here, we have chosen the energy itself as order parameter, i.e.  $f(H(q, t) = cH(E, t)$ , and  $T = 5 \text{ K}$ . The simulation goes over  $10^7$  MC updates. Besides the “physical” energy  $E$  we show also the paving term  $cH(E, t)$  (with  $c = 0.05$ ). As the simulation progresses, the system, driven by the low-temperature, may fall into a local minimum. But unlike a canonical low-temperature simulation it will not become trapped as the paving term  $cH(E, t)$  increases the longer the system stays in this local minimum. Hence, the effective energy barriers decrease and the system will finally escape continuing its search for different local energy configurations. As a consequence, the

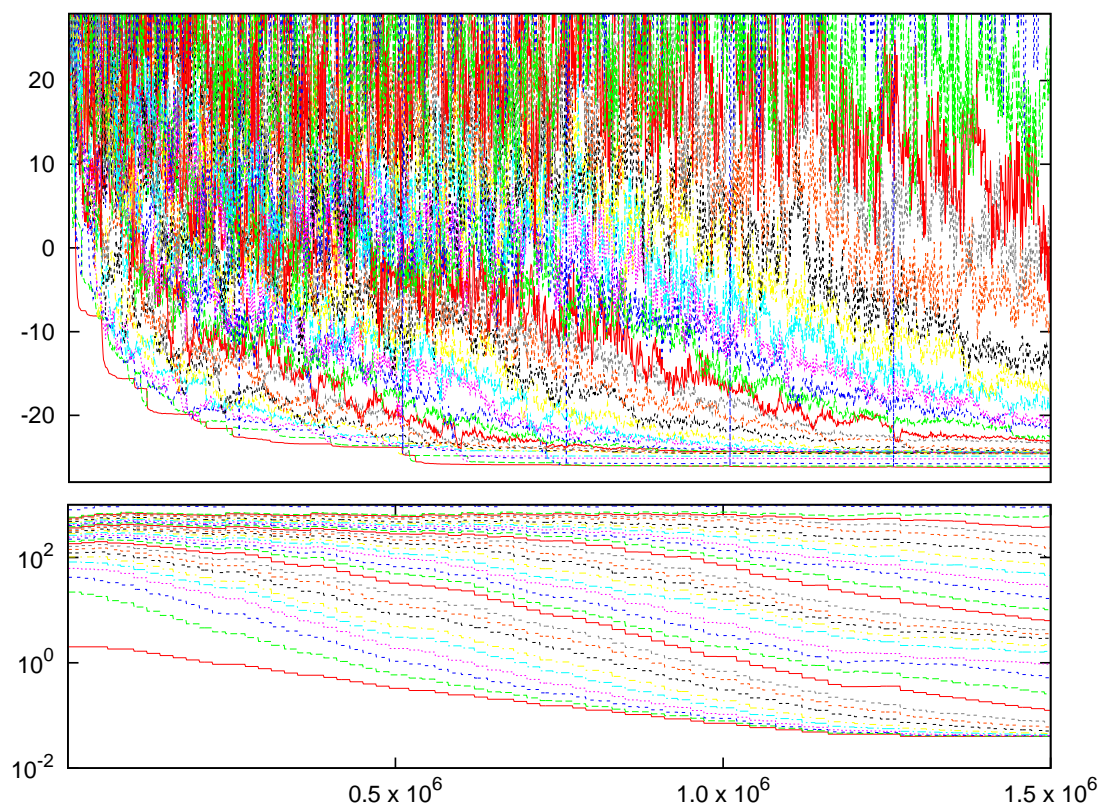


Abbildung 7.8: Energies (upper plot in kcal/mol) and temperatures (in K) of the 30 replicas adapted parallel tempering simulation of the trp-cage protein plotted against the number of steps. Each color stands for a different temperature level. The closeness of the energy-levels in the upper plot indicate high ratios of exchange. The vertical blue lines indicates the replication step. The lower plot demonstrates the change of temperatures to gain stable rates of exchange between neighbouring levels during the course of the simulation.

*effective* energy landscape of a protein is more smooth than the physical landscape. This can be seen in 7.10 where we have enlarged a small part of the above simulation displaying both physical energy  $E$  and effective energy  $E_{elp} = E + cH(E, t)$ . Within this smoother landscape the probability increases to find the native structure. We have shown in 7.11 both the NMR structure of the trp-cage protein (PDB-code 1L2Y) and the lowest-energy configuration ( $E = -18.4$  kcal/mol) found in this specific run. Both conformers differ by a RMSD-B of 3.2 Å which is due to a differences in the unstructured coil of both configurations: neglecting the last two amino acids in the floppy C-terminus reduces the RMSD-B to a value of 1.7 Å .

What is the optimal temperature in an energy landscape paving run? In order to answer this question we have performed elp simulations of the trp-cage protein at various temperatures. The results are shown in figure 7.12 where we display both the mean and the median of the distribution of minimal energies obtained at each temperature in 10 runs of  $10^6$  MC updates. Note that the minimal energies have a broad distribution (resulting in larger error bars for the mean) but decrease with temperature. The differences in energy are minimal and for temperatures above  $T = 50$  K within the error bars. Substantially lower energies are found below this temperature, but both mean and median seem to approach an optimal value of  $E_{min} \approx -15$  kcal/mol at  $T = 5$  K.

The performance of elp also depends on the factor  $c$  by which the histograms are weighted in the calculation of the effective energy  $\tilde{E} = E + cH(E, t)$ . The upper plot in figure 7.12 displays this relation for various factors  $c$  at the “optimal” temperature  $T = 5$  K. On average, lower minimal energies are found when  $c$  decreases until an optimal value of this factor at  $c = 0.05$  is reached. For even smaller factors, the performance of elp worsens again and the values of the minimal energies found increase. While we did not find an expression to determine a priori the optimal value of  $c$ , we note that its value is of order  $k_b T$ . This is reasonable as  $c$  sets the energy scale.

Our above results are obtained for elp runs in which local minima are differentiated only by their energy (type *A* simulations). However, many local minima exist that differ little by their energy. One can therefore expect a better performance of elp if one distinguishes between these minima. We have tested a number of different quantities such as helicity, end-to-end distance, or TPR-cage-distance  $d_{e-e}$ , but found the best improvement for the case where the minima are characterized by both their energy and helicity  $n_h$  (number of amino acids being part of a helical structure), i.e. where  $E_{elp} = E + cH(E, t)$  This result is not surprising as the helicity is a natural reaction coordinate for the trp-cage protein which has only helices as secondary structure elements. We display in figure 7.13 again both mean and median of lowest energies (obtained in 10 elp runs of  $10^6$  updates each) as a function of temperature. As in the previous case where local minima are only distinguished by their energies, the performance of elp increases with decreasing temperature. While the temperature dependence is much stronger an optimal value is again approached once the temperature is below  $T = 10$  K. We remark that the performance also depends on the factor  $c$  with  $c = 0.05$  as an optimal value (upper plot).

Comparing elp runs with the two different effective energy terms we find that differen-

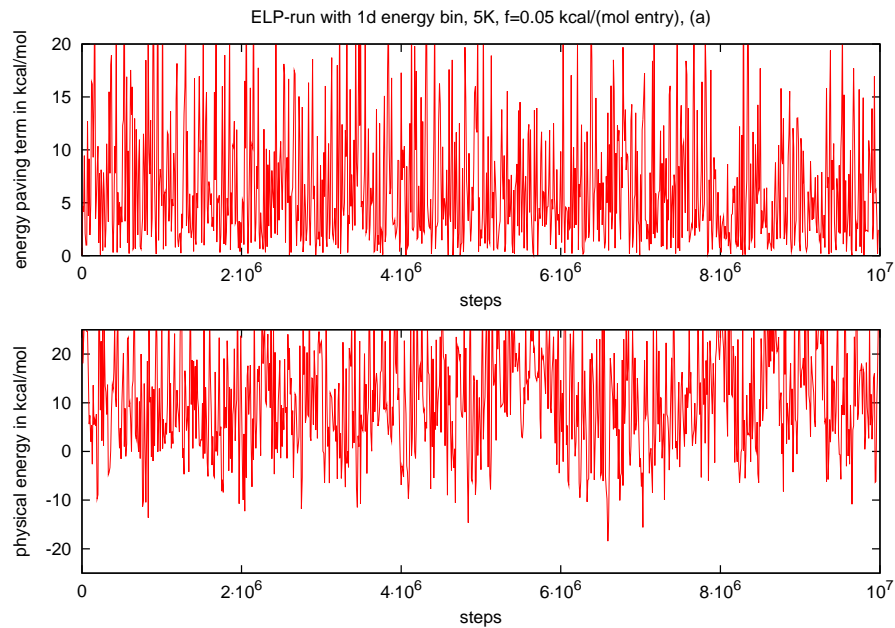


Abbildung 7.9: Typical elp-simulation of the trp-cage protein with PFF01. The temperature was  $T = 5K$ . The horizontal axis plots the number of steps during the simulation, the vertical axis the modified energy  $E_{elp}(E)$  (upper) and the physical energy  $E$  (lower).

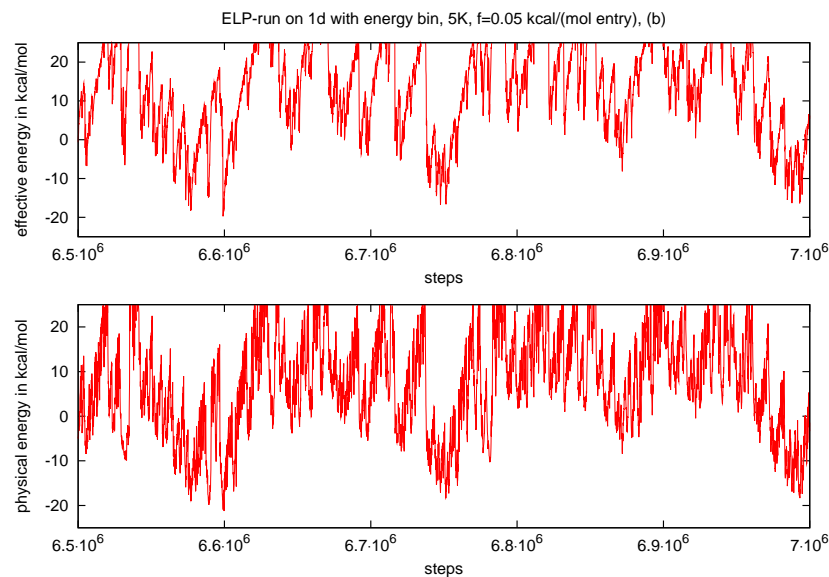


Abbildung 7.10: Here one part of the simulation in figure 7.9 is enlarged. The temperature was  $T = 5K$ . The horizontal axis plots the number of steps during the simulation, the vertical one the modified energy  $E_{elp}(E)$  (upper) and the physical energy  $E$  (lower).

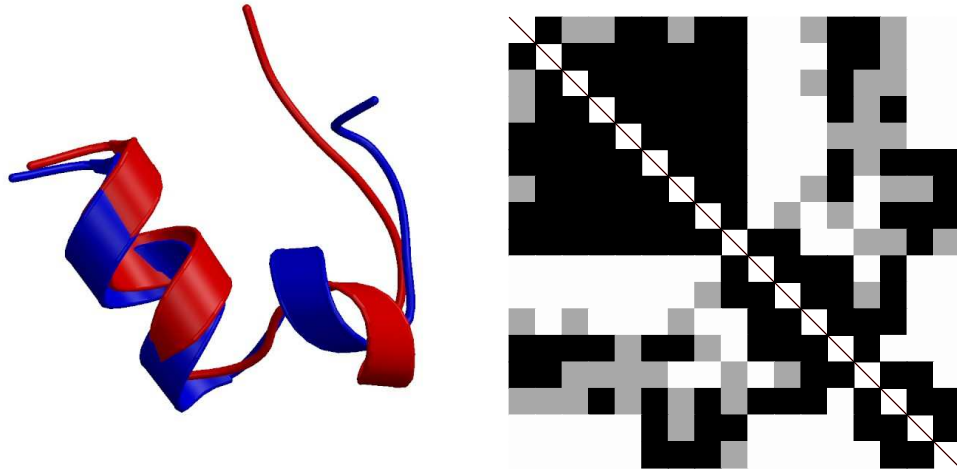


Abbildung 7.11: On the left side an overlay of the NMR-structure (red) and the lowest-energy conformation found in elp-simulation of figure 7.9 (blue) is shown. Both configurations differ by an RMSD-B of  $3.2 \text{ \AA}$  ( $1.7 \text{ \AA}$  when neglecting the last two residues in the sloppy C-terminus). The right side shows the according  $C_\beta$  matrix.

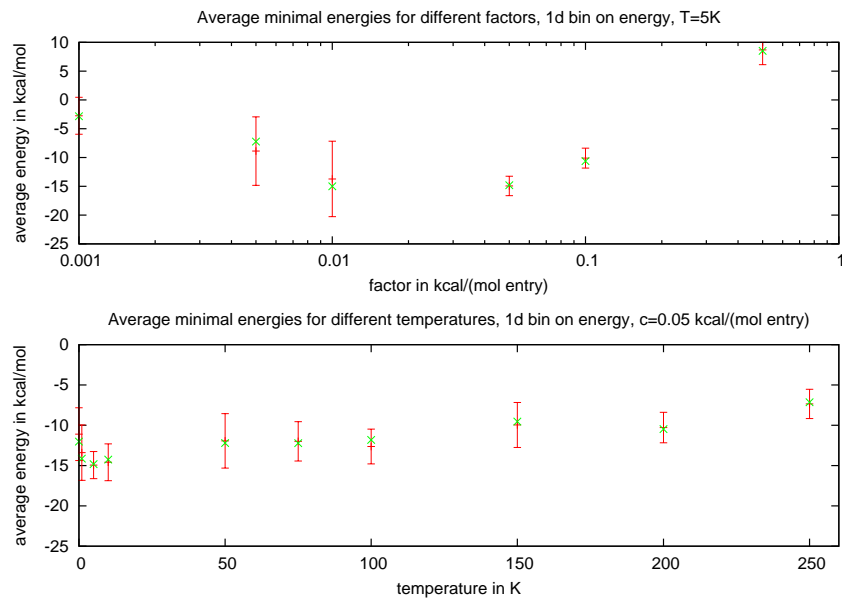


Abbildung 7.12: The upper plot displays the mean and the median of lowest energies measured in 10 elp runs as function of the factor  $c$  for a temperature  $T = 5K$ . The lower plot shows the same quantities but for a factor  $c = 0.05$  and  $E_{elp} = E + cH(E, t)$  as a function of temperature.



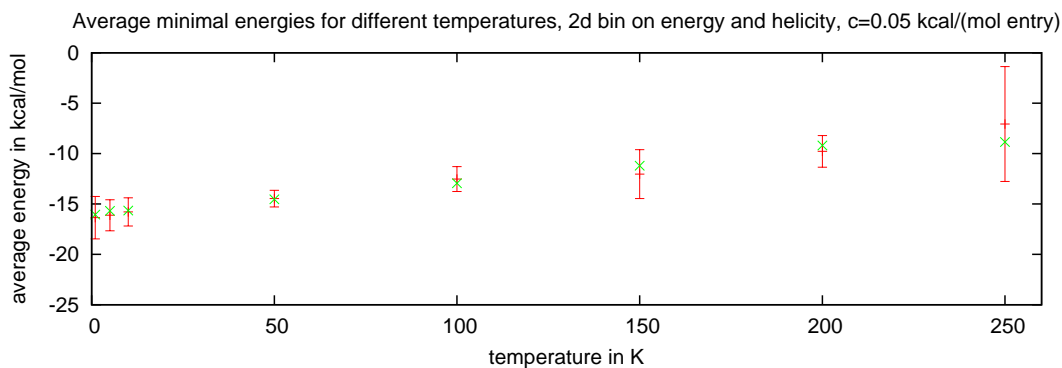


Abbildung 7.13: The plot shows the average energy  $\langle E \rangle$  as a function of temperature the mean and the median of lowest energies measured in 10 elp runs for a factor  $c = 0.05$  and  $\tilde{E} = E + cH(E, n_H, t)$ .

tiating local minima not only according to their energies (type *A* runs) but also according to their helicity (type *B* runs) leads to an improved sampling of low-energy configurations. This can be seen in table 7.5 that summarizes the mean and median of the distribution of lowest energies sampled in the various elp runs. Listed are also the lowest energies ever found in these runs.

elp-variant	$E_{min}$	Mean	Median
$E_{elp} = E + 0.05H(E, t), T = 250K$	-11.13	-7.35	-7.15
$E_{elp} = E + 0.05H(E, t), T = 5K$	-17.07	-14.94	-14.82
$E_{elp} = E + 0.05H(E, n_H, t), T = 250K$	-12.05	-7.07	-8.86
$E_{elp} = E + 0.05H(E, n_H, t), T = 5K$	-17.85	-16.12	-15.68
$E_{elp} = E + 0.05H(E, t), T = 0K$	-14.69	-12.48	-12.93
$E_{elp} = E + 0.05H(E, n_H, t), T = 0K$	-17.78	-14.99	-15.50

Tabelle 7.5: Lowest energy  $E_{min}$  ever found mean and median of the lowest energy distribution in 10 elp runs with the specified parameters.

However, not only are lower energies sampled in type *B* simulations than in such of type *A*, but the low-energy region is also sampled faster. In order to demonstrate this point we have measured in our elp runs the RMSD-B to the NMR structure.  $r_{min}(t)$  is the lowest value of this quantity found till time  $t$ . Its average  $\langle r_{min}(t) \rangle$  (evaluated over 10 elp runs) as a function of MC time is displayed in figure 7.14 for both elp runs with  $cH(E, T)$  (type ‘A’) and  $cH(E, n_H, t)$  (type ‘B’). Here we have chosen optimal values  $c = 0.05$  and  $T = 5K$ . The elp runs with the more discriminating paving term  $cH(E, n_H, t)$  have over the whole time range lower values  $\langle r_{min}(t) \rangle$  than the simple elp runs with  $f(q, t) = cH(E, t)$ . This indicates that on average native-like configurations are sampled earlier. In fact, configurations with a

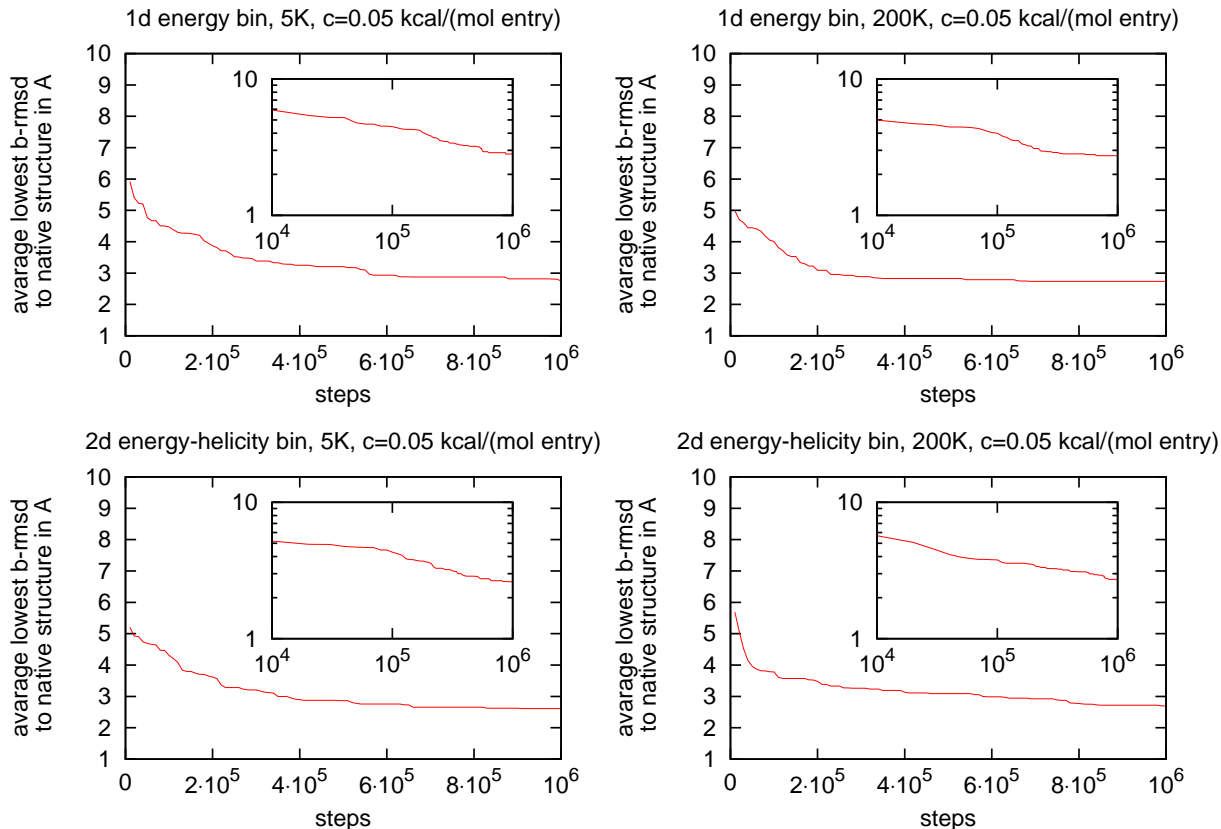


Abbildung 7.14: Investigation of the average minimal backbone RMSD  $\langle r_{min}(t) \rangle$  as calculated from 10 elp runs with  $E_{elp} = E + cH(E, t)$  and such with  $E_{elp} = E + cH(E, n_H, t)$ . In all runs, we have chosen  $c = 0.05$  and either  $T = 5K$  or the  $T = 200K$ . The inset displays the same quantities on a log-log scale.

backbone-RMSD of less than  $3\text{\AA}$  are found in runs of type A on average after  $5.6 \cdot 10^5$  sweeps, while only  $3.5 \cdot 10^5$  sweeps are needed in runs of type B. Note that the decrease in RMSD seems to follow a power law as the data points fall in a log-log plot on a roughly straight line (see inset of figure 7.14). However, the error bars of our data points are too large to allow measuring the exponents.

Besides the size of energy bins and the choice of order parameters to distinguish local minima, elp has two free parameters: the (low) temperature  $T$  in the Boltzmann-weight and the factor  $c$  by which the histogram entries are weighted (we have also investigated the effect of different bin-sizes but these are minor in effect as long as enough bins in physically interesting regions exist). Our analysis of the temperature dependence of elp above prompted us to consider the possibility of a zero-temperature version of elp. For  $T \rightarrow 0$  only moves with

Sequence & secondary structure	Weight	# Atoms	# Atoms in PFF01	RMSD-B
<i>TADNKFNKEQQNAFYEILHLPNLNEEQRNG CCCCCCCCCHHHHHHHHCCCCCHHHHHH FIQSLKDDPSQSANLLAEAKKLNDAPKA HHHHHHHCCCCCHHHHHHHHHHHHHHHHCCC</i>	6753 <i>D</i>	941	595 (*)	2.69 (*) Å

Tabelle 7.6: The basic data about Protein A (1BDD). The structure is given in 1-letter-code, the secondary structure in 3-state-code. (\*) indicates numbers for our simulations, in which part of the protein was cut off. These are shown in italics. Due to the length of the amino acid chain the first column was cut in half after the first 30 amino acids. The last column gives the RMSD-B from the native state to the best folded conformation. Please note that due to the narrow energy gap between the folded state and a dissimilar state by now we do not consider 1BDD folded ab-initio correctly.

$\Delta\tilde{E} \leq 0$  will be accepted. This leads to an acceptance criterion:

$$\Delta E + c\Delta H(q, t) \leq 0 \leftrightarrow c\Delta H(q, t) \leq -\Delta E \quad (7.1)$$

where  $E$  is the physical energy. Hence, within elp the system can even at  $T = 0$  overcome any energy barrier. The waiting time for such a move is proportional to the height of the barrier that needs to be crossed. Note that the factor  $c$  sets now only the time scale and in this sense the  $T = 0$  form of elp is parameter-free. We have plotted two examples for such 0K elp run in figure 7.15. ( $E_{elp} = E + cH(E, t)$ ) and figure 7.16 ( $E_{elp} = E + H(E, n_H, t)$ ). Energy  $E$  and helicity  $n_H$  are shown as a function of MC time. Both figures illustrate that even at  $T = 0$  K elp allows the crossing of energy barriers and the sampling of large parts of the configuration space.

As in the case of elp runs at finite temperature, lower energies are found when the paving term discriminates local minima not only according to their energy but also according to their helicity. The respective values for mean and median of the lowest energies are listed in table 7.5. While the minimal energies are slightly higher than the optimal ones found at finite temperatures, these differences are within the error-bars. Similarly we find little difference between the optimal finite temperature runs and the 200K-versions of elp in the plots of  $\langle r_{min}(t) \rangle$  in Fig. 7.14. We also find that 0 K - elp runs have a comparable efficiency with finite temperature runs (at optimal chosen temperatures) in locating native-like configurations. More important than the temperature appears to be the construction of the paving term that has to distinguish local minima according to coordinates describing the system best (as the helicity in our case).

### 7.3 Protein A

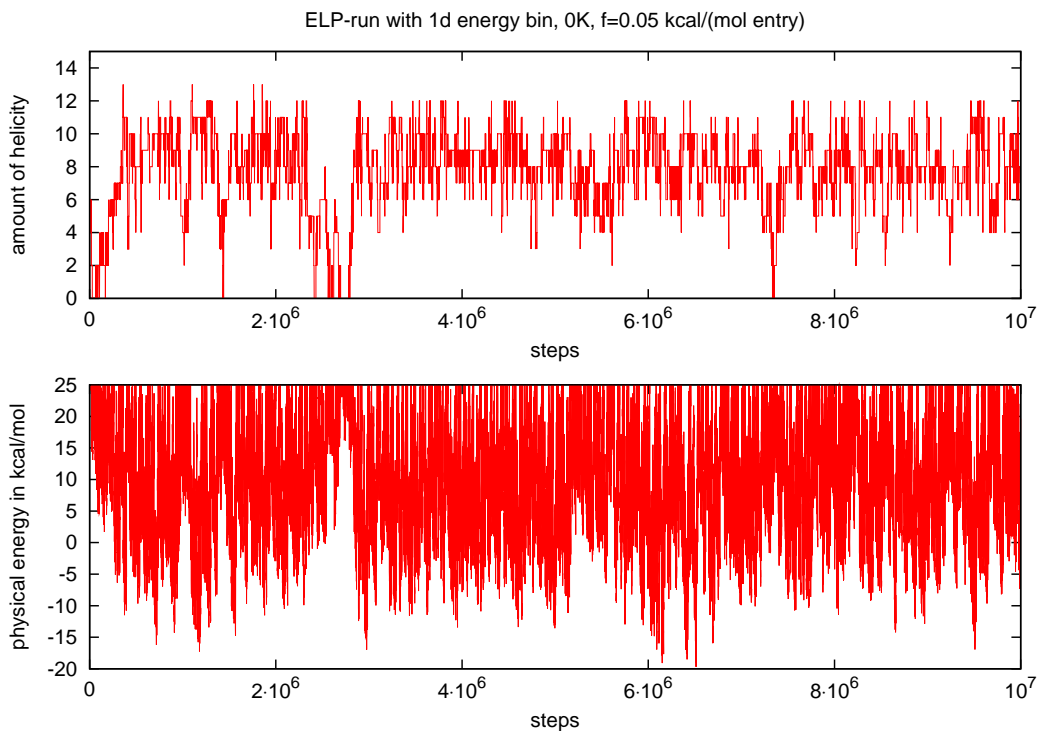


Abbildung 7.15: Simulations with the 0K-elp-variant of the *trp*-cage protein using the forcefield PFF01. Here we applied  $E_{elp} = E + cH(E, t)$ . Shown are the physical energy  $E$  and the helicity  $n_H$  as function of Monte Carlo steps.

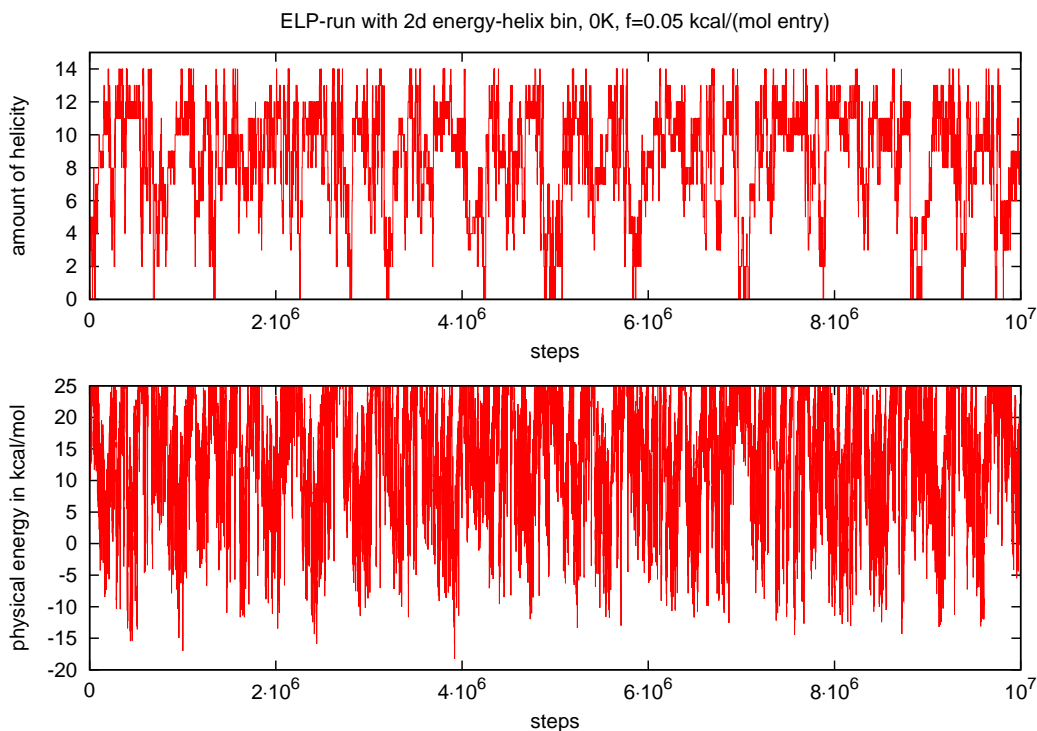


Abbildung 7.16: Simulations with the 0K-elp-variant of the *trp*-cage protein using the forcefield PFF01. Here we use  $E_{elp} = E + cH(E, n_H, t)$ . Shown are the physical energy  $E$  and the helicity  $n_H$  as function of Monte Carlo steps.

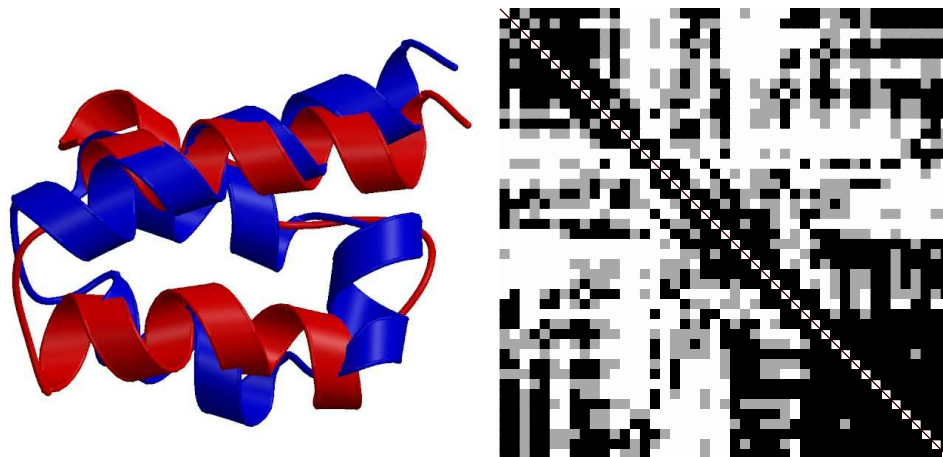


Abbildung 7.17: Shown is on the left an overlay of the native structure (red) and the best structure found during the simulations for Protein A (1BDD). For these runs the first amino acids of the N-terminus have been cut off as described in the text. The RMSD-B is 2.67 Å. On the right the according  $C_{\beta}$ -matrix is shown.

Protein A has proven to be a major challenge in prediction with the forcefield PFF01. To lower computational demands we cut off part of the sloppy N-terminus. Running simulations with different optimization methods Protein A proved highly expensive in computational time. The simulations provided structures both similar and dis-similar to the native state. One structure alike the experimentally measured conformation is shown in figure 7.17. It is the lowest energy structure we gained during these simulations. However we also found dis-similar structures within 1 kcal/mol which differ structurally strongly. The RMSD-B between these competing structures and the native structure is in the order of 9 Å [53]. The small energy difference presented does not allow a rational distinction between these states. Due to the limitation of stochastic optimization methods the energy gap must be sufficient large to allow judging the quality of prediction. Presently we do not consider the Protein A being reproducibly folded with PFF01 which may change as additional simulations are run.

## 7.4 HIV-Accessory Protein

We performed simulations on the HIV-Accessory Protein using its 40 amino acid head-piece. We removed part of the sloppy C-terminus and applied adaptive parallel tempering (apt)[106] as well as basin hopping[54]. We found in both cases lowest energy structures in excellent agreement with experimental data (fore example apt-RMSD-B of 2.46 Å). An over-

Sequence & secondary structure	Weight	# Atoms	# Atoms in PFF01	RMSD-B
QEKEAIERLKLALGFEEESLVIQAYFACEKNE CCHHHHCHHHHCCCCCHHHHHHHHHCCCCC NLAANFLLS <i>QNF</i> DDE HHHHHHHHHCCCCC	5162 <i>D</i>	713	(*) 392	(*) 2.46 Å

Tabelle 7.7: Key parameters about the headpiece of the HIV-Accessory Protein 1F4I. The structure is given in 1-letter-code, the secondary structure in 3-state-code. (\*) indicates parameters for our simulations, in which part of C-terminus was cut off. This is indicated by the italics in the sequence. Due to the length of the amino acid chain the first column was cut after the first 30 amino acids to a second line. The last column gives the RMSD-B between the native state and the best folded conformation.

Sequence & secondary structure	Weight	# Atoms	# Atoms in PFF01	RMSD-B
MLSDEDFKAVFGMTRSAFANLPLWKQQLK CCCHHHHCCCCCHHHHCCCCCHHHHHHHH KEKGLF HHCCCC	4172 <i>D</i>	596	364	3.56 Å

Tabelle 7.8: Key parameters for the Villin Headpiece 1VII. The structure is given in 1-letter-code, the secondary structure in 3-state-code. Due to the length of the amino acid chain the first column was cut after the first 30 amino acids to a second line. The last column gives the RMSD-B between the native state and the best folded conformation.

lay is shown in figure 7.18. All low-energy conformations gained during the simulation are close to the native state with RMSD-B values between 2 and 3 Å. The protein seems to be very stable in the forcefield PFF01. No competing structures appear within 5 kcal/mol of the best estimate of the global minimum which have an RMSD-B of greater than 4.

## 7.5 Villin Headpiece

The Villin headpiece was investigated in different protein studies. An overview about its basic properties is given in table 7.8. It is the protein the forcefield PFF01 was originally developed for[50, 49, 53]. In prior versions of PFF01 several non-native structures competed with the native state. We therefore pursued a rational decoy approach to improve our forcefield in several iterations. We generate a large set of good candidates that energetically compete with the native conformations. As long as one of these decoys has a better energy than the native conformation, the forcefield was modified to stabilize the native conformation in

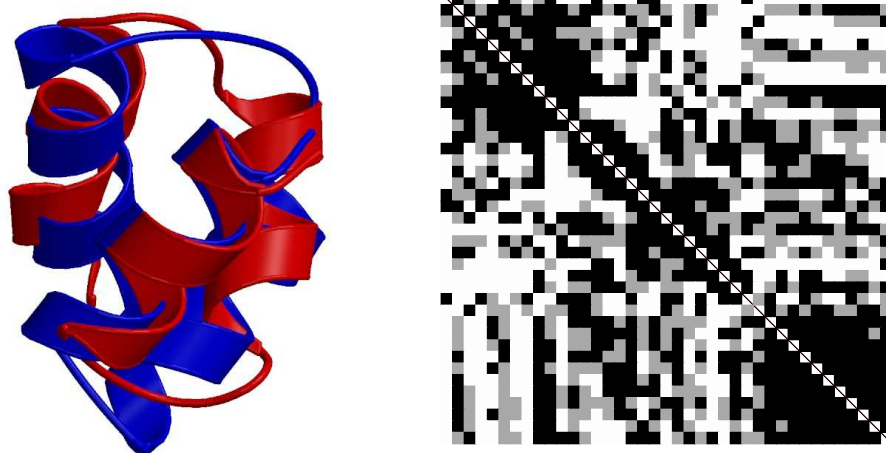


Abbildung 7.18: Shown is an overlay of the experimentally measured native structure (red) and the best structure found during the simulations of the HIV-Accessory Protein 1F4I (left picture). For these runs the first amino acids of the N-terminus have been cut off like described in the text. The RMSD-B is 2.46 Å. The right picture shows the according  $C_{\beta}$ -matrix.

comparison to all other decoys. When this was achieved we generated a new set of decoys by re-folding the Villin headpiece, which we again compared with the native conformation. After several iterations of optimizing the parameters we found a final set of parameters which provided a better (i.e. lower) energy for the native conformation than for all other competing structures found during the simulations. This one set of parameters is now applied to all other simulations presented here[49, 48, 53]. An overlay of the structure corresponding to the best estimate found in the simulation for the Villin headpiece is given in figure 7.19.

## 7.6 Bacterial Ribosomal Protein L20

The Bacterial Ribosomal Protein L20 (1GYZ) [98] is the biggest protein (60 amino acids) folded by an all-atom free-energy forcefield approach by now. Even with the simplifications PFF01 does for modeling protein structure in a free-energy forcefield the simulations needed about 40 accumulated CPU-years before no further improvement in free-energy was noticed. We applied our evolutionary method on a server-client based system for these simulations. Some basic information about the Bacterial Ribosomal Protein L20 is given in table 7.9. An overlay of the best structure found in our simulations with the experimentally determined native state is displayed in figure 7.20.

The Bacterial Ribosomal Protein L20 is by now the only protein for which we extensively tested the evolutionary algorithm[104] by a server-client model. We did this in three phases. In

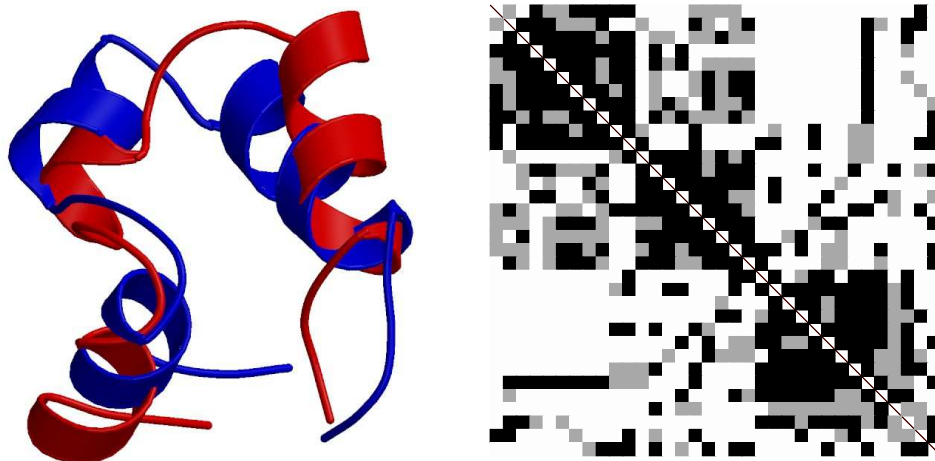


Abbildung 7.19: Shown is an overlay of the native structure (red) and the best structure found during the simulations for the Villin Headpiece 1VII (left picture). The RMSD-B is 3.65 Å. On the right is the according  $C_\beta$  matrix.

the first phase we created a starting set of conformation by running standard MC-simulations at high temperatures (400 K) on random starting conformations. These simulations were sorted by individual energy terms (Lennard-Jones, hydrogen bonding, solvent interaction and electrostatics) to find structures for further optimization. We ranked them by these energy terms, took for each the best 50 structures and eliminated duplicates. We gained a starting set of 266 structures which we optimized by normal sa-simulations with increasing numbers of steps (second phase). Whenever a client indicated available resources we randomly chose one of the set of structures and ran a sa-optimization simulation on the client for it. The resulting conformation was sorted in again by comparison with the existing set of structures. We chose the one closest in RMSD-B and, when this values was lower than 3 Å, discarded the structure with the higher energy. When no structure was close in RMSD-B we compared with the structure worst in energy of our present set and discarded one of these two after their energetic comparison. This scheme results in both conformational diversity and good optimization of the energy as shown in figure 7.21. Also conformations far from the native state are assumed during the evolutionary algorithm. In a third phase we selected the 50 structures best in energy for further refinement. We gained a resulting structure best in energy in high agreement with the experimentally measured native conformation. Additionally in the final set of conformations six out of the ten energetically lowest conformations represented the native state of the protein. We further noticed a significant increase of native content during the simulations underlining the existence of a folding funnel. We also tested different selection schemes for new simulations. We biased the starting conformations for further optimization



Sequence & secondary structure	Weight	# Atoms	# Atoms in PFF01	RMSD-B
WIARINAAVRAYGLNYSTFINGLKKAGIEL CCHHHHHHHCCCCCCHHHHHHHHHHCCCC DRKILADMAVRDPQAFEQVVNKVKEALQVQ CCCCCHHHHHHCHHHHHHHHHHHHHHHHCCC	6728 <i>D</i>	973	593	4.64 Å

Tabelle 7.9: The basic data about the Bacterial Ribosomal Protein L20 1GYZ. The structure is given in 1-letter-code, the secondary structure in 3-state-code. Due to the length of the amino acid chain the first column was cut after the first 30 amino acids to a second line. The last column gives the RMSD-B from the native state to the best folded conformation.

according to their energetic ranking. The different selection scheme provided no significant difference to a un-biased selection.

It showed that our ten energetically best conformations resulted from starting conformations selected by hydrogen bonding. This can be interpreted, that this energy term is highly correlated to secondary structure. Obviously it proved difficult to form secondary structure (i.e. here helices) when a compact structure has already formed. The resulting best structure started from a conformation with an RMSD-B of 11.65 Å to the native state indicating that no native content was included in this starting structure.

## 7.7 Discussion on the Efficiency of Optimization Techniques

To compare the different optimization techniques[123] we chose the simulations of the trp-cage protein for which we have the most data. Simple *Monte-Carlo* simulations failed in global optimization of the free-energy function. Using simulations with  $10^7$  steps at temperatures between 50K and 250K, low-lying energies were not found (lowest energies were higher than  $-10$  kcal/mol, simulations not shown). Also no near-native structures were found even at higher energies. Similarly standard *simulated annealing* (sa) simulations did not succeed in prediction of the native state (again no structures with total energies below  $-10$  kcal/mol, simulations not shown).

The application of a simple *genetic algorithm* (ga) proved difficult. The analysis of ga-simulations indicated two major protein problems. Often the crossing and mutation of individuals provided clashing conformations. Repairing these clashes was computationally highly expensive. Also the local relaxation of structures after each generation further raised these high computational demands. Though they have been successfully applied to protein folding[97] our simulations indicate ga being computationally too demanding in comparison with other techniques.

The techniques *stochastic tunneling* (stun), *evolutionary algorithm* (not shown on 1L2Y),

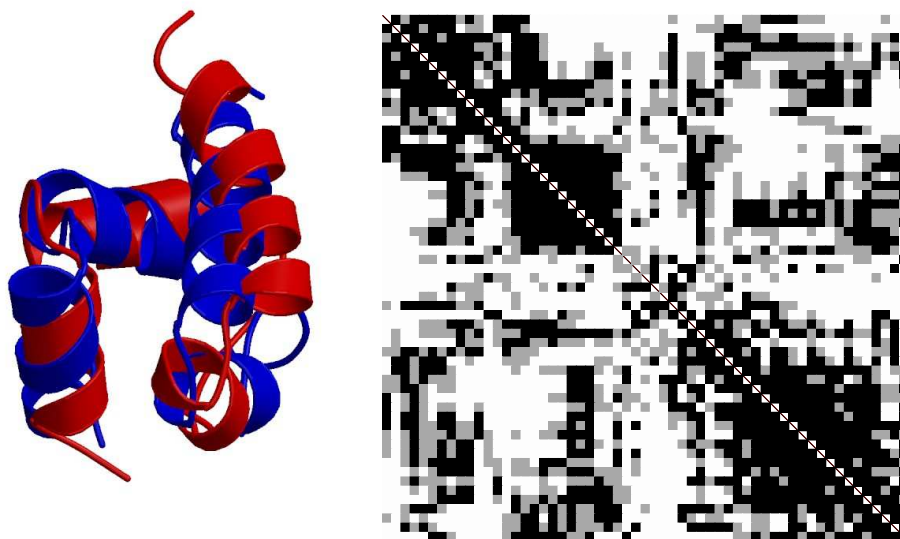


Abbildung 7.20: Shown is on the left picture an overlay of the experimentally measured native structure (red) and the structure corresponding to the best estimate of the global minimum found during the simulations for the Bacterial Ribosomal Protein L20 1GYZ. The RMSD-B is 4.64 Å. The right picture shows the according  $C_{\beta}$ -matrix.

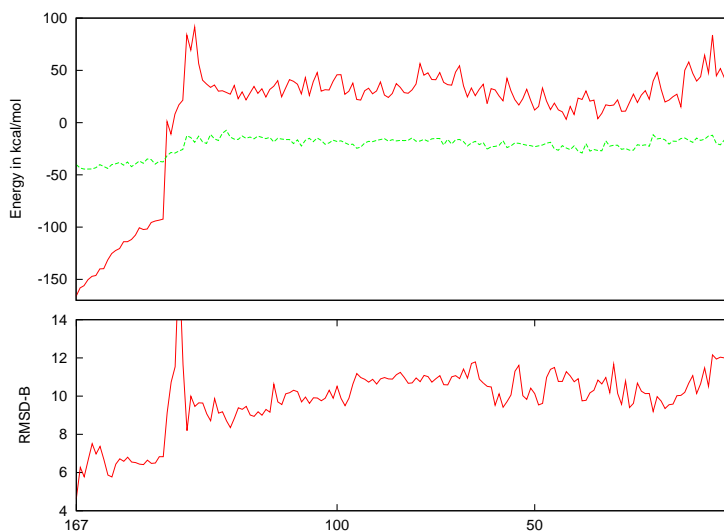


Abbildung 7.21: This graph gives the energy (in kcal/mol, upper plot) and the RMSD-B (in Å lower plot) of the lowest energetic structure of 1GYZ found by the evolutionary algorithm. The horizontal axis gives the number of accepted runs from the end (left) to the start (right). In the upper plot both the total energy (red) and the hydrogen bonding energy (green) are displayed. Later runs are longer sa-simulations to allow for further improvement in energy.

*basin hopping* (bh), *energy landscape paving* (elp) and *adapted parallel tempering* (apt) all succeeded in prediction of the native state from random initial conditions. In comparison the different techniques show each some advantages compared to the other techniques. The evolutionary algorithm is not very efficient in its use of computational power but can use non-homogeneous computer networks with unreliable network connections. It produced very good estimates of the global minimum ( $-27.3$  kcal/mol). Stun and apt use comparable amounts of CPU-time to generate low-lying energies ( $-25.79$  kcal/mol,  $-25.61$  kcal/mol respectively) but require a careful tuning of the various free parameters. In stun the temperature and the transformation parameter  $\gamma$  for its non-linear transformation are quite fickle. It is difficult to find their optimal values to run the technique at peak efficiency. Similarly, the efficiency of apt depends strongly on the parameters for adjusting the temperatures. However this automatic adjustment of temperatures showed as strong improvement compared to normal pt. elp generally generates worse estimates of the global minimum ( $-22.97$  kcal/mol). However its parameters for minimization are few and very robust. In the 0K-variant of elp only bin size and the order parameter(s) of the histogram remain. We experienced few runs which were locally entrapped. The 0K variant is esp. interesting since it is totally eliminating the artificial temperature from the simulations.

We also investigated whether the conformational space is explored thoroughly by the different techniques. All these methods, stun, apt, elp and apt created strongly dislike conformations during the simulations. RMSD-B values of the sampled conformations differed by more than  $6 \text{ \AA}$  even after the initial hydrophobic collapse.

From the compared techniques basing hoppin provided the best estimate of the global minimum ( $-29.22$  kcal/mol). Also the starting and ending temperature of each sa-simulation, the energy threshold for accepting such a simulation and the number of steps per simulations are few and robust parameters. The values do not need to be optimal to provide a careful exploration of conformational space.

The stability of the trp-cage protein proved remarkable. The above energy values all represent this protein in its native ensemble but differ in the quality of locally optimizing this structure. Therefore the quality of local relaxation seems to be the major difference between the applied optimization methods.

This robustness of the native conformations underlines the possible existence of a folding funnel in the energy landscape as illustrated in figure 7.22. The vast amount of low-lying structures are at some point in conformational space connected to the cluster containing the lowest energy structure.

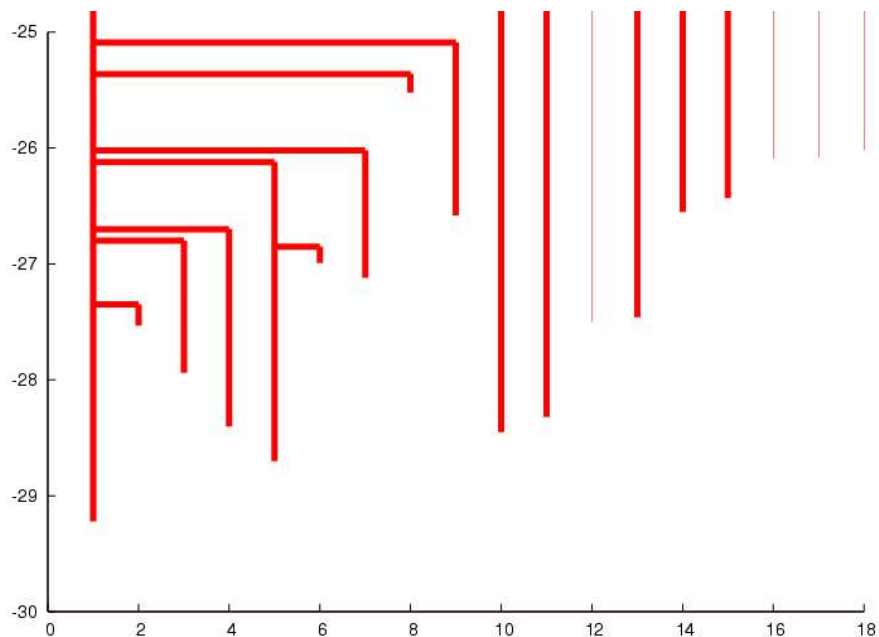


Abbildung 7.22: Illustration of the low-lying energy landscape of the trp-cage protein in the forcefield PFF01. The horizontal axis counts different clusters of conformations while the y-axis gives the lowest lying energies within such clusters in kcal/mol. In this picture only the conformations belonging to low-lying energies below -25 kcal/mol are shown which are around 1000. As the clusters accumulate structures they may unite. The thick lines designate clusters which unite at higher energies and are separated by barriers in between. The thin lines designate clusters which remain isolated in conformational space. This picture was created with about 6000 low-energy structures gained from different simulations. For each structure the RMSD-B values were generated against all other structures. Two structures belong to the same cluster if any energetically lower lying structure in this cluster has an RMSD-B to this structure of below 2 Å .

# Kapitel 8

## Summary

The work in this thesis was motivated by two central questions:

- Protein folding is ultimately governed by complicated quantum-mechanical effects, such as the formation of hydrogen bonds, Fermi-repulsion of electronic clouds and interaction of protein surface with a complex environment: *Can the folding of a protein be understood and represented by a classical free-energy-forcefield and, if yes, how can it be done in a computationally treatable way?*
- Proteins have many degrees of freedom and no exploitable symmetries. It is known that global minimization of rough and high-dimensional energy landscapes, like those in spin-glass theory, is very difficult. Therefore: *Due to the complexity of such a forcefield, are there optimization methods allowing to find the global minimum and what about the efficiency of these methods?*

PFF01[50, 49, 53, 48] combined with efficient stochastic optimization methods may be suitable to give one possible answer to these questions. We explored energy landscape paving (elp)[109], stochastic tunneling (stun) [105], temperature adopted parallel tempering (apt)[52, 107, 106, 104], distributed computing [108] and the basin hopping method[54]. The classical free-energy forcefield PFF01 models the interactions in a protein. Quantum-mechanical effects like hydrogen bonding which are relevant during the folding process of proteins are included in a way accurate enough to allow predictive folding of helical proteins but simple enough to be handled by present day computational resources. Much work and effort has been put to the task of finding the global minimum in such a high-dimensional conformational space by means of stochastic optimization methods. The results from the simulations combining the forcefield PFF01 with these optimization methods are presented in this thesis. They are in high agreement with experimental measurements of protein structure. We have predictively ab-initio in-silico folded the trp-Cage protein (1L2Y, 20 amino acids) [105, 107, 104, 107], the HIV-accessory protein (1F4I, 40 amino acids) [50, 54, 52, 51, 106, 104], the Villin headpiece (1VII 36 amino acids) [50, 49, 54] and the Bacterial Ribosomal Protein L20 (1GYZ, 60 amino acids) [104]. Thus both the *validation of the forcefield* for several non-

homologous proteins and the *practicability of the approach* of predicting protein structure by minimization of the free-energy forcefield PFF01 has succeeded.

Once we could establish that this forcefield correctly predicts the native states of several helical proteins, we can apply it to understanding the process of protein folding. The four major contributions to the free energy in PFF01 are

- Lennard-Jones
- Implicit solvent interaction
- Hydrogen bonding
- Electrostatics

During the simulations we noticed the major energy difference between different low-energy structures resulting from hydrogen bonding in competition with the implicit solvent interaction. The solvent interaction leads to the collapse of the whole structure to compact conformations while the hydrogen bonding is the stabilizing term for secondary structure. Lennard-Jones interactions mainly prevent the clash of atoms while electrostatics provides only minor effects in this model. The *quantitative model* arising from PFF01 confirms arguments concerning the importance of competition between hydrogen bonding and solvation effects for protein folding[25, 26]. One major advantage in the consideration of all-atom forcefields, such as PFF01, arises from the complete free-energy landscape, including the characterization of the native conformation in accordance with experimental measurements. The applicability of an implicit solvent model, which has also been subject to some debate, is justified by the results of the model. We note however, that the application of such a model appears to limit the overall resolution of the model to 2-4 Å, depending on the protein studied.

In addition these results also confirm the new view of protein folding, where a *folding funnel*[27, 44, 47, 89, 90, 109] characterizes the landscape in vicinity of the native state. A folding funnel embodies the concept that a major part in conformational space has a strong bias towards the native state. The existence of a folding funnel was postulated to overcome Levinthals paradox[68, 27, 55], which stipulated that the conformational space of a protein is too large to be searched in a reasonable time, as exemplified by the following short model calculation. The trp-cage protein with its 20 amino acids has in our model 63 degrees of freedom (angles allowing rotations). If we estimate that each angle can assume only three spatially allowed settings we gain a resulting conformational space of  $3^{63} \approx 10^{30}$  conformations, which is far too large to be explored by a random walk.

Translated into computational effort we note that our code presently evaluates about 1,000,000 energies per day for the trp-cage protein on standard off-the-shelf hardware. An enumerative search of the entire conformational space would require

$$\frac{3^{63} \text{Conformations}}{1,000,000 \frac{\text{Conformations}}{\text{day}}} \approx 3 \cdot 10^{21} \text{ years}$$

indicating that some guiding mechanisms must exist to permit the folding of the trp-cage protein with PFF01 in only about 1 CPU week. In this time only a very tiny portion of conformational space could have been explored. For the other, larger proteins investigated the size of the overall conformational space is even more daunting. *There must be a strong bias towards the region of conformational space representing the native state of the protein. In literature this bias is called the folding funnel.*

Because proteins are products of natural evolution designed both towards stability and functionality[9], one may speculate that they are selected by choosing versatility, stability and fold-ability, as implied by the concept of the funnel mechanism in conformational space. The existence of this folding funnel biases a development of native or native-like formations of parts of the protein starting from each possible starting structure. It was demonstrated in many experiments that protein folding is a reversible process. Therefore it might be possible that proteins get stuck in *kinetic traps* during the folding process[8], which are energetically low-lying and accessible regions in conformational space surrounded by high energetic barriers. To leave them a protein has to unfold out of the trap and re-fold again into the folding funnel. However the quick folding of many proteins suggests a single, smooth folding funnel for these proteins. It seems unlikely that random amino acid sequences of proteins have these properties. More likely they offer a multitude of metastable states mostly consisting of unstructured coils and not one selected native ensemble. This indicates that *the speed of protein folding is not a simple function of amino-acid length but is dependent on the topology of the energy landscape.* A smooth folding funnel allows the rapid development of native structure which results in short folding times. Kinetic traps or a general roughness in the energy landscape lengthen folding time.

To allow a representation of the folding funnel we chose the trp-cage protein for which we did a huge number of investigations. We sampled the conformational space in two experimentally measurable coordinates, the amount of helical content and the end-to-end distance of N- and C-terminus[109]. Figure 8.1 shows a funnel-like structure biased towards the native state in the free-energy landscape.

## Outlook

The results presented in this thesis lead to further questions in protein folding. First of all remains the task of also including  $\beta$ -sheets in the forcefield PFF01. This forcefield is now well tested for helical proteins but at the same time helical structure seems to be slightly overemphasized as secondary structure element in comparison with  $\beta$  structures. Another important point is testing the forcefield for other, larger proteins with different folds. Furthermore remains a possible analysis on the dynamics of protein folding. By now we only evaluated the final point of our simulations and compare it with the native state. However one could do a dynamics analysis of the low-energy decoy tree by assuming a diffusional process for related structures using the master-equations[20]. This would allow to answer questions related to two- or many-state folding or the existence of folding intermediates.

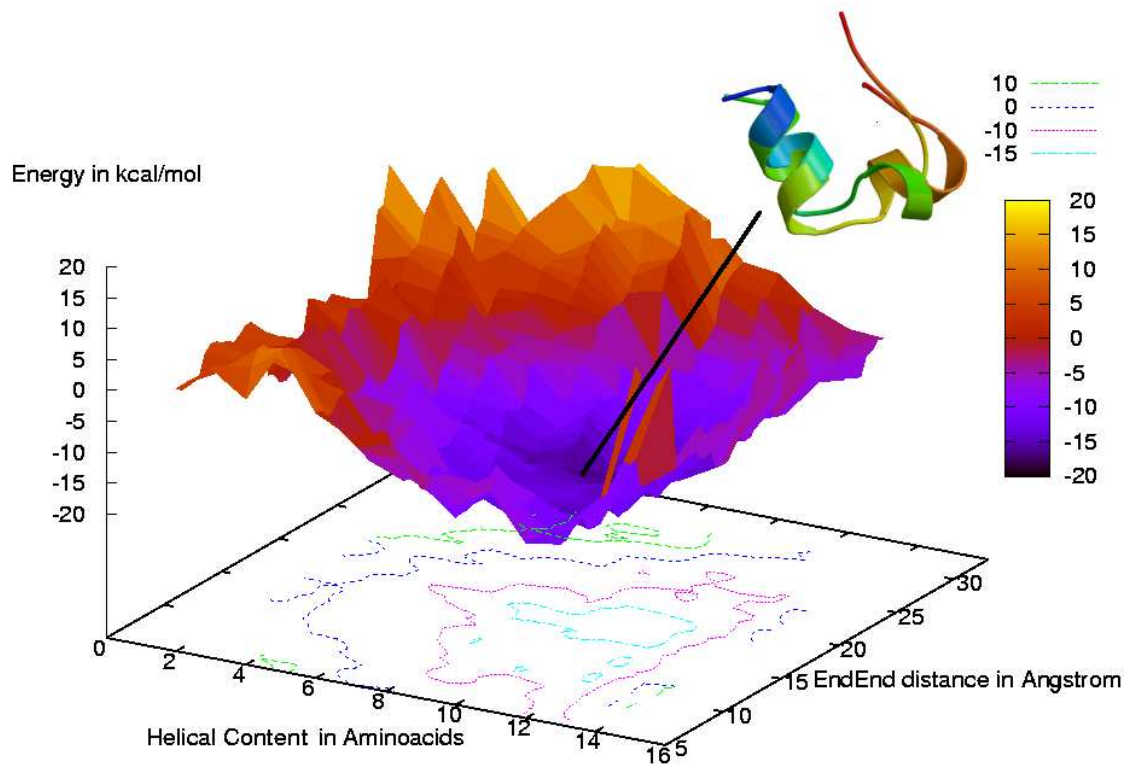


Abbildung 8.1: Energy landscape of the *trp*-cage protein as obtained from a  $0K$  ELP run with a random configuration as starting point. As inset the structure corresponding to the lowest found energy is displayed as overlay with the NMR-structure. The backbone RMSD is  $3.2 \text{ \AA}$  ( $2.2 \text{ \AA}$  when neglecting the last two residues in the floppy C-terminus)[109].



# Anhang A

## Used programs and definitions

### Absolute Contact Order

The absolute contact order (aco) is the mean separation in sequence by contacting atoms:

$$\text{aco} = \frac{1}{M} \sum_{\text{Contacts}} \|i - j\|$$

( $M$  is number of atomic contacts,  $i - j$  separation in sequence between residues  $i$  and  $j$  with distance between atoms less than  $0.6nm$ ). The relative contact order (rco) is the absolute contact order divided by the number of residues  $N$ :  $\text{rco} = \text{aco}/N$ . The contact order has a close relation with the folding time of a protein. However it is still subject of discussion whether the absolute or relative contact order is stronger correlated with folding times [59, 63, 19].

### $C_\beta$ -matrix

A  $C_\beta$ -matrix allows a quick optical comparison of two structures. To generate it the relative distances of all  $C_\beta$  atoms of the two compared structures are calculated for each of the structures. Then the difference for each entry in these two distance matrices is taken. If this difference is less than  $0.75 \text{ \AA}$  the according dot on the  $C_\beta$ -matrix is turned black, for differences between  $0.75 \text{ \AA}$  and  $1.5 \text{ \AA}$  it is turned grey and white for relative distances greater than  $1.5 \text{ \AA}$ .

This  $C_\beta$ -matrix gives a good overview about secondary structure relation between the two compared structures since the position of the  $C_\beta$  is equal if the secondary structure is equal. Also this matrix can serve as indication for the position of the sidechains (because we're taking the  $C_\beta$  and not  $C_\alpha$  atoms). Slightly problematic is the choose of the small intervals of  $0.75$  and  $1.5 \text{ \AA}$  since often experimental errors are larger. However these values also allow a close comparison between different structures gained from our simulations.

## DSSP[61]

DSSP was used to characterize the secondary structure of proteins. The secondary structure was afterwards further refined towards three-state characterization. *H* ( $\alpha$ -Helix), *G* (3/10-helix) and *I* ( $\pi$ -helix) are considered as helix *H*. *E* (extended  $\beta$ -strand) and *B* (residue in isolated  $\beta$ -bridge) are considered as  $\beta$ -strand *E*. The other, i.e. *T* (hydrogen bonded turn), *C* (coil) and *S* (extended strand), are considered as coil *C*.

## Generation of a Random Starting Structure

In order to generate a random starting structure the program package POEM can be used. One possibility is putting big artificial equal charges on the N- and the C-termini and running a short simulation. Another possibility is running a simulation at a very high (for example  $10^{10}$  K) which will totally randomize the structure since no energy bias except prevention of clashing conformations exists anymore. Yet another possibility is using the function randomize in POEM which will apply random changes to all degrees of freedom while still fulfilling the Fermi-repulsion.

## Gnuplot

All the graphs in this thesis were created with Gnuplot.

<http://www.gnuplot.info/index.html>

## L<sup>A</sup>T<sub>E</sub>X

The published format of this thesis was created using L<sup>A</sup>T<sub>E</sub>X.

<http://www.latex-project.org>

## Molscript

The overlay pictures were created with Molscript.

<http://www.avatar.se/molscript/>

## PDB-database

The PDB (Protein Data Bank) lists around 30,000 structurally measured proteins (March 2005). It is available for free in the Internet[12].

<http://www.rcsb.org/pdb/>

## POEM

The program package POEM (Prediction by Optimization of an Energy Model) was used for running all of the described simulations. It is a complex program allowing manipulation of protein structure, calculation of energies in the forcefield PFF01 and running different optimization methods. It is written in C. For compilation the gcc of Linux was used. It includes a cross-compiling version for windows which allows calculation as a screen saver using a simple server-client model.

## RMSD

The RMSD (Root Mean Square Deviation) between two proteins is

$$RMSD = \min\left(\sqrt{\frac{\sum \text{atoms} \Delta x^2 + \Delta y^2 + \Delta z^2}{\# \text{ Atoms}}}, *\right)$$

(\*) denotes all possible spatial translations and rotations of the two proteins against each another, which means an ideal superimposition of both structures. It is a common method to quantize the similarity of two structures. The sum over all atoms can be replaced by sums over specific atoms like only backbone-atoms in the RMSD-B value.

However the RMSD values can also be misleading. Although two proteins may incorporate very similar secondary structure, they can arrange them in strongly dislike topologies. One example would be a bundle of three helices. The third helix can be put top or bottom, which means that though the secondary structure is equal, the topological arrangements results in high RMSD values hiding this similarity.

## VMD[57]

VMD was used for manipulation and pictures of proteins.



# Literaturverzeichnis

- [1] B. J. Alder and T. E. Wainwright. Phase transition for a hard sphere system. *J. Chem. Phys.*, 27:1208–1209, 1957.
- [2] N. Alves and U. Hansmann. Helix formation and folding in an artificial peptide. *J. Chem. Phys.*, 117:2337–2343, 2002.
- [3] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.
- [4] N. Ashcroft and R. Mermin. *Condensed Matter Physics*. Willey & Sons, New York, 1980.
- [5] F. Avbelj. Use of a potential of mean force to analyze free energy contributions in protein folding. *Biochemistry*, 31:6290–6297, 1992.
- [6] F. Avbelj and J. Moult. Role of electrostatic screening in determining protein main chain conformational preferences. *Biochemistry*, 34:755–764, 1995.
- [7] R. Baldwin. Temperature dependence of the hydrophobic interaction in protein folding. *Proc. Natl. Acad. Sci.(USA)*, 83:8069–8072, 1986.
- [8] R. Baldwin. The nature of protein folding pathways: the classical versus the new view. *J. Biomol. NMR*, 5:103–109, 1995.
- [9] R. L. Baldwin. Matching speed and stability. *Nature*, 369:183–184, 1994.
- [10] B. Berg and U. Hansmann. Configuration space for random walk dynamics. *Eur. Phys. J. B*, 6:395–398, 1998.
- [11] J. M. Berg, J. L. Tymoczky, and L. Stryer. *Biochemistry, fifth edition*. Michelle Julet, 2002.
- [12] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.
- [13] R. Bonneau, J. Tsui, I. Ruczinski, D. Chivian, C. M. E. Strauss, and D. Baker. Rosetta in CASP4: progress in ab-initio protein structure prediction. *Proteins, SF&G*, 45:119–126, 2001.

- [14] J. D. Bryngelson and P. G. Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci.(USA)*, 84:7524–7528, 1987.
- [15] J. D. Bryngelson and P. G. Wolynes. Intermediates and barrier crossing in a random energy model with applications to protein folding. *J. Chem. Phys.*, 93:6902–6915, 1989.
- [16] G. Casari and M. J. Sippl. Structure derived hydrophobic potentials. a hydrophobic potential derived from x ray structures of globular proteins is able to identify native folds. *J. Molec. Biol.*, 224:725–732, 1992.
- [17] H. Chan. Kinetics of protein folding. *Nature*, 373:664–665, 1995.
- [18] H. S. Chan and K. A. Dill. Comparing folding codes for proteins and polymers. *Proteins, SF&G*, 24:335–344, 1996.
- [19] M. Cieplak. Cooperativity and contact order in protein folding. *Phy. Rev. E*, 69:031907, 2004.
- [20] M. Cieplak, M. Henkel, J. Karbowski, and J. R. Banavar. Master equation approach to protein folding and kinetic traps. *Phys. Rev. Lett.*, 89:3654–3657, 1998.
- [21] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins and nucleic acids. *J. Am. Chem. Soc.*, 117:5179, 1995.
- [22] C. J. Cramer and D. G. Truhlar. Implicit solvation models: Equilibria, structure, spectra, and dynamics. *Chem. Rev.*, 99:2161–2200, 1999.
- [23] Y. Cui, R. Chen, and W. Wong. Protein folding simulation with genetic algorithm and supersecondary structure constraints. *Proteins*, 31:247–57, 1998.
- [24] M. Daune. *Molecular Biophysics, Structures in Motion*. Oxford University Press, 1999.
- [25] K. Dill. Dominant forces in protein folding. *Biochemistry*, 29:7155–8133, 1990.
- [26] K. Dill, S. Bromberg, K. Yue, K. Fiebig, D. Yee, and P. Thomas. Principles of protein folding. a perspective from simple exact models. *Protein Science*, 4:561–6022, 1995.
- [27] K. Dill and H. Chan. From levinthal to pathways to funnels: The new view of protein folding kinetics. *Nature Structural Biology*, 4:10–19, 1997.
- [28] J. Doye. Physical perspectives on the global optimization of atomic clusters. <http://xxx.lanl.gov/abs/cond-mat/0007338>, 2005.
- [29] J. Doye and D. Wales. Thermodynamics of global optimization. *Phys. Rev. Letters*, 80:1357–1360, 1998.

- [30] Y. Duan and P. A. Kollman. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282:740–744, 1998.
- [31] D. Eisenberg and A. D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319:199–203, 1986.
- [32] F. Eisenmenger, U. H. Hansmann, S. Hayryan, and C.-K. Hu. [smmp] a modern package for simulations of proteins. *Comp. Phys. Comm.*, 138:192–212, 2001.
- [33] J. Fauchere and V. Pliska. Hydrophobic parameters  $\pi$  of amino-acid side chains from the partitioning of n-acetyl-amino-acid amides. *Eur. J. med. Chem.-Chim. ther.*, 18:369–375, 1983.
- [34] H. Frauenkorn, U. Bastolla, E. Gerstner, P. Grassberger, and W. Nadler. A new monte carlo algorithm for protein folding. *cond-mat/9705146*, 1997.
- [35] D. L. Freeman and J. D. Doll. Computational studies of clusters: Methods and results. *Ann. Rev. Phys. Chem.*, 47:43–80, 1996.
- [36] R. A. Friesner. *Computational Methods for Protein Folding*, volume 120. special volume of Adv. Chem. Phys., 2002.
- [37] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York, 1983.
- [38] J. Gasteiger and T. Engel. *Chemoinformatics*. Wiley-VCH, 2003.
- [39] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, 1995.
- [40] M. Gruebele. Protein folding: the free energy surface. *Curr. Op. Struc. Bio.*, 12:161–168, 2002.
- [41] D. Hadzi and D. Hadzi. *Theoretical Treatments of Hydrogen Bonding*. John Wiley & Son Ltd., 1997.
- [42] U. Hansmann. Protein folding in silico : An overview. *Comp. Sci. Eng.*, 5:64–69, 2003.
- [43] U. Hansmann and Y. Okamoto. Numerical comparison of three recently proposed algorithms in the protein folding problem. *J. Comput. Chem.*, 18:920–933, 1997.
- [44] U. H. Hansmann, M. Masuya, and Y. Okamoto. Characteristic temperatures of folding of a small peptide. *Proc. Natl. Acad. Sci. USA*, 94:10652–10656, 1997.
- [45] U. H. E. Hansmann. Global optimization by energy landscape paving. *Phys. Rev. Letters*, 88:068105, 2002.

- [46] U. H. E. Hansmann. Generalized ensemble simulations of the human parathyroid hormone fragment pth(1-34). *J. Chem. Phys.*, 120:417–422, 2004.
- [47] C. Hardin, M. Eastwood, M. Prentiss, Z. Luthey-Schulten, and P. Wolynes. Folding funnels: The key to robust protein structure prediction. *J. Comp. Chem.*, 23:138–146, 2003.
- [48] T. Herges. *Entwicklung eines Kraftfeldes zur Strukturvorhersage von Helixproteinen*. University of Dortmund, Doctorate Thesis, 2003.
- [49] T. Herges, A. Schug, B. Burghardt, and W. Wenzel. Exploration of the free energy surface of a three helix peptide with stochastic optimization methods. *Int. J. Quant. Chem.*, 99:854–893, 2004.
- [50] T. Herges, A. Schug, H. Merlitz, and W. Wenzel. Stochastic optimization methods for structure prediction of biomolecular nanoscale systems. *Nanotechnology*, 14:1161–1167, 2003.
- [51] T. Herges, A. Schug, and W. Wenzel. All atom folding and misfolding of the villin headpiece in a free-energy forcefield. (submitted), 2004.
- [52] T. Herges, A. Schug, and W. Wenzel. Protein structure prediction with stochastic optimization methods: Folding and misfolding the villin headpiece. *Lecture Notes in Computer Science*, 3045:454–464, 2004.
- [53] T. Herges and W. Wenzel. An all-atom force field for tertiary structure prediction of helical proteins. *Biophys. J.*, 87(5):3100–3109, 2004.
- [54] T. Herges and W. Wenzel. In silico folding of a three helix protein and characterization of its free energy landscape in an all-atom forcefield. *Phys. Rev. Letters*, 94:018101, 2005.
- [55] B. Honig. Protein folding: From the Levinthal paradox to structure prediction. *J. Molec. Biol.*, 293:283–293, 1999.
- [56] X. Huang. On global sequence alignment. *Computer Applications in the Biosciences*, 10:227–235, 1994.
- [57] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.
- [58] A. Irbäck, C. Peterson, F. Potthast, and O. Sommelius. Local interactions and protein folding : A 3d off-lattice approach. *J. Chem. Phys.*, 107:273–282, 1997.
- [59] D. N. Ivankov, S. O. Garbuzynskiy, E. Alm, K. W. Plaxco, D. Baker, and A. V. Finkelstein. Contact order revisited: Influence of protein size on the folding rate. *Protein Science*, 12:2057–2062, 2003.



- [60] W. L. Jorgensen, D. S. Maxwell, and J. J. Tiradorives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *JACS*, 118:11225–11236, 1996.
- [61] W. Kabsch and C. Sander. A dictionary of protein secondary structure. *Biopolymers*, 22:2577–2637, 1983.
- [62] S. Kirkpatrick, C. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.
- [63] J. Kubelka, J. Hofrichter, and W. A. Eaton. The protein folding speed limit. *Curr. Op. Struc. Biol.*, 14:76–88, 2004.
- [64] P. Larranaga, C. M. H. Kuijpers, R. H. Murga, I. Inza, and S. Dizdarevic. *Genetic Algorithms for the Travelling Salesman Problem: A Review of Representations and Operators*, volume 13. Artificial Intelligence Review archive, 1999.
- [65] E. Lattman. Casp4. *Proteins, SF&G*, 44:399, 2001.
- [66] B. K. Lee and F. M. Richards. The interpretation of protein structures: Estimation of static accessibility. *J. Molec. Biol.*, 79:379–400, 1971.
- [67] P. Leopold, M. Montal, and J. Onuchic. Protein folding funnels: A kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci.(USA)*, 89:8721–8725, 1992.
- [68] C. Levinthal. Are there pathways for protein folding ? *J. Chem. Phys.*, 65:44–45, 1968.
- [69] C. Lin, C. Hu, and U. Hansmann. Parallel tempering simulations of hp-36. *Proteins:Structure, Function and Genetics*, 52:436+, 2003.
- [70] A. P. Lyubartsev, A. A. Martinovski, S. V. Shevkunov, and P. Vorontsov-Velyaminov. New approach to monte carlo calculation of the free energy: Method of expanded ensembles. *J. Chem. Phys.*, 96:1776–1783, 1992.
- [71] J. Ma and J. Straub. Simulated annealing using the classical density distribution. *J. Chem. Phys.*, 101:533–541, 1994.
- [72] A. D. Mackerell. Empirical force fields for biological macromolecules: Overview and issues. *J. Comp. Chem.*, 25:1584–1604, 2004.
- [73] A. D. Mackerell, J. Wiorkiewicz-Kuczera, and M. Karplus. All-atom empirical energy function for the simulation of nucleic acids. *J. Am. Chem. Soc.*, 117:11946–11975, 1995.
- [74] J. R. Maple, M.-J. Hwang, K. J. Jalkanen, T. P. Stockfisch, and A. T. Hagler. Derivation of class ii force fields: V. quantum force field for amides, peptides, and related compounds. *J. Comp. Chem.*, 19:430–458, 1998.

- [75] J. R. Maple, M.-J. Hwang, T. P. Stockfisch, U. Dinur, M. Waldman, C. S. Ewig, and A. T. Hagler. Derivation of class ii force fields. i. methodology and quantum force field for the alkyl functional group and alkane molecules. *J. Comp. Chem.*, 15:162–182, 1994.
- [76] E. Marinari and G. Parisi. Simulated tempering: A new monte carlo scheme. *Europ. Phys. Letters*, 19:451–458, 1992.
- [77] I. K. McDonald and J. M. Thornton. Satisfying hydrogen bonding potentials in protein folding. *J. Molec. Biol.*, 238:777–793, 1994.
- [78] N. Metropolis, A. W. R. and M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.
- [79] M. Mezard, G. Parisi, and M. Virasoro. *Spin Glass Theory and Beyond*. World Scientific, New Jersey, 1989.
- [80] S. L. Miller. A production of amino acids under possible primitive earth conditions. *Science*, 117:528–529, 1953.
- [81] L. Mirny and E. Shakhnovich. Protein folding theory: From lattice to all-atom models. *Annual Review of Biophysics and Biomolecular Structure*, 30:361–396, 2001.
- [82] J. Moult, K. Fidelis, A. Zemina, and T. Hubbard. Critical assessment of methods of protein structure (casp): round iv. *Proteins, SF&G*, 45:2–7, 2001.
- [83] A. Nayeem, J. Vila, and H. Scheraga. A comparative study of the simulated-annealing and monte carlo-with-minimization approaches to the minimum-energy structures of polypeptides: [met]-enkephalin. *J. Comp. Chem.*, 12(5):594–605, 1991.
- [84] J. W. Neidigh, R. M. Fesinmeyer, and N. H. Anderson. Designing a 20-residue protein. *Nat. Struct. Biol.*, 9:425–430, 2002.
- [85] G. Nemethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paternali, A. Zagari, S. Rumsey, and H. Scheraga. Energy parameters in polypeptides 10. improved geometrical parameters and nonbonded interactions for use in the ecepp/3 algorithm. *J. Phys. Chem.*, 96:6472–6484, 1992.
- [86] G. Nemethy, M. S. Pottle, and H. A. Scheraga. Energy parameters in polypeptides. 9. updating of geometrical parameters, nonbonded interactions, and hydrogen bond interactions for the naturally occurring amino acids. *J. Phys. Chem.*, 87:1883–1887, 1983.
- [87] B. Øksendal. *Stochastic Differential Equations: an introduction with applications. 5th edition*. Berlin: Springer, 1998.

- [88] E. T. Ong, K. M. Lim, K. H. Lee, and H. P. Lee. A fast algorithm for three-dimensional potential fields calculation: fast fourier transform on multipoles. *J. Comp. Phys.*, 192(1):244–261, 2003.
- [89] J. N. Onuchic and A. E. Garcia. Folding a protein in a computer: An atomic description of the folding/unfolding of protein a. *Proc. Natl. Acad. Sci.(USA)*, 100:13898–13903, 2003.
- [90] J. N. Onuchic, Z. Luthey-Schulten, and P. Wolynes. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.*, 48:545–600, 1997.
- [91] G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes. Water in protein structure prediction. *Proc. Natl. Acad. Sci.(USA)*, 101:3352–3357, 2004.
- [92] D. Pearlman, D. Case, J. Caldwell, W. Ross, T. Cheatham, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman. Amber, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. *Comp. Phys. Commun.*, 91:1–41, 1995.
- [93] J. Ponder and D. Case. Force fields for protein simulations. *Adv. Prot. Chem.*, 66:27–85, 2003.
- [94] P. Privalov. Stability of proteins. *Adv. Prot. Sci.*, 33:167–241, 1979.
- [95] P. Privalov and S. Gill. Stability of protein structure and hydrophobic interaction. *Adv. Prot. Sci.*, 39:191–234, 1988.
- [96] L. Qiu, S. A. Pabit, A. E. Roitberg, and S. J. Hagen. Smaller and faster: The 20-residue trp-cage protein folds in 4 microseconds. *J. Am. Chem. Soc.*, 124:12952, 2002.
- [97] A. A. Rabow and H. A. Scheraga. Improved genetic algorithm for the protein folding problem by use of a cartesian combination operator. *Protein Science*, 5:1800–1815, 1996.
- [98] S. Raibaud, I. Lebars, M. Guillier, C. Chiaruttini, F. Bontems, A. Rak, M. Garber, F. Allemand, M. Springer, and F. Dardel. Nmr structure of bacterial ribosomal protein l20: Implications for ribosome assembly and translational control. *J. Molec. Biol.*, 323:143–151, 2002.
- [99] G. Ramachandran and V. Sasiskharan. Conformation of polypeptides and proteins. *Adv. Prot. Chem.*, 23:283–437, 1968.
- [100] F. Rao and A. Caffisch. Replica exchange molecular dynamics simulations of reversible folding. *J. Chem. Phys.*, 119:4035–4042, 2003.

- [101] Y. Rhee and V. Pande. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys. J*, 84:775–786, 2003.
- [102] A. Sali, E. Shakhnovich, and M. Karplus. How does a protein fold? *Nature*, 369:248–251, 1994.
- [103] J. Schneider, I. Morgenstern, and J. Singer. Bouncing towards the optimum: Improving the results of monte carlo optimisation algorithms. *Phys. Rev. E*, 58:5085–5095, 1998.
- [104] A. Schug, T. Herges, A. Verma, and W. Wenzel. Investigation of the parallel tempering method for protein folding. *Phys. Cond. Matter, special issue: Structure and Function of Biomolecules (in press)*, 2005.
- [105] A. Schug, T. Herges, and W. Wenzel. Reproducible protein folding with the stochastic tunneling method. *Phys. Rev. Letters*, 91:158102, 2003.
- [106] A. Schug, T. Herges, and W. Wenzel. All atom folding of the three helix hiv accessory protein with an adaptive parallel tempering method. *Proteins*, 57(4):792–798, 2004.
- [107] A. Schug and W. Wenzel. All-atom folding of the trp-cage protein in an all-atom forcefield. *Europhysics Lett.*, 67:307–313, 2004.
- [108] A. Schug and W. Wenzel. Reproducible folding of a four helix protein in an all-atom forcefield. *J. Am. Chem. Soc.*, 126(51):16736–16737, 2004.
- [109] A. Schug, W. Wenzel, and U. Hansmann. Energy paving simulations on the trp-cage protein. *J. Chem. Phys. (in press)*, 2005.
- [110] W. R. P. Scott, P. H. Hunenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Kruger, and W. F. van Gunsteren. The gromos biomolecular simulation program package. *J. Phys. Chem. A*, 103:3596–3607, 1999.
- [111] K. A. Sharp, A. Nicholls, R. Friedman, and B. Honig. Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models. *Biochemistry*, 30:9686–9697, 1991.
- [112] C. Simmerling, B. Strockbine, and A. Roitberg. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.*, 124:11258–11259, 2002.
- [113] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Molec. Biol.*, 286:209–225, 1997.
- [114] M. J. Sippl, G. Nemethy, and H. A. Scheraga. Intermolecular potentials from crystal data. 6. determination of empirical potentials for o-h···o=c hydrogen bonds from packing configurations. *J. Phys. Chem.*, 88:6231–6233, 1984.

- [115] D. Sitkoff, K. A. Sharp, and B. Honig. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.*, 98:1978–1988, 1994.
- [116] J. Skolnick and A. Kolinski. Computer simulations of globular protein folding and tertiary structure. *Annu. Rev. Phys. Chem.*, 40:207–235, 1989.
- [117] D. A. Smith. *Modeling the Hydrogen Bond*. ACS Symposium Series, 1994.
- [118] M. J. Sternberg. *Protein Structure Prediction: A practical Approach*. Oxford University Press, 1996.
- [119] W. T. Sullivan, D. Werthimer, S. Bowyer, J. Cobb, D. Gedye, and D. Anderson. Astronomical and biochemical origins and the search for life in the universe. *Proc. of the Fifth Intl. Conf. on Bioastronomy*, 161, 1997.
- [120] J. Tomasi and M. Persico. Molecular interactions in solution: An overview of methods based on continuous distributions of the solvent. *Chem.Rev.*, 94:2027–2094, 1994.
- [121] W. van Gunsteren, F. J. Luque, D. Timms, and A. Troda. From structure to function, taking account of solvation. *Ann. Rev. Biophys. and Biomol. Struct.*, 23:847, 1994.
- [122] W. F. van Gunsteren and H. J. C. Berendsen. *Groningen Molecular Simulation (GROMOS) Library Manual (Biomos)*. 9747 AG Groningen, The Netherlands, 1987. Nijenborgh 4.
- [123] A. Verma, A. Schug, T. Herges, and W. Wenzel. Comparison of stochastic optimization methods for all-atom protein folding. submitted, 2004.
- [124] D. Voet, J. G. Voet, C. W. Pratt, A. G. Beck-Sickinger, U. Hahn, and A. G. Beck-Sickinger. *Lehrbuch der Biochemie*. Wiley-Vch, 2002.
- [125] D. J. Wales and J. P. Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *J. Phys. Chem.*, 101:5111, 1997.
- [126] W. Wenzel and K. Hamacher. Stochastic tunneling approach for global optimization of complex potential energy landscapes. *Phys. Rev. Lett.*, 82:3003–3007, 1999.
- [127] Y. Xiang, H. Jiang, W. Cai, and X. Shao. An efficient method based on lattice construction and the genetic algorithm for optimization of large lennard-jones clusters. *J. Phys. Chem. A*, 108:3586–3592, 2004.
- [128] B. Zagrovic, C. D. Snow, M. R. Shirts, and V. S. Pande. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *Journal of Molecular Biology*, 323:927–937, 2002.

