Protein Structure Prediction using global optimization by basin-hopping

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Falk Hoffmann

aus Frankfurt(Oder), Deutschland

Jülich, 7. August 2014

aus dem Institute of Complex Systems (ICS-6): Strukturbiochemie des Forschungszentrums Jülich

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Jun.-Prof. Dr. Birgit Strodel
Koreferent: Prof. Dr. Dieter Willbold

Tag der mündlichen Prüfung: 8. Juli 2014

Statement of Author

I, Falk Hoffmann, hereby certify that the work presented here is, to the best of my knowledge and belief, original and the result of my own investigations. I have fully acknowledged and referenced the ideas and work of others, whether published or unpublished, in my thesis. My thesis contains no material published elsewhere or extracted in whole or in part from a thesis submitted for a degree at this or any other university. Where the results are produced in collaboration with others, I have clearly mentioned my contributions.

Zusammenfassung

Proteine sind die Hauptakteure in Zellen. Ihre Funktion hängt mit ihrer atomaren Struktur zusammen, weshalb deren Vorhersage sehr wichtig ist. Die Strukturvorhersage von Proteinen wurde in den letzten Jahrzehnten untersucht, bleibt aber immer noch eine der größten Herausforderungen in der Biochemie. Eine Hypothese der Proteinstrukturvorhersage besagt, dass die natürliche Struktur eines Proteins mit dem globalen Minimum der Energie in der Energielandschaft des Proteins verknüpft ist. Wir benutzen den *basin-hopping* Ansatz zur globalen Optimierung, um die Struktur von unterschiedlichen Proteinen zu untersuchen. Wir finden heraus, dass die Benutzung von chemischen Verschiebungen des Proteinrückgrats und der Cβ-Seitenkettenatome als strukturelle Randbedingung die Proteinstrukturvorhersage signifikant verbessern kann. Weiterhin werden Untersuchungen mit einer unvollständigen Anzahl an chemischen Verschiebungen durchgeführt, die zeigen, dass sogar die Benutzung von nur einer Art chemischer Verschiebungen ausreichend ist, um die richtigen Sekundärstrukturelemente eines Proteins zu finden. Zusätzlich werden verschiedene Monte Carlo-Schritte untersucht. Wir entwickeln einen Ansatz, der basierend auf der Sekundärstrukturvorhersage der Aminosäuren die richtigen Tertiärstrukturkontakte findet. Dazu werden unterschiedliche maximale Dehidralwinkeländerungen getestet, welche zeigen, dass deren beste Wahl zu einer Verbesserung der Simulationszeit im Vergleich mit früheren Änderungen führt. Wir studieren den Einfluss von Rückgrat- und Seitenketten-dehidralwinkeländerungen und stellen fest, dass beide Änderungen wichtig für die Verbesserung der Struktur des Proteins sind. Wissensbasierte Monte Carlo-Schritte können die Genauigkeit und Geschwindigkeit von Simulationen erhöhen. Wie führen Monte Carlo-Schritte ein, welche auf der statistischen Verteilung der Dehidralwinkel von Proteinen im Ramachandranplot ihrer Aminosäuren basieren. Die Monte Carlo-Schritte erlauben uns, Proteine mit α Helices zu falten, während Proteine mit β Faltblättern eine Herausforderung bleiben. Wir vergleichen moderne β Faltblatt-Vorhersageprogramme bezüglich ihrer Effektivität für die Vorhersage von Restkontakten zur Erzeugung struktureller Nebenbedingungen, die innerhalb der *basin-hopping*-Simulationen zur Erzeugung von β-Faltblättern angewandt werden. Es kann erwartet werden, dass die Kombination dieser verschiedenen Ansätze die Proteinstrukturvorhersage mit dem *basin-hopping* Ansatz zur globalen Optimierung signifikant verbessern kann.

Abstract

Proteins are the main actors in the cell and their function is associated with their atomistic structure. The prediction of protein structures has been investigated over the last decades, but still remains one of the big challenges in biochemistry. One of the hypothesis in protein structure prediction is that the native configuration of a protein is connected to the global minimum of the energy in the energy landscape of the protein. We use the basin-hopping approach to global optimization to investigate the structure of different proteins. We find that the usage of backbone and Cβ chemical shifts as structural constraints can significantly improve the prediction of protein structures. Furthermore, studies with incomplete backbone chemical shift information are performed and show that even the usage of one type of chemical shifts is sufficient to find the correct secondary structure of proteins. Furthermore, several Monte Carlo moves are studied. We introduce an approach that derives tertiary structures from the secondary structure assignments of individual residues. Different maximum dihedral angle changes are tested and reveal that the best choice leads to a remarkable increase in the simulation time compared with previous moves. We study the influence of backbone and side chain dihedral angle moves and show that both moves have an important influence on the refinement of the structure of a protein. It has been shown that knowledge-based Monte Carlo moves can increase the accuracy and speed of simulations. We introduce a Monte Carlo move set which is based on the statistical distribution of the dihedral angles of proteins in the Ramachandran plot of their amino acids. The moves allow us to fold proteins with α helices while proteins with β sheets still remain a challenge. We compare state-of-the-art β sheet predictors on their efficiency in terms of their prediction of residue contacts to create structural constraints which are used in basin hopping simulations for the production of β sheets. The combination of these different approaches can be expected to significantly improve the prediction of protein structures with the basin-hopping approach to global optimization.

Acknowledgement

I want to express special thanks to my supervisor, Jun.-Prof. Dr. Birgit Strodel, for giving me the opportunity to work in her group and do the research. She constantly was speaking with me about all open questions and I could learn many things under her guidance.

I also want to thank the coworkers in my group, especially Ken, Sanjay, Ioan, Jide, Chetan, Qinghua, Philipp, Rashmi, Nam, Bogdan, Martin, Oliver, Marianne, Michael and Dusan for their nice working atmosphere and useful scientific discussions.

The project was funded by grants from the International Helmholtz Research School of Biophysics and Soft Matter (IHRS BioSoft) and the Forschungszentrum Jülich which I want to thank. Special thanks to the Jülich Supercomputing Centre and the Jülich Aachen Research Alliance (JARA) between RWTH Aachen University and Forschungszentrum Jülich for providing and maintaining the computing resources applied to the project.

Especially I want to thank all my friends which I met in Jülich, Düren, Aachen, Köln and around. You made my time unforgettable.

This thesis would have been impossible without the support of my parents who always believed in me during the time in Jülich and before. Danke Mama und Papa.

I want to thank Gabriela Goytia who always stands behind me when I need support and supported me in every single second. Te amo.

Contents

List of Abbreviations

Ω_{ROT}	Rotation angle around the axis of the two rotamers in an amino acid
BB	Backbone
BFGS	Broyden–Fletcher–Goldfarb–Shanno
BH	Basin-Hopping
C'	Carbon atom in carbonyl group of peptide bond
CASP	Critical Assessment of Techniques for Protein Structure Prediction
COM	Center of mass
dD	d dimensions (d is a positive integer)
DSS	4,4-dimethyl-4-silapentane-1-sulfonic acid
EOM	Equation of motion
GPU	Graphics Processing Unit
L-AA	Left-handed amino acid
L-BFGS	Limited Broyden–Fletcher–Goldfarb–Shanno
LJ	Lennard-Jones
MC	Monte Carlo
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
PDB	RCSB Protein data bank
PDF	Probability density function

PES	Potential energy surface
ppm	Parts per million
PS	Primary Structure
QS	Quaternary structure
R	Side Chain
RMSD	Root mean square displacement
RMSF	Root mean square force
SC	Side chain
SS	Secondary structure
TMS	Tetramethylsilane
TS	Tertiary structure

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Proteins

Cells are the basic unit of all living organisms. Proteins play a role in nearly every process of the cell and are essential for the life of organisms. Many proteins are enzymes. Enzymes are highly selective catalysts which accelerate metabolic reactions.

1.1.1 Structure

Amino acids

Proteins are linear polymers whose basic units are amino acids. All amino acids in living organisms are α amino acids. The amino group in an α amino acid is connected to the first carbon atom after the one in the carbonyl group. Figure 1.1 shows a general scheme of an α amino acid.

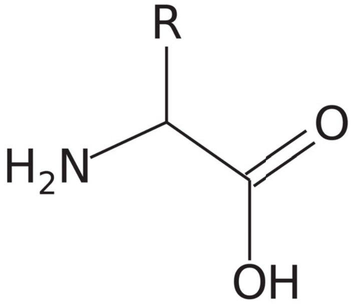


Figure 1.1: Structural formula of an α amino acid. The different amino acids are characterized by their side chains R. The structure was created with GChemPaint [1].

The carbon atom which connects the backbone of an amino acid with its side chain R is a chiral carbon atom, except for glycine where the side chain R contains only one hydrogen atom. An object is chiral if there is no sequence of translations and/or rotations which can transform the amino acid to its mirror image. There are two different chiral forms or enantiomers in amino acids: left-handed and right-handed amino acids. The two different forms for alanine with a CH3 group as the side chain are shown in figure 1.2.

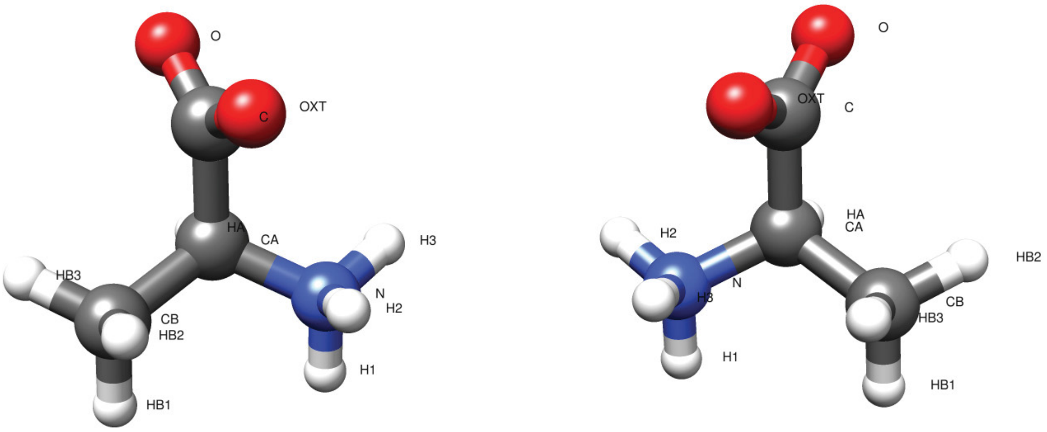


Figure 1.2: Left-(left side) and right-handed(right side) enantiomer of alanine in balls and stick representation. Carbon atoms are shown in grey, hydrogen atoms in white, nitrogen atoms in blue and oxygen atoms in red. The Hα atom is behind the central Cα atom. The structure was created with Chimera [2].

If one places the Hα atom behind the Cα atom on a line which is perpendicular to the plane of the viewer and hits this plane in the Cα atom like it is shown in figure 1.2, a left- and right-handed helix are defined by the CORN rule: If the carboxyl group (CO), the side chain (R) and the nitrogen atom (N) are orientated (in this order) counterclockwise around the central $C\alpha$ atom, the amino acid is a left-handed helix. If these groups are orientated clockwise around the central Cα atom, the amino acid is a right-handed helix. The chirality of a protein is very important as it has a big influence on the dihedral angles of the proteins (see below). In nature, 19 left-handed (L-AA) and one nonchiral (glycine) amino acids are present. 19 of them are proteinogenic amino acids which have a chemical structure like in figure 1.1 and differ in their side chain R. Proline has a different structure with a ring involving the N-terminal amine group. Table 1.1 shows the structures for all naturally occurring amino acids.

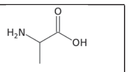
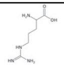
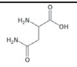
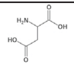
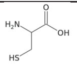
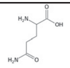
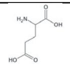
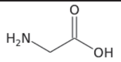
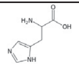
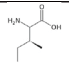
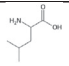
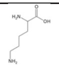
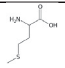
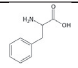
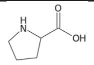
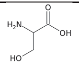
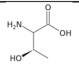
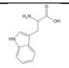
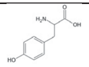
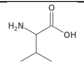
amino acid	3 letter code	1 letter code	structure
Alanine	ALA	A	
Arginine	ARG	R	
Asparagine	ASN	N	
Aspartic acid	ASP	D	
Cysteine	CYS	C	
Glutamine	GLN	Q	
Glutamic acid	GLU	E	
Glycine	GLY	G	
Histidine	HIS	H	
Isoleucine	ILE	I	
Leucine	LEU	L	
Lysine	LYS	K	
Methionine	MET	M	
Phenylalanine	PHE	F	
Proline	PRO	P	
Serine	SER	S	
Threonine	THR	T	
Tryptophan	TRP	W	
Tyrosine	TYR	Y	
Valine	VAL	V	

Table 1.1: Naturally occurring acids with their names (first column), three letter abbreviation (second column), one letter abbreviation (third column) and structure (fourth column). The structures were created with GChemPaint [1].

Polypeptide chain

The amino acids are linked together by a peptide bond to create the structure of a polypeptide chain. The general structure of a polypeptide chain is shown in figure 1.3.

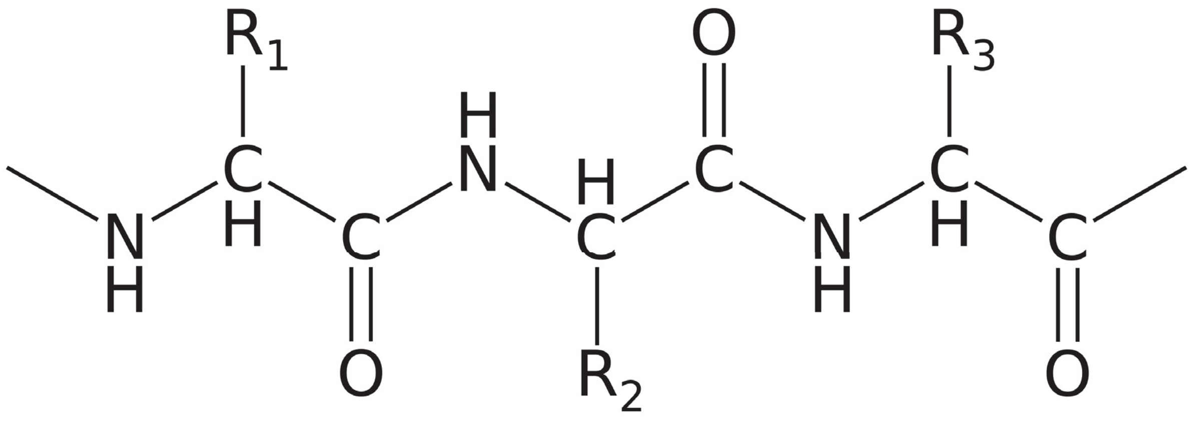


Figure 1.3: Structural formula of three connected amino acids in a polypeptide chain. The structure was created with GChemPaint [1].

A peptide is a short polypeptide chain with less than 100 amino acids while a protein is a longer chain with more than 100 amino acids. However, often polypeptide chains between 50 and 100 amino acids are also called proteins.

The distribution of the π electrons in the double bound of an amino acid in a protein is not fixed to the carbonyl group, but is partially located along the carbonyl C-O and along the amide C-N bonds of the peptide. The peptide bond has therefor two resonance forms which are shown in figure 1.4.

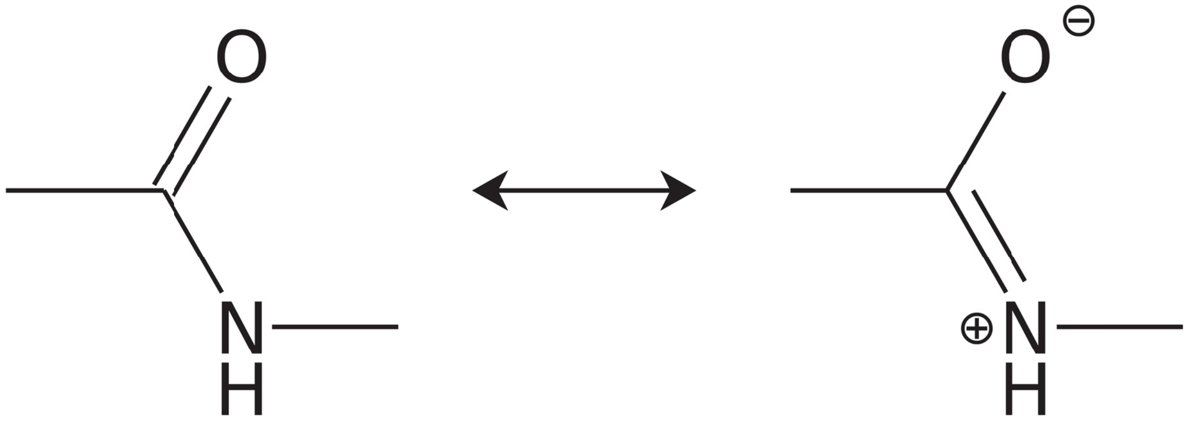


Figure 1.4: The two resonance forms of the peptide bond. The partial double bond creates a planarity around the peptide bond. The structure was created with GChemPaint [1].

All atoms which do not belong to the side chain R are called backbone (BB) atoms [3]. That are the carbon (C') and oxygen of the carbonyl group, the nitrogen and hydrogen of the amide group, and the Cα and Hα atoms. The Cα atom connects the backbone with the side chain. The Hα atom is also bonded to the Cα atom. All atoms which belong to R are called side chain (SC) atoms.

A dihedral angle is an angle between two planes. Two planes in the three-dimensional (3D) space have a dihedral angle if they are nonparallel and nonidentical. In this case, the two planes have a line of intersection. A line can be identified by two nonidentical points while a plane is identified by three nonidentical points which are not on a line. A dihedral angle can be therefor identified by four nonidentical atoms: two atoms at the line of intersection (which belong to both planes) and one atom per plane outside the line of intersection which describes the 3D orientation of the plane. Dihedral angles play an important role in the structure of a polypeptide chain. There are three different dihedral angles in the BB: Φ which includes the BB atoms C'-N-Cα-C' and defines the C'-C' distance, Ψ which includes the BB atoms N-Cα-C'-N and defines the N-N distance and Ω which includes the BB atoms Cα-C'-N-Cα and defines the Cα-Cα distance. Here, the first three of the four atoms belong to the first plane while atoms 2–4 belong to the second plane. The double bound character of the central C'-N bond, in which is the central bond in Ω, prevents rotations around its bond. The dihedral angle Ω is mostly in a planar conformation with values around 180° which is called the trans conformation of the peptide bond. The other two dihedral angles Φ and Ψ determine the local structure of the BB [3].

Global structure

Primary Structure The linear sequence of amino acids is the primary structure (PS) of a protein. It has an amino terminal and a carboxyl terminal. Without loss of generality, the PS starts at the amino terminal and finishes at the carboxyl terminal.

The primary structure determines the 3D shape of a protein. However, the prediction of the correct 3D structure is still a big challenge in biochemistry. The first step from the PS to the correct 3D structure is the determination of the secondary structure (SS).

Secondary Structure Neighbouring amino acids can form local structure segments which define the secondary structure of a protein. The SS is stabilized by hydrogen bonds between backbone atoms. Hydrogen bonds are attractive interactions between a hydrogen atom with a positive partial charge which is bonded to a donor and an electronegative atom with a negative partial charge which is called acceptor. The donor is an electronegative atom. Examples for donors and acceptors are nitrogen, oxygen or fluorine. In the protein backbone, hydrogen bonds are formed between the hydrogen atom of the amide group of one amino acid and the oxygen atom of the carbonyl group of another amino acid. Hydrogen bonds stabilize the local structure. The pattern of hydrogen bonds describes the type of SS element. The most common SS elements are α helices and β sheets.

Hydrogen bonds between the carbonyl group of amino acid i and the amide group of amino acid $i + 4$ or $i - 4$ stabilize α helices. In an α helix, 3.6 amino acids form a turn and 13 atoms are involved in the formation of the full turn. That is why an α helix is also called a 3.6_{13} helix. As the hydrogen bonds involve only local residues, α helices can generally be detected faster than β sheets. Figure 1.5 shows the α helix of the residues 2–8 of the peptide with PDB code 1L2Y.

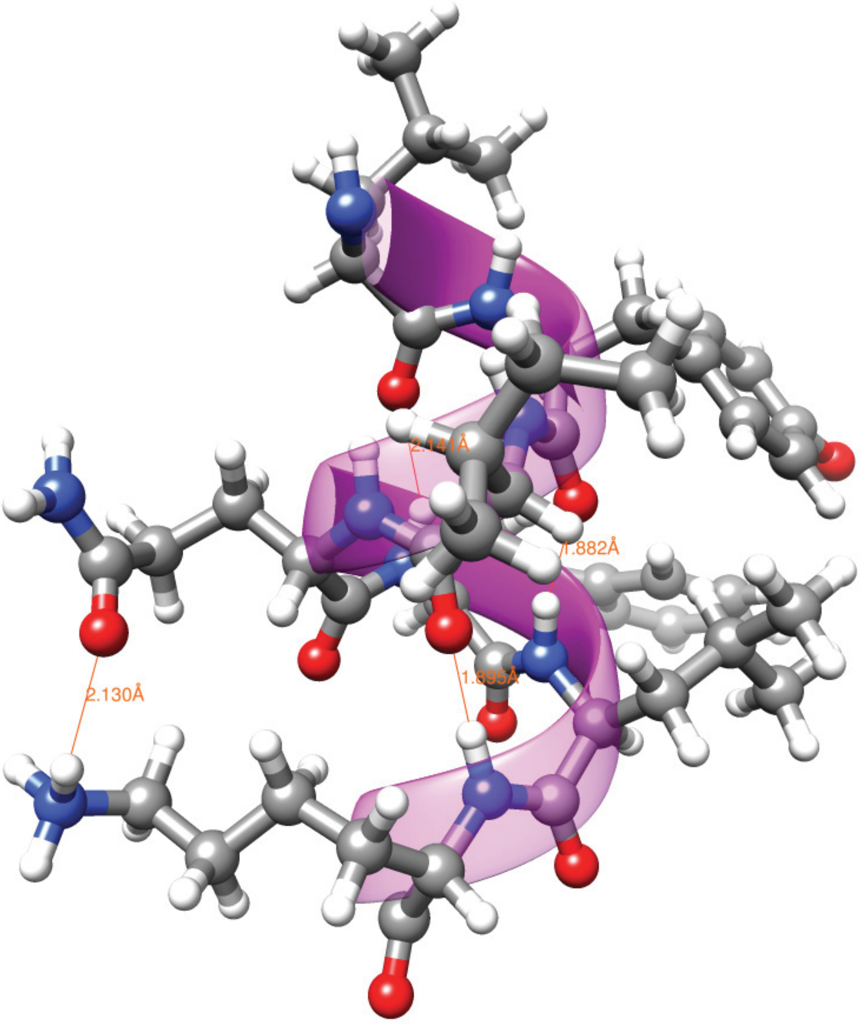


Figure 1.5: Backbone atoms of the residues 2–8 of peptide 1L2Y in ball and stick representation and secondary structure in purple and NewCartoon representation showing the helical structure of the α helix. Oxygen atoms are shown in red, nitrogen atoms in blue, carbon atoms in grey and hydrogen atoms in white. Hydrogen bonds are shown in orange and labeled with their length in Å. The picture was created with Chimera [2].

Other helices are also common in proteins: 3_{10} helix, π helix and polyproline helix. They are not so often present in natural proteins like α helices, but have energetically favorable hydrogen bonds. The π helix is a 4.4_{16} helix while the polyproline helix occurs in the presence of repeating proline residues. In general, the rotation angle Ω_{ROT} of a polypeptide helix in trans formation can be calculated by formula 1.1 [4]:

$$3\cos\Omega_{ROT} = 1 - 4\cos^2\left(\frac{\Phi + \Psi}{2}\right). \tag{1.1}$$

The rotation angle Ω_{ROT} is negative for left-handed and positive for right-handed helices. The right-handed α helix is more common than the left-handed α helix.

The second most common SS elements are β sheets. A β sheet consists of β strands. Every β strand is usually between 3 and 10 amino acids long. The structure of a β strand is nearly fully extended. The orientation of β strands gives the possibility to align two or more strands parallel or antiparallel to each other and form a β sheet. The strands are stabilized by hydrogen bonds between neighbouring carbonyl and amide groups of amino acids from different β strands. Figure 1.6 shows the formation of hydrogen bonds in a parallel and antiparallel β strand pair.

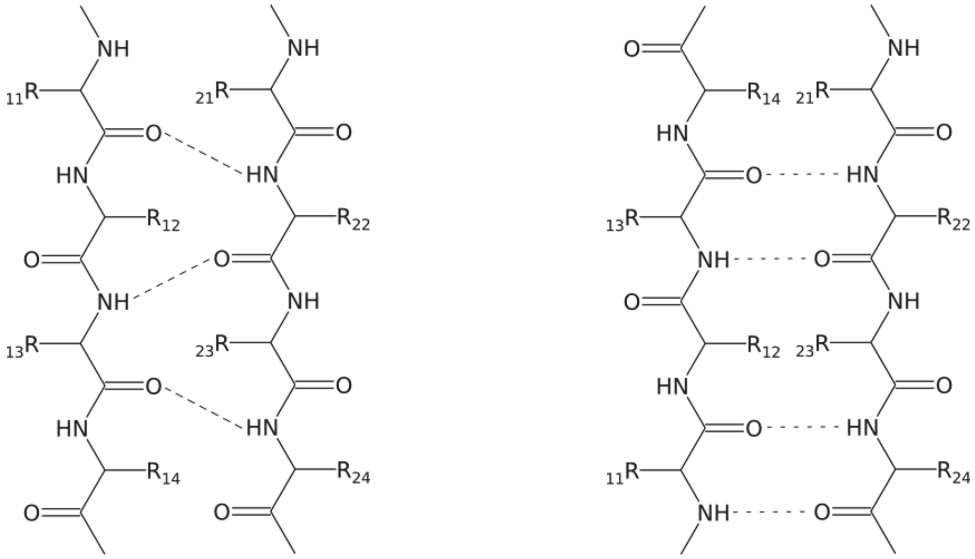


Figure 1.6: Structure of a parallel (left) and antiparallel (right) β sheet. Hydrogen bonds are shown with a dashed line. Structures were created with GChemPaint [1].

Hydrogen bonds in an antiparallel β sheet are linear. For this reason, antiparallel β sheets are more stable than parallel β sheets. The correct determination of β sheets is in general a more difficult task than the determination of α helices because neighbouring strands can be separated by many residues. Thus, long-range contacts need to be established during β sheet formation.

Other secondary structure elements like turns or ω loops connect helix and strand elements with each other.

Tertiary structure A protein can have several SS elements, while the global shape of a protein with different SS elements is called protein tertiary structure (TS). The arrangement is influenced by the interactions of the SC atoms of the amino acids. The TS involves interactions between distant residues with respect to the sequence and can be characterized by their hydrophobicity: In the center of the protein is a hydrophobic core while there are more hydrophilic amino acids at the surface of a protein which leads to the formation of hydrogen bonds with the surrounding water molecules of the aqueous solution.

Quarternary Structure A protein which consists of more than one domain arranges the tertiary structure of all chains in an energetically favorable way. In this case, the tertiary structure of a polypeptide chain is called a protein subunit. The global shape of different protein subunits in the protein is the quarternary structure (QS) of the protein. In this thesis, the focus lies on the determination of the correct TS from the PS as the TS is the basis for the determination of the QS.

Ramachandran plots

The dihedral angles of the protein backbone characterize the local shape and the SS of a polypeptide chain. As described above, the dihedral angle Ω is nearly always close to 180°, i.e. in the trans conformation because of the properties of the peptide bond. The pair of the other dihedral angles Φ and Ψ (Φ, Ψ) characterizes the orientation of neighbouring residues. A Ramachandran plot is a diagram where Ψ is plotted against Φ. The dihedral angles can have values in the interval $\Phi, \Psi \in [-180°, 180°)$ because of their periodicity. The rotation angle Ω_{ROT} for helices in trans conformation can be calculated with equation 1.1. Ω_{ROT} does not change so much in the same helix which means that the sum $\Phi + \Psi$ is more or less a constant value. For example, in case of an α helix $\Phi + \Psi \approx -105°$, for a 3_{10} helix $\Phi + \Psi \approx -75°$ and for a π-helix $\Phi + \Psi \approx -130°$. All dihedral angles of the trans polyproline II helix are around the point $(\Phi, \Psi) \approx (-75°, 150°)$ because only one type of amino acid (proline) is present. Most of the dihedral angles of the α helix are on a diagonal between $(-90°, -15°)$ and $(-35°, -70°)$ with their center at $(\Phi, \Psi) \approx (-60°, -35°)$. The centers of the 3_{10} helix at $(\Phi, \Psi) \approx (-49°, -26°)$ and the π helix at $(\Phi, \Psi) \approx (-55°, -70°)$ are mostly covered by the more populated region of α helices while the polyproline II dihedral angles are in the β sheet region (see below). The left-handed helices are also present in the Ramachandran plot by switching the signs of Φ

and Ψ. They do not occur so often in nature like the right-handed counterparts resulting in a smaller region in the Ramachandran plot.
Amino acids in β sheets are part of a β strand which is an extended structure. However, the different amino acids in a β strand have their own chirality leading to a wide range of dihedral angles Φ and Ψ with their center at $(\Phi, \Psi) \approx (-135°, 135°)$ in the Ramachandran plot [5]. In general, one can find the β sheet regions in the quadrant with negative Φ and positive Ψ values although parts of the region have negative Ψ values close to $-180°$. α sheets have alternating dihedral angles in the right- and left-handed helical regions. They are only rarely observed and may play a role in amyloid diseases [6, 7].
Turns and loops have dihedral angles outside the regions of helices and sheets. A β hairpin connects two antiparallel β sheets, but includes many elements of different turns.
In summary, the most populated regions in a Ramachandran plot of proteins are, in decreasing order of population:

1. right-handed helical region around $(\Phi, \Psi) \approx (-60°, -35°)$
2. β sheet region around $(\Phi, \Psi) \approx (-135°, 135°)$
3. left-handed helical region around $(\Phi, \Psi) \approx (60°, 35°)$

The Ramachandran plot of the protein with PDB code 1ACJ from the RCSB Protein data bank (PDB) is shown in figure 1.7. It shows all three described regions.

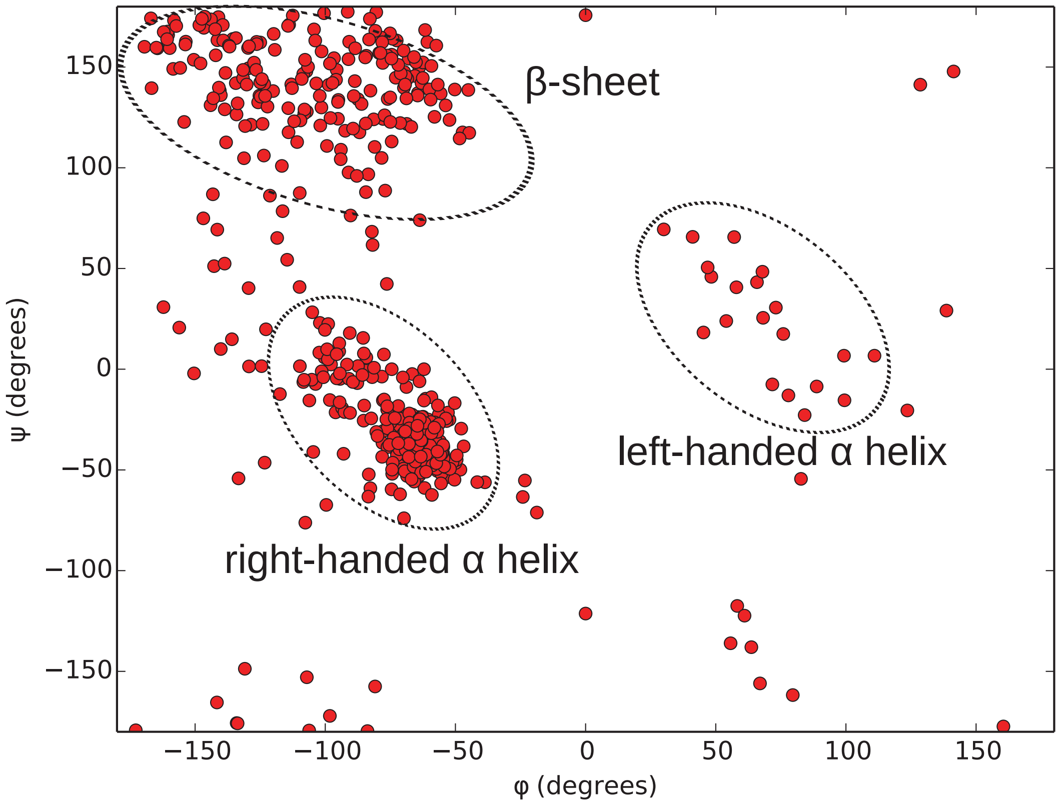


Figure 1.7: Ramachandran plot of the protein with PDB code 1ACJ. All Φ and Ψ dihedral angles present in this protein are plotted. The most important SS elements are highlighted by black circles.

1.1.2 Protein Structure Prediction

Knowledge of the structure of a protein is key for understanding its function. The importance of protein functions makes the prediction of the structure of a protein to one of the main goals in biochemistry. The task is to predict the 3D structure of a protein from its amino acid sequence. The quality of the protein structure prediction is tested every two years in the Critical Assessment of Techniques for Protein Structure Prediction [8]. The prediction for a single protein focuses on the prediction of its secondary and tertiary structure while the prediction of protein complex formation is the prediction of the quarternary structure. The prediction of protein complexes is called protein-protein docking and is based on the interactions between two proteins which mostly already have correct predefined secondary and/or tertiary structures. The focus of this thesis is on the correct prediction of the secondary and tertiary structure of a protein from its primary structure.

Secondary Structure Determination and Prediction

Secondary Structure Determination As described in section 1.1.1, the secondary structure elements can be defined by their hydrogen bond pattern. Methods which determine the secondary structure of a protein, calculate the hydrogen bonds from a given 3D structure of a protein. Based on these hydrogen bonds, the methods assign a secondary structure element to every amino acid in the polypeptide chain. Examples for secondary structure packages are DSSP [9], STRIDE [10] and DEFINE [11].

Secondary Structure Prediction SS prediction is the assignment of a secondary structure element to the residues of a polypeptide sequence. The first SS predictors [12–16] could only distinguish between α helical and coiled structures as α helices are most easily predictable and the most common SS elements (see section 1.1.1). These methods are mainly based on models which are specific for α helices. Later, the methods used a more general approach and were extended to the prediction of β sheets [17, 18]. However, these previous methods did not reach a high enough accuracy to be called reliable.

With increasing protein structure data, the most successful SS predictors at the moment are based on machine learning techniques. They use the information from protein data banks as training sets to adjust their parameters and/or functions. Neural network techniques like Psipred [19] or Porter [20] or support vector machines [21, 22] reach accuracies of 80% or more [23, 24].

Tertiary Structure Prediction

The prediction of the tertiary structure from the protein primary structure, with or without the help of secondary structure prediction tools, is more difficult than the prediction of the secondary structure because the correct relative position of the protein atoms has to be determined. While the accuracy of a SS predictor is measured by the comparison of the result with the assignment from a SS determination program to a structure saved in a data bank, the accuracy of a TS predictor is measured with some structural properties like the root mean square displacement (RMSD) between the result of the predictor and the structure in the data bank. This has the advantage that there is, in principle, no upper limit for the accuracy of a TS predictor as the result is not compared with an assignment, which limits the best accuracy of SS predictors to 90%. However, there is still an experimental limit in both cases as the experimental methods used to save the structures in the data bank can only resolve structures up to a spacial limit under experimental conditions, e.g. for x-ray up to 1 Å for temperatures of more than 100 K. These limitations are described in more detail in section 1.2. Tertiary structure prediction methods can be divided into three groups: *ab initio* modelling, comparative modelling and refinement. The aim of methods of the first two groups is to find the 3D structure of a protein with no or just limited knowledge of the structure at the beginning. Refinement methods, which are often also used as the last step in methods of the first two groups, aim to optimize a structure which is largely correctly folded but needs some improvement. In most of the cases, the geometry of the side chains has to be modified as they are the most flexible groups. Comparative modelling methods like homology modelling or protein threading use already solved protein structures from a protein data bank to find new protein structures. In homology modelling, sequences of proteins are compared in order to find similarities based on the assumption that proteins with high sequence homology share a similar structure. Protein threading or fold recognition tries to find the correct fold of a protein, which does not have a homolog in the data bank, by scanning the unknown protein sequence and comparing with sequences of already known folds.

Ab initio or *de novo* protein structure prediction does not use any information from a data bank. These methods predict the structure of a protein just from their PS. Two classes of *de novo* protein structure methods are common: evolutionary covariation methods and energy- or fragment-based methods. In evolutionary covariation methods, protein sequence alignments are used to predict correlated mutations and these coevolved residues are used to predict the 3D structure of a protein. One example is the program EVfold [25]. Energy- or fragment-based methods are mainly based on stochastic methods like global optimization (see section 3.7). These methods were successfully applied to small proteins. Folding of bigger proteins needs a lot of computational power and is applied in distributed computing projects like Folding@home [26] or Rosetta@Home [27].

1.2 NMR

1.2.1 Experimental methods in protein structure determination

91,414 protein structures have been solved and saved in the RCSB protein data bank [28] at March 24, 2014. Among them, 89.28% were resolved by X-ray crystallography, 9.93% by Nuclear Magnetic Resonance (NMR) spectroscopy, 0.56% by electron microscopy and 0.23% by other experimental methods. The numbers show the importance of X-ray crystallography and NMR spectroscopy for the determination of protein structures. In X-ray crystallography, the sample has to be prepared as a well ordered crystal. The crystal is probed by a X-ray beam and the electromagnetic radiation is scattered by the atoms of the protein, creating a diffraction pattern which depends on the distribution of the atoms and the local electron density. The diffraction pattern is used to determine the position of the atoms of the molecule and to construct a 3D model of the sample. The advantage of X-ray crystallography is its high spatial resolution. However, well ordered crystals have to be prepared which limits its applicability for dynamic processes and temperatures above 100 K.

On the other hand, NMR spectroscopy can be used to resolve the structure of proteins in solid state [29] and in solution [30]. NMR spectroscopy uses the properties of nuclear magnetic moments, which are induced by nuclear spins [31]. Nuclear spins in an external magnetic field react to this field and interact with nuclear spins of their environment. These interactions are characteristic for an atom and its environment and can be measured by a specific resonance frequency which is described by its chemical shift. As there is no crystal needed for the preparation of the sample, NMR measurements can be done in solution and the dynamics of biomolecules can be investigated. In contrast to X-ray, NMR measurements are applied to an ensemble of structures undergoing thermal fluctuations. The measured observable is an average over all ensembles and are used to build a 3D model. Energy minimization methods are typically used to refine the “averaged” structure in case of the presence of an unphysical situation like high-energy structures or atom clashes. Solid-state NMR can be used for the determination of big protein structures while liquid-state NMR is limited to small proteins as big proteins cannot be prepared in a sufficiently concentrated solution for liquid-state NMR measurements.

1.2.2 NMR chemical shifts

All nuclei with an odd mass number (e.g. ^{1}H, ^{13}C, ^{15}N) and all nuclei with an even mass number and an odd charge (e.g. ^{2}H) have a spin angular momentum $\mathbf{S}$. The spin angular momentum is quantized and the different states are ordered according to the spin quantum number I of the nuclei [31]. The total momentum of a nuclear spin with spin quantum number I is

$$\mid \mathbf{S} \mid = \hbar\sqrt{I\left(I+1\right)}. \tag{1.2}$$

Here, $\hbar = h/\left(2\pi\right)$ with h as Planck’s constant. Without any external magnetic field, the nuclear spins orientate randomly. After applying an external magnetic field $\mathbf{B}_0 = B_0\mathbf{e}_z$ in one direction (without loss of generality the z-direction), the nuclear spins orientate in the direction of the magnetic field. The projection of the direction of the spin angular momentum on the z-axis can only have values which are separated by $\hbar$ and which range from $-I$ to I. The magnetic moment in z-direction μ_z of a nuclear spin is proportional to its spin angular momentum in the same direction:

$$\mu_z = \gamma\mathbf{I}_z. \tag{1.3}$$

The gyromagnetic ratio γ is characteristic for every type of nuclei.
Once a protein is put in an external magnetic field $\mathbf{B_0}$, the magnetic moments μ_i of the nuclei of the protein interact with the field. The corresponding Hamiltonian is given by

$$\hat{H} = -\hat{\mu}_i \cdot \hat{\mathbf{B}} \tag{1.4}$$

The actual magnetic field $\mathbf{B}_0$ is normally shielded by the electron density around a nucleus, yielding a new magnetic field $\mathbf{B}$,

$$\mathbf{B} = (1 - \sigma)\,\mathbf{B_0}, \tag{1.5}$$

which interacts with the nuclei according to equation 1.4. In equation 1.5, σ is the strength of the shielding. The energy E of a state whose projection of the magnetic moment on the z-axis is called μ_z can then be calculated by

$$E = -\mu_z B_z. \tag{1.6}$$

With the help of equation 1.3, equation 1.6 can be transformed to

$$E = -\gamma B \mid \mathbf{I}_z \mid . \tag{1.7}$$

The energy difference ΔE between two neighbouring energy states E_i and E_{i+1} (whose spin angular momentum on the z-axis have values which are separated by $\hbar$) is therefor

$$\Delta E = E_{i+1} - E_i = \gamma \hbar B. \tag{1.8}$$

This energy difference is very low resulting in a nearly equivalent population of the lower (ground state) and upper state (excited state). In equilibrium, the population ratio between the excited state P_{ex} and the ground state P_{gr} can be calculated by the Boltzmann distribution:

$$\frac{P_{ex}}{P_{gr}} = \mathrm{e}^{-\frac{\gamma \hbar B}{k_B T}}. \tag{1.9}$$

Applying a second oscillating magnetic field $\mathbf{B}_1$ with Larmor frequency $\omega_L = \gamma B$ perpendicular to the first magnetic field will create transitions between the ground and the excited state, whose intensities can be measured. Due to the fact, that the difference between measured frequencies is very small in comparison with the operating frequency of the spectrometer, the frequencies are usually represented by the frequency difference to a reference structure which is typically tetramethylsilane (TMS) or 4,4-dimethyl-4-

silapentane-1-sulfonic acid (DSS):

$$\delta = \frac{\omega - \omega_{ref}}{\omega_{ref}} \tag{1.10}$$

In equation 1.10, ω is the measured frequency, ω_{ref} is the frequency of the reference structure and δ is the chemical shift. The chemical shift is measured in parts per million (ppm) and every chemical shift gives a characteristic peak in the spectrum. Chemical shifts are very sensitive to their local environment [32–37]. The sensitivity makes it possible to reconstruct the local surrounding of a nucleus through their shielding effect and allows to solve structures [38].

Quantum mechanical approaches give the highest accuracy in the prediction of chemical shifts. However, these methods are limited to very few amino acids [39, 40]. Most state-of-the-art chemical shift predictors can be divided into two groups: sequence-based chemical shift predictors like CS-Rosetta [38], CHESHIRE [41] or SPARTA [42] which assign chemical shifts to atoms based on a homology modelling with data bank fragments, and structure-based chemical shift predictors like ShiftX [43] and CAMSHIFT [44] which are based on the atomic positions and calculate chemical shifts with the help of empirical functions. Hybrid methods like TALOS+ [45], TALOS-N [46] and SHIFTX2 [47] can significantly improve the accuracy.

Among the referenced chemical shift predictors, just CAMSHIFT [44] calculates the chemical shifts with the help of differentiable functions from the atomic coordinates. This allows to calculate forces and to include chemical shifts as structural restraints in molecular dynamics (MD) [48] and basin-hopping [49] simulations.

Chapter 2

Aims

The prediction of protein structures is one of the most fundamental challenges in biochemistry. Every year several thousands of new protein structures are solved and saved in the RCSB protein data bank [28]. However, the experimental elucidation of protein structures is expensive and the resolution is limited. In principle, computational methods can reach atomistic resolution. Nowadays, MD simulations are able to fold proteins on time scales of μs and even ms using supercomputers, graphics processing units (GPU) and distributed computing projects. However, the applicability is still limited to small proteins and many folding processes happen on even longer time scales like seconds or minutes. In many of these situations, it is not possible to follow the folding process during a MD simulation in order to understand it. Instead, Monte Carlo simulations can be used. Under the assumption, that the global energy minimum of a protein corresponds to the native state [50], global optimization methods can be applied to determine the structure. The basin-hopping approach to global optimization is a very effective method to explore the energy landscape within a limited number of MC steps.

The aim of this thesis is to increase the efficiency of the basin-hopping method for protein structure prediction. To this end, three improvements to the method have been introduced: First, chemical shifts are exploited by combining the basin-hopping approach with chemical shift restraints using a penalty function. The approach is parametrized and applied to three peptides with complete and incomplete information of backbone and Cβ chemical shifts. Second, secondary structure assignments of individual residues are used to derive tertiary structures. The method is benchmarked for three peptides and successfully applied to proteins with more than 50 residues. Third, a new dihedral angle move set is introduced. The moves are based on the Ramachandran plots of the different amino acids. Moreover, state-of-the-art β sheet predictors are compared and their combination with the basin-hopping approach is presented.

In Chapter 3 the methods used in this work are presented. The three new developments for the basin-hopping approach are described in Chapter 4 together with their results. This chapter is divided into three subsections corresponding to three manuscripts, of which one is published, another one is accepted for publication and the third one is in preparation. Finally, the results are summarized in Chapter 5. The studies demonstrate that the basin-hopping approach to global optimization with improved MC moves and/or experimental restraints is on the route to become a powerful tool for *ab initio* protein structure prediction.

Chapter 3

Methods

3.1 Statistical ensembles

The probability density function (PDF) ρ in the phase space describes a statistical ensemble in classical mechanics. Quantum effects where the ensemble is defined by a density matrix $\hat{\rho}$ are ignored (see below). The microscopic state $\mathbf{X} = (\mathbf{q}_1, ..., \mathbf{q}_N, \mathbf{p}_1, ..., \mathbf{p}_N)$ is defined by the generalized coordinates $\mathbf{q} = (\mathbf{q}_1, ..., \mathbf{q}_N)$ and the momenta $\mathbf{p} = (\mathbf{p}_1, ..., \mathbf{p}_N)$ of the N particles in the ensemble. The PDF of a microstate can be fully described with this information: $\rho(t) = \rho(\mathbf{q}_1(t), ..., \mathbf{q}_N(t), \mathbf{p}_1(t), ..., \mathbf{p}_N(t))$.

3.1.1 Microcanonical ensemble

The thermodynamic microcanonical ensemble is thermally isolated and described by a constant number of particles N, constant volume V and constant energy E. The PDF for a microcanonical ensemble can be calculated by

$$\rho = \frac{1}{h^N C} \frac{1}{\Omega} f\left(\frac{H - E}{\omega}\right). \tag{3.1}$$

Here, C is the Boltzmann counting factor which takes into account that particles of the same kind are indistinguishable and exchangeable:

$$C = N_1! N_2! ... N_s! \tag{3.2}$$

with s as the number of particles of same kind, Ω the number of accessible microstates in the ensemble, H the total energy of the system (as an eigenvalue of the Hamiltonian), ω

the width of function f, and f a function which describes the range of energies in which to include energies. The number of microstates Ω can be calculated by

$$\Omega = \int_{\mathbf{X}} \frac{1}{h^N C} f\left(\frac{H-E}{\omega}\right) \mathrm{d}\mathbf{X} \tag{3.3}$$

and the probability $P(E)$ to find the system in a specific microstate with energy E is

$$P(E) = \frac{1}{\Omega(E)}. \tag{3.4}$$

3.1.2 Canonical ensemble

The thermodynamic canonical ensemble can exchange heat with a surrounding heat bath and is described by a constant number of particles N, constant volume V and constant temperature T. The PDF for a canonical ensemble can be calculated by

$$\rho = \frac{1}{h^N C} \exp\left(\frac{A-E}{k_B T}\right). \tag{3.5}$$

Here, k_B is Boltzmann's constant and A a normalization factor which ensures that ρ is a normalized function:

$$\exp\left(-\frac{A}{k_B T}\right) = \int_{\mathbf{X}} \frac{1}{h^N C} \exp\left(-\frac{E}{k_B T}\right) \mathrm{d}\mathbf{X}. \tag{3.6}$$

The probability $P(E_m)$ to find the system in a specific microstate m with energy E_m can be calculated by replacing the integral with a sum via

$$P(E_m) = \frac{\mathrm{e}^{-E_m/k_B T}}{\sum_m \mathrm{e}^{-E_m/k_B T}} \tag{3.7}$$

3.1.3 Grand canonical ensemble

The thermodynamic grand canonical ensemble can exchange heat and particles with the environment and is described by constant chemical potential μ, constant volume V and constant temperature T. The probability $P(E_m, N_m)$ of a microstate with energy E_m and number of particles N_m is

$$P(E_m, N_m) = \mathrm{e}^{\left(\Omega_{grand} + N_m \mu - E_m\right)/k_B T} \tag{3.8}$$

where $\Omega_{grand}(\mu, V, T)$ is the grand canonical potential. It is a function of μ, V and T, but is constant for every microstate. It ensures the normalization and allows an easy calculation of thermodynamic properties. $\Omega_{grand}(\mu, V, T)$ can be generally calculated via

$$\Omega_{grand}(\mu, V, T) = -k_B T \ln\left(\sum_m \mathrm{e}^{\frac{\mu N_m - E_m}{k_B T}}\right) \tag{3.9}$$

The sum goes over all accessible microstates in the system and is different for bosons (Bose-Einstein distribution), fermions (Fermi-Dirac distribution) and indistinguishable classical particles. The probability $P(E_m, N_m)$ to find the system in a specific microstate with energy E_m and number of particles N_m is thus

$$P(E_m, N_m) = \frac{\mathrm{e}^{(N_m\mu - E_m)/k_B T}}{\sum_m \mathrm{e}^{(N_m\mu - E_m)/k_B T}} \tag{3.10}$$

With the help of the fugacity, which is a measure of how easy it is to add a particle to or remove a particle from the system,

$$z = \mathrm{e}^{\frac{\mu}{k_B T}}, \tag{3.11}$$

the grand canonical partition function

$$\Xi = \sum_m \mathrm{e}^{(N_m\mu - E_m)/k_B T} \tag{3.12}$$

can be calculated from the canonical partition function

$$Z = \sum_m \mathrm{e}^{-E_m/k_B T} \tag{3.13}$$

via

$$\Xi(T, V, \mu) = \sum_{N=0}^{\infty} z^N Z(T, V, N). \tag{3.14}$$

3.2 Principle of stationary action

The equations of motions (EOM) for all systems in mechanics can be derived from the principle of stationary action. The principle says that a system which develops in time between two time steps t_0 and t_1 takes the path with an action S which is stationary to first order,

$$\delta S = 0, \tag{3.15}$$

where δ is a small change in this formalism. S is a functional of the general coordinates $\mathbf{q} = (\mathbf{q}_1, ..., \mathbf{q}_N)$ and can be written as an integral over the Lagrangian L between the time steps t_0 and t_1:

$$S\left[\mathbf{q}\left(t\right)\right] = \int_{t_0}^{t_1} L\left(\mathbf{q}\left(t\right), \dot{\mathbf{q}}\left(t\right), t\right) \mathrm{d}t, \tag{3.16}$$

with N being the number of particles in the system. Putting equations 3.15 and 3.16 together we get

$$\delta \int_{t_0}^{t_1} L\left(\mathbf{q}, \dot{\mathbf{q}}, t\right) \mathrm{d}t = 0. \tag{3.17}$$

If we change our coordinates $\mathbf{q}$ by an infinitesimal small variation $\epsilon\eta\left(t\right)$ with $\eta\left(t_0\right) = 0$ and $\eta\left(t_1\right) = 0$ and a constant, but small factor ϵ we obtain

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} \int_{t_0}^{t_1} L\left(\mathbf{q}\left(t\right) + \epsilon\eta\left(t\right), \dot{\mathbf{q}}\left(t\right) + \epsilon\dot{\eta}\left(t\right), t\right) \mathrm{d}t = 0. \tag{3.18}$$

This leads to

$$\int_{t_0}^{t_1} \left(\frac{\partial L}{\partial \mathbf{q}}\eta + \frac{\partial L}{\partial \dot{\mathbf{q}}}\dot{\eta}\right) \mathrm{d}t = 0 \tag{3.19}$$

and after partial integration to

$$\int_{t_0}^{t_1} \left(\frac{\partial L}{\partial \mathbf{q}} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\mathbf{q}}}\right) \eta\left(t\right) \mathrm{d}t = 0. \tag{3.20}$$

This is only valid for all possible small variations $\eta\left(t\right)$ if the Euler-Lagrange-Equation is fulfilled:

$$\frac{\partial L}{\partial \mathbf{q}} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\mathbf{q}}} = 0 \tag{3.21}$$

3.3 Conservative force field

The principle of stationary action can be applied to calculate the EOM for a particle in a conservative force field. In such a force field, the curl of force $\mathbf{F}$ and the net work by moving a particle in this force field with same start and end point vanishes:

$$\nabla \times \mathbf{F} = 0 \tag{3.22}$$

$$\oint_C \mathbf{F} \cdot \mathrm{d}\mathbf{q} = 0. \tag{3.23}$$

The independence of the path C allows us to write the force as a negative gradient of a potential $V(\mathbf{q})$:

$$\mathbf{F} = -\nabla V(\mathbf{q}) \,. \tag{3.24}$$

We can use this potential to write the Lagrangian as the difference of kinetic and potential energy:

$$\begin{aligned} L(\mathbf{q}, t) &= T(\dot{\mathbf{q}}) - V(\mathbf{q}) && (3.25) \\ &= \frac{m}{2}\dot{\mathbf{q}}^2 - V(\mathbf{q}) \,. && (3.26) \end{aligned}$$

With the help of the Euler-Lagrange-Equation 3.21 we get Newton's second law or Newton's EOM:

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{q}} &= -\frac{\partial V}{\partial \mathbf{q}} = \mathbf{F} && (3.27) \\ \frac{\partial L}{\partial \dot{\mathbf{q}}} &= m\dot{\mathbf{q}} = \mathbf{p} && (3.28) \\ \frac{\partial L}{\partial \mathbf{q}} &= \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\mathbf{q}}} && (3.29) \\ \mathbf{F} &= \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(m\dot{\mathbf{q}}) \,. && (3.30) \end{aligned}$$

Here, $\mathbf{p}$ is the momentum of the particle. In case of a nonrelativistic particle, where the mass is independent of the time, the force acting on a particle is proportional to the acceleration:

$$\mathbf{F} = m\ddot{\mathbf{q}}. \tag{3.31}$$

3.4 Born-Oppenheimer approximation

Electrons are quantum particles. Equation 3.31 cannot be used to describe the movement of electrons. Here, electronic wave functions play a fundamental role, which are described by the Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}\chi(\mathbf{q}, t) = \hat{H}\chi(\mathbf{q}, t) \,. \tag{3.32}$$

$\hat{H}$ is the Hamilton operator and χ the wave function of the quantum system. If the Hamiltonian does not depend explicitly on time, one can write the solution of Equation

3.32 as a product of a time-dependent part Ψ_t and a time-independent part Ψ_{tind},

$$\chi\left(\mathbf{q}, t\right) = \Psi_t \Psi_{tind}. \tag{3.33}$$

The time-independent wave function Ψ_{tind} (which will be called the wave function Ψ from this point) can be calculated from the time-independent Schrödinger equation

$$\hat{H}\Psi = E\Psi. \tag{3.34}$$

Here, E is the eigenvalue and Ψ the eigenstate of the Hamilton operator, i.e. E is the energy of eigenstate Ψ. In case of a nonrelativistic particle in a potential $V\left(\mathbf{q}, \mathbf{Q}\right)$, equation 3.34 can be written as

$$E\Psi\left(\mathbf{q}, \mathbf{Q}\right) = \left(-\frac{\hbar}{2m}\nabla^2 + V\left(\mathbf{q}, \mathbf{Q}\right)\right)\Psi\left(\mathbf{q}, \mathbf{Q}\right). \tag{3.35}$$

Here, $\mathbf{q}$ and $\mathbf{Q}$ are the electronic and the nuclear coordinates, respectively. The kinetic energy of the nuclei which contain heavy protons and neutrons is much smaller than the kinetic energy of the electrons because the mass of the nucleus m_N is much larger than the mass of an electron m_e:

$$\frac{\mathbf{P}^2}{2m_N} << \frac{\mathbf{p}^2}{2m_e}. \tag{3.36}$$

Here, we assume that we are in the center of mass (COM) system where the momenta $\mathbf{P}$ of nuclei and $\mathbf{p}$ of electrons are equal and opposite. In this case, we can solve the electronic Schrödinger equation in which the kinetic energy of the nuclei is subtracted from the full Hamiltonian,

$$\hat{H}_e\left(\mathbf{q}, \mathbf{Q}\right)\Psi_e\left(\mathbf{q}; \mathbf{Q}\right) = E_e\left(\mathbf{Q}\right)\Psi_e\left(\mathbf{q}; \mathbf{Q}\right) \tag{3.37}$$

with

$$\Psi\left(\mathbf{q}, \mathbf{Q}\right) = \Psi_e\left(\mathbf{q}; \mathbf{Q}\right)\Psi_N\left(\mathbf{Q}\right), \tag{3.38}$$

where Ψ_N and Ψ_e are the wave functions of nuclei and electrons, respectively. The electronic Hamiltonian $\hat{H}_e$ is the sum of the kinetic energy of the electrons $\hat{T}_e$, the electron-electron interaction $\hat{V}_{ee}$ and the interaction of electrons with the nuclei $\hat{V}_{eN}$:

$$\hat{H}_e = \hat{T}_e + \hat{V}_{ee} + \hat{V}_{eN}. \tag{3.39}$$

The electronic energies $E_e\left(\mathbf{Q}\right)$ can be calculated by solving equation 3.37 repeatedly for small changes in the nuclear positions $\mathbf{Q}$. $E_e\left(\mathbf{Q}\right)$ is also called the potential energy surface

(PES). It is identical to the potential $V(\mathbf{Q})$ from equation 3.26. In the second step, the energy of the full system can be calculated by

$$\left(\hat{T}_N + E_e(\mathbf{Q})\right)\Psi_N(\mathbf{Q}) = E\Psi_N(\mathbf{Q}). \tag{3.40}$$

E in Equation 3.40 is the total energy of the molecule and Ψ_N the nuclear wave function. Equation 3.40 is a quantum mechanical EOM. However, assuming the movement of the faster electrons as instantaneous (Born-Oppenheimer approximation) and the nuclei as point particles allows the approximation that the nuclei follow Newton's nonrelativistic EOM, equation 3.31.

3.5 Potential energy functions

For molecular simulations, the exact PES from equation 3.37 is approximated by statistical functions. Here, one empirically models the potential energy contributions in equation 3.39 without explicitly representing electrons. The potential energy is a sum of bonded (V_{bonded}) and nonbonded ($V_{nonbonded}$) interactions:

$$V_{total} = V_{bonded} + V_{nonbonded}. \tag{3.41}$$

The bonded interactions are the result of covalent bonds between the atoms. They are a superposition of a term for bonds (V_{bond}), an angle term (V_{angle}), a dihedral angle term ($V_{dihedral}$) and an improper dihedral angle term ($V_{improper}$). Improper dihedral angles are used to ensure the planarity of aromatic groups, the amide group and to enforce the correct chirality (see section 1.1.1):

$$V_{bonded} = V_{bond} + V_{angle} + V_{dihedral} + V_{improper}. \tag{3.42}$$

The nonbonded interactions describe interactions between atoms which are separated by at least 3 covalent bonds. Without external fields (e.g. magnetic fields) and constraints (e.g. chemical shift constraints, see below), nonbonded interactions are the sum of Lennard-Jones (LJ) interactions (V_{LJ}) and electrostatic interactions (V_{elec}):

$$V_{nonbonded} = V_{LJ} + V_{elec}. \tag{3.43}$$

The different terms are described in detail in the following sections.

3.5.1 Bond potential

Vibrational motions occur between two covalently bonded atoms i and j. The potential energy V_{bond} is a function of the distance d between the two atoms and has a minimum at the equilibrium distance d_{equi}. If the distance between d and d_{equi} is small, the function can be approximated as an harmonic oscillator around this minimum. The approximation is usually used for all bonds in molecular mechanics force fields as more advanced potentials like the Morse potential are computationally more demanding.

$$\begin{aligned}
V_{bond}\left(d\right) = \sum_{bonds} & \left[V_{bond}\left(d_{equi}\right) + \left[\frac{\partial}{\partial d} V_{bond}\left(d_{equi}\right)\right]\left(d - d_{equi}\right)\right. && (3.44)\\
& + \frac{1}{2}\left[\frac{\partial^2}{\partial d^2} V_{bond}\left(d_{equi}\right)\right]\left(d - d_{equi}\right)^2 && (3.45)\\
& \left. + \frac{1}{6}\left[\frac{\partial^3}{\partial d^3} V_{bond}\left(d_{equi}\right)\right]\left(d - d_{equi}\right)^3 + ...\right] && (3.46)\\
\approx \sum_{bonds} & \left[V_{bond}\left(d_{equi}\right) + \left[\frac{\partial}{\partial d} V_{bond}\left(d_{equi}\right)\right]\left(d - d_{equi}\right)\right. && (3.47)\\
& \left. + \frac{1}{2}\left[\frac{\partial^2}{\partial d^2} V_{bond}\left(d_{equi}\right)\right]\left(d - d_{equi}\right)^2\right]. && (3.48)
\end{aligned}$$

The first term is just a constant which can be set to 0: $V_{bond}\left(d_{equi}\right) = 0$. The first derivative of V_{bond} vanishes at the minimum d_{equi}. The second derivative is the spring constant of the harmonic oscillator:

$$k = \frac{\partial^2}{\partial d^2} V_{bond}\left(d_{equi}\right). \tag{3.49}$$

The spring constant is specific to every atom type pair (e.g. C-C, O-O, C-H). The bond term can thus be simplified to

$$V_{bond}\left(d\right) = \sum_{bonds} \frac{1}{2} k_{ij}\left(d_{ij} - d_{ij}^0\right)^2. \tag{3.50}$$

Here k_{ij} is the spring constant for the bond between atoms i and j, d_{ij} is the distance between them, and d_{ij}^0 their equilibrium distance. The sum goes over all covalent bonds in the molecule.

3.5.2 Angle potential

The relative position of three atoms i, j and k is important, when atoms i and j and atoms j and k are covalently bonded while atoms i and k are not covalently bonded. The vector from atom j to atom i is called $\mathbf{v}_{ji}$. Similarly, the vector from atom j to atom k is called $\mathbf{v}_{jk}$. The angle between $\mathbf{v}_{ji}$ and $\mathbf{v}_{jk}$ is called θ_{ijk}. The potential energy function is harmonically approximated around the equilibrium angle θ^0_{ijk} like the distance around the equilibrium distance for the bonds. The angle part of the potential energy V_{angle} can then be calculated with Hookes law:

$$V_{angle} = \sum_{angles} \frac{1}{2} k_{ijk} \left(\theta_{ijk} - \theta^0_{ijk}\right)^2 . \tag{3.51}$$

The function is a superposition of all valence angles of the molecule. It is easier to distort an angle θ_{ijk} from its equilibrium θ^0_{ijk} than to distort the bond d_{ij} from its equilibrium distance d^0_{ij}, which means that the force constants k_{ijk} for angle bending are typically smaller than the force constants k_{ij}.

3.5.3 Torsion angle potential

Dihedral angles involve four atoms and are described in section 1.1.1. Let Φ_{ijkl} be the dihedral angle between atoms i, j, k and l and let Φ^0_{ijkl} be its equilibrium value. The potential energy for dihedral angle changes is a sum over cosine functions of the deviations of the dihedral angle from its equilibrium:

$$V_{dihedral} = \sum_{dihedrals} \frac{V_{ijkl}}{2} \left[1 + \cos\left(n\Phi_{ijkl} - \Phi^0_{ijkl}\right)\right] . \tag{3.52}$$

The periodicity allows values for the dihedral angle Φ_{ijkl} in the interval $[-180°, 180°)$ (see section 1.1.1) where $\Phi_{ijkl} = 0°$ is the cis- and $\Phi_{ijkl} = -180°$ the trans-configuration of the dihedral angle Φ_{ijkl}. V_{ijkl} describes the energy barrier for the torsional motion and n its periodicity, an integer which describes the number of minima/maxima in the interval $[-180°, 180°)$. The energy needed to distort a dihedral angle Φ_{ijkl} from its equilibrium angle Φ^0_{ijkl} is typically smaller than the energy needed for the distortion of an angle θ_{ijk} from its equilibrium θ^0_{ijk} and for the distortion of a bond length d_{ij} from its equilibrium value d^0_{ij}. While bonds and angles only change slightly, structural changes are mostly expected from changes of dihedral angles.

3.5.4 Improper dihedral potential

Improper torsion angles are used to ensure the correct geometry and chirality of a specific configuration. The improper torsion angle has a similar functional form than the bond and angle parts of the potential energy:

$$V_{improper} = \sum_{improper} \frac{1}{2} k_{ijkl} \left(\Psi_{ijkl} - \Psi^0_{ijkl} \right)^2 . \tag{3.53}$$

Here Ψ_{ijkl} is the improper dihedral angle involving atoms i, j, k and l, and Ψ^0_{ijkl} is the equilibrium value. Assuming atoms i, k and l are covalently bonded to atom j, the improper dihedral angle is defined as the angle between the line defined by atoms i and j and the plane defined by atoms j, k and l. Improper dihedral angles are used to ensure a desired planarity of the four atoms O, C', Cα and N because it is computationally cheaper to use equation 3.53 with a high spring constant k_{ijkl} than to use equation 3.52. Improper dihedrals are also used for other planar structures like ester or aromatic ring structures.

3.5.5 Lennard Jones potential

The Lennard Jones potential V_{LJ} describes the interaction between atoms, which are not covalently bonded and which are separated by at least three covalent bonds. The interaction has a repulsive part for small distances and a weak attractive part for longer distances. The interactions between two particles i and j are usually described by the 12-6-Lennard Jones potential V_{LJ}:

$$V_{LJ} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] . \tag{3.54}$$

Here, σ_{ij} is the distance where the attractive and repulsive part of the LJ potential cancel out and ϵ_{ij} is the energy at the minimum of this potential. The preferred distance corresponds to the minimum of the LJ potential at $r_{ij,min} = 2^{\frac{1}{6}} \sigma_{ij}$. The sum in equation 3.54 goes over all $N(N-1)/2$ atom pairs in the system.

The attractive term r_{ij}^{-6} is proportional to the dispersion interaction V_{ij}^{disp} between two atoms i and j:

$$V_{ij}^{disp} \approx \frac{3}{2} \frac{I_i I_j}{I_i + I_j} \frac{\alpha_i \alpha_j}{r^6} . \tag{3.55}$$

Here, α_i and α_j are the dipole polarizibilities of atoms i and j and I_i and I_j their first ionization potentials.

The repulsive term r_{ij}^{-12} has its origin in the Pauli exclusion principle and also models the internuclear repulsion.

3.5.6 Electrostatic potential

The electric field $\mathbf{E}$ for a charge density ρ in vacuum can be calculated from the (second) Maxwell equation:

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0}. \tag{3.56}$$

For a protein in solution, we can assume that the charges are point charges. In this case we can transform equation 3.56 for a (homogeneous) material with a dielectric constant ϵ_r to

$$\mathbf{E} = \sum_{i=1}^{N} \frac{q_i}{4\pi\epsilon_0\epsilon_r r_i}\mathbf{e}_r, \tag{3.57}$$

where $\mathbf{e}_r$ is the normal vector in the direction of the electric field $\mathbf{E}$ and ϵ_0 the electric permittivity in vacuum. The sum goes over all N charged particles i with charges q_i at positions r_i. This is the electric field a particle j with charge q_j would feel, creating a force $\mathbf{F}_{el}$ on this particle in absence of a magnetic field $\mathbf{B}$ according to Lorentz's law:

$$\mathbf{F}_{el} = q_j\left(\mathbf{E} + \mathbf{v} \times \mathbf{B}\right) = \sum_{i=1}^{N} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}}\mathbf{e}_r, \tag{3.58}$$

where r_{ij} is the distance between particles i and j. Equation 3.58 is the Coulomb force and the electrostatic interaction between point charges is called Coulomb interaction. Integrating equation 3.58 according to equation 3.24 and summing over all N charges gives the potential energy for the electrostatic interactions V_{elec}:

$$V_{elec} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}}. \tag{3.59}$$

The slow r_{ij}^{-1} decay is the reason why Coulomb interactions are long-range interactions. The calculation of the double sum in the nonbonded interactions V_{LJ} and V_{elec} is the time consuming part in a simulation with many particles. Usually, smoothing functions and cutoff distances are used to reduce this problem.

3.6 Monte Carlo simulations

Protein folding is a complex mechanism. In principle, a full description of the folding pathway can be obtained with MD simulations. MD simulations of proteins are mainly performed by solving the EOM in a canonical system which is coupled to its environment via a thermostat. However, MD simulations of complex systems are extremely slow. For most proteins, it is not possible to simulate folding events under experimental conditions (time scale of seconds) with currently available computers.

However, it is not needed to know the full trajectory of a folding pathway explicitly to understand the folding mechanism. Monte Carlo (MC) simulations allow a stochastic sampling of the relevant conformational phase space in order to get approximations for statistical properties. In a MC step, a microstate $\mathbf{X}$ is globally or locally modified to a new microstate $\mathbf{Y}$. The new microstate is accepted or rejected with the transition probability $T\left[\mathbf{X} \rightarrow \mathbf{Y}\right]$. A MC move corresponds to a discrete time step $\Delta\tau_s$ in a MD simulation. Usually, a statistical property Π is measured every a MC steps with $a > 1$ because of the existence of unsuccessful moves. A sweep $\Delta\tau$ is called the time between two updates of the measurement of a statistical quantity: $\Delta\tau = a\Delta\tau_s$. The ergodicity theorem allows then to calculate the statistical ensemble average $\langle\Pi\rangle$ of Π for a high enough number of measurements M starting with the first measurement at τ_0:

$$\bar{\Pi} = \lim_{M\to\infty} \frac{1}{M} \sum_{a=1}^{M} \Pi\left(\tau_0 + a\Delta\tau\right) \equiv \langle\Pi\rangle = \int \Pi\left(\mathbf{X}\right) \rho\left(\mathbf{X}\right) \mathrm{d}\mathbf{X} \tag{3.60}$$

with $\rho\left(\mathbf{X}\right)$ as the microstate probability distribution of microstate $\mathbf{X}$. The probability distribution should also be time independent, which means that the probability to move from $\mathbf{X}$ to $\mathbf{Y}$ does not depend on the history which led the system progress to state $\mathbf{X}$. Such processes without explicit memory are called Markov processes. They can be described by the master equation:

$$\frac{\Delta\rho\left(\mathbf{X}\right)}{\Delta\tau_s} = \sum_{\mathbf{Y}} \left[\rho\left(\mathbf{Y}\right) T\left(\mathbf{Y} \rightarrow \mathbf{X}; \Delta\tau_s\right) - \rho\left(\mathbf{X}\right) T\left(\mathbf{X} \rightarrow \mathbf{Y}; \Delta\tau_s\right)\right]. \tag{3.61}$$

The ensemble is in a stationary state if $\Delta\rho\left(\mathbf{X}\right)/\Delta\tau_s = 0$. If the new state $\mathbf{Y}$ is not allowed to be extremely different than the previous state $\mathbf{X}$, the inner brackets of the previous equation have to vanish which is the condition for detailed balance:

$$\frac{T\left(\mathbf{X} \rightarrow \mathbf{Y}; \Delta\tau_s\right)}{T\left(\mathbf{Y} \rightarrow \mathbf{X}; \Delta\tau_s\right)} = \frac{\rho\left(\mathbf{Y}\right)}{\rho\left(\mathbf{X}\right)} \tag{3.62}$$

which is independent of the time step $\Delta\tau_s$. For a canonical system, equation 3.62 results into

$$\frac{T\left(\mathbf{X} \rightarrow \mathbf{Y}\right)}{T\left(\mathbf{Y} \rightarrow \mathbf{X}\right)} = \mathrm{e}^{-\beta\Delta E} \tag{3.63}$$

with $\beta = 1/(k_B T)$ and $\Delta E = E(\mathbf{Y}) - E(\mathbf{X})$. The transition probability $T(\mathbf{X} \to \mathbf{Y})$ is usually a product of the selection probability $S(\mathbf{X} \to \mathbf{Y})$ and the acceptance probability $A(\mathbf{X} \to \mathbf{Y})$. Here, S is the probability for a specific move being chosen and A is the probability for this move being accepted. The acceptance probability is typically written as

$$A(\mathbf{X} \to \mathbf{Y}) = \min\left[1, \frac{S(\mathbf{Y} \to \mathbf{X})}{S(\mathbf{X} \to \mathbf{Y})} \frac{\rho(\mathbf{Y})}{\rho(\mathbf{X})}\right] \tag{3.64}$$

Usually, a protein is most of the time in one of the stable states which means that the number of relevant states is limited. The probability density ρ outside these states is neglectable. A Markov chain Monte Carlo simulation sampling these more populated states is called importance sampling. The most famous importance sampling method is the Metropolis method. Here, the probability density $\rho(\mathbf{X})$ is set to the canonical microstate probability at a given temperature T. In this case, the acceptance probability can be written as

$$A(\mathbf{X} \to \mathbf{Y}) = \min\left[1, \frac{S(\mathbf{Y} \to \mathbf{X})}{S(\mathbf{X} \to \mathbf{Y})} \mathrm{e}^{-\beta[E(\mathbf{Y}) - E(\mathbf{X})]}\right] \tag{3.65}$$

For methods, in which detailed balance is not needed, e.g. in basin-hopping (see section 3.8) the ratio of forward and backward selection probabilities can be neglected and the Metropolis acceptance criterion simplifies to

$$A(\mathbf{X} \to \mathbf{Y}) = \min\left[1, \mathrm{e}^{-\beta[E(\mathbf{Y}) - E(\mathbf{X})]}\right] \tag{3.66}$$

Thus, a move is always accepted if the energy of the new microstate $\mathbf{Y}$ is smaller than the energy of the previous microstate $\mathbf{X}$. The new microstate can also be accepted if its energy is higher than the energy of the previous microstate, but this probability decays exponentially with the energy difference between the new and the previous microstate.

3.7 Global optimization

The aim of global optimization is to find the optimum of a (or a set of) function(s) f according to some criteria. A typical example is the task to find the global minimum of a function f. Global optimization methods have a broad range of applications like curve fitting, the travelling salesman problem [51], flight planning, travel circuit or protein structure prediction. In protein structure prediction, mainly stochastic optimization methods are used which use random variables to explore the energy landscape of the protein where the energy is typically given by equation 3.41. Examples are simulated annealing [52], parallel tempering [53] or stochastic tunneling [54].

3.8 Basin-Hopping

In the basin-hopping (BH) approach to global optimization [55–57] moves are proposed by perturbing the current geometry, and are accepted or rejected based upon the energy difference between the local minimum obtained by minimization from the instantaneous configuration and the previous minimum in the chain. In effect the potential energy surface is transformed into the basins of attraction [58, 59] of all the local minima, so that the energy for configuration $\mathbf{q}$ is

$$\widetilde{V}(\mathbf{q}) = \min\{V(\mathbf{q})\}, \tag{3.67}$$

where min denotes minimization. Large steps can be taken to sample this transformed landscape, since the objective is to step between local minima. Furthermore, there is no need to maintain detailed balance when taking steps, because the BH approach attempts to locate the global potential energy minimum and is not intended to sample thermodynamic properties. The BH algorithm has been implemented in the GMIN program [60] and has already been employed to find the global minimum of peptides and peptide complexes in previous work [49, 61–69].

3.9 Limited Broyden–Fletcher–Goldfarb–Shanno algorithm

In order to find the local minimum of a biological molecule, minimization methods are used with the aim to find the shortest way from the current position $\mathbf{q}_c$ to the next local minimum $\mathbf{q}_{min}$ in terms of minimization steps. $\mathbf{q}_{min}$ is a local minimum if there exists an $\epsilon > 0$ for which all $\mathbf{q}_{env}$ with $||\mathbf{q}_{env} - \mathbf{q}_{min}|| < \epsilon$ have an energy with $V(\mathbf{q}_{min}) < V(\mathbf{q}_{env})$. Gradient descent, Newton's method and conjugate gradient techniques are among the most used minimization methods. Gradient descent (also known as steepest descent) methods calculate the next minimization step proportional to the negative of the gradient of the current position at the PES. This ensures that the function decreases fastest for small enough step sizes. Newton's methods and conjugate gradient techniques are computationally more expensive than gradient descent methods. On the other hand, they find the minimum within fewer minimization steps [70].

In Newton's method, the search for the optimal step $\Delta\mathbf{q}$ is based on a Taylor approximation of the PES function $V(\mathbf{q})$ around the actual position $\mathbf{q}_c$ until second order:

$$V_T(\mathbf{q}_c + \Delta\mathbf{q}) = V(\mathbf{q}_c) + \nabla V(\mathbf{q}_c)\Delta\mathbf{q} + \frac{1}{2}\left[\mathbf{H}V(\mathbf{q}_c)\right](\Delta\mathbf{q})^2 . \tag{3.68}$$

Here, $\mathbf{H}$ is the Hessian matrix. Newton's method minimizes the function $V_T(\mathbf{q}_c + \Delta\mathbf{q})$ with respect to the step $\Delta\mathbf{q}$:

$$\nabla V(\mathbf{q}_c) + \left[\mathbf{H}V(\mathbf{q}_c)\right]\Delta\mathbf{q} = 0 \tag{3.69}$$

yielding the step

$$\Delta\mathbf{q} = -\left[\mathbf{H}V(\mathbf{q}_c)\right]^{-1}\nabla V(\mathbf{q}_c) . \tag{3.70}$$

$\left[\mathbf{H}V(\mathbf{q}_c)\right]^{-1}$ is the inverse Hessian matrix of the PES function V at point $\mathbf{q}_c$. The new position can be calculated with

$$\mathbf{q}_{new} = \mathbf{q}_c + \Delta\mathbf{q}. \tag{3.71}$$

The calculation of the inverse of the Hessian matrix $\left[\mathbf{H}V(\mathbf{q}_c)\right]^{-1}$ is computationally very expensive, especially in a high-dimensional space like for a molecule with $3N-6$ degrees of freedom. It is faster to calculate $\mathbf{y} = \left[\mathbf{H}V(\mathbf{q}_c)\right]^{-1}\nabla V(\mathbf{q}_c)$ as the solution of the system of linear equations

$$\left[\mathbf{H}V(\mathbf{q}_c)\right]\mathbf{y} = \nabla V(\mathbf{q}_c) . \tag{3.72}$$

However, the Hessian has still to be calculated.
In quasi-Newton methods, an approximation $\mathbf{A}$ for the Hessian matrix $\left[\mathbf{H}V(\mathbf{q}_c)\right]$ from the gradient $\nabla V(\mathbf{q}_c)$ is calculated. The approximation should fulfill the Taylor expansion for the gradient ∇V around $\mathbf{q}_c$:

$$\nabla V(\mathbf{q}_c + \Delta\mathbf{q}) = \nabla V(\mathbf{q}_c) + \mathbf{A}\Delta\mathbf{q}. \tag{3.73}$$

The quasi-Newton methods differ in the way how the Hessian matrix approximation $\mathbf{A}$ is calculated. In the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, the new approximation $\mathbf{A}_{new}$ is calculated from the previous one $\mathbf{A}_{old}$ by

$$\mathbf{A}_{new} = \mathbf{A}_{old} + \frac{\mathbf{g}\mathbf{g}^T}{\mathbf{g}^T\mathbf{d}} - \frac{\mathbf{A}_{old}\mathbf{d}\mathbf{d}^T\mathbf{A}_{old}}{\mathbf{d}^T\mathbf{A}_{old}\mathbf{d}}. \tag{3.74}$$

Here, $\mathbf{g}$ is the gradient difference,

$$\mathbf{g} = \nabla V(\mathbf{q}_c + \Delta\mathbf{q}) - \nabla V(\mathbf{q}_c) , \tag{3.75}$$

and $\mathbf{d} = \alpha\mathbf{y}$ the step with $\mathbf{y}$ being the direction of the step as the solution of

$$\mathbf{A}_{old}\mathbf{y} = -\nabla V\left(\mathbf{q}_c\right) \tag{3.76}$$

and α an acceptable step size:

$$\mathbf{q}_{new} = \mathbf{q}_c + \alpha\mathbf{y}. \tag{3.77}$$

The superscript T indicates the transpose of the matrix. Usually the inverse matrix is updated

$$\mathbf{A}_{new}^{-1} = \mathbf{A}_{old}^{-1} + \frac{\left(\mathbf{d}^T\mathbf{g} + \mathbf{g}^T\mathbf{A}_{old}^{-1}\mathbf{g}\right)\left(\mathbf{d}\mathbf{d}^T\right)}{\left(\mathbf{d}^T\mathbf{g}\right)^2} - \frac{\mathbf{A}_{old}^{-1}\mathbf{g}\mathbf{d}^T + \mathbf{d}\mathbf{g}^T\mathbf{A}_{old}^{-1}}{\mathbf{d}^T\mathbf{g}}. \tag{3.78}$$

The limited BFGS (L-BFGS) algorithm does not store the full approximated matrix $\mathbf{A}$ or its inverse $\mathbf{A}^{-1}$, but just a few vectors which represent the approximation. It only requires a linear memory and is used for optimization tasks with a large number of degrees of freedom like proteins. It is implemented in GMIN and used for the minimization following a Monte Carlo move in the basin-hopping method.

3.10 Monte Carlo moves for proteins

In GMIN [60], two kinds of MC moves (dihedral angle moves and Cartesian moves) and three move sets (standard move set, neighbour moves and loop modelling) are implemented for proteins. At every MC step, the structure of the molecule with the current Markov energy is perturbed by changing the positions of atoms from randomly chosen residues according to the MC move type.

In Cartesian moves, the position of an atom is changed from its previous position $\mathbf{q}_{old} = \left(q_{x,old}, q_{y,old}, q_{z,old}\right)^T$ to a new position $\mathbf{q}_{new} = \left(q_{x,new}, q_{y,new}, q_{z,new}\right)^T$ where all projections $q_{x,new}, q_{y,new}, q_{z,new}$ onto the corresponding Cartesian axis are in an interval around their projections of the previous position, e.g. $q_{i,new} \in [q_{i,old} - \alpha, q_{i,old} + \alpha]$ for $i \in \{x, y, z\}$. α is the maximum allowed change per Cartesian direction which is usually increased or decreased by a factor 1.05 after every 50 MC step in order to adjust the number of accepted moves in the direction of a predefined value. Cartesian moves can be used for testing some small local perturbations. However, as changes of atomic Cartesian coordinates tend to disrupt the bonded network of a protein (see section 1.1.1), Cartesian moves are not the best choice to study protein structure prediction or protein folding.

Typical MC moves for proteins are dihedral angle moves as dihedral angles describe the structure between neighbouring residues and also side chains (see section 1.1.1). Similarly to Cartesian moves, the backbone dihedral angles Φ and Ψ of randomly chosen residues are changed from their previous values Φ_{old} and Ψ_{old} to new values Φ_{new} and Ψ_{new} in the interval $\Phi_{new} \in [\Phi_{old} - \alpha, \Phi_{old} + \alpha]$ (same for Ψ). Here, α is the maximum dihedral angle change for the dihedral angles Φ and Ψ. The dihedral angle Ω for the peptide bond could also be changed, but is usually not perturbed because of its little variance. Small changes in Ω are supposed to occur from the minimization of the structure. The side chain dihedral angles χ_i are usually also changed. One option to change side chain dihedral angles are group rotation moves [71]. In group rotation moves, the side chain dihedral angles χ_1 (around the Cα-Cβ axis), χ_2 (around the Cβ-Cγ axis) and χ_3 (around the Cγ-Cδ axis) are perturbed, if possible, in three steps: First, all side chain dihedral angles are changed with a probability which is specific for this dihedral angle. Second, a new dihedral angle χ_{new} is calculated from the previous value χ_{old} with a maximum dihedral angle specific change α by choosing a random value from the interval $\chi_{new} \in [\chi_{old} - \alpha, \chi_{old} + \alpha]$. Third, the bond vector which connects the rotated group with the rest of the molecule is calculated and normalized. Backbone and side chain moves can be explicitly turned on and off which is one of the approaches described and analyzed in section 4.2. Dihedral angles allow to perform large configurational changes, especially for high values of α, as a dihedral angle change of an amino acid in the centre of a bigger protein has an influence on the positions of all atoms until the protein's termini. By doing so, the energy landscape can be analyzed rapidly over a huge conformational space which is a big advantage in comparison to MD simulations. Dihedral angle moves were successfully applied in basin-hopping simulations of proteins [49, 61–69].

The standard MC move set by Mortenson et al. [63] sets the probability to change the dihedral angles of a randomly chosen residue according to the position of this residue in the polypeptide chain. A maximum probability P_{max} is set to the residues at both termini of the chain while a minimum probability P_{min} is set to the center of the chain. The probability increases linearly from the center of the chain to the termini from P_{min} to P_{max}. This move set ensures that more residues from the termini are perturbed during the simulation while changes in the center, which have a bigger influence on the overall structure of the protein, are less frequently changed. Most of the previous BH studies [63–65, 67, 68] are based on this standard move set.

In neighbour moves, the dihedral angle changes of Φ and Ψ are not just performed for a randomly chosen residue, but also for the neighbouring residues (1 or 2 residues in both directions of the polypeptide chain). This move set is suitable for segment moves where not just single residues shall be changed, but full segments. The method is based on the assumption that the formation of local segments like α helices or β sheets is accomplished fastest when more than one residue contributing to this segment are perturbed in a cooperative process.

In loop modeling, the user can predefine regions of amino acids which should not be changed. In the (changeable) intermediate regions with residues $(N_1, ..., N_n)$, the probability to change a dihedral angle pair (Φ, Ψ) of an amino acid is maximal for the amino acid in the centre of this region $((N_1 + N_n)/2)$ and decreases linearly to its borders N_1 and N_n. This loop is specifically useful if the position of SS elements like α helices or β sheets are known before the start of the simulation. A further description and application of this method follows in section 4.2.

Chapter 4

Results

The aim of this thesis is to develop and test new tools for the usage of the basin-hopping approach to global optimization in order to predict the structures of proteins. The tools were included in the GMIN program and can be divided into three parts:

1) **Protein structure prediction using global optimization by basin-hopping with NMR shift restraints**

In this study, a modified version of the chemical shift predictor CAMSHIFT [44] is included in the GMIN program [60]. Hereby, the calculation of the total energy of the system and the force on the atoms is the sum of the force field (equation 3.41) and CAMSHIFT energies and forces for which the ratio can be modified by a factor α:

$$E = E_{FF} + \alpha E_{CAM} \tag{4.1}$$

$$\mathbf{F} = \mathbf{F}_{FF} + \alpha \mathbf{F}_{CAM}. \tag{4.2}$$

With the help of the peptides with PDB codes 1LE0 and 1L2Y, the ratio α and the tolerance parameter n of the CAMSHIFT penalty function are parametrized. Furthermore, a function for n is developed which allows to effectively adjust the parameter depending on the root mean square force (RMSF) threshold. Simulations for these peptides are performed and it is shown that the simulations with chemical shift restraints outperforms the simulations without restraints.

J. Chem. Phys., 138:025102, 2013 (impact factor (IF) 3.164). Complete execution of BH simulations and GMIN programming and 90% of the manuscript.

2) **Protein structure prediction: assembly of secondary structure elements by basin-hopping**

In this study, we focus on the improvement of the MC steps. MC steps are very important in basin-hopping simulations as they are the key factor for the effectiveness of the energy landscape exploration. Different SS predictors are compared to their accuracy and the best among them, Porter [20] is used to apply structural constraints on dihedral angles in the polypeptide chain. We further improve the efficiency of basin-hopping by introducing an approach that derives tertiary structures from the secondary structure assignments of individual residues. We term this approach secondary-to-tertiary basin-hopping and benchmark it for three miniproteins, trpzip, trp-cage and ER-10. Different step sizes are tested for their effectiveness to fold these peptides. Near-native structures with a RMSD of less than 2Å can be found for these three peptides. For the refinement of the protein structures, different ratios between backbone and side chain dihedral angle moves are used and compared. The structures can be improved significantly to a RMSD of lower than 1.8 Å for 1ERP and RMSDs lower than 0.5 Å for 1LE0 and 1L2Y. We also demonstrate that with this approach we can also fold bigger proteins with up to 79 residues.

ChemPhysChem, DOI: 10.1002/cphc.201402247, 2014 (impact factor (IF) 3.349). Complete analysis of all simulations, half of the BH simulations and 80% of the manuscript.

3) **Protein structure prediction using basin-hopping with knowledge-based Monte Carlo moves**

New MC moves are introduced in this study. The moves are based on the Ramachandran plots of the randomly chosen residues, which should be modified at a given MC step. A probability selection function based on the statistics of 488 proteins is developed and used to choose new dihedral angles during basin-hopping simulations. Sequence moves are developed to fold local protein sequences in one MC move. State-of-the-art programs for β predictions are compared and used for the application of β contact restraints in basin-hopping simulations.

Manuscript under preparation. GMIN programming, complete execution of BH simulations and 100% of the manuscript.

In the following the results of these three studies are presented as manuscripts, two of them published, and one in preparation. References given in the following sections refer to the individual manuscripts (and not to the references given at the end of this thesis).

4.1 Protein structure prediction using global optimization by basin-hopping with NMR shift restraints

THE JOURNAL OF CHEMICAL PHYSICS **138**, 025102 (2013)

Protein structure prediction using global optimization by basin-hopping with NMR shift restraints

Falk Hoffmann[1] and Birgit Strodel[1,2,a)]

[1]*Institute of Complex Systems: Structural Biochemistry, Research Centre Jülich, 52425 Jülich, Germany*
[2]*Institute of Theoretical and Computational Chemistry, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany*

(Received 30 September 2012; accepted 12 December 2012; published online 11 January 2013)

Computational methods that utilize chemical shifts to produce protein structures at atomic resolution have recently been introduced. In the current work, we exploit chemical shifts by combining the basin-hopping approach to global optimization with chemical shift restraints using a penalty function. For three peptides, we demonstrate that this approach allows us to find near-native structures from fully extended structures within 10 000 basin-hopping steps. The effect of adding chemical shift restraints is that the α and β secondary structure elements form within 1000 basin-hopping steps, after which the orientation of the secondary structure elements, which produces the tertiary contacts, is driven by the underlying protein force field. We further show that our chemical shift-restraint BH approach also works for incomplete chemical shift assignments, where the information from only one chemical shift type is considered. For the proper implementation of chemical shift restraints in the basin-hopping approach, we determined the optimal weight of the chemical shift penalty energy with respect to the CHARMM force field in conjunction with the FACTS solvation model employed in this study. In order to speed up the local energy minimization procedure, we developed a function, which continuously decreases the width of the chemical shift penalty function as the minimization progresses. We conclude that the basin-hopping approach with chemical shift restraints is a promising method for protein structure prediction. © *2013 American Institute of Physics*. [http://dx.doi.org/10.1063/1.4773406]

I. INTRODUCTION

The determination of protein structures is one of the most important challenges in biochemistry. Computational techniques can help find the three-dimensional arrangement of atoms. However, the exact determination of native structures from denatured or unfolded proteins is still a challenge. The usage of structural restraints obtained from experiments such as nuclear magnetic resonance (NMR) measurements shows significant improvement in this field of research.[1–11] About 12% of the structures saved in the RCSB protein data bank[12] are produced from NMR data. Chemical shifts are the most readily and accurately measurable NMR observables in solution and in the solid state,[5] and can be used to predict the molecular structure,[4–9,13–18] including the structure of a low-populated, on-pathway folding intermediate.[19]

Many of the simulations for NMR based structure determination use sequence homology information.[4,5,8,9,16,20] In such approaches structural motifs are selected from databases of existing protein structures based on NMR data, such as chemical shifts, residual dipolar couplings (RDCs), *J*-couplings, or nuclear Overhauser effect (NOE) data.[21] However, the usage of molecular fragment replacement approaches with chemical shift information depends on the structural model and cannot be easily used to calculate conformational changes or combined with RDC, *J*-couplings, or NOE data. Applying chemical shift restraints using a penalty function avoids these problems. Here, the Hamiltonian is applied such that it reduces the conformational search to structures with small shift restraints. This approach was used successfully to perform structural refinements of proteins.[6,7]

In the works by Vendruscolo and co-workers[6,7] the CamShift method[22] was used for the incorporation of chemical shift restraints. CamShift is a tool recently introduced for the rapid prediction of NMR chemical shifts from protein structures based on an approximation of the chemical shifts as polynomial functions of interatomic distances.[22] This chemical shift predictor is combined with a tunable soft-square harmonic well as a penalty function to compute the differences between calculated and experimental chemical shifts.[6,7] Furthermore, the chemical shifts are differentiable functions of the atomic coordinates, which enables the calculation of forces. Vendruscolo and co-workers were able to find the structures of a set of proteins with 56–108 residues with a resolution of 0.8–2.2 Å using CamShift molecular dynamics (MD) simulations of previously partially folded proteins.[7] The determination of peptide structures from unfolded conformations using a Monte Carlo (MC) approach was also possible with a simulated annealing (SA) protocol.[6]

In this study, we combine the basin-hopping (BH)[23,24] approach to global optimization with NMR chemical shift restraints using CamShift. The BH method, which is a generalization of the Monte Carlo-minimization approach,[25] has

[a)] Author to whom correspondence should be addressed. Electronic mail: b.strodel@fz-juelich.de.

0021-9606/2013/138(2)/025102/7/$30.00

been successfully used to identify the global minimum of peptides and proteins,[26–31] including structures of peptide complexes.[32–34] The availability of forces in CamShift enables us to combine it with the BH method. In this work we demonstrate that this approach allows us to find near-target structures from fully extended peptide conformations. We present the results from chemical shift-restrained BH simulations of three peptides with the PDB[12] codes 1LE0,[35] 1L2Y,[36] and 1YRF.[37] We show that we are able to find the folded structures within 10 000 BH steps, while the unrestrained BH simulations of same run length fail to locate near-native structures.

II. METHODS

A. Structural models

The structures for 1LE0, 1L2Y, and 1YRF were downloaded from the RCSB protein data bank[12] and used as target structures for the BH simulations. 1LE0 is a 12 amino acid β-hairpin;[35] 1L2Y is a 20 amino acid peptide with a short α-helix, a 3_{10}-helix, and a polyproline II helix at the C-terminus;[36] and 1YRF is a 35-residue subdomain of the villin headpiece consisting of three α-helices.[37] These minipeptides have been used as test cases in previous folding studies.[38–56] We employed CamShift[22] to calculate ^{1}Hα, amide ^{1}H, ^{13}Cα, ^{13}Cβ, carbonyl ^{13}C, and amide ^{15}N chemical shifts from the target structures and used them as target chemical shifts for the definition of the restraint function. These are denoted δ_{exp} in the following. Fully extended structures of the peptides were generated from their structural sequence using VMD[57] and employed as starting structures for the BH simulations (Figure S1 of the supplementary material[58]).

We used the CHARMM22 force field parameters[59,60] to model the peptides. To model the aqueous solvent we employed the generalized Born model FACTS.[61] For the calculation of the nonbonded interactions, the cutoff scheme suggested in the FACTS documentation was employed, i.e., truncation of both long-range electrostatics at 14 Å using a shift function and the van der Waals energy with a polynomial switching function applied between 10 and 12 Å.

B. Basin-hopping

The BH approach to global optimization[23,24] is analogous in principle to the Monte Carlo-minimization approach.[25] Moves are proposed by perturbing the current geometry and are accepted or rejected on the basis of the Metropolis criterion,[62] which uses the energy difference between the local minimum obtained by minimization from the instantaneous configuration and the previous minimum in the Markov chain. In effect, the potential energy surface is transformed into the basins of attraction of all the local minima so that the energy for configuration $\mathbf{r}$ is

$$\widetilde{E}(\mathbf{r}) = \min\{E(\mathbf{r})\}, \tag{1}$$

where "min" denotes local minimization. Large steps can be taken to sample this transformed landscape, since the objective is to step between local minima. Furthermore, there is no need to maintain detailed balance when taking steps because the BH approach attempts to locate the global potential energy minimum and is not intended to sample thermodynamic properties. The BH algorithm is implemented in the GMIN program.[63]

Basin-hopping has been employed successfully to find the global minimum of peptides and proteins,[26–31,64] including structures of peptide complexes.[32–34] In our study, we performed BH simulations using between 100 and 10 000 BH steps. The moves for perturbing the current geometry of the peptides were taken in backbone and sidechain dihedral angle space.[28] At each BH step, on average 30% of these dihedrals were randomly chosen and then twisted by an angle of maximally 60°. Dihedral angles which define planar structures, such as rings, were considered non-twistable to keep their planarity.[65] In all BH runs the temperature was set to 300 K. We use a limited-memory variation of the Broyden-Fletcher-Goldfarb-Shanno update by Nocedal[66] (LBFGS) for energy minimization.

C. Chemical shift restraints

We implemented chemical shift restraints into the GMIN program with a modified version of the program CamShift.[22] CamShift calculates chemical shifts using distance dependent functions of the atomic coordinates for the influence of backbone atoms, sidechain atoms, and nonbonded atoms. Furthermore, it also includes a dipole approach for the influence of aromatic rings and a parametrized function for dihedral angles. CamShift enables us to calculate chemical shifts quickly and accurately.[22] Furthermore, it allows to calculate forces from chemical shifts.

We use a soft-square harmonic potential as introduced by Vendruscolo and co-workers[6] to define the chemical shift penalty function E_{CS}, which restrains the structures to conformations in agreement with the chemical shifts of the target structure. Figure 1 shows that E_{CS} is split into three regions: a flat-bottomed region that takes into account inaccuracies in the chemical shift predictions, a harmonic region that penalizes statistically significant deviations between the computed and experimental shifts, and a linear region that prevents large deviations of individual chemical shifts from dominating the magnitude of E_{CS} and thus frustrating the conformational search.[6] The width of the potential well E_{CS} is governed by the parameter n.

The CamShift energy E_{CS} and force $\mathbf{F}_{\text{CS}}$ are added to the CHARMM22 energy E_{FF} and force $\mathbf{F}_{\text{FF}}$, respectively,

$$E = E_{\text{FF}} + \alpha E_{\text{CS}}, \tag{2}$$

$$\mathbf{F} = \mathbf{F}_{\text{FF}} + \alpha \mathbf{F}_{\text{CS}}. \tag{3}$$

Here, the adjustable parameter α defines the contribution of the chemical shift restraints to the total energy E. If the value of α is too high, the forces resulting from CamShift are too large, creating instabilities during the energy minimization process. If the value of α is too low or the tolerance parameter n is too large, the influence of CamShift is too weak to provide an improvement over unrestrained simulations. If the

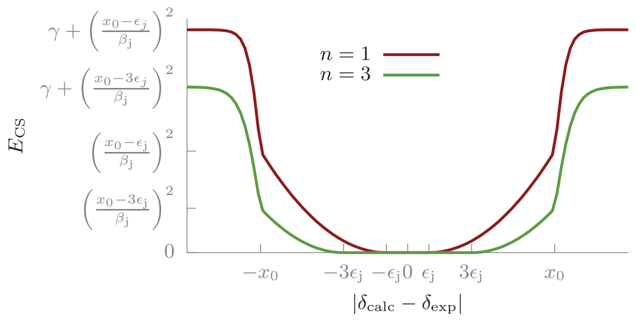


FIG. 1. Chemical shift penalty energy E_{CS} as a function of the difference between the chemical shifts of the simulated (δ_{calc}) and the target structure (δ_{exp}) for $n = 1$ (red) and $n = 3$ (green). The width of the flat-bottomed part is $2n\epsilon_j$, which is adjustable by modifying n with ϵ_j being the accuracy of the chemical shift predictions for atom type j. x_0 is the cutoff of the harmonic part of the energy function, β_j is a scaling parameter determining the magnitude of the energy penalty, and γ influences how large the energy penalty can grow beyond x_0.

value of n is too small, small deviations from the target structure will generate chemical shifts that result in large penalties, thus creating a rugged energy landscape. It will therefore be more difficult to locate the global minimum as the system can easily become trapped in deep local minima.[6] In the first part of our study we identified the optimal values of α and n for the combination of GMIN and CamShift as described in Sec. III A. It should be noted that the unit of E_{FF} is kcal/mol in the CHARMM force field,[59] while E_{CS} is a dimensionless quantity.[6]

III. RESULTS AND DISCUSSION

A. Optimization of the parameters α and n

We prepared the systems as described in Sec. II A. First, we performed two types of short chemical shift-restrained simulations with 100 BH steps for 1LE0 and 1L2Y, one using value pairs with $\alpha = 1$ and varying n from $n = 0.5$ to $n = 4$, and the second with constant $n = 1$ but varying α from $\alpha = 0$ to $\alpha = 3$. For the latter we chose $n = 1$ because from the runs with varying n only the one with $n = 1$ could find parts of the β-hairpin for 1LE0, as the structural results in Figure S2 of the supplementary material[58] show. In the runs with varying α, the unrestrained simulation ($\alpha = 0$) was not able to produce a structure resembling the β-hairpin, while the other values for α were more successful. Figure S2 of the supplementary material[58] shows that the simulations with $(\alpha, n) = (1, 1)$ and $(\alpha, n) = (2, 1)$ find structures fitting best to the β-hairpin within the short 100 step-BH runs. To test if this choice of values for α and n is universal or peptide specific, we performed 100 step-BH simulations of 1L2Y using the various (α, n) pairs. Figure S3 of the supplementary material[58] shows that only the simulations with $(\alpha, n) = (1, 0.5)$ and $(1, 1)$ could find parts of the α-helix. In a previous chemical shift-restrained MC study using a SA protocol, Robustelli *et al.* also chose $n = 1$ yet in connection with higher values for α.[6] The ideal value of α depends on the absolute value of the force field energy E_{FF}: the larger this value, the larger α has to be chosen.

During our systematic test of the interplay between α and n, we further observed that n had to be optimized for the LBFGS minimizer, while keeping $\alpha = 1$ constant throughout each BH simulation. For n we found that the local minimization at a given BH step is more successful in terms of robustness and speed if n is decreased while the minimization progresses. We use the root mean square force (RMSF) during the minimization as progression variable to determine n:

$$n = \begin{cases} 3 & \text{RMSF} > 1, \\ 3 + \frac{2}{3}\log(\text{RMSF}) & 10^{-3} < \text{RMSF} \leq 1, \\ 1 & \text{RMSF} \leq 10^{-3}, \end{cases} \tag{4}$$

where the unit of RMSF is computed for the change of the total energy E. We start with the relatively large value $n = 3$ to make sure that the first few minimization steps after changing the dihedral angles are mainly force-field driven. Figure 1 shows that for large values of n, the calculated chemical shifts of a wide range of conformations fall near the flat-bottomed region of E_{CS} and thus generate relatively small energetic penalties. Once the minimization has sufficiently progressed, the conformation is increasingly forced towards the target structure by decreasing n, i.e., by increasing the penalties for the calculated shifts of atoms j, which deviate by more than $n\epsilon_j$ from the experimental chemical shifts. We reduce the value of n continuously to the previously determined $n = 1$.

B. Results for 1LE0, 1L2Y, and 1YRF

We performed chemical shift restrained and unrestrained BH simulations using 1000 and 10 000 BH steps for the peptides 1LE0 and 1L2Y, and 1000 and 5000 BH steps for 1YRF. We did not continue the BH run for 1YRF up to 10 000 BH steps because we did not observe a significant improvement during the last 3000 BH steps, and the result after 2000 BH steps is already very convincing. The best structures, which we obtained for the three peptides within 1000 steps and full-length BH simulations, are shown in Figure 2. Here, the definition of the best structures is with respect to the total energy $E = E_{FF} + E_{CS}$, which is lowest for these structures. This allows us to test, if by using the total energy as criterion, structures with low E correspond to structures with low backbone root mean square displacement (RMSD) from the target structure. This can only be the case when the force field correctly predicts the target structure as the global minimum. Thus, we included a β-sheet structure and helical structures in our test set in order to check if both structural elements are correctly supported by the CHARMM22 potential in connection with the FACTS solvation model.

1. *1LE0*

The structures for 1LE0 in Figure 2 show that the β-hairpin can be determined with very high accuracy. Within 1000 BH steps the β-sheet is already correctly identified, while the turn region still needs improvement. After 10 000 steps this deficiency was resolved, so that the RMSD of the best structure is only 0.86 Å from the target structure. The

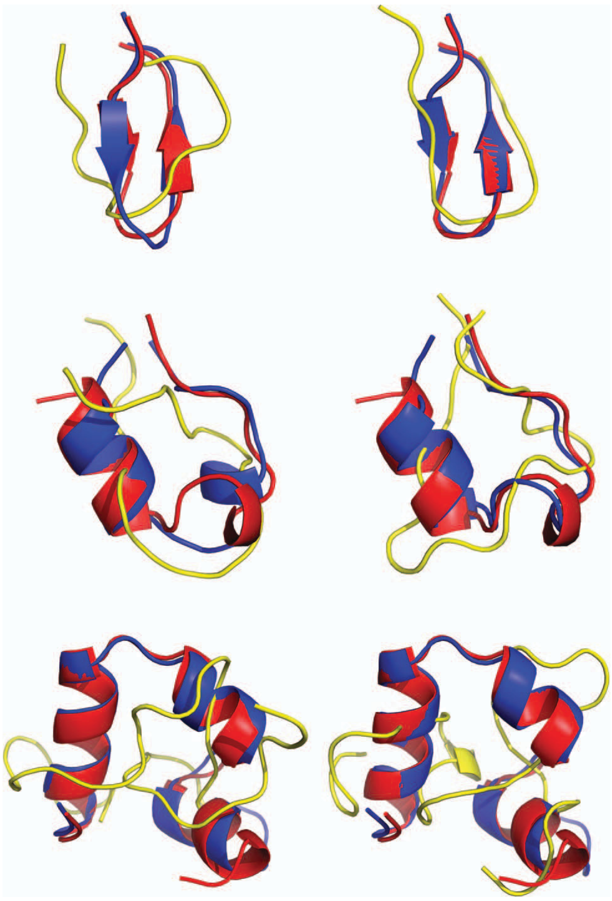


FIG. 2. Target structures (red), structures of unrestrained (yellow), and chemical shift restrained (blue) simulations after 1000 (left) and 10 000 (right) BH steps for 1LE0 (top) and 1L2Y (center), and after 1000 (left) and 5000 (right) BH steps for 1YRF (bottom).

unrestrained BH run was not able to produce the β-sheet within 10 000 steps.

2. 1L2Y

The correct structure of 1L2Y was also identified within 10 000 BH steps using chemical shift restraints, while the unrestrained BH simulation did not even find the α-helix for the first nine residues. Imposing chemical shift restraints, the α-helix was found quickly (within 1000 BH steps) and accurately. The biggest deviations from the target structure are seen for the termini and for the transition between the α-helix and the 3_{10} helix (residues 10 and 11). The middle part of the peptide needed longest before its correct structure was located. The RMSD for the best structure after 10 000 BH steps is 2.18 Å. In order to pinpoint the origin of the deviation between the predicted and the target structure, we plotted the deviation between computed and target chemical shifts, $\delta_{\mathrm{calc}} - \delta_{\mathrm{exp}}$, for each C$\alpha$ atom of 1L2Y (Figure 3). This analysis reveals that for residues 3–9 and 12–17 the predicted shifts are restrained to their target shifts since $|\delta_{\mathrm{calc}} - \delta_{\mathrm{exp}}|/\epsilon < 1$, which for $n = 1$ corresponds to the flat-bottomed region of the chemical shift penalty function (Figure 1) leading to $E_{\mathrm{CS}} = 0$ for these atoms.

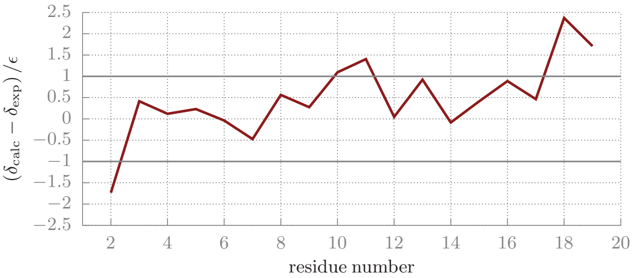


FIG. 3. Deviation between the Cα chemical shifts of the predicted and the target structures for 1L2Y. $\delta_{\mathrm{calc}} - \delta_{\mathrm{exp}}$ is shown for each residue apart from residues 1 and 20, because CamShift does not provide chemical shifts for the first and last residue. The chemical shift deviation is given in units of the CamShift accuracy $\epsilon_{\mathrm{C}\alpha}$ for the prediction of Cα chemical shifts.

Figure 3 shows that residues 2, 18, and 19 produce the largest deviations from the target structure. This is because CamShift does not calculate the chemical shifts for the first and last amino acids in the chain, because the CamShift prediction for a given atom relies on the distances to atoms in the two neighboring residues. Therefore, the structures for the first and last residues have to be predicted without chemical shift restraints. In general, this means that the largest structural deviations come from the terminal residues, as the predicted structure for 1L2Y in Figure 2 supports. The wrong structures for the first and last residues give rise to wrong interatomic distances, which are needed for the chemical shift calculations for residues 2 and $N_{\mathrm{res}} - 1$, with N_{res} being the total number of residues in the chain. In turn, this leads to inaccurate chemical shifts δ_{calc} for residues 2 and $N_{\mathrm{res}} - 1$, hampering the structure prediction for these residues. This effect propagates to residues 3 and $N_{\mathrm{res}} - 2$ before eventually leveling off. For small peptides such as 1L2Y the deviation of only a few residues leads to an appreciably increased RMSD from the target structure. This effect will decrease for larger peptides.

3. 1YRF

The 5000 step-BH run with chemical shift restraints produced a structure for 1YRF with a RMSD of 3.81 Å. The best structure, which was found within 1000 BH steps looked already very good and could only slightly improved during the subsequent 4000 BH steps. From the structures in Figure 2 it is visible that, as discussed above for 1L2Y, the largest deviations originate from the terminal residues. If we exclude residues 1 and 36 from the RMSD calculation we obtain a RMSD of 2.44 Å, which further decreases to 1.88 Å by excluding residues 1, 2, 35, and 36, and to 1.39 Å for the RMSD between residues 4 and 33. Excluding more residues does not further improve the RMSD. As for the other two peptides, the unrestrained BH run did not produce a structure resembling the target structure. None of the helices were found during this run.

In order to better understand the interplay between the force field energy and the chemical shift penalty, and their influence on folding the helical peptide 1YRF, we plotted the total energy $E = E_{\mathrm{FF}} + E_{\mathrm{CS}}$, the CamShift penalty energy

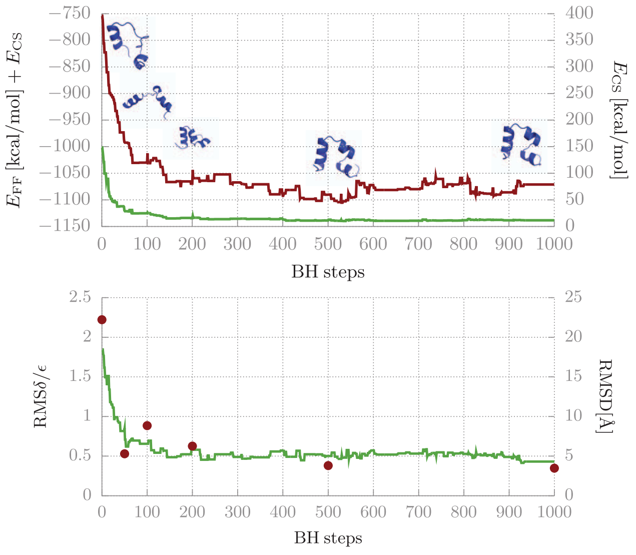


FIG. 4. Folding of 1YRF during the first 1000 BH steps. (Top) Energy for each state of the Markov chain. The red line represents the total energy (E_{FF} + E_{CS}) and the green line represents the CamShift energy (E_{CS}). Structures after 50, 100, 200, 500, and 1000 BH steps are shown in blue. (Bottom) RMSδ is shown for Cα atoms for each state of the Markov chain (green line). It was normalized with regard to the atom type specific CamShift accuracy $\epsilon_{C\alpha}$. For the structures after 0, 50, 100, 200, 500, and 1000 BH steps the RMSD is provided as well (red dots).

E_{CS}, the structural RMSD, and the root mean square chemical shift deviation RMSδ (for Cα atoms) from the target chemical shifts during the first 1000 BH steps (Figure 4). All four quantities reach a plateau in less than 200 BH steps, which are sufficient for the structure to find the secondary structure elements, i.e., the three α-helices (see blue structure after 200 steps). Identifying the α-helices is accompanied by a marked decrease of E_{CS}. During the following 800 BH steps the structure improved by finding the correct arrangement of the α-helices with respect to each other and local refinements. These improvements are mainly force field driven as E_{FF} decreases more strongly than E_{CS} for the near-target structures. The penalty energy plateaus at $E_{CS} \approx 11.5$, implying that within 1000 BH steps not all predicted chemical shifts fall into the flat-bottomed region of the chemical shift penalty function (Figure 1). As discussed above, the largest deviations originate from the amino acids in neighborhood to the terminal residues. The improvement of the structure is confirmed by the RMSD. It decreases from ≈7 Å at BH step 200 to ≈4 Å at BH step 1000, which is already close to the final RMSD of 3.81 Å for the best structure after 5000 BH steps.

4. Incomplete chemical shift assignments

It is often not possible to measure and assign all chemical shifts in a NMR experiment. To test the robustness of our approach with respect to incomplete chemical shift assignments, we performed BH simulations of 1YRF where only one of the six chemical shift types, ^{1}Hα, amide ^{1}H, ^{13}Cα, ^{13}Cβ, carbonyl ^{13}C, or amide ^{15}N chemical shifts were used in the restraining function. The number of chemical shift restraints applied is given by $N_{\text{shift}} \times (N_{\text{res}} - 2)$, with N_{shift} being the number of chemical shift types considered (for the first and last residue no chemical shifts are calculated). In the simulations above we set $N_{\text{shift}} = 6$, while in the simulations with only one chemical shift type $N_{\text{shift}} = 1$. Additionally, we performed one simulation with $N_{\text{shift}} = 3$, where restraints for ^{1}H, ^{13}Cα, and ^{15}N chemical shifts were included. We chose these three shift types as these are the most frequently measured chemical shifts for proteins as the statistics derived from a total of about 5.6 × 10^{6} chemical shifts in the Biological Magnetic Resonance Data Bank (http://www.bmrb.wisc.edu/) reveals (see Figure S4 of the supplementary material[58]). Figure 5 shows the structures obtained after 1000 BH steps with reduced chemical shift restraints. Apart from the simulation with only ^{13}Cβ chemical shift restraints, the other simulations with $N_{\text{shift}} = 3$ and $N_{\text{shift}} = 1$ are able to fold parts of the peptide into α-helices. The predicted structures from these simulations are much closer to the target structure than the structure from the unrestrained 1000 step BH simulation (Figure 2).

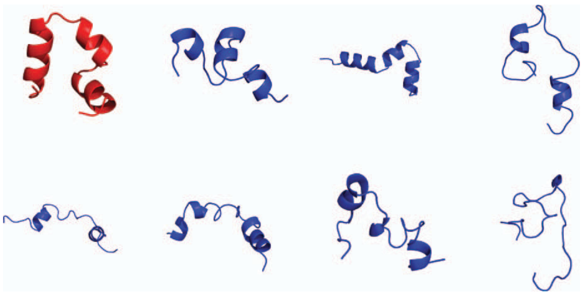


FIG. 5. Structures for 1YRF after 1000 BH steps with reduced chemical shift restraints. Top (from left to right): Target structure (red) and predicted structures (blue) with chemical shift assignments for (^{1}H, ^{13}Cα, ^{15}N), ^{13}C, and ^{15}N. Bottom: Predicted structures with chemical shift assignments for ^{1}Hα, ^{1}H, ^{13}Cα, and ^{13}Cβ. The results are sorted according to decreasing prediction quality.

For a more detailed analysis of the performance of the BH simulations with reduced chemical shift restraints, we determined the secondary structure of each residue in the structures given in Figure 5 using STRIDE[67] (Figure 6). The simulation with only carbonyl ^{13}C chemical shift restraints succeeded to predict all three α-helices at almost identical positions to the target structure. This can be explained by

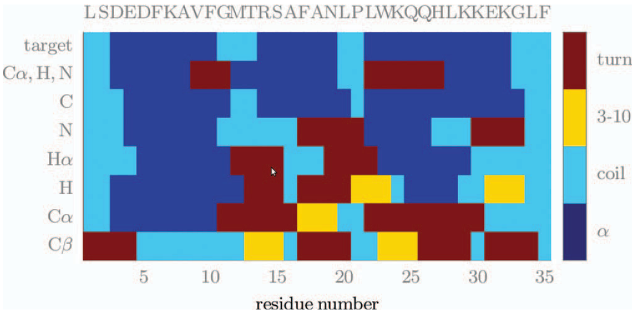


FIG. 6. Secondary structure per residue in the structures shown in Figure 5. On the top line the one letter code of each residue is given, on the bottom line the residue number. The left column designates the chemical shift restraints applied in the simulations.

considering that the backbone torsional angles Φ and Ψ are strong determinants of the ^{13}C chemical shifts. Their influence on this chemical shift is about 50%, while on ^{13}Cα and ^{13}Cβ chemical shifts their effect is only 25% and 10%, respectively.[1] Therefore, ^{13}Cα and ^{13}Cβ chemical shifts are less effective as restraints for secondary structure prediction. With ^{13}Cα chemical shift restraints one of the three α-helices can still be predicted within 1000 BH steps, while ^{13}Cβ chemical shift restraints fail to fold any of the helices. ^{13}Cβ chemical shifts are shifted downfield by about 2.5 ppm in β-sheets, but have nearly random coil values in helices.[1] Thus, it is not surprising that the information from merely ^{13}Cβ chemical shifts is not sufficient for the identification of the helices of 1YRF. Individual ^{15}N, ^{1}Hα, and ^{1}H chemical shift restraints are successful in the prediction of two of the three helices within 1000 BH steps. While both ^{15}N and ^{1}H chemical shifts are not very good predictors of dihedral angles or indicators of secondary structure, they are very sensitive to hydrogen bonding[1] and are therefore helpful as restraints in protein folding simulations. ^{1}Hα chemical shifts are known as a reliable indicator of secondary structure, and backbone dihedral torsional effects are the most important contribution to ^{1}Hα chemical shift deviations. This explains the good performance of the BH run with only ^{1}Hα chemical shift restraints. The combined application of ^{1}H, ^{13}Cα, and ^{15}N chemical shift restraints leads to the prediction of all three helices in less than 1000 BH steps. Here, even the length of the coil sequences between the second and third helix (residues 21 and 22), and at the N- and C-termini (residues 1–2 and 33–35, respectively) are correctly predicted.

IV. CONCLUSION

Computational methods that utilize chemical shifts to produce protein structures at atomic resolution have recently been introduced. These methods use the information contained in experimental chemical shifts together with structural homology of proteins in structural databases such as the RCSB protein data bank to generate new structures,[4,5,8,9,16,20] or directly incorporate chemical shifts as restraints in molecular simulations with an energetic penalty function analogous to those used in standard NMR structure calculations.[6,7] In the current work, we applied the latter idea and combined the basin-hopping (BH) approach to global optimization[23,24] with chemical shift restraints by using the chemical shift penalty function introduced by Vendruscolo and co-workers.[6,7] For the calculation of NMR chemical shifts from protein structures we used the CamShift method, which approximates chemical shifts as polynomial functions of interatomic distances.[22]

For the proper implementation of chemical shift restraints into the BH approach we determined the optimal weight of the chemical shift penalty energy with respect to the CHARMM22 force field[59,60] employed in conjunction with the solvation model FACTS.[61] Furthermore, we developed a function, which continuously decreases the width of the chemical shift penalty function during each local energy minimization procedure, which thereby becomes more robust. We demonstrated for three peptides that the BH approach with chemical shift restraints is able to find near-native structures from fully extended structures within 10 000 BH steps. The conformational searches were able to fold α and β secondary structure elements in less than 1000 BH steps, and correctly orient their tertiary contacts in subsequent BH steps. The unrestrained BH runs, on the other hand, failed to fold any of the secondary structure elements within 10 000 BH steps. Much longer unrestrained BH runs would be needed for the conformational searches to succeed without guidance from chemical shift restraints. In another study we tested whether or not the CHARMM22/FACTS potential supports the target structures of 1LE0, 1L2Y, and 1YRF as global minima. We found that the RMSD values of the global minima from the respective targets are between 1.5 and 3 Å. Our conclusion therefore is that it is rather inefficient sampling and not the CHARMM22/FACTS potential that precluded the generation of near-native structures in the unrestrained BH simulations of the current study.

We tested our approach for incomplete chemical shift assignments, where the information from only one chemical shift type was included in each of the chemical shift-restraint BH simulations. Apart from the simulation with ^{13}Cβ chemical shift restraints, these simulations succeeded to predict secondary structure elements within 1000 BH steps. For each of the chemical shift types, the success (and failure) can be explained based on the relation between structure and chemical shifts in proteins.[1] The usage of fewer chemical shifts speeds up the restrained BH simulations as the computational overhead compared to unrestrained BH simulations scales linearly with N_{shift}. However, in order to obtain as good prediction results as from the runs with more chemical shift restraints, more BH steps have to be conducted. The full-length BH simulations with $N_{\text{shift}} = 6$, for which the results are shown in Figure 2, required 10 CPU days for 1LE0, 12 CPU days for 1L2Y, and 16 CPU days for 1YRF. All BH simulations were run on a single 2.93 GHz Intel Xeon Processor X5570. For the folding of proteins of comparable length using chemical shift restrained Monte Carlo simulations with a simulated annealing protocol Robustelli *et al.*[6] needed between 380 and 473 CPU days. This comparison reveals that it is more effective to apply chemical shift restraints via both energy and energy gradients, as it is realized in BH and molecular dynamics,[7] than considering only the energy as in simulated annealing based on Monte Carlo simulations.[6] Like Robustelli *et al.*,[6] we found that the major bottleneck of the chemical shift restrained simulations is the computation of chemical shifts with each call to the energy function. The unrestrained BH simulations of the same length required less than a CPU day for 1LE0 and 1L2Y, and 2.5 CPU days for 1YRF. We currently work on a relief of this computational cost.

We conclude that the BH approach with chemical shift restraints is a promising method for protein structure prediction. The approach is an addition to existing methods based on chemical shift restrained Monte Carlo simulations using a simulated annealing protocol,[6] molecular dynamics simulations with chemical shift restraints,[7] and various molecular fragment replacement approaches with chemical shift information.[4,5,8,9,16,20] The three proteins that we considered as test cases contain fewer than 50 amino acids, and have relatively simple topologies. It is expected that the amount of computational time required to achieve convergence will significantly increase for larger proteins with more complex topologies, which will probably limit the application of the current implementation of the BH approach with chemical shift restraints to proteins not much larger than 50 to 60 residues. Therefore, we are currently implementing knowledge-based Monte Carlo moves into the GMIN program, which should speed up the folding of secondary structure elements for BH runs with and without chemical shifts restraints. Additionally, the BH approach could easily be combined with restraints traditionally used in NMR structure calculations such as NOEs, *J*-couplings, and RDCs, which, in connection with chemical shift restraints, will open the possibility for the BH approach to become a valuable tool in structural biology.

ACKNOWLEDGMENTS

The authors thank Dr. Andrea Cavalli for helpful discussions and Professor David Wales for proofreading the manuscript.

[1] D. S. Wishart and D. A. Case, Method Enzymol. **338**, 3 (2002).
[2] C. D. Schwieters, J. J. Kuszewski, N. Tjandra, and G. M. Clore, J. Magn. Reson. **160**, 65 (2003).
[3] A. T. Brunger, Nat. Protoc. **2**, 2728 (2007).
[4] A. Cavalli, X. Salvatella, C. M. Dobson, and M. Vendruscolo, Proc. Natl. Acad. Sci. U.S.A. **104**, 9615 (2007).
[5] P. Robustelli, A. Cavalli, and M. Vendruscolo, Structure (London) **16**, 1764 (2008).
[6] P. Robustelli, A. Cavalli, C. M. Dobson, M. Vendruscolo, and X. Salvatella, J. Phys. Chem. B **113**, 7890 (2009).
[7] P. Robustelli, K. Kohlhoff, A. Cavalli, and M. Vendruscolo, Structure (London) **18**, 923 (2010).
[8] Y. Shen *et al.*, Proc. Natl. Acad. Sci. U.S.A. **105**, 4685 (2008).
[9] D. S. Wishart *et al.*, Nucleic Acids Res. **36**, W496 (2008).
[10] M. Berjanskii *et al.*, Nucleic Acids Res. **37**, W670 (2009).
[11] R. Das *et al.*, Proc. Natl. Acad. Sci. U.S.A. **106**, 18978 (2009).
[12] F. C. Bernstein *et al.*, J. Mol. Biol. **112**, 535 (1977).
[13] J. Kuszewski, A. M. Gronenborn, and G. M. Clore, J. Magn. Reson. Ser. B **107**, 293 (1995).
[14] J. G. Pearson, J.-F. Wang, J. L. Markley, H.-B. Le, and E. Oldfield, J. Am. Chem. Soc. **117**, 8823 (1995).
[15] P. Luginbühl, T. Szyperski, and K. Wüthrich, J. Magn. Reson. Ser. B **109**, 229 (1995).
[16] H. Gong, Y. Shen, and G. D. Rose, Protein Sci. **16**, 1515 (2007).
[17] R. Montalvao, A. Cavalli, and X. Salvatella, J. Am. Chem. Soc. **130**, 15990 (2008).
[18] Y. Shen, R. Vernon, D. Baker, and A. Bax, J. Biomol. NMR **43**, 63 (2009).
[19] P. Neudecker *et al.*, Science **336**, 362 (2012).
[20] F. Delaglio, G. Kontaxis, and A. Bax, J. Am. Chem. Soc. **122**, 2142 (2000).
[21] G. M. Clore and A. M. Gronenborn, Proc. Natl. Acad. Sci. U.S.A. **95**, 5891 (1998).
[22] K. J. Kohlhoff, P. Robustelli, A. Cavalli, X. Salvatella, and M. Vendruscolo, J. Am. Chem. Soc. **131**, 13894 (2009).
[23] D. J. Wales and J. P. K. Doye, J. Phys. Chem. A **101**, 5111 (1997).
[24] D. J. Wales and H. A. Scheraga, Science **285**, 1368 (1999).
[25] Z. Li and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **84**, 6611 (1987).
[26] P. Derreumaux, J. Chem. Phys. **106**, 5260 (1997).
[27] P. Derreumaux, J. Chem. Phys. **107**, 1941 (1997).
[28] P. N. Mortenson and D. J. Wales, J. Chem. Phys. **114**, 6443 (2001).
[29] P. N. Mortenson, D. A. Evans, and D. J. Wales, J. Chem. Phys. **117**, 1363 (2002).
[30] J. M. Carr and D. J. Wales, J. Chem. Phys. **123**, 234901 (2005).
[31] A. Verma, A. Schug, K. H. Lee, and W. Wenzel, J. Chem. Phys. **124**, 044515 (2006).
[32] B. Strodel and D. J. Wales, J. Chem. Theor. Comput. **4**, 657 (2008).
[33] B. Strodel, J. Lee, C. Whittleston, and D. Wales, J. Am. Chem. Soc. **132**, 13300 (2010).
[34] O. O. Olubiyi and B. Strodel, J. Phys. Chem. B **116**, 3280 (2012).
[35] A. G. Cochran, N. J. Skelton, and M. A. Starovasnik, Proc. Natl. Acad. Sci. U.S.A. **98**, 5578 (2001).
[36] J. Neidigh, R. Fesinmeyer, and N. Andersen, Nat. Struct. Biol. **9**, 425 (2002).
[37] T. K. Chiu *et al.*, Proc. Natl. Acad. Sci. U.S.A. **102**, 7517 (2005).
[38] C. Simmerling, B. Strockbine, and A. E. Roitberg, J. Am. Chem. Soc. **124**, 11258 (2002).
[39] S. Chowdhury, M. C. Lee, G. Xiong, and Y. Duan, J. Mol. Biol. **327**, 711 (2003).
[40] A. Schug, T. Herges, and W. Wenzel, Phys. Rev. Lett. **91**, 158102 (2003).
[41] A. Schug, T. Herges, A. Verma, K. H. Lee, and W. Wenzel, ChemPhysChem **6**, 2640 (2005).
[42] A. Schug, W. Wenzel, and U. Hansmann, J. Chem. Phys. **122**, 194711 (2005).
[43] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, Science **334**, 517 (2011).
[44] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, Biophys. J. **100**, L47 (2011).
[45] J. Maupetit, P. Derreumaux, and P. Tufféry, Nucleic Acids Res. **37**, W498 (2009).
[46] P. Thévenet *et al.*, Nucleic Acids Res. **40**, W288 (2012).
[47] J. W. Pitera and W. Swope, Proc. Natl. Acad. Sci. U.S.A. **100**, 7587 (2003).
[48] J. Juraszek and P. G. Bolhuis, Proc. Natl. Acad. Sci. U.S.A. **103**, 15859 (2006).
[49] D. Paschek, H. Nymeyer, and A. E. García, J. Struct. Biol. **157**, 524 (2007).
[50] K. Klenin and W. Wenzel, Int. J. Comput. Commun. **1**, 1 (2007).
[51] I. H. Radford, A. R. Fersht, and G. Settanni, J. Phys. Chem. B **115**, 7459 (2011).
[52] J. Kubelka, T. K. Chiu, D. R. Davies, W. A. Eaton, and J. Hofrichter, J. Mol. Biol. **359**, 546 (2006).
[53] T. Cellmer, M. Buscaglia, E. R. Henry, J. Hofrichter, and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. **108**, 6103 (2011).
[54] Y. Tang, M. J. Grey, J. McKnight, A. G. Palmer III, and D. P. Raleigh, J. Mol. Biol. **355**, 1066 (2006).
[55] P. L. Freddolino and K. Schulten, Biophys. J. **97**, 2338 (2009).
[56] D. R. Ripoll, J. A. Vila, and H. A. Scheraga, J. Mol. Biol. **339**, 915 (2004).
[57] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[58] See supplementary material at http://dx.doi.org/10.1063/1.4773406 for a graphical presentation of the starting structures, the performance of chemical shift restrained BH runs for various (α, n) pairs, and a graphical presentation showing the statistics of how frequently the different chemical shifts are measured in proteins as derived from a total of about 5.6×10^6 chemical shifts in the Biological Magnetic Resonance Data Bank.
[59] B. R. Brooks *et al.*, J. Comput. Chem. **4**, 187 (1983).
[60] A. D. MacKerell, Jr. *et al.*, J. Phys. Chem. B **102**, 3586 (1998).
[61] U. Haberthür and A. Caflisch, J. Comput. Chem. **29**, 701 (2008).
[62] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[63] D. J. Wales, *GMIN: A Program for Basin-Hopping Global Optimisation*, see http://www-wales.ch.cam.ac.uk/software.html.
[64] M. A. Miller and D. J. Wales, J. Chem. Phys. **111**, 6610 (1999).
[65] M. Bauer, B. Strodel, S. Fejer, E. Koslover, and D. Wales, J. Chem. Phys. **132**, 054101 (2010).
[66] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, SIAM J. Sci. Stat. Comput. **16**, 1190 (1995).
[67] D. Frishman and P. Argos, Proteins: Struct., Funct., Bioinf. **23**, 566 (1995).

Supplementary Information

Protein structure prediction using global optimization by Basin-Hopping with NMR shift restraints

Falk Hoffmann[1] and Birgit Strodel[1,2]*[*]

[1]Institute of Complex Systems: Structural Biochemistry,
Research Centre Jülich, 52425 Jülich, Germany

[2]Institute of Theoretical and Computational Chemistry,
Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

[*] corresponding author: b.strodel@fz-juelich.de

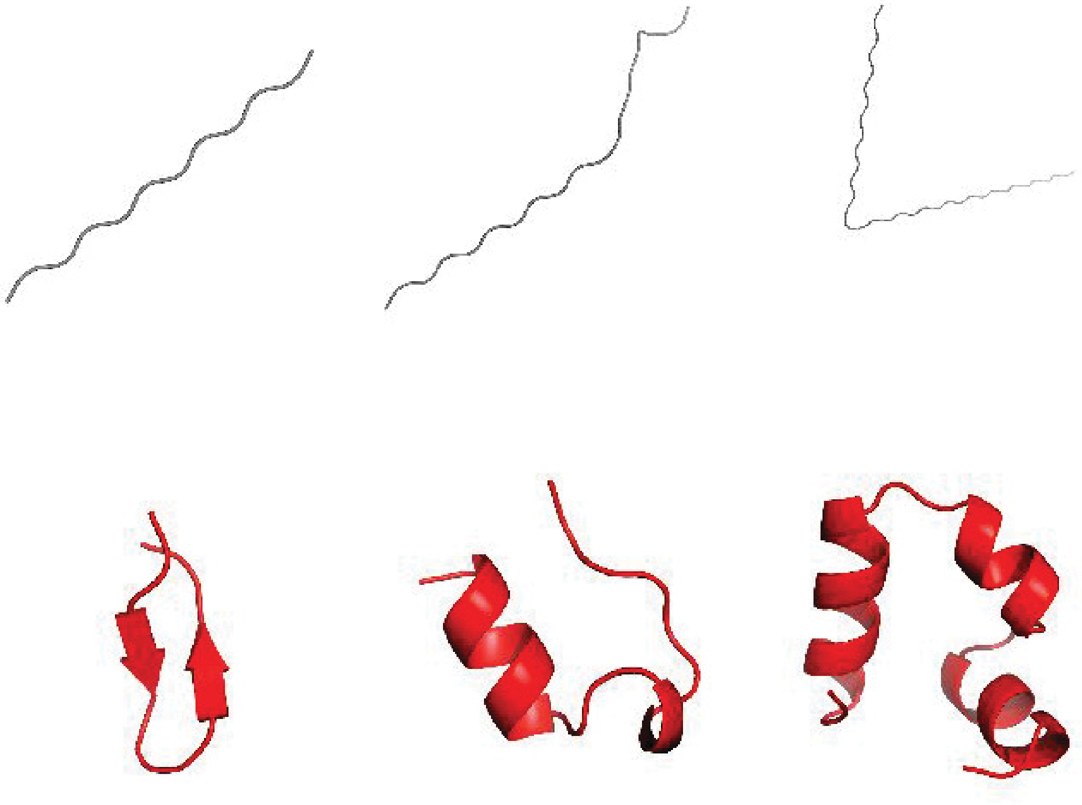


Figure S1: Starting structures (grey) and target structures (red) of 1LE0 (left), 1L2Y (center) and1YRF (right).

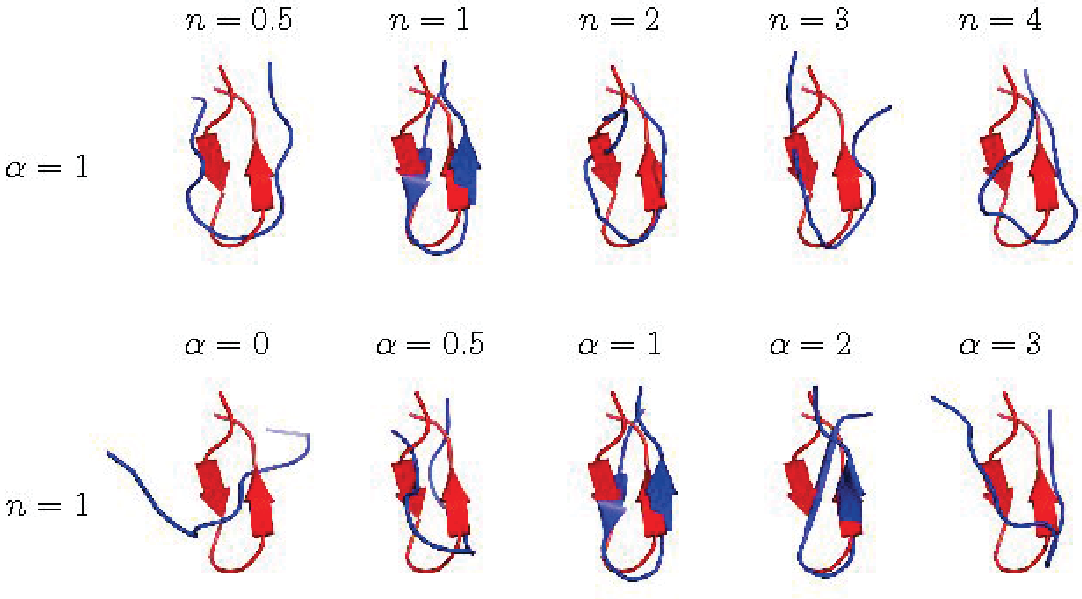


Figure S2: Target structures (red) and simulated structures (blue) for 0.5 ≤ n ≤ 4.0 for α=1 (top) and 0 ≤ α ≤ 3 for n=1 (bottom) for 1LE0 after 100 BH steps.

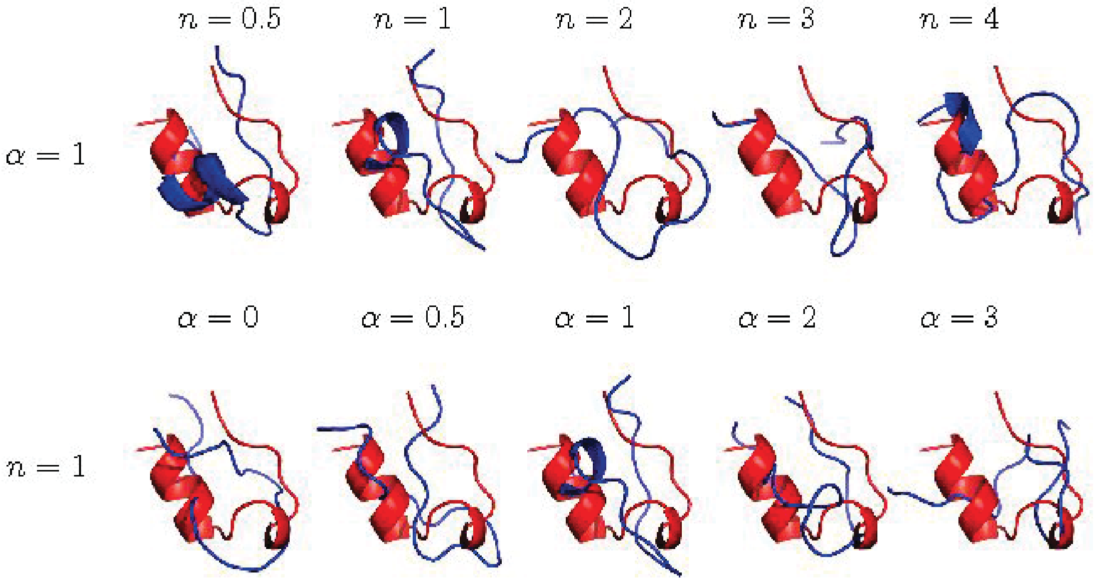


Figure S3: Target structures (red) and simulated structures (other colors) for 0.5 ≤ n ≤ 4.0 for α=1 (top) and 0 ≤ α ≤ 3 for n=1 (bottom) for 1L2Y after 100 BH steps.

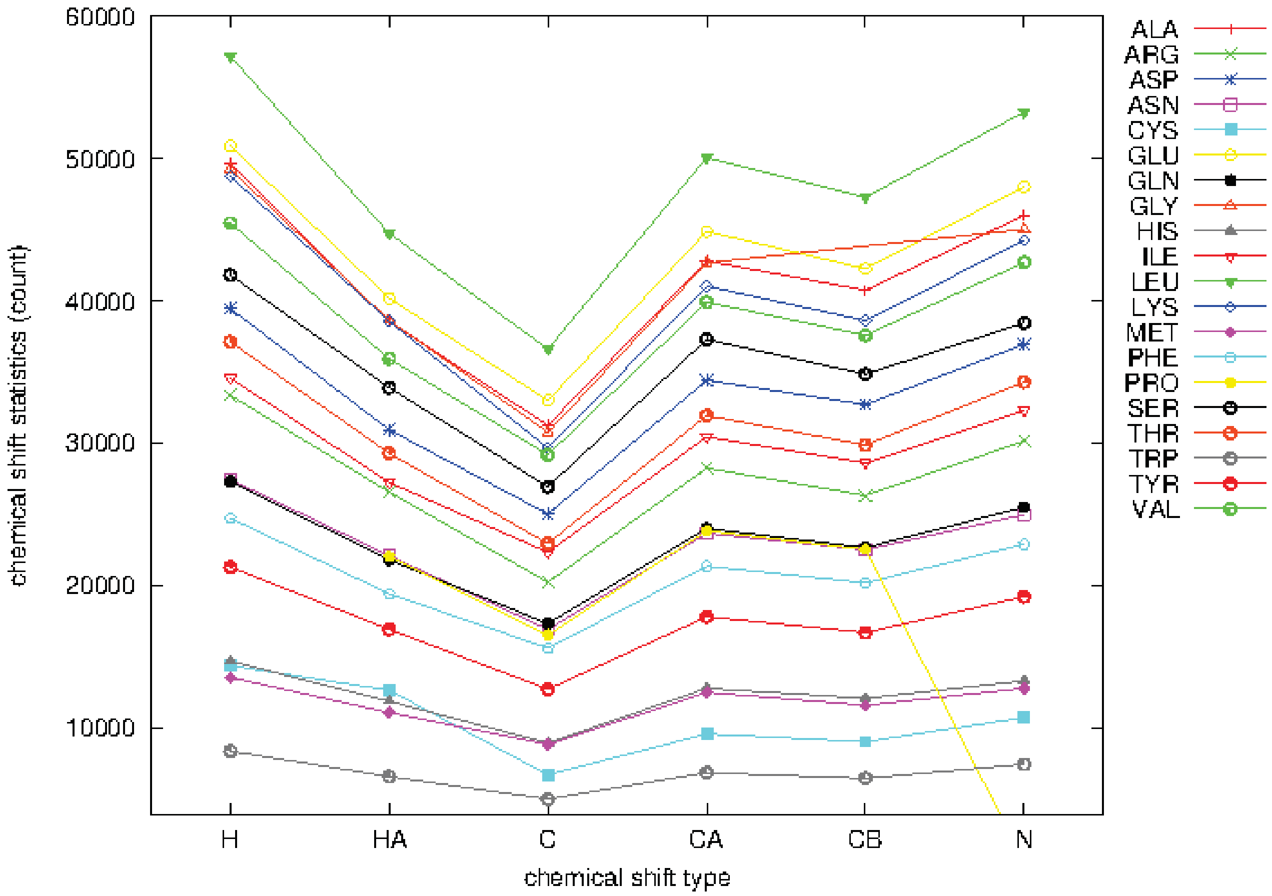


Figure S4: The statistics presented in this figure were calculated from the full BMRB database (http://www.bmrb.wisc.edu/) as of 29 November, 2012. The calculated statistics are derived from a total of 5,648,668 chemical shifts.

4.2 Protein structure prediction: assembly of secondary structure elements by basin-hopping

DOI: 10.1002/cphc.201402247

Protein Structure Prediction: Assembly of Secondary Structure Elements by Basin-Hopping

Falk Hoffmann,[b] Ioan Vancea,[c] Sanjay G. Kamat,[b] and Birgit Strodel*[a, b]

The prediction of protein tertiary structure from primary structure remains a challenging task. One possible approach to this problem is the application of basin-hopping global optimization combined with an all-atom force field. In this work, the efficiency of basin-hopping is improved by introducing an approach that derives tertiary structures from the secondary structure assignments of individual residues. This approach is termed secondary-to-tertiary basin-hopping and benchmarked for three miniproteins: trpzip, trp-cage and ER-10. For each of the three miniproteins, the secondary-to-tertiary basin-hopping approach successfully and reliably predicts their three-dimensional structure. When it is applied to larger proteins, correctly folded structures are obtained. It can be concluded that the assembly of secondary structure elements using basin-hopping is a promising tool for de novo protein structure prediction.

1. Introduction

The prediction of protein structure from an amino acid sequence is one of the most important computational problems in bioinformatics and one of the great challenges in structural biology. Knowledge of the three-dimensional structure of proteins gives invaluable insights into the molecular basis of their function, and might therefore facilitate finding treatments and cures for many diseases. It is generally assumed that a protein folds into a native conformation, or ensemble of conformations, that is at or near the global free-energy minimum.[1] Thus, protein structure prediction can be understood as the search for an energy minimum in the conformational space of the protein. From a computational point of view, the problem of finding native-like conformations for a given primary structure—referred to as de novo protein structure prediction—can be decomposed into two tasks: 1) developing an accurate energy function for which the native protein folding and energy minimum coincide, and 2) developing an efficient protocol for searching the energy landscape.

The focus of the current study is on the latter task. The extensive exploration of the whole conformational space of a protein is generally not possible, as it would be a time-prohibitive endeavor. Approaches based on the Metropolis Monte Carlo (MC) method offer the possibility to efficiently explore the conformational space or at least specific regions of it. Searching for the conformational space using MC methods is usually a two-step process—a trial conformation move followed by an energy evaluation. In this work, we use the basin-hopping (BH) global optimization algorithm,[2,3] which is analogous in principle to the MC-minimization approach.[4] Global optimization can be defined as the procedure of finding the lowest value of a given function. The BH algorithm is a stochastic global optimization method, which uses MC moves on a transformed potential energy surface, where a structural perturbation is followed by energy minimization. BH has been used successfully to find the global minimum of peptides and proteins,[5–12] including peptide complexes.[13–15]

Several possibilities to improve the efficiency of MC sampling exist, including the optimization of trial moves for proteins[16] and applying experimental restraints during an MC simulation.[17,18] The aim of this study is the improvement of the trial moves. For proteins, a typical MC move consists of moving often contiguous residues randomly in a single MC step. The efficiency of the trial moves can be increased by incorporating residue-specific structural preferences derived from experimental structures.[19,20] It is well known that the Φ and Ψ angles of the protein backbone are more densely centered around particular regions, with the distribution of the (Φ,Ψ) densities depending on the identity of the amino acid.[21] Likewise, protein side chains tend to exist in a limited number of low-energy conformations called rotamers.[22] Instead of considering the full geometrically possible conformational space, only populated (Φ,Ψ) regions and a small number of rotamers can be used for designing MC moves that describe the most frequently occurring amino acid conformations.

Another way to incorporate database-driven information into an MC scheme can be realized by basing the protein structure prediction on secondary structure assignments of the

[a] Dr. B. Strodel
Institute of Theoretical and Computational Chemistry
Heinrich Heine University Düsseldorf
40225 Düsseldorf (Germany)
E-mail: b.strodel@fz-juelich.de

[b] Dr. F. Hoffmann,+ S. G. Kamat, Dr. B. Strodel
Institute of Complex Systems: Structural Biochemistry
Forschungszentrum Jülich, 52425 Jülich (Germany)

[c] Dr. I. Vancea+
European Molecular Biology Laboratory c/o DESY
Notkestrasse 85, 22603 Hamburg (Germany)

[+] These authors contributed equally to this work.

Supporting Information for this article is available on the WWW under http://dx.doi.org/10.1002/cphc.201402247.

residues.[23,24] The secondary structure is the three-dimensional form of local segments of proteins, which consists of local inter-residue interactions mediated by hydrogen bonds. Amino acids vary in their ability to form secondary structure elements. The dominating secondary structures are α helices (henceforth denoted H) and sheets consisting of β strands (E). These regular secondary structure elements are linked by either tight turns or loose, flexible loops. Furthermore, other types of helices, such as the 3_{10} helix and π helix exist. These structural elements are collectively denoted C for "coil" in the following sections. It should be noted that random coil is not a true secondary structure, but is the class of conformations that indicate an absence of regular secondary structure. Thus, the secondary structure of a protein is characterized by a sequence of letters over the alphabet {E,H,C}, with one letter per amino acid. Most secondary structure prediction methods are evolution-based methods (also known as homology-based methods), which either exploit neural network-based approaches (e.g. Porter[25] and Psipred[26]), hidden Markov models (e.g. SAM[27]), or the frequency analysis of amino acid conformational states (e.g. Gor IV[28]). In this work, we used Porter for the prediction of secondary structure as it was identified as the best performing prediction method.[29] Miceli et al. compared the performance of nine secondary structure prediction tools applied to two protein data sets.[29] In this study, we confirmed that Porter is superior to the other methods that are based on performance criteria other than those used by Miceli et al.

The secondary structure assignment was followed by the actual folding simulation using BH. Here, we applied MC moves only to the intervening amino acids in the C conformation and connecting the H or E secondary structure elements, allowing them to establish their tertiary contacts. We term this approach secondary-to-tertiary BH. It is similar in idea to fragment-assembly approaches, which are applied in the de novo methods of Rosetta[30–32] and Chunk-Tasser.[33] We showed for the three peptides trpzip (PDB[34] ID: 1LE0),[35] trp-cage (1L2Y),[36] and ER-10 (1ERP),[37] that this secondary-to-tertiary BH implementation allows the reliable prediction of correctly folded structures within 2500 BH steps. Furthermore, we demonstrated that, for larger proteins with up to 79 residues, our approach can be used to predict correctly folded protein structures. These developments make the BH approach to global optimization a promising tool for de novo protein structure prediction, which is computationally less demanding compared to other prediction methods, and which will be applied to future studies.

2. Results and Discussion

2.1. Testing of Secondary Structure Predictors

We evaluated the precision of three secondary structure predictors—Porter,[25] Psipred,[26] and SAM[27]—by counting the number of H↔E, H↔C and E↔C mix-ups for the PDB25Select database. Table 1 shows the results for all proteins of the database, which were further split into small proteins with less than or equal to 100 residues and large proteins of greater than 100 residues, and into α and β proteins that contain only α helices and β sheets, respectively.

Table 1. Performance analysis of secondary structure prediction methods. The numbers of secondary structure mix-ups are provided for Porter, Psipred and SAM for all proteins of the PDB25Select database, proteins with ≤100 amino acids, proteins with >100 amino acids, and α and β proteins. For the first three protein sets, the percentage of mix-ups relative to the total number of amino acids in the set in question is given in parentheses.

Set	Predictor	H↔E		H↔C		E↔C	
all	Porter	9	(0.1 %)	511	(2.9 %)	484	(2.7 %)
	Psipred	186	(1.1 %)	1565	(8.8 %)	1765	(10.0 %)
	SAM	210	(1.2 %)	1645	(9.3 %)	1765	(10.0 %)
≤100	Porter	2	(0.1 %)	42	(2.8 %)	54	(3.5 %)
	Psipred	12	(0.8 %)	126	(8.2 %)	184	(12.0 %)
	SAM	19	(1.2 %)	174	(11.4 %)	178	(11.6 %)
>100	Porter	6	(0.0 %)	469	(2.9 %)	430	(2.7 %)
	Psipred	174	(1.1 %)	1439	(8.9 %)	1581	(9.8 %)
	SAM	191	(1.2 %)	1471	(9.1 %)	1587	(9.8 %)
α	Porter	0		44		3	
	Psipred	8		160		29	
	SAM	12		139		14	
β	Porter	0		3		63	
	Psipred	0		2		108	
	SAM	6		18		117	

The most striking result is that in most cases Porter performed much better than Psipred and SAM. The number of mix-ups was a factor of three to $\gtrsim 30$ larger for Psipred and SAM in comparison to Porter in almost all cases (i.e. type of mix-up and protein set). The only exceptions are the Psipred predictions for β proteins. For this set, neither Porter nor Psipred wrongly predicted H instead of E (SAM has six such mix-ups), whereas there were only three and two H assignments instead of C for Porter and Psipred, respectively. With regard to E↔C, Porter again performed significantly better than Psipred. For all three prediction methods, the number of H↔E mix-ups was by at least one order of magnitude lower than the number of H↔C and E↔C mix-ups, independent of protein length and type of fold. This indicates that helices and β sheets can be distinguished from one another by Porter, Psipred and SAM; this is important, as the basis for the BH approach used here is the accurate assignment of secondary structure for the subsequent assembly of the tertiary structure. This assumption is especially justified for Porter, which had only 0.1 % H↔E mix-ups for the total database, that only affect α/β proteins (i.e. proteins that contain both α helices and β sheets) as there were no H↔E mix-ups for α and β proteins. Compared to H↔E, the numbers of H↔C and E↔C mix-ups are somewhat higher, but generally below 3 % for Porter. For small proteins, the correct prediction of β sheets seems to be slightly more difficult, with 3.5 % E↔C mix-ups demonstrated by Porter. For both α and β proteins, E was predicted instead of C and H was predicted instead of C for only three residues in each case. This again showed that Porter is highly capable of distinguishing α and β folds. These findings led us to use Porter as the starting point for our BH simulations for the prediction of secondary structures. The main task of the subsequent BH simulations was to identify the correct tertiary contacts and to correct wrongly assigned secondary structures, which mainly involved those in which C was wrongly assigned instead of H or E.

2.2. First BH Round: From Secondary to Tertiary Structure

Porter was used to determine the secondary structure of the residues of trp-cage and ER-10, while they were manually assigned in the case of trpzip based on its target structure, as this peptide is too short to be treated by Porter. In Table 2, we present the assignments together with the secondary structure of the targets. The Porter predictions for the helix lengths in trp-cage and ER-10 were often one residue short, whereas all other predictions were correct. The 3_{10} helix in trp-cage (indicated by the letter H for residues 11–13 in the target) is by default not considered by Porter as only α and β structures were assigned. Thus, the 3_{10} helix had to be found by the BH approach. For trpzip, we assigned residues 5 and 8 to be in the coil state in order to evaluate if the BH methodology was able to identify the full β sheet.

Table 2. Secondary structure assignments along with the target secondary structure. For trp-cage and ER-10 the secondary structure assignments were obtained using Porter, whereas they were manually assigned for trpzip. The letter H for residues 11–13 in the trp-cage target denotes a 3_{10} helix.

Peptide	Secondary structure	
trpzip	assignment:	CEEECCCCEEEC
	target:	CEEEECCEEEEC
trp-cage	Porter:	CHHHHHHHCCCCCCCCCCCC
	target:	CHHHHHHHHCHHHCCCCCCC
ER-10	Porter:	CHHHHHHHCCCHHHHHHCCCCCHHHHHHHHHCCCCCCC
	target:	CHHHHHHHHCCHHHHHHHCCCHHHHHHHHHHHCCCCCC

The BH runs in this round used the information given in Table 2 in a manner described under Computational Details. For each peptide, high-temperature molecular dynamics simulations were used to generate 20 different unfolded structures, which were taken as starting structures for the BH runs. We considered three different maximum dihedral twisting angles of 30°, 60° and 90° per starting structure. Furthermore, 10 independent BH runs were performed for each starting structure and twisting angle, using different seeds for the random number generation. This amounted to $20\times10\times3=600$ BH runs per peptide. BH runs were conducted for 1000, 2000 and 5000 MC steps (also known as BH steps) for trpzip, trp-cage and ER-10, respectively. As an example, the BH input file for trpzip with step size 60° is provided in the Supporting Information.

2.2.1. *Energy versus RMSD Plots*

The performance of each BH run was measured in terms of the energy and C_α root-mean-square deviation (RMSD) from the target structure, and the three best structures per run as determined by both energy and RMSD were considered for analysis. In the following discussion, these sets of structures are denoted as "low-energy" and "low-RMSD" structures, respectively. In the ideal case, the energy function ranks the native structure in first place with respect to energy (lowest energy), that is, the sets of low-energy and low-RMSD structures are identical or at least overlap to a large extent. Figure 1 shows the energy versus RMSD plots for low-energy (blue) and low-RMSD structures (red) for the 600 BH runs per peptide, along with the structures of overall lowest energy and lowest RMSD. The results for the maximum twisting angles of 30°, 60° and 90° are displayed together. Detailed results for the individual step sizes are provided in Figures S1–S3 of the Supporting Information.

The results for the energy-minimized target structures (i.e. the energy-minimized PDB structures) using the CHARMM22/FACTS energy function are displayed as yellow dots in Figure 1. The changes to the RMSD as a result of the minimization procedure are small (< 0.5 Å). In the following discussion, we will use the structure of the energy-minimized target as a reference for the RMSD calculations as within the BH procedure one cannot expect to get closer to the PDB structure than the minimized target structure. Thus, the yellow dots in Figure 1 occur at an RMSD of zero. The energy of the energy-minimized target structures is higher than the energy of many of the low-energy structures. The target structures are NMR solution structures, which were determined by minimizing the distance or dihedral angle violations resulting from experimental constraints. It is important to note that the ensemble of structures obtained is an "experimental model", which is not necessarily the best solution when modeled with an empirical force field, such as CHARMM22/FACTS. We therefore subjected the three target structures to further optimization by performing BH runs of 1000 steps with maximum dihedral angle changes of 20°, which were applied to both backbone and side chains of three to five randomly selected contiguous residues. This procedure generated energy-optimized structures at the cost of the RMSD, which increases. The energy and RMSD values of these structures, which we call "optimized target structures", are $-307.2\ \text{kcal}\,\text{mol}^{-1}$ and 1.33 Å for trpzip, $-510.7\ \text{kcal}\,\text{mol}^{-1}$ and 1.82 Å for trp-cage, and $-907.6\ \text{kcal}\,\text{mol}^{-1}$ and 2.58 Å for ER-10, respectively. These results are shown as orange dots in Figure 1.

Figure 1 shows the success of the secondary-to-tertiary BH procedure, that is, the application of MC moves at residues between secondary structure elements H and E to obtain tertiary structure from the secondary structure data. For all three peptides, native-like structures were found, where a threshold for the C_α RMSD of 2.0 Å from the target for defining native-like conformations was used. For trpzip, trp-cage and ER-10, we identified 198, 512 and two native-like conformations, respectively. The lowest-RMSD structures shown in Figure 1 have

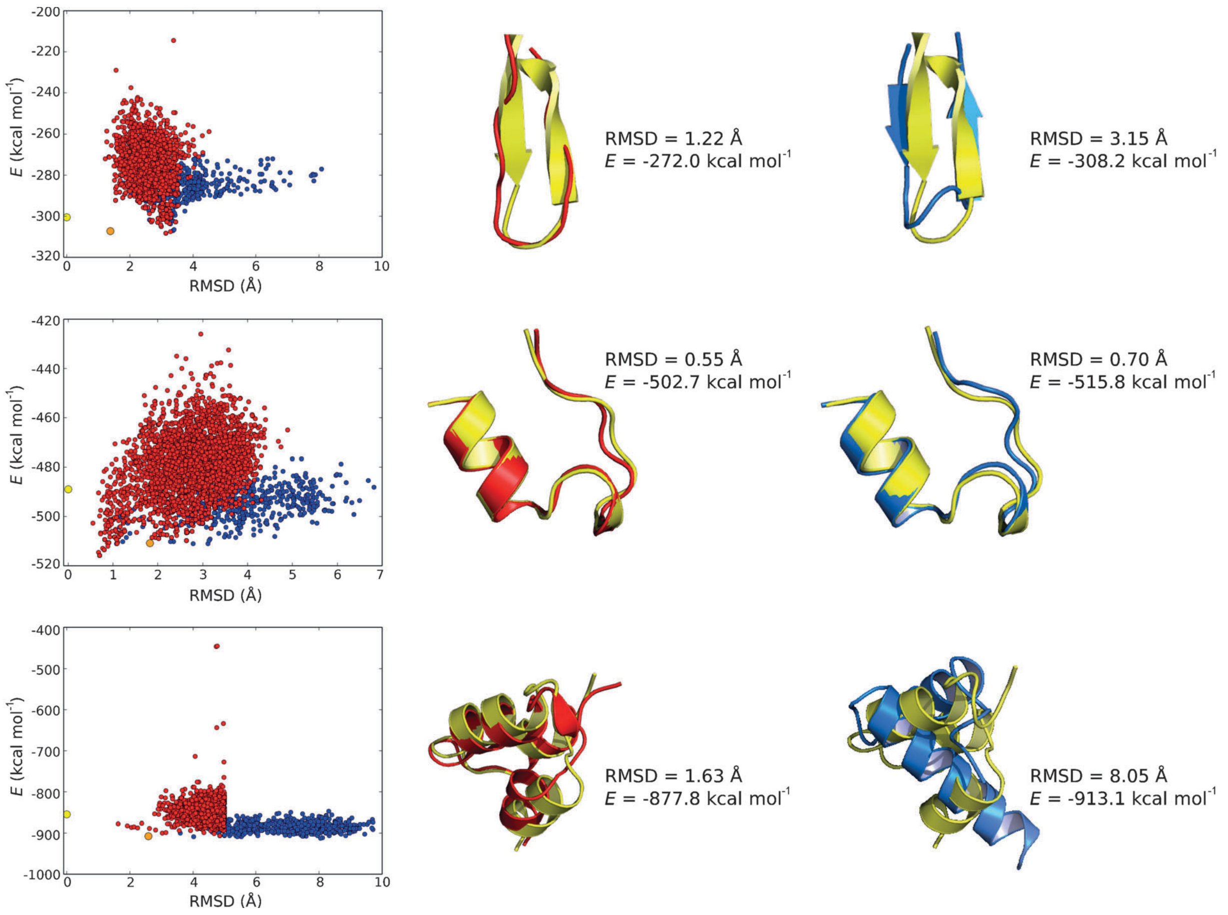


Figure 1. Results from the first BH round for trpzip (top), trp-cage (middle) and ER-10 (bottom). Left: Energy versus RMSD plots for the low-energy (blue) and low-RMSD (red) structures obtained from 600 BH runs for each peptide. For the low-RMSD structures only conformations with an RMSD value <5 Å are included, explaining the sharp cut at RMSD ≈ 5 Å for the red dots for ER-10. The minimized and the optimized target structures are represented by a yellow and an orange dot, respectively. Middle and right: Lowest-RMSD structure (red) and lowest-energy structure (blue) along with the target structure (yellow). The RMSD and energy values of these structures are shown.

RMSD values of 1.22 Å (trpzip), 0.55 Å (trp-cage), and 1.63 Å (ER-10), whereas the lowest-energy structures have RMSD values of 3.15 Å, 0.70 Å and 8.05 Å, respectively. The results for trp-cage indicate that for this peptide the CHARMM22/FACTS potential can distinguish the native structure from unfolded structures. This conclusion is supported by the funnel shape of the energy versus RMSD plot for trp-cage. Furthermore, the secondary-to-tertiary BH approach samples native-like structures of lower RMSD and lower energy than obtained for the optimized target structure (orange dot in Figure 1).

For trpzip the best structure obtained was the optimized target structure. The lowest-RMSD structure has an energy which is 35 kcal mol^{-1} higher than the energy of the optimized target structure, and the RMSD of the latter structure is almost 2 Å below the RMSD of the lowest-energy structure. The lowest-RMSD structure exhibits the hairpin, yet no β sheet formed due to the missing H-bonds between the two strands. Figure S4 shows the various structures for trpzip obtained in this study, with the H-bonds indicated in these structure plots, and an analysis of the interaction energies between the residues in these structures. The higher energy of the lowest-RMSD structure compared to that of the optimized target is due to the missing H-bonds and the electrostatic stabilization between Glu5 and Lys8. Moreover, the tryptophan residues Trp2, Trp4, Trp9, and Trp11 in the lowest-RMSD structure are not oriented as they are in the target, in which they are stacked and T-shaped with respect to each other, which further destabilizes the lowest-RMSD structure. On the other hand, in the lowest-energy structure the β sheet partially formed but the turn region deviates from the target structure. The turn folds towards the β sheet and is stabilized by a H-bond between the side chains of Asn7 and Thr10. However, the largest energetic stabilization of this structure compared to the target results from electrostatic attraction between the N- and C-terminal residues despite C-terminal amidation. The tryptophan residues that are in a stacked orientation also stabilize this structure, although the Trp2–Trp4 and Trp2–Trp11 interactions are not as strong as in the target (Figure S4). This analysis reveals how subtle the interplay between atomic positions and overall energy in all-atom energy functions is, making protein structure prediction with all-atom models a challenge. Furthermore, it has been demonstrated that implicit solvent models tend to over-weight nonnative states, which are stabilized by nonnative electrostatic attractions.[38]

Nonetheless, for both trpzip and trp-cage we found an overlap between the sets of structures of low RMSD and low energy (i.e. between the red and blue dots in Figure 1), which indicates that the CHARMM22/FACTS energy function is able to predict native-like structures as low-energy structures for both peptides. Among the low-energy structures, there were 14 native-like structures of trpzip and 40 of trp-cage. However, for ER-10 we observed a clear separation between the red and blue dots shown in Figure 1 and found no native-like structure among the low-energy structures. We identified two native-like structures at energies of approximately $-880\,\text{kcal}\,\text{mol}^{-1}$, which is approximately $30\,\text{kcal}\,\text{mol}^{-1}$ higher than the energy of the low-energy structures at high RMSD. The comparison between the RMSD and energy values for the lowest-RMSD and lowest-energy structures shown in Figure 1 highlights this observation. Nevertheless, there are some ER-10 conformations with RMSD values below 4 Å among the low-energy structures. This result indicates that the secondary-to-tertiary BH approach is also able to identify native-like structures for ER-10. The question rather is whether the considered energy function can distinguish between near-native and nonnative conformations for ER-10, which is addressed in Section 2.3.

2.2.2. The Dependence of Prediction Efficiency on Step Size

Apart from reliably identifying near-native structures as discussed above, the aim was also to find them quickly. To this end, we determined for each step size and peptide the average RMSD and energy values of the low-RMSD and low-energy structures, respectively. In addition, we analyzed how many BH steps were needed to locate the lowest-RMSD and lowest-energy structures in each BH run. These quantities allowed us to deduce which of the maximum twisting angles of 30°, 60° or 90° yields the best and fastest predictions. The results of this analysis are presented in Figure 2.

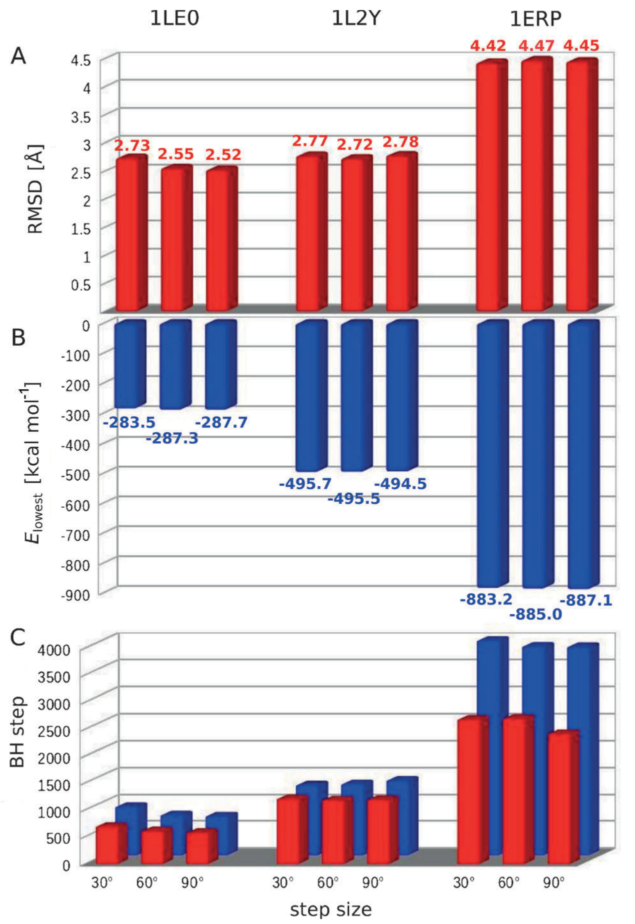


Figure 2. Results from the first BH round are shown for the maximal step size of 30°, 60° and 90°. A) The mean of the RMSD values of the low-RMSD structures and B) the mean of the energies of the low-energy structures, averaged over the 200 BH runs per peptide, and step size, are shown. C) The average numbers of BH steps needed to locate the structures of lowest RMSD (red) and lowest energy (blue) in each of the BH runs are shown.

The data in Figure 2 A and B allow us to conclude that the step size has no large influence on the identification of low-RMSD and low-energy structures. For all three peptides, the averaged RMSD and energy values were similar for the step sizes considered and none of the step sizes consistently outperformed the others. The energy versus RMSD plots for the different step sizes (Supporting Information) demonstrate that the identification of similar structures is independent of the maximum twisting angle. The final RMSD and energy values do not depend on the RMSD and energy values of the starting structures (Supporting Information). That is, near-native structures were identified not only when the BH run was initiated from somewhat folded conformations, but also from completely unfolded conformations.

Figure 2 C shows that the use of a maximum step size of 60° or 90° enabled near-native structures to be located more quickly than when the maximum twisting angle was only 30°. Low-RMSD conformations were generally produced faster than low-energy structures, implying that the RMSD did not further improve upon decreasing the energy. This is due to the above-mentioned problem of assuming the native structure as a global energy minimum and the fact that atom-based potentials are particularly sensitive to the precise position of the interacting atoms, hampering the detection of native-like geometries. As already pointed out for ER-10, the CHARMM22/FACTS potential does not identify the native structure as the global minimum on the potential energy surface. Therefore, for this peptide, an average of 1000 more BH steps were needed to find the lowest-energy structure than was required to find the location of the lowest-RMSD structure. This problem was less serious for trpzip and trp-cage, for which native-like structures correspond to low-energy structures.

In summary, less than 1000, 2000 and 5000 BH steps were generally sufficient to detect near-native (or low-energy) structures for trpzip, trp-cage and ER-10, respectively. The average computational time required for each BH run was 1.7 h for trpzip, 7.4 h for trp-cage and 54.1 h for ER-10 with a single 2.93 GHz Intel Xeon X5570 processor. Whereas a smaller twisting angle did not prevent the identification of near-native structures, a larger step size of 60° or 90° helped to find them faster. Thus, we conclude that the secondary-to-tertiary BH approach works. The aim of the following BH round was to test whether the low-RMSD and low-energy conformations identified thus far could be further optimized in unconstrained BH simulations.

2.3. Second BH Round: Refinement of Tertiary Contacts

From the structures obtained in the previous BH round, we randomly selected 31 conformations for trpzip, 38 for trp-cage, and 54 for ER-10 with low RMSD, and 56 conformations for trpzip, 77 for trp-cage, and 70 for ER-10 with low energy. We applied the following upper cutoffs for the selection of structures based on either RMSD or energy: 3.0 Å and −295 kcal mol^{-1} for trpzip, 2.5 Å and −505 kcal mol^{-1} for trp-cage, and 5.0 Å and −900 kcal mol^{-1} for ER-10. For each starting structure we performed three independent BH runs of 5000 steps for trpzip and trp-cage, and 7000 steps for ER-10. In this round, we released all constraints and applied dihedral angle changes to three, four or five randomly selected contiguous residues. All residues were considered independent from the initial secondary structure prediction, thereby enabling wrongly predicted structures to be corrected during the BH optimization procedure. We tested three different ratios of dihedral angle changes for the backbone (BB) and side chains (SC): 1) alternating BB and SC moves; 2) an SC move every fifth BH step, otherwise BB moves; 3) a BB move every fifth BH step, otherwise SC moves. The different BB/SS frequency schemes are denoted as 1:1, 4:1 and 1:4. Hence, the number of BH runs was $(31+56)\times3\times3=783$ for trpzip, $(38+77)\times3\times3=1035$ for trp-cage and $(70+54)\times3\times3=1116$ for ER-10. The performance of each BH run was measured in terms of energy and RMSD, considering the three best structures for both values. In addition, we monitored whether a BH run was started from a low-RMSD or a low-energy structure from the previous BH round. The maximum dihedral angle change in each run was 30° with group rotation moves[39] applied to the side chains. The small step size was chosen as, in this BH round, the aim was to further optimize near-native (or low-energy) structures, and not to generate completely different structures. An example input file for such a BH run for trpzip is provided in the Supporting Information.

The simulations of this round were analyzed in the same manner as the BH simulations of the first round. We produced energy versus RMSD plots (Figure 3) together with the structures of lowest RMSD and lowest energy detected for each

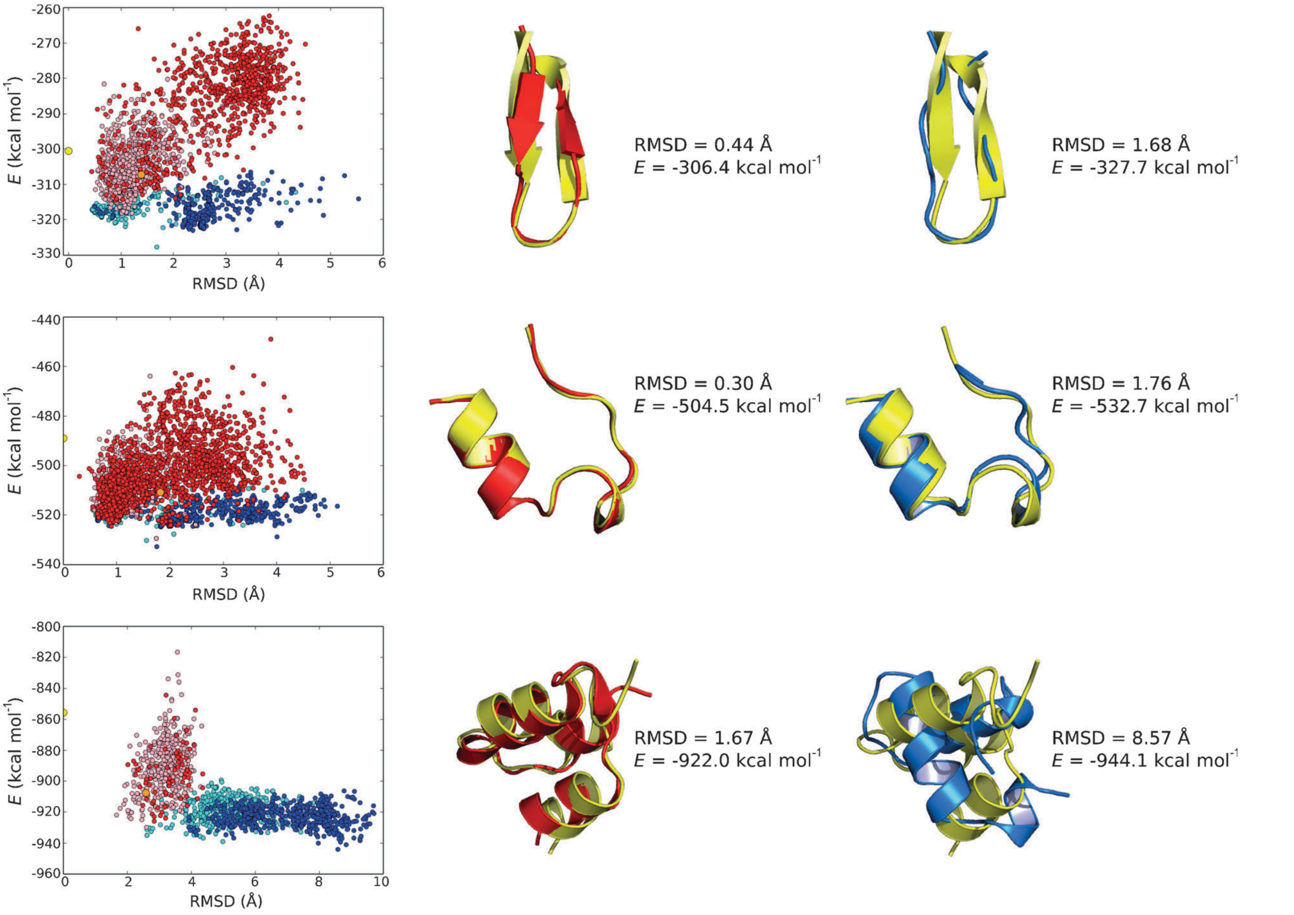


Figure 3. Results from the second BH round are shown for trpzip (top), trp-cage (middle) and ER-10 (bottom). Left: Energy versus RMSD plots for the low-energy (blue) and low-RMSD (red) structures obtained from 783 BH runs for trpzip, 1035 BH runs for trp-cage and 1116 BH runs for ER-10. The darkness of the colors indicates whether a BH run was started with a structure of low RMSD (light red or blue) or of low energy (dark red or blue), obtained in the first BH round. The minimized and the optimized target structures are represented by a yellow and an orange dot, respectively. Middle and right: Lowest-RMSD structure (red) and lowest-energy structure (blue) along with the target structure (yellow). The RMSD and energy values of these structures are shown.

peptide. We calculated the average RMSD and energy of all low-RMSD and low-energy structures, respectively, taking into account whether a BH run was started from a low-RMSD or a low-energy conformation from the first BH round. In order to be able to decide which of the BB/SC perturbation ratios worked best, we monitored the average number of BH steps needed before the best structure with respect to RMSD or energy was detected. Below we present the combined results for all BB/SC move ratios, whereas results are shown separately for the 1:1, 4:1 and 1:4 move combinations in Figures S5–S7.

2.3.1. Energy versus RMSD Plots

The first and obvious result is that unconstrained remodeling of the structures identified in the first secondary-to-tertiary BH round led to a considerable decrease in both energy and RMSD. As before, the energy-minimized target structure was used as reference for the calculation of the RMSD. The average energy decreased by approximately 30 kcal mol^{-1} for trpzip, approximately 25 kcal mol^{-1} for trp-cage, and by approximately 35 kcal mol^{-1} for ER-10. For all three peptides the optimized target structure (orange dot in the energy vs. RMSD plots) no longer belongs to the best structures in terms of RMSD or energy. For trpzip and trp-cage, many near-native structures were detected: of all the saved structures, 25.6% and 44.9% have an RMSD $\leq$2 Å for trpzip and trp-cage, respectively. Especially for trp-cage, it is almost unimportant whether the successful runs were initiated from low-RMSD or low-energy structures from the previous BH round. Information on the starting structure is provided in the energy versus RMSD plots in Figure 3, in which light and dark colors correspond to low-RMSD and low-energy starting structures, respectively. The ratio of light- and dark-colored dots below 2 Å is 5.6:1 for trpzip and 1:1.2 for trp-cage.

For trpzip, many of the low-energy structures have an RMSD $<$2 Å, which means that the CHARMM22/FACTS potential can distinguish between native-like and nonnative structures for the β hairpin. However, it has to be noted that the structure of lowest RMSD (RMSD = 0.44 Å) has an energy of 20 kcal mol^{-1} more than the value for the lowest-energy conformation with an RMSD of 1.68 Å. In the latter, the hairpin is properly formed, yet it lacks the β sheet. It has fewer backbone H-bonds compared to the target structure as the two strands are not perfectly aligned for β sheet formation. Instead, this structure is mainly stabilized by an H-bond between the N- and C-terminal residues, leading to an energy decrease of more than 60 kcal mol^{-1} compared to the same inter-residue interaction in the target (Figure S4). Another appreciable stabilization of approximately 30 kcal mol^{-1} originates from another H-bond between the side chains of Glu5 and Asn7. The tryptophan residues are not perfectly oriented with respect to each other, leading to higher interaction energies compared to the target. In the lowest-RMSD structure the β sheet is partially formed. The largest deviation from the target structure occurs around Glu5 and Lys8, which are not in the β state and have their side chains oriented differently than in the target.

For trp-cage the findings are similar to those of trpzip: a structure of rather low RMSD (0.30 Å) was detected, which has an energy of approximately 28 kcal mol^{-1} more than that of the lowest-energy conformation. However, the latter is also a near-native structure with an RMSD of 1.76 Å, which has the α helix and 3_{10} helix correctly formed. Only the C-terminal residues in the coil conformation are arranged slightly differently than in the target structure. This can be explained by the formation of H-bonds involving the side chains of the last five residues, creating a turn that is not present in the target structure (Figure S8). In conclusion, the CHARMM22/FACTS potential leads to a funnel shape of the energy versus RMSD plots for both trpzip and trp-cage, enabling the prediction of native-like structures for both peptides based on energy ranking.

The situation is different for ER-10. As was seen in the first BH round, we observed a separation between low-RMSD and low-energy structures. There is almost no overlap between these two sets of conformations, that is, between the red and blue dots in the energy versus RMSD plot for ER-10 in Figure 3. The low-RMSD set contains only six native-like structures (RMSD $\leq$2 Å), while the majority of the low-energy structures have an RMSD value $>$5 Å. The reason for this discrepancy is that the three helices in ER-10 are held together by three disulfide bridges, which are between residues Cys3 and Cys19, Cys10 and Cys37, and Cys15 and Cys27.[37] These disulfide bridges are not present in the lowest-energy structure, where the S—S distances are 16.7 Å for Cys3–Cys19, 4.9 Å for Cys10–Cys39, and 7.5 Å for Cys15–Cys27, while the disulfide bond length is 2.0 ± 0.2 Å. Instead, a salt bridge between Asp23 and Lys24 is formed in the lowest-energy structure giving rise to an interaction energy of −84.2 kcal mol^{-1} between these two residues. Thus, this salt bridge is stable and prevents the formation of the correct turn between the second and third helix of this structure. In the CHARMM force field, disulfide bonds between cysteine residues have to be defined by the user during the setup of the protein model, that is, in this case there is no possibility for a disulfide bond to form during the simulation. This shortcoming could be addressed as in the sOPEP coarse-grained force field, which permits the formation of S—S bonds based on the distance between the cysteine side-chain centroids.[40,41]

2.3.2. The Dependence of Prediction Efficiency on Move Set

The statistical analysis of the simulation results in Figure 4 highlights that for trpzip and trp-cage the low-RMSD structures have a considerably lower RMSD when the BH runs were started from low-RMSD instead of the low-energy structures obtained in the first BH round (Figure 4A). For ER-10, the differences between the average RMSD values of structures obtained when starting from low-energy or low-RMSD conformations are rather small ($<$0.2 Å). Interestingly, the energy of the low-energy structures is not affected by the choice of the starting structures for any of the peptides (Figure 4B). This allows us to conclude that BH remodeling of structures obtained from the initial secondary-to-tertiary approach is robust with respect to

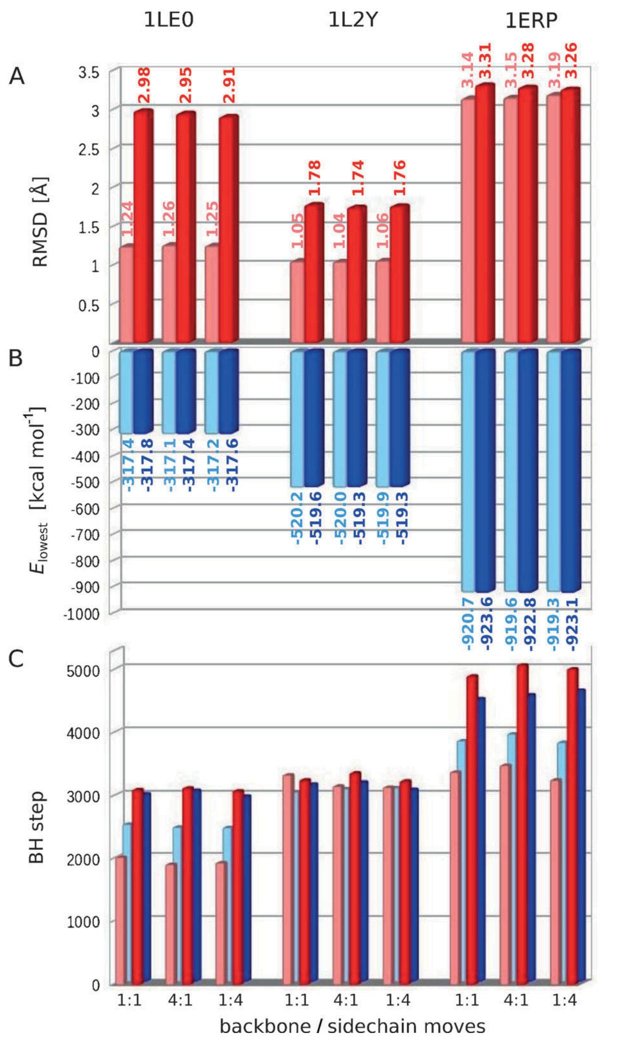


Figure 4. Results from the second BH round are shown for the BB/SC move combinations of 1:1, 4:1 and 1:4. A) The mean of the RMSD values of the low-RMSD structures and B) the mean of the energies of the low-energy structures, averaged over the BH runs per peptide, and move combination. The results shown in light colors were obtained from initial structures taken from the low-RMSD set from the first BH round, while the darker colors indicate that BH runs started from low-energy structures. C) The average numbers of BH steps needed to locate the structures of lowest RMSD (red) and lowest energy (blue) in each of the BH runs are shown.

energy minimization, whereas the improvement of the RMSD can depend on the starting configuration.

A marginal influence of the BB/SC move ratio is observed for the number of BH steps needed before the lowest-RMSD and lowest-energy structure in each BH run is detected. On average, the 1:4 move set needs fewer BH steps than the 1:1 and 4:1 move sets, although the improvement is minor. Thus, it seems to be of some advantage to have more side-chain moves (in contiguous residues) compared to backbone moves for the efficient lowering of energy and RMSD. This observation underpins the importance of side-chain packing for the native protein structure, as both side-chain–side-chain and side-chain–backbone interactions are important for in the stabilization of folded protein structures.[42] Nonetheless, we can conclude that the BH approach is robust with respect to the step-taking scheme and step size, when the results from the previous BH round are also taken into account.

As in the first BH round, we observed that often fewer BH steps are required to find low-RMSD structures compared to low-energy conformations (i.e. compare red vs. blue bars in Figure 4C). However, this result is not as clear as in the previous BH round and is also not universal. In case of trpzip and ER-10, the BH steps needed to locate low-RMSD and low-energy structures are smaller when started from low-RMSD instead of the low-energy structures from the first BH round (i.e. compare light-colored vs. dark-colored bars in Figure 4C). However, for trp-cage the average number of BH steps needed to encounter the structures of lowest energy and RMSD is independent of the starting structure. Furthermore, it is also independent of whether a lowest-RMSD or a lowest-energy structure was identified. This finding for trp-cage can be explained with the overlap between the two sets of low-RMSD and low-energy structures, in other words, low-RMSD and low-energy structures are often identical.

In summary, this statistical analysis confirmed the above conclusion that refinement of structures in this BH round was most successful when started from structures that are already close to the target structure. This finding justifies our two-step approach with a first secondary-to-tertiary BH round, followed by structure refinement of the best candidates in a second BH round. The average computational time required for each BH run in this round was 8.3 h for trpzip, 13.1 h for trp-cage and 63.6 h for ER-10, on a single 2.93 GHz Intel Xeon X5570 processor. The longer simulation times compared to those of the first BH round can be explained by the larger number of BH steps applied in this round. The computational time could be reduced by decreasing the number of BH steps, which is justified by the results presented in Figure 4C. This shows that only 2000–3000 BH steps were necessary for trpzip and trp-cage and <5000 steps for ER-10, that is, about 2000 more BH steps were performed than actually needed. Another possibility to reduce the wall-clock time is to implement a parallel version of the limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) minimization method used in this work.[43]

2.4. Protein Structure Prediction of Larger Proteins

The next step was to test whether the secondary-to-tertiary BH approach also works for larger proteins. To this end, we considered three proteins: the 35-residue villin, protein B with 53 amino acids, and a 79-residue DnaJ-like protein known as PyJ. The folding of two of these proteins has recently been studied using molecular dynamics (MD)[44] and MC-simulated annealing simulations.[45] The secondary structure predictions produced by Porter along with the secondary structure of the targets are shown in Table 3. It can be seen that Porter yielded good predictions for the secondary structure. In some cases the predicted helices were one residue too short. Only for protein B did Porter overestimate the helicity of the N-terminal residues next to the first helix.

We selected unfolded conformations produced by high-temperature MD runs of these three proteins as starting structures

Table 3. Secondary structure assignments along with the target secondary structure.

Peptide	Secondary structure	
villin	Porter:	CCHHHHHHHHCCCHHHHHCCCHHHHHHHHHHCCCC
	target:	CCHHHHHHHHCCCHHHHHHCCHHHHHHHHHHHCCC
protein B	Porter:	CCHHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHCCCHHHHHHHHHHHHHHCC
	target:	CCCCCCCCCHHHHHHHHHCCCCCCCHHHHHHHHHCCCHHHHHHHHHHHHHHCC
PyJ	Porter:	CCCCCCHHHHHHHHHHCCCCCCCCCCHHHHHHHHHHHHHHCCCCCCCCHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCC
	target:	CCCCCCHHHHHHHHHHHCCCCCCCCCHHHHHHHHHHHHHHCCCCCCCCHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCC

for the BH simulations. The current BH simulations were performed only to demonstrate that our approach also worked for larger proteins. Thus, we considered only two starting structures per protein with initial RMSD values from the target ranging from 8 Å for villin to >30 Å for protein B and PyJ. For each starting structure we performed eight independent secondary-to-tertiary BH simulations using a maximum twisting angle of 60°. After 5000 BH steps for the smaller proteins and 10 000 BH steps for PyJ the following structures of lowest RMSD were obtained: 2.9 Å for villin, 2.4 Å for protein B, and 6.4 Å for PyJ. Visual inspection of these structures revealed that the helical structures adopted the correct tertiary topology. In the second BH round, we aimed at refining the best structures from the first BH round by performing 56 BH runs for the lowest-RMSD structure for each protein. Here, we considered combinations of different maximal step sizes (30° or 40°) and different BB/SC move ratios (1:1, 4:1 and 1:4) and applied 3000 BH steps in each run. We were able to further improve the structure predictions and obtained the following RMSD values: 1.9 Å for villin, 2.1 Å for protein B, and 4.9 Å for PyJ. The corresponding native-like structures are shown in Figure 5. Given the larger size of PyJ, we assumed a C_α RMSD value below 5 Å to be sufficient to define native-like structures. We also tested our secondary-to-tertiary BH approach for a 36-residue WW domain (PDB ID: 2KCF). We obtained a reasonable RMSD value of 4.6 Å within 3000 BH steps. However, the β-sheet structure was not fully established and further refinement over another 3000 BH steps did not considerably improve this structure. Longer simulations and further methodological developments are needed for improving the prediction of long-range β-contacts with BH.

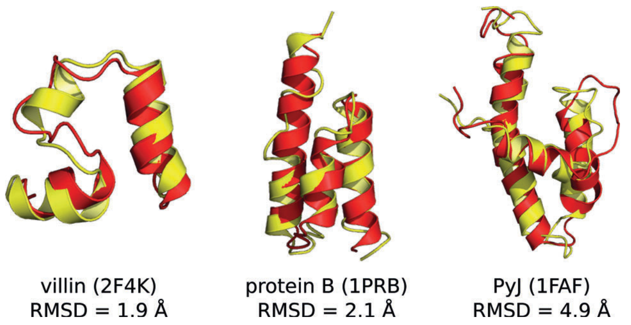


Figure 5. Native-like structures produced by BH for larger proteins. For each protein the lowest-RMSD structure (red) and target structure (yellow) are shown. The C_α RMSD values between the two structures are shown, along with the PDB accession code of the target structure.

For villin and protein B, our prediction results compare well to the results obtained by Lindorff-Larsen et al.[44] and Adhikari et al.[45] The RMSD values for the best structures for these two proteins are lowest for the MD predictions,[44] followed by our predictions, and then those from the MC simulated annealing runs.[45] Though it should be noted that whereas the RMSD for our villin structure is lower than that obtained by Adhikari et al.,[45] in their structure the middle helix is better predicted than by our BH approach. For protein B, we mainly overestimate the helicity of the N-terminal helix, which originates from the Porter prediction. In the protein B structure obtained by Adhikari et al., the helicity is also overestimated.[45] Although the microsecond-long MD simulations involving explicitly represented solvent molecules produce the best native-like structures,[44] they are the most computationally demanding.[45] The calculations by Adhikari et al. took around 600 CPU hours on an Intel 2.6 GHz Sandy Bridge Xeon E5-2670 processor for each protein. Using the same processor running NAMD, a single 10 μs MD trajectory would take around 3 000 000 CPU hours per protein.[45] The BH simulations presented here accumulated to less than 35 h for villin, around 40 h for protein B and 50 h for PyJ on a single 2.93 GHz Intel Xeon X5570 processor, counting the first and second BH round together. Thus, given the reduced computational demand of our BH approach and the good results for helical proteins, it might become a promising alternative to existing methods for protein structure prediction.

3. Conclusions

In this study, we have used the MC-based BH approach to global optimization, as previous studies have shown that BH is an effective tool for predicting global minima of peptides[5–12] and peptide assemblies.[13–15] In order to further improve the efficiency of the BH approach to protein structure prediction, we have implemented knowledge-based MC moves by incorporating secondary structure information from secondary structure prediction. We refer to this approach as secondary-to-tertiary BH. We have evaluated the performance of the secondary-to-tertiary BH scheme for three peptides: trpzip (PBD ID: 1LE0), trp-cage (1L2Y) and pheromone ER-10 (1ERP). To perturb the conformation of selected residues, we applied dihedral angle moves, as simple Cartesian moves usually perform poorly because they tend to disrupt the bonded structure of molecules. To change the dihedral angles of the side chains, we used group rotation moves, which were recently introduced to the BH scheme and shown to be effective.[39]

Based on the primary structure, each residue of the sequence was assigned a local secondary structure, which was either helix, extended or coil. We have compared the performance of three secondary structure predictors: Porter, Psipred and SAM. We found that Porter clearly provided the best prediction, independently of protein fold and length, which supports the findings of an earlier study.[29] Thus, we used Porter for secondary structure assignment as a starting point for subsequent BH simulations in which only the conformation of the residues predicted to be coil are perturbed. In doing this, we enabled the secondary structure elements to be assembled into their tertiary structure. In the case that Porter wrongly predicted coil instead of helix or strand, this could be corrected by random trial moves applied to the residues assumed to be coil. This secondary-to-tertiary BH approach was successful for the three peptides under study, as native-like structures with an RMSD of less than 2 Å from the target were found within 1000 steps for trpzip and trp-cage, and within 2500 steps for ER-10. We have benchmarked random dihedral angle moves applied to the coil residues with a maximum change of 30°, 60° or 90° and found that larger step sizes of 60° or 90° fold the proteins more efficiently.

To refine the structures predicted by the secondary-to-tertiary BH approach, we performed further BH simulations of the low-energy and RMSD structures that had been found. In order to account for the possibility that Porter wrongly assigns helix or strand instead of coil, trial moves were applied to all residues in the refinement BH runs. To do this, we used dihedral angle moves for the backbone, affecting Φ and Ψ, and group rotation moves for the side chains, perturbing the conformation of three to five randomly chosen, yet contiguous residues. We have benchmarked alternative backbone and side-chain moves with different relationships (1:1, 1:4 and 4:1) using a maximal dihedral angle change of 30° for both backbone and side chains. This rather small perturbation was chosen because the goal of these BH simulations was to refine the already folded structures. This approach is successful as both energy and RMSD were considerably improved for all three peptides, leading to the identification of more native-like structures than in the initial secondary-to-tertiary BH approach. We did not observe a strong dependence on the ratio of backbone and side-chain moves, underpinning the importance of both backbone and side chains and their interrelation for the protein structure.

In conclusion, we have introduced secondary-to-tertiary BH optimization and benchmarked this approach for three peptides. We have demonstrated that this approach reliably and effectively identifies native-like structures, which can be further refined in subsequent BH runs without restraints placed on the trial moves. Our test runs for larger proteins have produced promising results, especially for helical proteins. In future, we will apply the secondary-to-tertiary BH approach to more proteins with more than 50 amino acid residues and aim to provide a benchmark for larger proteins as we have in this study for three miniproteins. Prior to this, further methodological developments are necessary for improving the prediction of long-range residue contacts in β sheets. Moreover, we will validate our methodology for larger proteins of mixed secondary structure in a blind test, such as the critical assessment of techniques for protein structure (CASP) experiment. The current study has demonstrated that the BH approach to global optimization with improved MC moves is on the way to become a promising and computationally low-demanding tool for ab initio protein structure prediction.

Computational Details

Secondary Structure Prediction

Miceli et al. compared different secondary structure predictors and found that the neuronal-network-based predictors Porter[25] and Psipred[26] and the hidden-Markov-chain-based predictor SAM are the three most reliable prediction methods with Porter being by far the best.[29] As quality parameters, they used the average performance accuracy (Q3)[47] and the segment overlap (SOV),[48] where Q3 is a measure of the percentage of correctly guessed secondary structures of single amino acids, and SOV is obtained by computing per-segment overlaps. Neither Q3 nor SOV test which of the secondary structures (i.e. H, E and C) are mistaken for one another in case of misprediction. However, for this study, which aims to predict tertiary protein structure by assembling segments of defined secondary structure, it is significant whether H and E are interchanged, or whether H or E is interchanged with C. While the latter type of false prediction can be easily corrected in the assembly process, the mix-up of H and E structural elements would hamper the tertiary structure prediction. Therefore, we compared the performance of Porter,[25] Psipred[26] and SAM[27] in terms of secondary structure mix-ups considering the cases of H↔E, H↔C and E↔C. We collected the mix-up statistics for the PDB25Select database,[49] which was used by Miceli et al.[29]

Protein Models

The structures for trpzip, trp-cage and ER-10 were downloaded from the RCSB Protein Data Bank[34] and used as target structures. Trpzip (PDB ID: 1LE0) is a 12-residue β hairpin known as a tryptophan zipper;[35] trp-cage (1L2Y) a 20 residue peptide with a short α helix, a 3_{10} helix, and a polyproline II helix at the C terminus, which is known as tryptophan–cage miniprotein;[36] and ER-10 (1ERP) a 38-residue pheromone ER-10 from the ciliated protozoan *Euplotes raikovi* consisting of three α helices.[37] These miniproteins have been used as test cases in previous folding studies.[41,44,50–63] We used the CHARMM22 force field[64,65] to model the peptides, and the generalized Born model FACTS[66] to describe the aqueous solvent. For the calculation of the nonbonded interactions, the cutoff scheme suggested in the FACTS documentation was used, that is, truncation of both long-range electrostatics at 12 Å using a shift function and the van der Waals energy with a polynomial switching function applied between 10–12 Å. We performed 20 ns MD simulations at an elevated temperature of $T=500$ K using a Langevin thermostat with a frictional coefficient of 5 ps^{-1} to produce 20 unfolded starting structures per peptide for the subsequent folding simulations. The RMSD values of the C_α atoms between the starting structures and the corresponding target structure were 5.6–10.9 Å for trpzip, 5.7–9.7 Å for trp-cage, and 8.6–

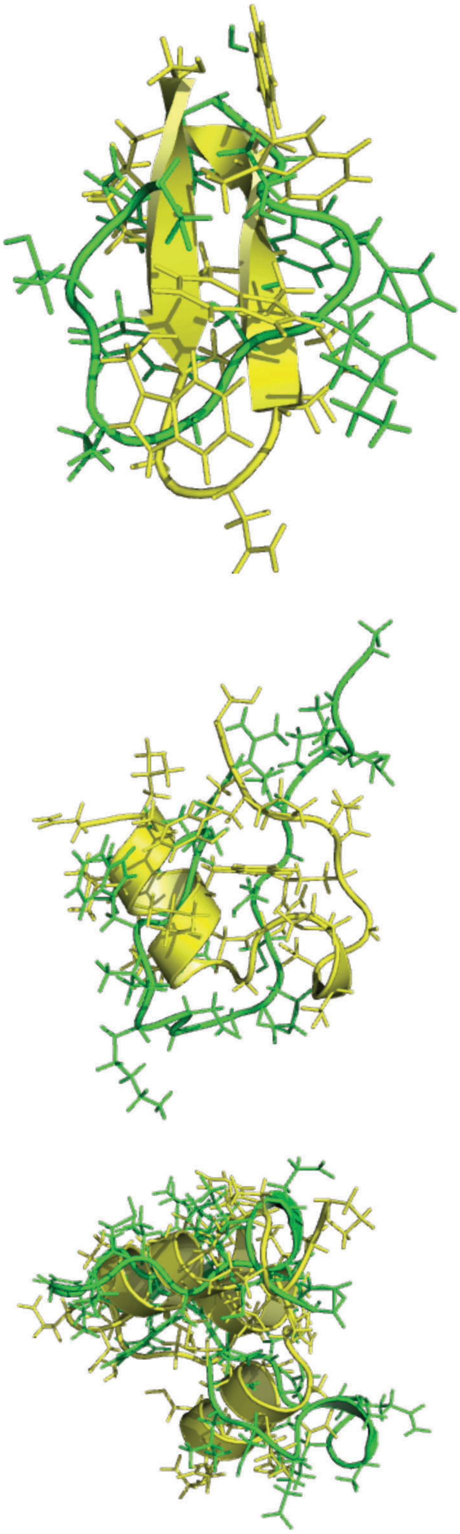


Figure 6. Representative initial structure (green) and target structure (yellow) for trpzip (top), trp-cage (middle) and ER-10 (bottom).

14.1 Å for ER-10. In Figure 6 the target structure and one representative starting structure are shown for each peptide.

For the testing of larger proteins, the structures of a 35-residue subdomain of the chicken villin headpiece (villin, 2F4K), the 53-residue protein-G-related albumin-binding module of an albumin-binding domain (protein B, 1PRB), and the N-terminal, DnaJ-like domain with 79 residues of murine polyomavirus tumor antigens (PyJ, 1FAF) were downloaded from the RCSB Protein Data Bank. For the BH simulations, the structures were prepared and modeled in the same way as the miniproteins. More simulation details for these proteins are provided in Section 2.4.

Basin-Hopping

In the BH approach to global optimization,[4,2,3] moves are proposed by perturbing the current geometry, and are accepted or rejected based upon the energy difference between the local minimum obtained by minimization from the current configuration and the previous minimum in the chain. In effect the potential energy surface is transformed into the basins of attraction[67,68] of all the local minima, so that the energy E for configuration $\mathbf{r}$ is [Eq. (1)]:

$$\tilde{E}(\mathbf{r}) = \min\{E(\mathbf{r})\} \tag{1}$$

where min denotes minimization. Large steps could be taken to sample this transformed landscape, as the objective was to step between local minima. Furthermore, there is no need to maintain detailed balance when taking steps, as the BH approach attempts to locate the global potential energy minimum and is not intended to sample thermodynamic properties. The BH algorithm has been implemented in the GMIN program[69] and has been used to find the global minimum of peptides and peptide complexes in previous studies.[5,6,8–15,18] In GMIN, local minimization is facilitated by using a modified version of the LBFGS procedure described by Liu and Nocedal.[43]

To perturb the existing geometry, we have the option of taking steps in dihedral angle space for the backbones and side chains of the peptides,[8] in which we consider the dihedral angles defining planar structures, such as rings, as rigid in order to maintain the planar geometry.[70] In earlier work, we selected a certain number of the rotatable dihedral angles for the backbone and side chains with different twisting probabilities depending on the position of the residue along the peptide chain[8] and twisted them up to a maximum angle, which can be initially set by the user and is normally in the range of 20°–50°. In this study, we used different approaches for trial dihedral moves. First, we developed a secondary-to-tertiary methodology, which uses the information from secondary structure prediction to determine the tertiary structure of the proteins. This approach is described in the next paragraph. Second, we introduced the possibility of applying trial moves to contiguous residues along the chain. Third, we applied generalized rotation moves to sample the rotameric states of protein side chains.[39] This scheme allows arbitrary groups of atoms to be rotated about an axis defined by a bond vector, maintaining maximum flexibility without introducing reliance on standard topologies. For instance, for a lysine side chain, three such rotatable groups are defined, in which atoms are rotated about the C_α–C_β, C_β–C_γ and C_γ–C_δ bonds.

Combination of Basin-Hopping with Secondary Structure Predictions

We then used the information from secondary structure prediction for the determination of the tertiary structure of a protein using the BH approach. Based on the initial secondary structure assignments, we set the Ramachandran angles (Φ,Ψ) to (−57°,−47°) and (−135°,135°) for α helices (H) and β strands (E), respectively. In the

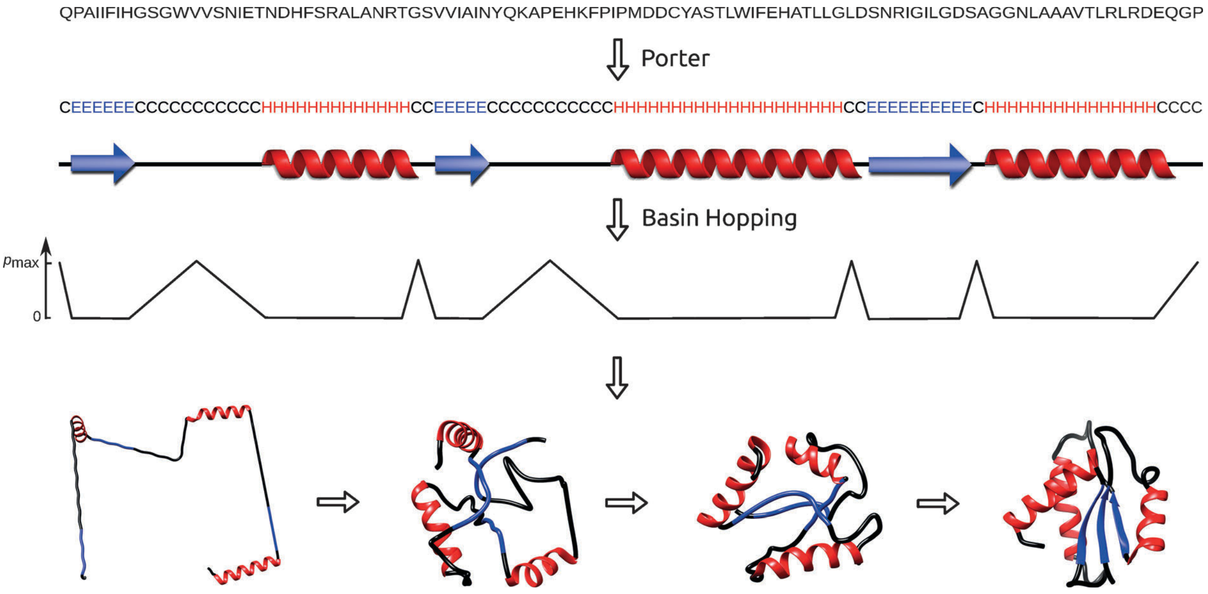

QPAIIFIHGSGWVVSNIETNDHFSRALANRTGSVVIAINYQKAPEHKFPIPMDDCYASTLWIFEHATLLGLDSNRIGILGDSAGGNLAAAVTLRLRDEQGP
Porter
CEEEEEECCCCCCCCCCHHHHHHHHHHHHHCCEEEEECCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHCCEEEEEEEEEECHHHHHHHHHHHHHHHCCCC
Basin Hopping
$p$max
0

Figure 7. Secondary structure prediction combined with BH. Based on the primary structure of the protein, the secondary structure is predicted using Porter. The starting structure for the BH run is modified by setting the (Φ,Ψ) angles to (−57°,−47°) and (−135°,135°) for residues predicted to belong to an α helix (H) and β strand (E), respectively. In the BH run, trial moves are only applied to residues that are not in the H or E state, with the twisting probability being highest (p_{max}) in the center of a segment of successive amino acids in the C state and decreasing linearly to zero at the ends of such a segment. During a BH run the tertiary contacts between secondary structure sequences become established, as illustrated at the bottom of this figure.

subsequent BH run we fixed these angles by 1) not allowing backbone dihedral angle moves for the amino acids for which H or E is being predicted, and 2) imposing constraints with a force constant of 1000 $\mathrm{kcal\,mol^{-1}\,Å^{-2}}$ on these dihedral angles during the energy minimization procedure. The constraints are necessary as otherwise the secondary structure elements would be lost during the energy minimization before the tertiary fold has been determined. This is especially true for amino acids in the E state, as β strands are often only stable as part of a β sheet. In this phase of the BH simulation, we also conserved the side-chain rotamers for the amino acids assigned to H and E. Instead, we concentrated on the amino acids predicted to be in a coil, as the tertiary contacts between the H and E segments arise from the correct structure of the protein regions in the C state. Thus, dihedral angle moves are applied only to the residues in the C state by twisting backbone and side-chain dihedral angles. Here, the twisting probability is set to be highest in the center of a segment of successive C assignments and decreases linearly until the ends of such a segment is reached, becoming zero for the amino acids assigned to H and E that are connected to the C segment. For the N- and C-terminal residues the coil state is likely. At the N terminus, the twisting probability is highest for the first residue and decreases linearly to zero until the first residue in the H or E state is reached. At the C terminus, the twisting probability increases linearly from zero for the last residues in the H or E state to its maximum for the terminating residue in the sequence. The approach used to include secondary structure information in the BH methodology is depicted in Figure 7.

For the prediction of secondary structure we use Porter[25] because we (see Section 2.1.) and Miceli et al.[29] have found that Porter provides the most reliable secondary structure prediction.

Simulation Outline

The aim of this study was to evaluate the secondary-to-tertiary BH scheme for the prediction of the tertiary structure of proteins. To this end, we limited our test set to three well-tested miniproteins with either α- or β-only structures. We started each folding simulation from 20 different initial structures per peptide. The BH simulations were divided into two rounds. First, BH runs were performed with constraints on the amino acids in the H and E conformational state according to the secondary structure prediction. From this round, the low-energy and low-RMSD (RMSD with respect to the target) structures were identified. In the ideal case, in which the force field produces the lowest energy for the native structure, these two structure sets would be identical. Unfortunately, these two sets were often different from each other as the physical, albeit empirical, force fields are not perfect. For the three peptides under consideration the potential of the CHARMM22/FACTS to identify the native structure as lowest energy structure is discussed in this article. A second round of BH runs was then performed for the low-energy and low-RMSD structures but without any constraints. Dihedral angle moves were applied to all amino acids in the chain. For the side chains, we used group rotation moves as described in ref. [39]. Unlike in previous work,[8–11,13–15,18] where the dihedral angles of randomly chosen residues along the chain were perturbed, we applied dihedral angle changes to three to five contiguous residues. Furthermore, in the first BH round, we tested different step sizes in the intervals (−30°,+30°), (−60°,+60°) and (−90°,+90°), whereas in the second BH round we tested whether alternating BB and SC moves, SC moves only at every fifth BH step, or BB moves only at every fifth BH step perform best. To benchmark each move set, we repeated each simulation ten times in the first and three times in the second BH round, using different seeds for random number generation.

Acknowledgements

F.H. and B.S. gratefully acknowledge the computing time granted by the JARA-HPC Vergabegremium and provided on the JARA-HPC Partition part of the RWTH Compute Cluster in Aachen (Grant No.JARA0018). We thank Dr. Michael Owen for critical reading of the manuscript.

Keywords: basin-hopping · molecular modeling · Monte Carlo moves · protein structure prediction · secondary structure assembly

[1] C. B. Anfinsen, *Science* **1973**, *181*, 223 – 230.

[2] D. J. Wales, J. P. K. Doye, *J. Phys. Chem. A* **1997**, *101*, 5111 – 5116.

[3] D. J. Wales, H. A. Scheraga, *Science* **1999**, *285*, 1368 – 1372.

[4] Z. Li, H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 6611 – 6615.

[5] P. Derreumaux, *J. Chem. Phys.* **1997**, *106*, 5260 – 5270.

[6] P. Derreumaux, *J. Chem. Phys.* **1997**, *107*, 1941 – 1947.

[7] M. A. Miller, D. J. Wales, *J. Chem. Phys.* **1999**, *111*, 6610 – 6616.

[8] P. N. Mortenson, D. J. Wales, *J. Chem. Phys.* **2001**, *114*, 6443 – 6454.

[9] P. N. Mortenson, D. A. Evans, D. J. Wales, *J. Chem. Phys.* **2002**, *117*, 1363 – 1376.

[10] J. M. Carr, D. J. Wales, *J. Chem. Phys.* **2005**, *123*, 234901.

[11] A. Verma, A. Schug, K. H. Lee, W. Wenzel, *J. Chem. Phys.* **2006**, *124*, 044515.

[12] M. T. Oakley, R. L. Johnston, *J. Chem. Theory Comput.* **2013**, *9*, 650 – 657.

[13] B. Strodel, D. J. Wales, *J. Chem. Theory Comput.* **2008**, *4*, 657 – 672.

[14] B. Strodel, J. W. L. Lee, C. S. Whittleston, D. J. Wales, *J. Am. Chem. Soc.* **2010**, *132*, 13300 – 13312.

[15] O. O. Olubiyi, B. Strodel, *J. Phys. Chem. B* **2012**, *116*, 3280 – 3291.

[16] M. R. Betancourt, *J. Chem. Phys.* **2011**, *134*, 014104.

[17] P. Robustelli, A. Cavalli, C. M. Dobson, M. Vendruscolo, X. Salvatella, *J. Phys. Chem. B* **2009**, *113*, 7890 – 7896.

[18] F. Hoffmann, B. Strodel, *J. Chem. Phys.* **2013**, *138*, 025102.

[19] W. W. Chen, J. S. Yang, E. I. Shakhnovich, *Proteins Struct. Funct. Bioinf.* **2007**, *66*, 682 – 688.

[20] S. Liang, N. V. Grishin, *Protein Sci.* **2002**, *11*, 322 – 331.

[21] G. Ramachandran, V. Sasisekharan, *Adv. Protein Chem.* **1968**, *23*, 283 – 438.

[22] J. W. Ponder, F. M. Richards, *J. Mol. Biol.* **1987**, *193*, 775 – 791.

[23] J. Meiler, D. Baker, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 12105 – 12110.

[24] M. Karakas, N. Woetzel, R. Staritzbichler, N. Alexander, B. E. Weiner, J. Meiler, *PLoS One* **2012**, *7*, e49240.

[25] G. Pollastri, A. McLysaght, *Bioinformatics* **2005**, *21*, 1719 – 1720.

[26] L. J. McGuffin, K. Bryson, D. T. Jones, *Bioinformatics* **2000**, *16*, 404 – 405.

[27] R. Hughey, A. Krogh, *SAM: Sequence Alignment and Modeling Software System.*, Technical Report UCSC-CRL-95-97, University of California, Santa Cruz, CA, **1995**.

[28] J. Garnier, J. F. Gibrat, B. Robson, *Methods Enzymol.* **1996**, *266*, 540 – 553.

[29] L. Miceli, L. Palopoli, S. E. Rombo, G. Terracina, G. Tradigo, P. Veltri in *9th International Conference Baton Rouge, LA, USA, May 25 – 27, 2009 Proceedings, Part I* (Eds.: G. Allen, J. Nabrzyski, E. Seidel, G. D. van Albada, J. Dongarra, P. M. A. Sloot), Springer, Heidelberg, **2009**, pp. 848 – 857.

[30] K. T. Simons, C. Kooperberg, E. Huang, D. Baker, *J. Mol. Biol.* **1997**, *268*, 209 – 225.

[31] P. Bradley, D. Chivian, J. Meiler, K. Misura, C. A. Rohl, W. R. Schief, W. J. Wedemeyer, O. Schueler-Furman, P. Murphy, J. Schonbrun, C. E. M. Strauss, D. Baker, *Proteins Struct. Funct. Genet.* **2003**, *53*, 457 – 468.

[32] P. Bradley, L. Malmström, B. Qian, J. Schonbrun, D. Chivian, D. E. Kim, J. Meiler, K. M. S. Misura, D. Baker, *Proteins Struct. Funct. Bioinf.* **2005**, *61*, 128 – 134.

[33] H. Zhou, S. B. Pandit, J. Skolnick, *Proteins Struct. Funct. Bioinf.* **2009**, *77*, 123 – 127.

[34] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, *J. Mol. Biol.* **1977**, *112*, 535 – 542.

[35] A. G. Cochran, N. J. Skelton, M. A. Starovasnik, *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 5578 – 5583.

[36] J. W. Neidigh, R. M. Fesinmeyer, N. H. Andersen, *Nat. Struct. Biol.* **2002**, *9*, 425 – 430.

[37] L. R. Brown, S. Mronga, R. A. Bradshaw, C. Ortenzi, P. Luporini, K. Wüthrich, *J. Mol. Biol.* **1993**, *231*, 800 – 816.

[38] R. Zhou, B. J. Berne, *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12777 – 12782.

[39] K. Mochizuki, C. S. Whittleston, S. Somani, H. Kusumaatmaja, D. J. Wales, *Phys. Chem. Chem. Phys.* **2014**, *16*, 2842 – 2853.

[40] J. Maupetit, P. Tuffery, P. Derreumaux, *Proteins Struct. Funct. Bioinf.* **2007**, *69*, 394 – 408.

[41] P. Thévenet, Y. Shen, J. Maupetit, F. Guyon, P. Derreumaux, P. Tufféry, *Nucleic Acids Res.* **2012**, *40*, W288 – W293.

[42] V. Z. Spassov, L. Yan, P. K. Flook, *Protein Sci.* **2007**, *16*, 494 – 506.

[43] D. Liu, J. Nocedal, *Math. Prog.* **1989**, *45*, 503 – 528.

[44] K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, *Science* **2011**, *334*, 517 – 520.

[45] A. N. Adhikari, K. F. Freed, T. R. Sosnick, *Phys. Rev. Lett.* **2013**, *111*, 028103.

[46] M. V. Berjanskii, M. I. Riley, A. Xie, V. Semenchenko, W. R. Folk, S. R. Van Doren, *J. Biol. Chem.* **2000**, *275*, 36094 – 36103.

[47] B. Rost, C. Sander, R. Schneider, *J. Mol. Biol.* **1994**, *235*, 13 – 26.

[48] A. Zemla, Č. Venclovas, K. Fidelis, B. Rost, *Proteins Struct. Funct. Bioinf.* **1999**, *34*, 220 – 223.

[49] U. Hobohm, C. Sander, *Protein Sci.* **1994**, *3*, 522 – 524.

[50] C. Simmerling, B. Strockbine, A. E. Roitberg, *J. Am. Chem. Soc.* **2002**, *124*, 11258 – 11259.

[51] S. Chowdhury, M. C. Lee, G. Xiong, Y. Duan, *J. Mol. Biol.* **2003**, *327*, 711 – 717.

[52] A. Schug, T. Herges, W. Wenzel, *Phys. Rev. Lett.* **2003**, *91*, 1 – 4.

[53] A. Schug, T. Herges, A. Verma, K. H. Lee, W. Wenzel, *ChemPhysChem* **2005**, *6*, 2640 – 2646.

[54] A. Schug, W. Wenzel, U. Hansmann, *J. Chem. Phys.* **2005**, *122*, 194711.

[55] S. Piana, K. Lindorff-Larsen, D. E. Shaw, *Biophys. J.* **2011**, *100*, L47 – L49.

[56] J. Maupetit, P. Derreumaux, P. Tufféry, *Nucleic Acids Res.* **2009**, *37*, W498 – W503.

[57] J. W. Pitera, W. Swope, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 7587 – 7592.

[58] J. Juraszek, P. G. Bolhuis, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 15859 – 15864.

[59] D. Paschek, H. Nymeyer, A. E. García, *J. Struct. Biol.* **2007**, *157*, 524 – 533.

[60] K. Klenin, W. Wenzel, *Int. J. Comput. Commun.* **2007**, *1*, 1 – 3.

[61] I. H. Radford, A. R. Fersht, G. Settanni, *J. Phys. Chem. B* **2011**, *115*, 7459 – 7471.

[62] T. Cellmer, M. Buscaglia, E. R. Henry, J. Hofrichter, W. A. Eaton, *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 6103 – 6108.

[63] Y. Tang, M. J. Grey, J. McKnight, A. G. Palmer III, D. P. Raleigh, *J. Mol. Biol.* **2006**, *355*, 1066 – 1077.

[64] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, M. Karplus, *J. Comput. Chem.* **1983**, *4*, 187 – 217.

[65] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, M. Karplus, *J. Phys. Chem. B* **1998**, *102*, 3586 – 3616.

[66] U. Haberthür, A. Caflisch, *J. Comput. Chem.* **2008**, *29*, 701 – 715.

[67] P. G. Mezey, *Potential Energy Hypersurfaces*, Elsevier, Amsterdam, **1987**.

[68] D. J. Wales, *J. Chem. Soc. Faraday Trans.* **1992**, *88*, 653 – 657.

[69] D. J. Wales, *GMIN: A Program for Basin-Hopping Global Optimisation*, http://www-wales.ch.cam.ac.uk/software.html.

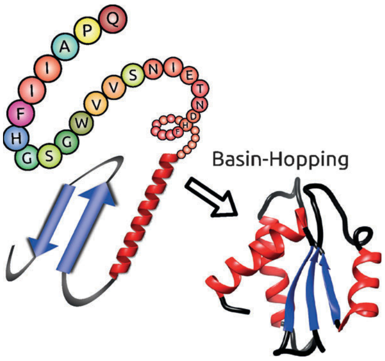
[70] M. S. Bauer, B. Strodel, S. N. Fejer, E. F. Koslover, D. J. Wales, *J. Chem. Phys.* **2010**, *132*, 054101.

Received: April 15, 2014
Published online on ■■ ■■, 0000

*F. Hoffmann, I. Vancea, S. G. Kamat, B. Strodel**

■■–■■

Protein Structure Prediction: Assembly of Secondary Structure Elements by Basin-Hopping



From secondary to tertiary structure: The basin-hopping approach to global optimization is used for protein structure prediction. The efficiency of basin-hopping is improved by introducing a methodology that derives tertiary structures from the secondary structure assignments of individual residues. It is demonstrated that this secondary-to-tertiary basin-hopping approach successfully and reliably predicts three-dimensional protein structures.

CHEMPHYSCHEM

Supporting Information

Protein Structure Prediction: Assembly of Secondary Structure Elements by Basin-Hopping

Falk Hoffmann,[b] Ioan Vancea,[c] Sanjay G. Kamat,[b] and Birgit Strodel*[a, b]

cphc_201402247_sm_miscellaneous_information.pdf

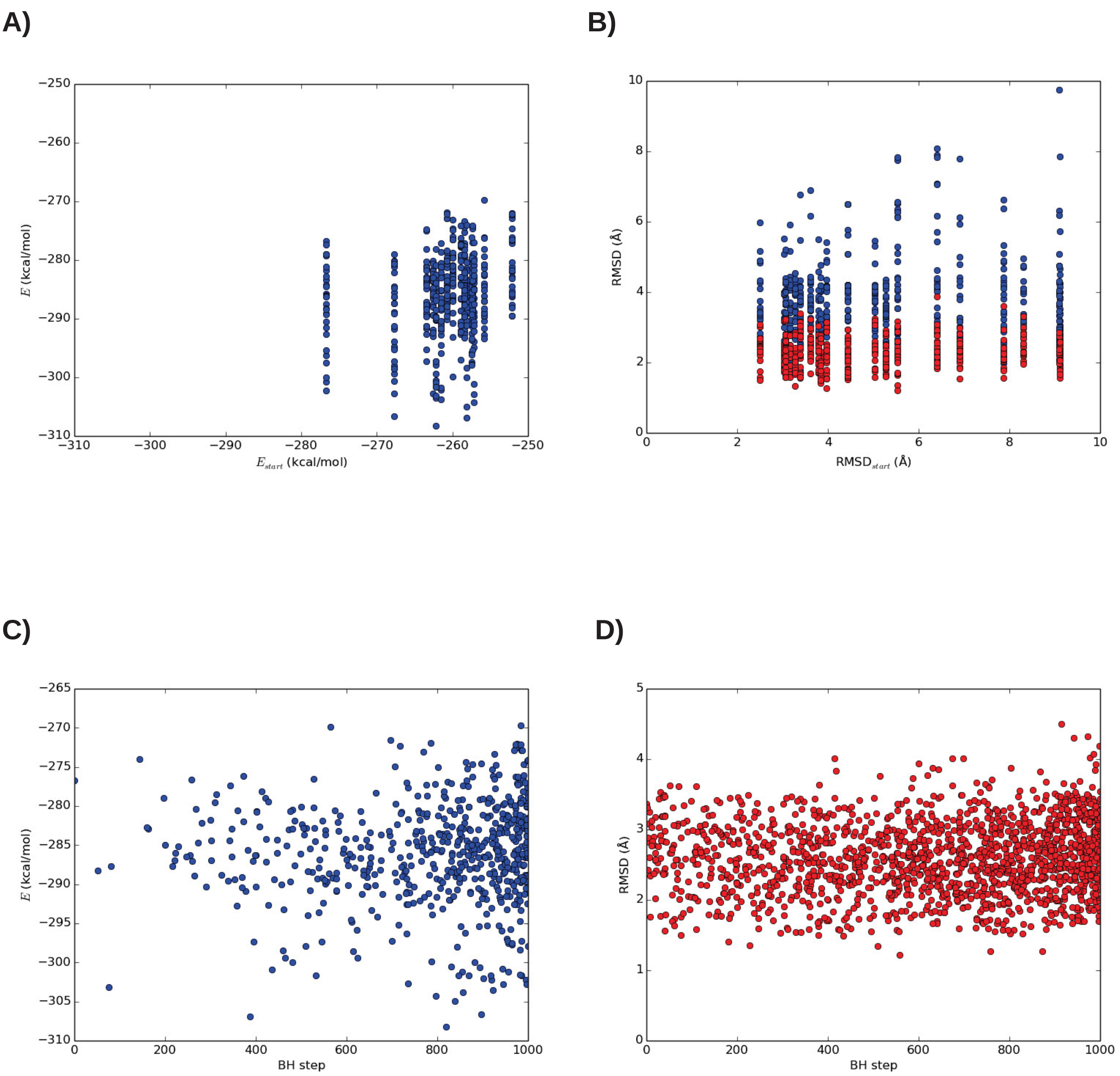


Figure S1: Results for trpzip from the first GMIN round. For all BH runs we show (A) lowest energy vs. start energy, (B) lowest RMSD vs. start RMSD, (C) lowest energy and (D) lowest RMSD vs. the BH steps needed to find the corresponding structures. Red dots correspond to low-RMSD and blue dots to low-energy structures.

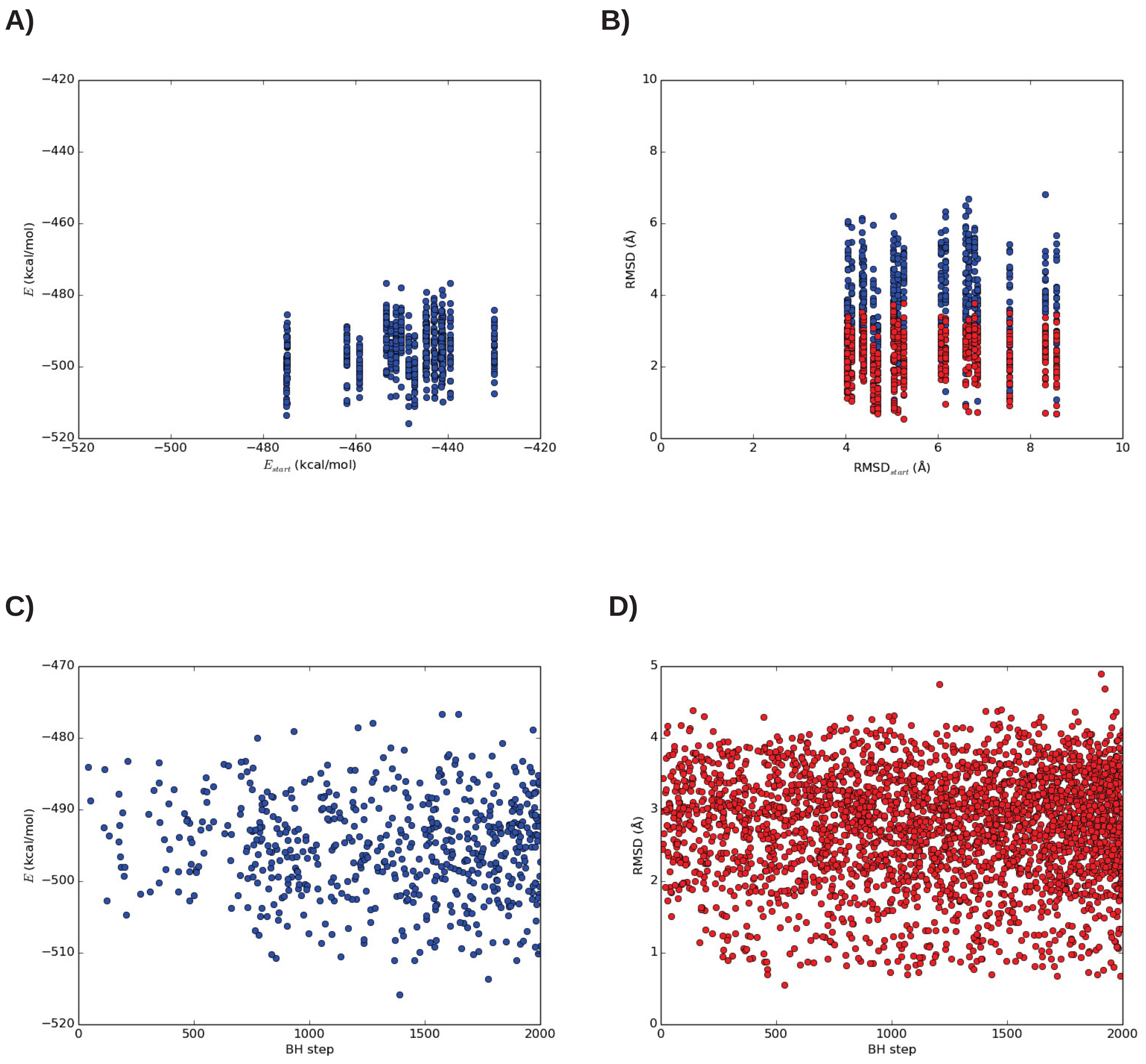


Figure S2: Results for trp-cage from the first GMIN round. For all BH runs we show (A) lowest energy vs. start energy, (B) lowest RMSD vs. start RMSD, (C) lowest energy and (D) lowest RMSD vs. the BH steps needed to find the corresponding structures. Red dots correspond to low-RMSD and blue dots to low-energy structures.

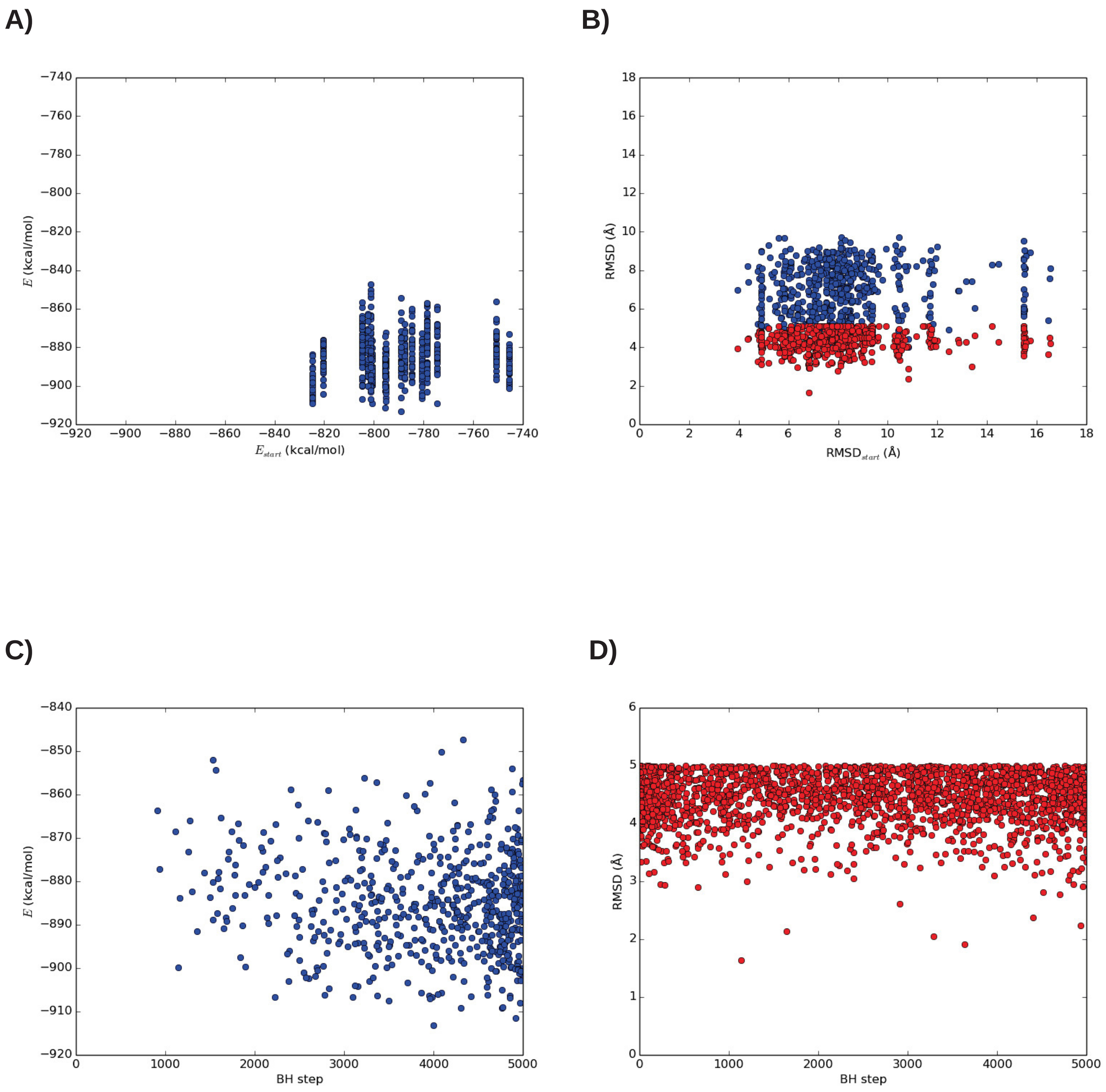
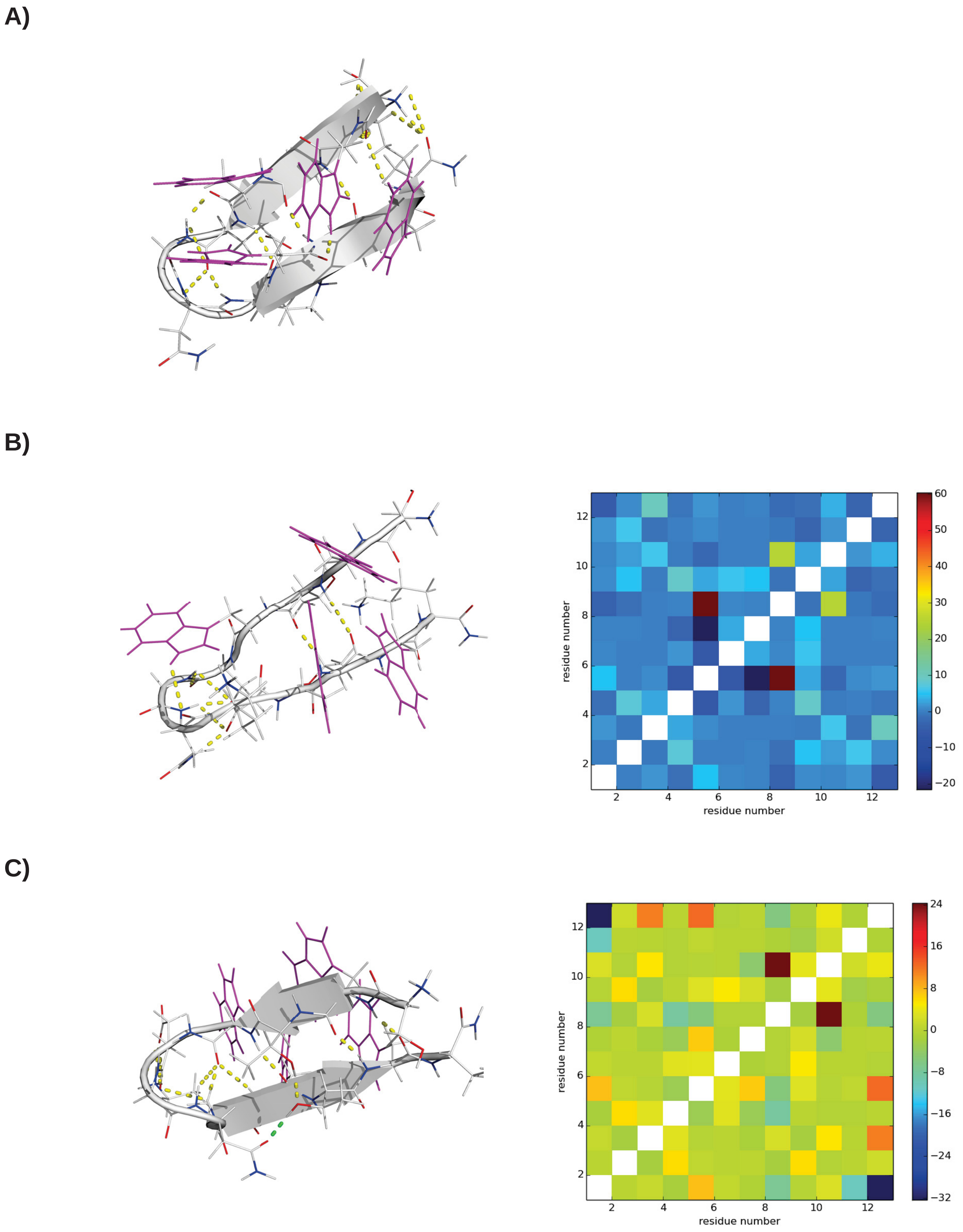


Figure S3: Results for ER-10 from the first GMIN round. For all BH runs we show (A) lowest energy vs. start energy, (B) lowest RMSD vs. start RMSD, (C) lowest energy and (D) lowest RMSD vs. the BH steps needed to find the corresponding structures. Red dots correspond to low-RMSD and blue dots to low-energy structures.


A)
B)
C)
residue number
residue number
2
4
6
8
10
12
60
50
40
30
20
10
0
−10
−20
24
16
8
0
−8
−16
−24
−32

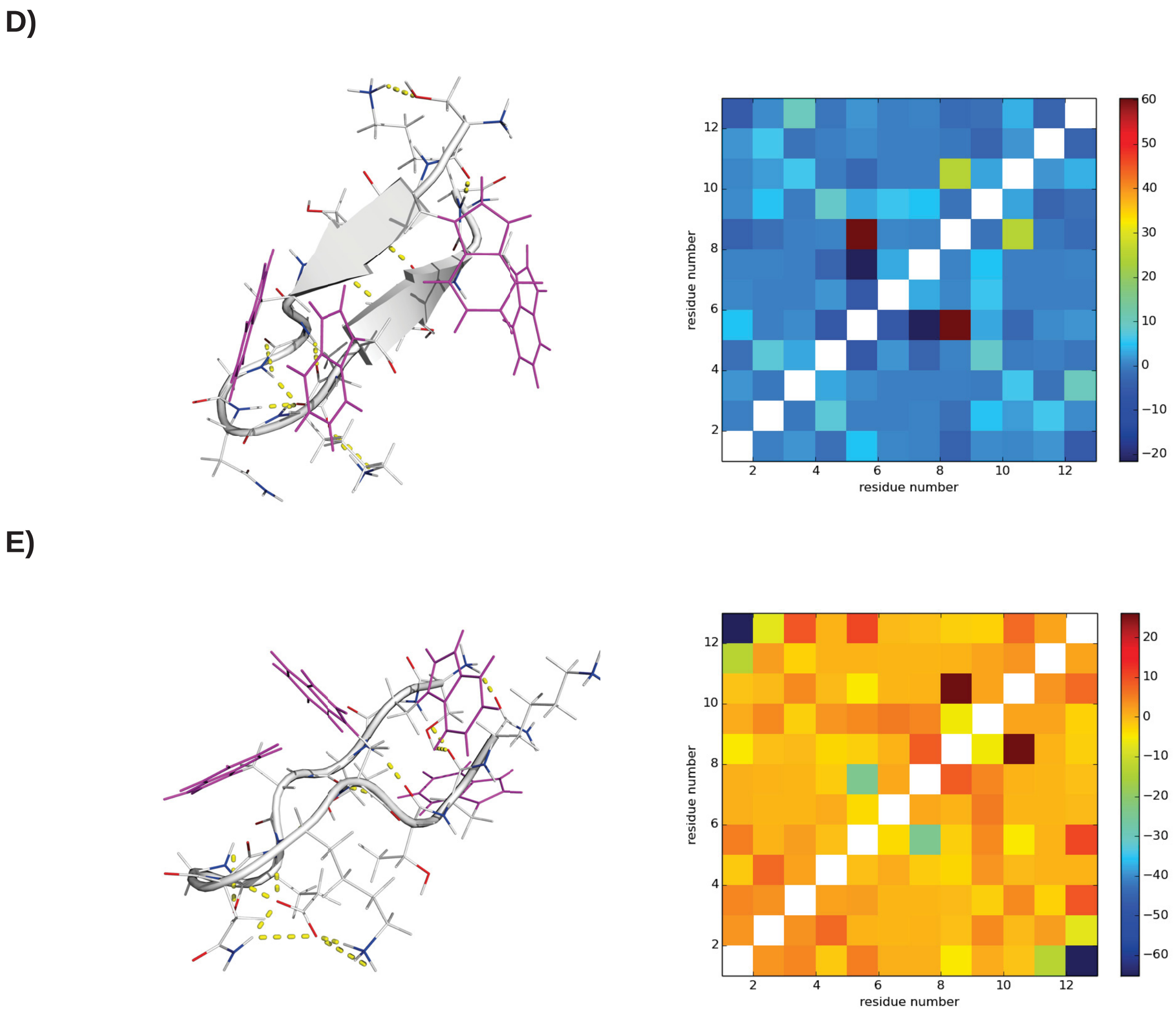


Figure S4: (Left) Target structure (A), lowest RMSD structures (B and D) and lowest energy structures (C and E) for trpzip from the first (B and C) and second (D and E) GMIN round. Nitrogen atoms are colored in blue, oxygen atoms in red and carbon atoms in white. Trp rings are marked in violet and interstrand hydrogen bonds are shown as dashed yellow lines. If hydrogen bonds between side chains are of importance, they are shown as well. (Right) Residue-residue interaction energies provided as difference between the energies of the structure shown on the left and the target structure in (A). The energy values (in kcal/mol) were calculated using the CHARMM22/FACTS potential. Note that the color scales are different between the plots due to different energy scales.

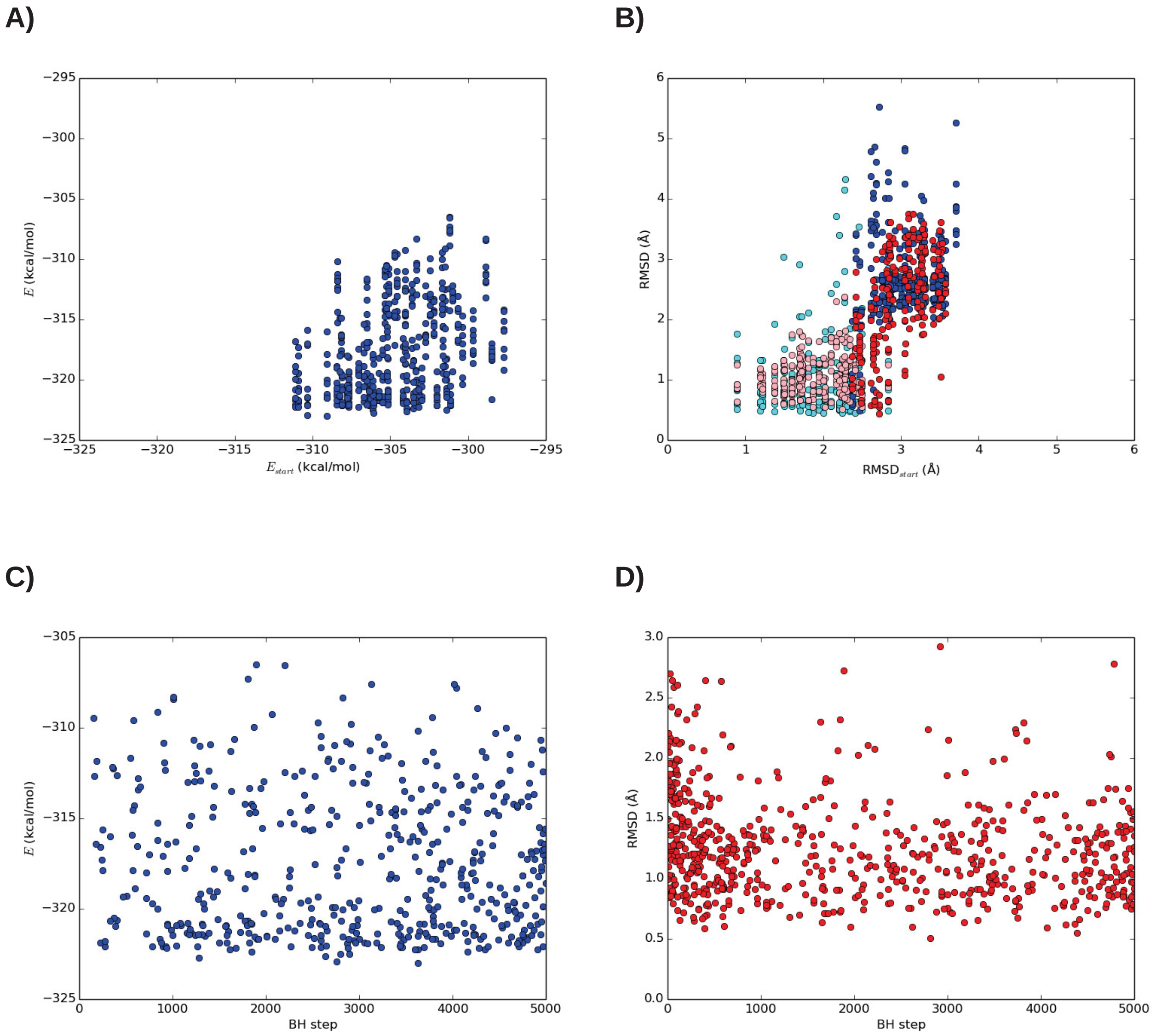


Figure S5: Results for trpzip from the second GMIN round. For all BH runs we show (A) lowest energy vs. start energy, (B) lowest RMSD vs. start RMSD, (C) lowest energy and (D) lowest RMSD vs. the BH steps needed to find the corresponding structures. Red dots correspond to low-RMSD and blue dots to low-energy structures. The darkness of the colors indicates whether a BH run was started with a structure of low RMSD (light red or blue) or of low energy (dark red or blue) obtained in the first BH round.

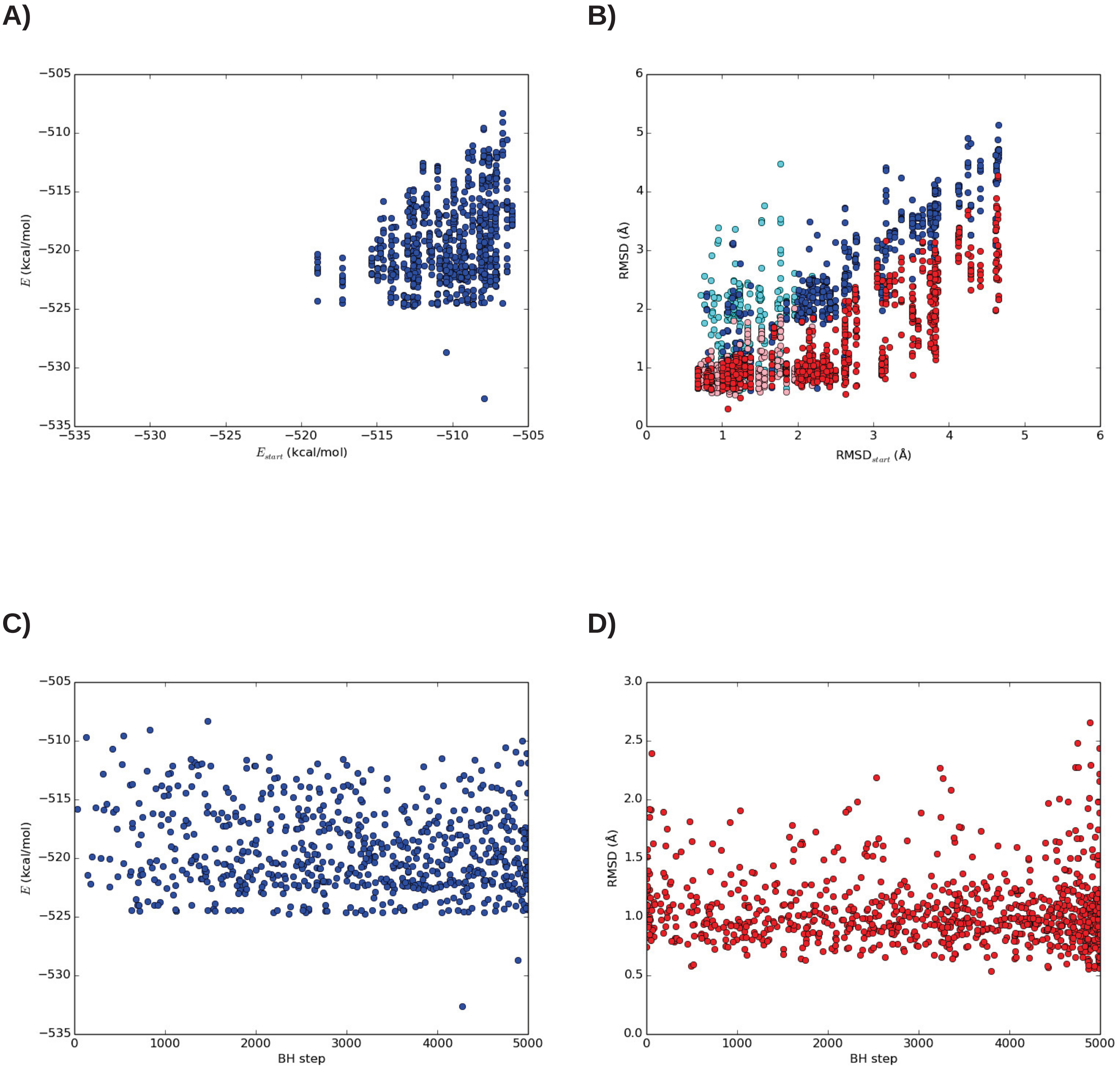


Figure S6: Results for trp-cage from the second GMIN round. For all BH runs we show (A) lowest energy vs. start energy, (B) lowest RMSD vs. start RMSD, (C) lowest energy and (D) lowest RMSD vs. the BH steps needed to find the corresponding structures. Red dots correspond to low-RMSD and blue dots to low-energy structures. The darkness of the colors indicates whether a BH run was started with a structure of low RMSD (light red or blue) or of low energy (dark red or blue) obtained in the first BH round.

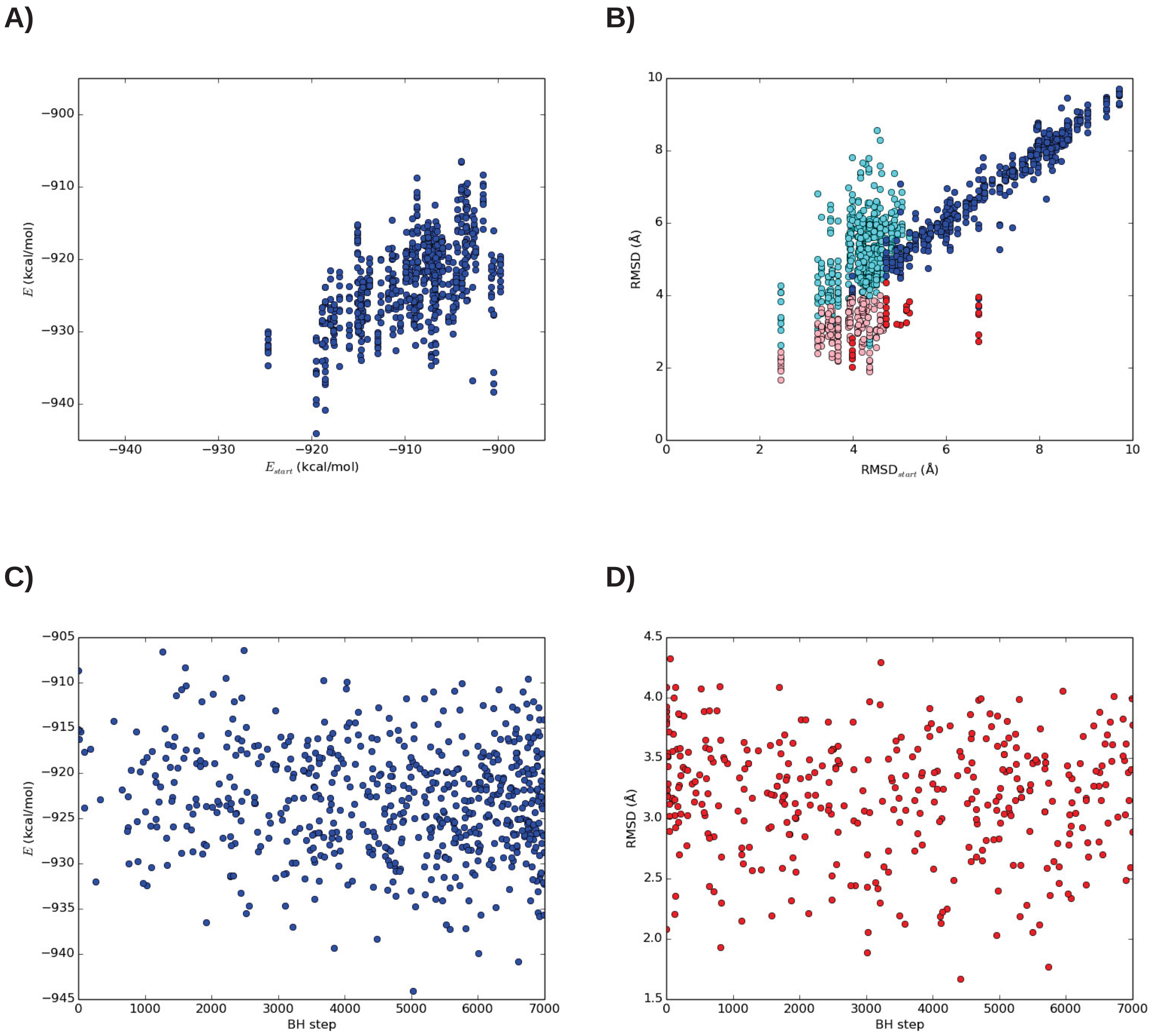


Figure S7: Results for ER-10 from the second GMIN round. For all BH runs we show (A) lowest energy vs. start energy, (B) lowest RMSD vs. start RMSD, (C) lowest energy and (D) lowest RMSD vs. the BH steps needed to find the corresponding structures. Red dots correspond to low-RMSD and blue dots to low-energy structures. The darkness of the colors indicates whether a BH run was started with a structure of low RMSD (light red or blue) or of low energy (dark red or blue) obtained in the first BH round.

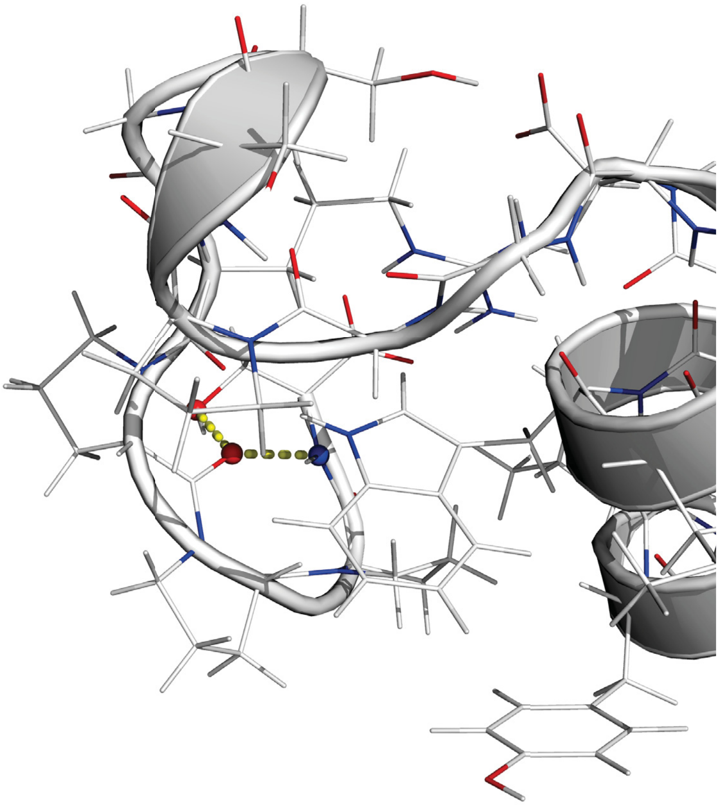


Figure S8: The C-terminal of the lowest energy structure of trp-cage obtained in the second GMIN round. For coloring explanation see Figure S4. Two stabilizing hydrogen bonds affecting the last five residues are displayed.

Sample GMIN input for trpzip: first basin-hopping round

Remarks: The secondary structure prediction from Porter is read from the external file *secstr.dat*. The keyword `CHMOVE` in conjuction with `LOOPMODEL` takes care that Monte Carlo moves are only applied to residues in the coil state as determined by Porter. The constraints in the subsequent CHARMM input are set to keep the corresponding dihedral angles constant during energy minimization. For more information on the different keywords see http://www-wales.ch.cam.ac.uk/GMIN.doc/node7.html.

```
RMS 5.0 0.1 3 1 0
TEMPERATURE 0.59
SLOPPYCONV 0.01
TIGHTCONV 0.0001
TRACKDATA
EDIFF 0.1
UPDATES 500
MAXIT 5000 5000
STEPS 1000 1.0
STEP 60 0.0
SAVE 3
SECPRED secstr.dat
CHPMAX  0.8
CHPMIN  0.2
CHMOVE LOOPMODEL
CHARMMTYPE top_all22_prot.inp par_all22_prot.inp
CHARMM
! Everything below the CHARMM line above is part of a CHARMM input file
set pardir "$CHARMM/toppar"

! Read standard topology and parameter files. These paths will need setting!
OPEN READ CARD UNIT 1 NAME @pardir/@top
READ RTF CARD UNIT 1
CLOSE UNIT 1

OPEN READ CARD UNIT 2 NAME @pardir/@par
READ PARAMETER CARD UNIT 2
CLOSE UNIT 2

! Generate the PSF
READ SEQU CARD
*
12
SER TRP THR TRP GLU GLY ASN LYS TRP THR TRP LYS
GENE A FIRS NTER LAST CT2 SETUp

OPEN UNIT 20 NAME input.crd READ CARD
READ COOR UNIT 20 CARD FREE
CLOSE UNIT 20

! Build the internal coordinate tables
IC FILL PRESERVE
IC PARAMETERS
IC PURGE
```

```
! set Phi and Psi of the helical residues (as predicted by Porter)
IC EDIT
DIHE  1 C  2 N   2 CA  2 C  -135.0
DIHE  2 N  2 CA  2 C   3 N   135.0
DIHE  2 C  3 N   3 CA  3 C  -135.0
DIHE  3 N  3 CA  3 C   4 N   135.0
DIHE  3 C  4 N   4 CA  4 C  -135.0
DIHE  4 N  4 CA  4 C   5 N   135.0
DIHE  8 C  9 N   9 CA  9 C  -135.0
DIHE  9 N  9 CA  9 C   10 N   135.0
DIHE   9 C 10 N  10 CA 10 C  -135.0
DIHE  10 N 10 CA 10 C  11 N   135.0
DIHE  10 C 11 N  11 CA 11 C  -135.0
DIHE  11 N 11 CA 11 C  12 N   135.0
END

! Build the Cartesian coordinates from the internal coordinates
COOR INIT
IC SEED 1 N 1 CA 1 C
IC BUILD

! set the constraints for the helical residues
CONS DIHE   A 1 C   A 2 N   A 2 CA  A 2 C  FORCE 1000.0  MIN -135.0  PERIOD 1
CONS DIHE   A 2 N   A 2 CA  A 2 C   A 3 N  FORCE 1000.0  MIN  135.0  PERIOD 1
CONS DIHE   A 2 C   A 3 N   A 3 CA  A 3 C  FORCE 1000.0  MIN -135.0  PERIOD 1
CONS DIHE   A 3 N   A 3 CA  A 3 C   A 4 N  FORCE 1000.0  MIN  135.0  PERIOD 1
CONS DIHE   A 3 C   A 4 N   A 4 CA  A 4 C  FORCE 1000.0  MIN -135.0  PERIOD 1
CONS DIHE   A 4 N   A 4 CA  A 4 C   A 5 N  FORCE 1000.0  MIN  135.0  PERIOD 1
CONS DIHE   A 8 C   A 9 N   A 9 CA  A 9 C  FORCE 1000.0  MIN -135.0  PERIOD 1
CONS DIHE   A 9 N   A 9 CA  A 9 C   A 10 N  FORCE 1000.0  MIN  135.0  PERIOD 1
CONS DIHE   A  9 C  A 10 N  A 10 CA A 10 C  FORCE 1000.0  MIN -135.0  PERIOD 1
CONS DIHE   A 10 N  A 10 CA A 10 C  A 11 N  FORCE 1000.0  MIN  135.0  PERIOD 1
CONS DIHE   A 10 C  A 11 N  A 11 CA A 11 C  FORCE 1000.0  MIN -135.0  PERIOD 1
CONS DIHE   A 11 N  A 11 CA A 11 C  A 12 N  FORCE 1000.0  MIN  135.0  PERIOD 1

! Set up the FACTS solvent model
! epsilon=1.0 and gamma=0.015
set diele 1.0

nbond nbxmod 5 atom cdiel eps @diele shift vatom vdistance vswitch -
      cutnb 14.0 ctofnb 12.0 ctonnb 10.0 e14fac 1.0 wmin 1.5

scalar wmain = radius

facts tcps 22 teps @diele gamm 0.015
```

Sample GMIN input for trpzip: second basin-hopping round

Remarks: The keyword `CHMOVE` in conjuction with `NEIGHBOURS` effects that Monte Carlo moves are applied to 3, 4 or 5 contiguous residues, which are randomly selected at each basin-hopping step. `CHFREQ` regulates the ratio between backbone and side chain moves (1:5 in this example). `GROUPROTATION` invokes group rotation moves for the side chains.

```
RMS 5.0 0.1 3 0 0
TEMPERATURE 0.59
SLOPPYCONV 0.01
TIGHTCONV 0.0001
TRACKDATA
EDIFF 0.1
UPDATES 500
MAXIT 5000 5000
STEPS 5000 1.0
STEP 30.0 0.0
SAVE 3
CHPMAX 1.0
CHPMIN 0.2
CHFREQ 1 1 5
CHMOVE NEIGHBOURS
GROUPROTATION 1
CHARMMTYPE top_all22_prot.inp par_all22_prot.inp
CHARMM
! Everything below the CHARMM line above is part of a CHARMM input file
set pardir "$CHARMM/toppar"

! Read standard topology and parameter files. These paths will need setting!
OPEN READ CARD UNIT 1 NAME @pardir/@top
READ RTF CARD UNIT 1
CLOSE UNIT 1

OPEN READ CARD UNIT 2 NAME @pardir/@par
READ PARAMETER CARD UNIT 2
CLOSE UNIT 2

! Generate the PSF
READ SEQU CARD
*
12
SER TRP THR TRP GLU GLY ASN LYS TRP THR TRP LYS
GENE A FIRS NTER LAST CT2 SETUp

OPEN UNIT 20 NAME input.crd READ CARD
READ COOR UNIT 20 CARD FREE
CLOSE UNIT 20

! Build the internal coordinate tables
IC FILL PRESERVE
IC PARAMETERS
IC PURGE
IC BUILD
```

```
! Set up the FACTS solvent model
! epsilon=1.0 and gamma=0.015
set diele 1.0

nbond nbxmod 5 atom cdiel eps @diele shift vatom vdistance vswitch -
      cutnb 14.0 ctofnb 12.0 ctonnb 10.0 e14fac 1.0 wmin 1.5

scalar wmain = radius

facts tcps 22 teps @diele gamm 0.015
```

4.3 Protein structure prediction with basin-hopping and Ramachandran moves

Protein structure prediction using basin-hopping with knowledge-based Monte Carlo moves

Falk Hoffmann* Birgit Strodel*†‡

August 7, 2014

1 Abstract

In this study, we extend the basin-hopping approach to global optimization to new Monte Carlo moves which are based on the statistical distribution of dihedral angles in the Ramachandran plots of amino acids. We show that α helices can be found faster with this new move set than with previously applied random dihedral angle moves. We compare state-of-the-art β sheet predictors and apply their forecasts as structural constraints in the basin-hopping approach. We conclude that the new move set with the β predictions is an aspiring method for protein structure prediction.

*Institute of Complex Systems: Structural Biochemistry, Forschungszentrum Jülich, 52425 Jülich, Germany

†Institute of Theoretical and Computational Chemistry, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

‡corresponding author: b.strodel@fz-juelich.de

2 Introduction

The prediction of protein structures is one of the most important challenges in biochemistry. The task of protein structure prediction is to identify the correct three dimensional fold of the protein (tertiary structure) from the amino acid sequence (primary structure). One very important step for finding the correct tertiary structure is the identification of α helices and β sheets, which are the most common secondary structures formed by hydrogen bonds between the amino acids. The arrangement of these secondary structure elements defines the tertiary structure.

Given a protein with N amino acids and its primary structure sequence $\mathbf{R} = \{R_1, R_2, \ldots, R_N\}$ it is relatively easy to predict short-range contacts. According to the definition of the Critical Assessment of Techniques for Protein Structure Prediction (CASP) [1], there is a contact between two residues R_i and $R_j \in \mathbf{R}$ if their C_β distance (C_α for glycine) is smaller than 8 Å. A contact is called a short-range contact if $6 \leq |i-j| \leq 11$, a medium-range contact if $12 \leq |i-j| \leq 23$ and a long-range contact if $|i-j| \geq 24$. Contacts between residues which are separated by less than 6 residues are dense and can be easily predicted from the secondary structure. α helices are formed by such dense contacts because they are stabilized by hydrogen bonds between residue R_i and R_{i+4}. β sheets are mainly formed by medium- or long-range contacts because the strands forming a sheet are typically separated by other secondary structure elements, such as turns or α helices. The prediction of α helices is therefor much easier than the prediction of β sheets and β bridges. The focus of the current study is on the identification of β sheets within the basin-hopping approach to protein structure prediction.

The manuscript is organized as follows: In section 2.1 we describe the protein dataset which we use. The functionality of state-of-the-art β sheet predictors are the topic of section 3. We then use two of these predictors for establishing β contacts in our simulations. The way how we do this is described in section 4. In section 5 we describe the BH simulation method and introduce Ramachandran-based moves, together with their comparison to random dihedral Monte Carlo moves. Finally, we summarize our results in section 6.

2.1 Dataset

2.1.1 BetaSheet916 dataset

For the performance analysis of the β sheet predictors we use the BetaSheet916 dataset from Cheng et al. [2] The dataset consists of all structures with at least 50 amino acids from the Protein Data Bank [3], which were resolved until May 2004 by X-ray with a resolution smaller than 2.5 Å and which do not contain non-standard amino acids and backbone interruptions, but do contain 10–100 β-residues where 90% of these residues have at least one β partner. The sequence identity is less than 15–20% to the UniqueProt databank [4] as of May 2004. Additionally, all chains with non-bidirectional β contacts are removed

```
GKITFYEDRGFQGRHYECSSDHSNLQPYFSRCNSIRVDSGCWMLYEQPNFQGPQYFLRRGDYPDYQQWMGLNDSIRSCRLIPHTGSH
CCEEEEECCCCCEEEEEECCCCCCCCCCCCCCEEEEECCCEEEEEECCCCCCCEEEECCCCCCCCCCCCCCCCCEEEEEEECCCCCC
*EEEEEEETTTEEEEEEE*S*BS**TTT*S**SEEEEEESEEEEESSGGG*S*EEEE*SEEESSTTTTT*SSS***EEEEE***S**

RLRIYEREDYRGQMVEITEDCSSLHDRFHFSEIHSFNVLEGWWVLYEMTNYRGRQYLLRPGDYRRYHDWGATNARVGSLRRAVDFY
EEEEEECCCCCCCEEEEECCCCCHHCCCCCCCCEEEEEEECCEEEECCCCCEEEEEEECCCCCCCCCCCCCCCCCCCEEEECCCCC
EEEEESSGGG*S*EEEE*S*BS*STTTSS*****EEEEEES*EEEESSSSS*S*EEEE*SEEE*SGGGGT*SS****EEEE*****
```

Figure 1: Amino acid sequence (top), secondary structure prediction using PsiPred [6] (center) and DSSP [7] assignments (bottom) for the protein with PDB code 1A45. Predicted or assigned residues in β sheets or β bridges are marked in blue. The first three lines belong to residues 1–87 and the last three lines to residues 88–173 of the protein.

which means that if DSSP assigns that residue R_j is a β residue partner of residue R_i then R_i has also to be assigned as a β residue partner of R_j. The BetaSheet916 dataset contains 916 chains with 187,516 residues. Among them, 48,996 are β residues forming 31,638 interstrand residue pairs. One can find 10,745 β strands with an average length of 4.6 residues and 8,172 β strand pairs with 4,519 antiparallel pairs, 2,214 parallel pairs and 1,439 pairs in isolated β bridges in the dataset. The number of β sheets in the dataset is 2,533. The average sequence separation between residue pairs and strand pairs is 43 and 40, respectively.

2.1.2 Example protein 1A45

We choose the protein with PDB code 1A45 [5] and a length of 173 resdiues from the BetaSheet916 dataset [2] as example to explain the functionality of the β predictors because this protein has a high β content. The amino acid sequence of 1A45 is shown in Figure 1. According to the secondary structure determination program DSSP [7], 69 of the 173 residues of the protein are in the β conformation. Thus the β content is 39.9%. The structure of this protein is shown in Figure 2.

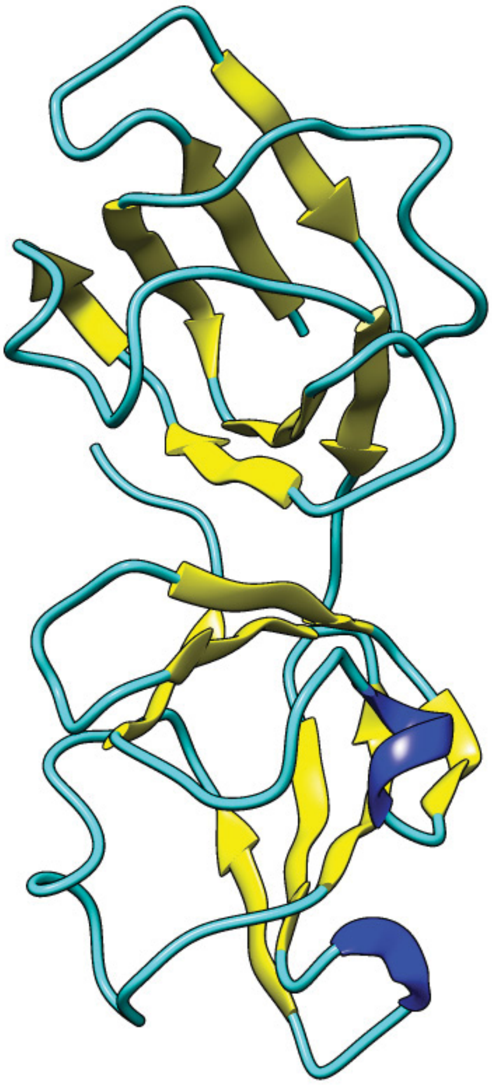


Figure 2: Structure of the protein with PDB code 1A45. The protein is shown as Ribbon with yellow color for β sheets, blue color for a 3-10 helix, cyan color for turn and white color for coil.

3 β contact Prediction

Various steps are performed to predict the contacts between β sheets. The first step is a secondary structure prediction (SSP). In two of the three approaches described below, the next step is to perform a multiple sequence alignment (MSA). The information from SSP and MSA is used in different ways to calculate the contacts between β residues. In BetaPro [2] two dimensional recursive neural networks (2D RNNs), in BCov [8] sparse inverse covariance estimation and integer programming, and in CMM [9] maximum-based correlated mutation measures are used to calculate these contacts.

3.1 Secondary Structure Prediction

With SSP the sequence of secondary structure elements $\mathbf{S} = \{S_1, S_2, \ldots, S_N\}$ for a given residue sequence $\mathbf{R} = \{R_1, R_2, \ldots, R_N\}$ is being predicted. Here, $S_i \in \{\text{H,E,C}\}$ where H stands for helix, E for extended configuration and C for coil. Miceli et al. [10] and Hoffmann et al. [11] have shown that Porter [12] is the most accurate method among the SSP methods available in 2009. Despite this finding, we choose PsiPred [6] because of its availability as a standalone version and its application in PSICOV [13], which is used in the BCov approach [8] that we want to evaluate. As an example, we show the SSP output from PsiPred [6] and the DSSP [7] assignment for the amino acid sequence of 1A45 in Figure 1. In the DSSP assignment, H stands for α helix, B for residues in isolated β bridges, E for extended strand, G for 3-10 helix, I for π helix, T for hydrogen bonded turn, S for bend and * for coil. B and E in DSSP correspond to E in the PsiPred output. If we compare both, we see that PsiPred [6] has 55 true positive (TP), 9 false positive (FP), 14 false negative (FN) and 95 true negative (TN) β sheet or β bridge predictions for this protein with a very high β content. Usually, one evaluates the performance for the full sequence, including helix and coil residues. However, here we focus on β predictions only because we aim to establish β contacts.

In our example the precision

$$P = 100\frac{TP}{TP + FP}$$

is 85.9, the recall

$$R = 100\frac{TP}{TP + FN}$$

is 79.7, the $F1$ score

$$F1 = \frac{2 \cdot P \cdot R}{P + R}$$

is 82.7, and the Matthews correlation coefficient

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

is 0.73. These values are typically used to evaluate the performance of such predictors.

3.2 Multiple Sequence Alignment

A MSA is a sequence alignment of at least three biological sequences. It is used to search for evolutionary relationships between a new or unknown sequence and sequences in a database. The information of the MSA can be used to find the evolutionary origin of a sequence and to search for similar mutations. In the case of β predictions, the MSA information is used to identify similar structures. The known information about these structures helps to specify the regions of β contacts in an unknown protein. A MSA works as follows:

1. Remove low-complexity regions and repeats.
2. Perform a n-letter search where n is the number of letters/residues.
3. Filter matching words with high score.
4. Order the words into a search tree.
5. Repeat the previous two steps with every n-letter word.
6. Compare database with remaining words.
7. Extend exact matches to high-scoring segment pairs (HSP).
8. Eliminate HSPs with values lower than empirical cutoff C.
9. Evaluate the significance of the HSP score described by an expect score EXP.
10. Align HSP regions to a longer sequence, if possible.
11. Show all matched database sequences.
12. Consider only sequences whose expect score EXP is lower than a threshold E.

As result of a MSA one obtains all sequences from a database, which match to the input sequence in the limit of the E threshold.

Here, we perform a sequence alignment between the sequences of the proteins in the BetaSheet916 [2] dataset and all proteins in the UNIREF100 database [14]. We use 5 iterations with Jackhmmer [15], which performs an iterative MSA sequence search against a protein database. Hereby, we set the E-value threshold to 0.01.

In Figure 3 one can see the MSA length distribution (top), the relative MSA sequence length distribution normalized to the protein length (center) and the cumulative distribution of the number of sequences per peptide (bottom). Here, the length refers to the number of residues of the found homologous in the databank. We normalized this number to the full protein length of the corresponding

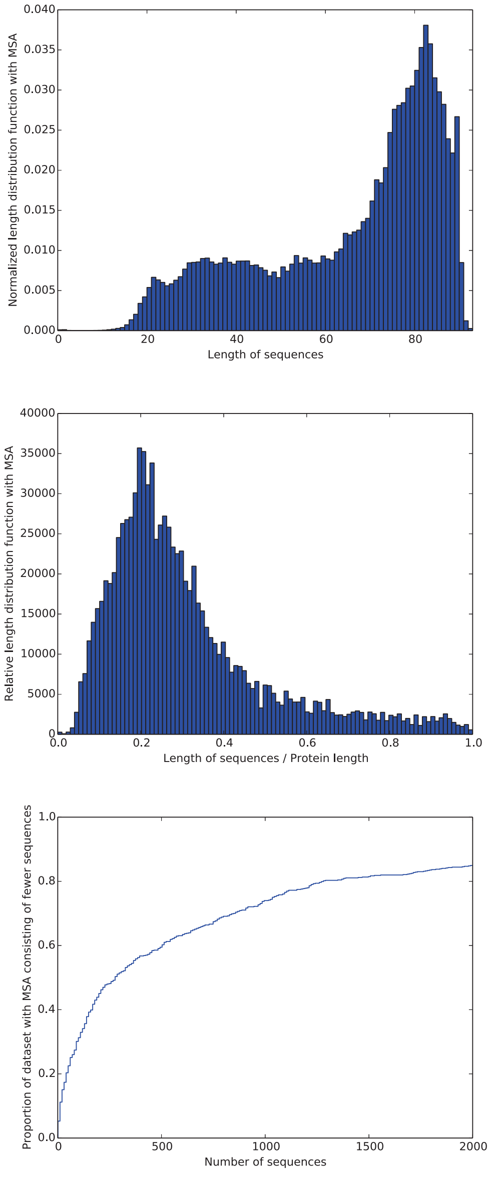


Figure 3: Normalized length distribution (top), unnormalized relative length distribution (center) and cumulative distribution of the number of sequences per peptide (bottom) of the MSA.

search protein from the BetaSheet916 [2] dataset in the second plot to see the proportion of the full alignment information we can get from the MSA. There is a huge variety in the MSA as there are structures with just one sequence and structures with more than 2000 homologous. The length of these sequences is in the range from 6 to 100. This means that independent from the size of our search proteins we cannot find homologous of these proteins in the databank with a size of more than 100 residues. An increase of the MSA threshold would increase this number, but also increases the part of sequences among the results which are not homologous enough to give reliable results about mutations or β contacts. Most of the sequences have an alignment of about 20% to the sequence of the protein target. This allows relatively accurate predictions for parts of the proteins, but shows the necessity of the SSP information to predict all β alignments as they are not always covered by MSA.

3.3 β contact prediction

There are many methods available for the prediction of β contacts. One of the first methods by Hubbard [16] uses statistical potentials. This approach predicts β strand alignments with an accuracy of $35 - 40\%$. Pairwise statistical potentials of β residues are used by Asogawa [17] and by Zhu and Braun [18], which can detect 35% of native strand alignments from other alignments. Baldi et al. [19] predict β residue contacts with neural networks, but their method is restricted to a contact prediction and is not extended to strand prediction. Steward and Thornton [20] could increase the accuracy to $45 - 48\%$ for strand alignments by using an information theoretic approach.

Cheng and Baldi [2] use a three-stage approach to predict β residue pairings, β strand pairings, β strand alignments and β sheet topology. Here, 2D RNNs are used to predict interstrand β residue pairs. Dynamic programming techniques convert the probabilities into pairing pseudoenergies and β strand alignments. Weighted graph matching algorithms are applied to create a β sheet architecture from the energies. More details about this prediction method BetaPro [2] are given in subsection 3.3.1.

Markov logic networks are used by Lippi and Frasconi [21] to predict β sheet topologies. In their method, logical formulas are applied and the weights of the formulas are trained from examples. Rajgaria et al. [22] use integer linear optimization to predict the three dimensional structure of a protein based on β contact predictions. Aydin et al. [23] use the output of BetaPro [2] and a Bayesian probabilistic model [24] to test it on a subset of the BetaSheet916 [2] dataset.

A maximum entropy-based correlated mutation measure (CMM) has been used by Burkoff et al. [9] They included a global probabilistic model for β contact prediction in the CMM and reach similar performance like BetaPro [2]. Their method is described in subsection 3.3.2.

Recently, Savojardo et al. [8] introduced BCov, an approach which predicts β sheet topologies with the help of the sparse inverse covariance estimation to calculate β strand partner scores and linear integer programming to convert

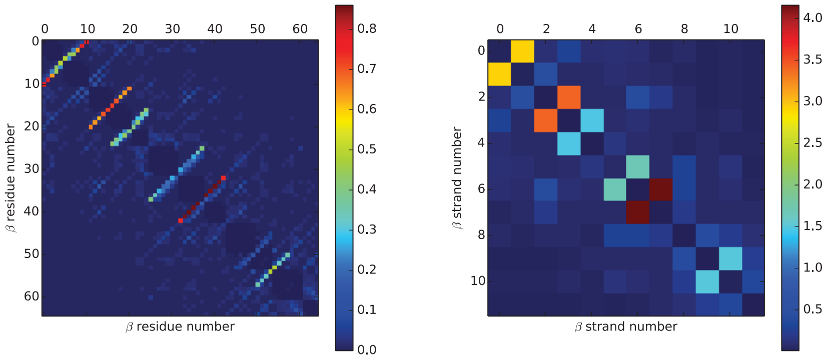


Figure 4: β residue-contact probability (left) and β strand pseudoenergy matrix (right) for 1A45 from BetaPro [2]. Only residues for which Psipred [6] predicted a β contact are included in the β residue-contact probability plot. All strands predicted by BetaPro [2] are included in the β strand pseudoenergy matrix.

the scores and to predict the β sheet topology. The approach is described in subsection 3.3.3.

3.3.1 BetaPro

BetaPro [2] determines β sheet topologies in three steps:

1. Interstrand β residue pairs are calculated with a neural network.
2. Residue pairing scores are translated into pairing pseudoenergies using dynamic programming techniques to predict β strand alignments.
3. Weighted graph algorithms are used to predict β sheets.

The three steps are explained briefly in the following.

β residue pair prediction using 2D-RNNs 2D-RNNs are used to adjust the parameters of the neural network. The two dimensions represent two residues which possibly interact with each other. The matrix with the correct values given by the PDB [3] structure is used to compare the output parameter of the output layer and train the function parameters of the hidden layer.
The output is a residue contact map which is shown for 1A45 in Figure 4. One can see the β contacts and even the strand alignments. In general it is more difficult to get the β strand alignment from the β contact prediction. Cheng and Baldi [2] used pairing constraints in their input and output matrix. By doing

pair number	strand 1	strand 2
1	3-7	1-5
2	13-18	6-11
3	34-38	12-16
4	42-46	17-21
5	54-57	22-25
6	75-81	26-32
7	88-93	33-38
8	101-105	39-43
9	121-127	44-50
10	130-133	51-54
11	139-145	55-61
12	165-168	62-65

Table 1: β strand pair prediction for 1A45 from BetaPro [2]. The first and the last residue number of every strand is shown. Strands in the same line create a β sheet contact.

so, the residue contact map tends to form line segments suggesting parallel or antiparallel β strands.

β strand alignment Cheng and Baldi [2] calculated a pseudoenergy between two strands S_i and S_j

$$E_{ij} = max_A E(A[S_i, S_j])$$

as the maximum of all pseudoenergies $E(A[S_i, S_j])$ of all possible strand alignments $A[S_i, S_j]$ of strands S_i and S_j which can be reached by sliding one strand along the other. A single pseudoenergy is the sum of all residue pair contact scores of all contact pairs involved in alignment $A[S_i, S_j]$. The maximum of these single pseudoenergies ensures that the best alignment is calculated. The result for the unnormalized strand pairing probabilities (pseudoenergies) for 1A45 is shown in Figure 4. Here, one can clearly see which strands pair with each other. However, these β strand contact map gives still no complete information about the direction of pairing (parallel or antiparallel).

β strand pair and β sheet topology prediction Assuming β strands as rectangles and including constraints (every β residue can have at most two partners as one strand can only pair with either side in a parallel or antiparallel manner, thus all strands have between 1 and 2 strand partners), Cheng and Baldi [2] transformed the pseudoenergy matrix in a connected and weighted strand pairing graph and searched for the subset of connections with the highest pseudoenergy sum of all included connections which fulfills the constraints. The result for 1A45 is presented in Table 1.

3.3.2 Correlated Mutation Measure

The method combines two parts:

1. maximum entropy-based correlated mutation measure,
2. β topology model.

Maximum entropy-based CMM The MSA gives the occurrence frequency $f_i(R_i)$ to have a specific amino acid R_i at position i in the sequence. The CMM method [9] searches among all probabilities $P(R_i)$ which fulfill

$$P(R_i) = f_i(R_i)$$

to low-order moments for the probability with the maximum entropy

$$\begin{aligned} P^* &= max_P \left(- \sum_{R_1, R_2, \ldots, R_N} P(R_1, R_2, \ldots, R_N) \log P(R_1, R_2, \ldots, R_N) \right) \\ &\propto \exp \left(- \sum_{1 \leq i < j \leq N} e_{ij}(R_i, R_j) + \sum_{1 \leq i \leq N} h_i(R_i) \right) \end{aligned}$$

where $e_{ij}(R_i, R_j)$ is the pair interaction energy between residues R_i at position i and R_j at position j, and $h_i(R_i)$ is the local field describing the preference for amino acid R_i at position i. It then calculates the CMM via

$$CMM(i,j) \propto \sum_{R_i, R_j} P_{ij}^{CMM}(R_i, R_j) \log \left(\frac{P_{ij}^{CMM}(R_i, R_j)}{f_i(R_i) f_j(R_j)} \right)$$

with

$$P_{ij}^{CMM}(R_i, R_j) \propto f_i(R_i) f_j(R_j) \exp\left(-e_{ij}(R_i, R_j)\right).$$

β topology model Let I be the set of β contacts in the system and $N_{\beta,con}$ the number of β contacts in I. If the residues at positions i and j are in β contact then $(i,j) \in I$. According to Bayes theorem [24] one can calculate the probability $P(\mathbf{S}, I|\mathbf{R})$ to find the secondary structure $\mathbf{S}$ and all β contacts I for a given amino acid sequence $\mathbf{R}$ via

$$P(\mathbf{S}, I|\mathbf{R}) \propto P(\mathbf{R}|\mathbf{S}, I) P(\mathbf{S}, I).$$

Burkoff et al. [9] assume that the secondary structure is known because they use DSSP [7] to calculate it. They neglect all dependencies on $\mathbf{S}$. However, for an unresolved protein structure the secondary structure is not known. We assume that we do not know the secondary structure, but use secondary structure predictors for the assignment. We further assume that these assignments are correct. This assumption can only be done in the limit of the accuracy of secondary structure predictors. Based on these assumptions, we can neglect the dependency on $\mathbf{S}$ and get

$$P(I|\mathbf{R}) \propto P(\mathbf{R}|I)P(I).$$

The contact prediction is calculated via

$$P(I) = P(\phi) \prod P(d_{ij})P(a_{ij}|d_{ij})P(b_{ij}|a_{ij}, d_{ij})$$

where ϕ is the permutation of all involved β strands, d_{ij} the direction of two strands i and j (1, -1 or 0 for parallel, antiparallel or isolated β bridge), a_{ij} the residue shift between strands i and j, and b_{ij} gives the bulge residue number if there is a bulge in strand i or j. Restricting the four factors to constraints occurring frequently in β regions according to the statistics in their tested datasets, Burkoff et al. [9] were able to determine $P(I)$. The likelihood $P(\mathbf{R}|I)$ was determined by

$$P(\mathbf{R}|I) \propto {N_{\beta,con}}^{aN_\beta - 1} \exp\left(-bN_{\beta,con}\right) \prod_{(i,j,d_{ij})\in I} L(R_i, R_j|d_{ij})$$

where the joint likelihood $L(R_i, R_j|d_{ij})$ is given by statistics, N_β is the total number of β residues occurring in all β strands, and a and b are constants to be determined by the model.

Combination of CMM and β topology model Burkoff et al. [9] combined the CMM and the β topology model via

$$\begin{aligned} P_{comb}(I|\mathbf{R}, CMM) &\propto P(I|\mathbf{R})P(I|CMM) \\ &\propto P(I|\mathbf{R}) \frac{\exp\left(\omega(CMM, I)\right)}{\sum_i \exp\left(\omega(CMM, I_i)\right)} \end{aligned}$$

where $\exp\left(\omega(CMM, I)\right)$ is the correlation calculated as

$$\exp\left(\omega(CMM, I)\right) = \log M \sum_i Z(i, I).$$

Here, M is the number of sequences in the MSA and $Z(i, I)$ the standardized score

$$Z(i, I) = \frac{\xi(i, I) - \mu_i}{\sigma_i}$$

of $\xi(i, I)$ with the mean μ_i and standard deviation σ_i. $\xi(i, I)$ is the mean of the set $\{CMM(i, j) : j \in I_i\}$ for which j is a β contact partner of residue i and I_i are all β contacts which involve residue i.

For our example protein 1A45 the requirement for using the MSA information in the CMM is not fulfilled. There have to be at least two sequences in the databank which have an alignment of at least one third of the protein length

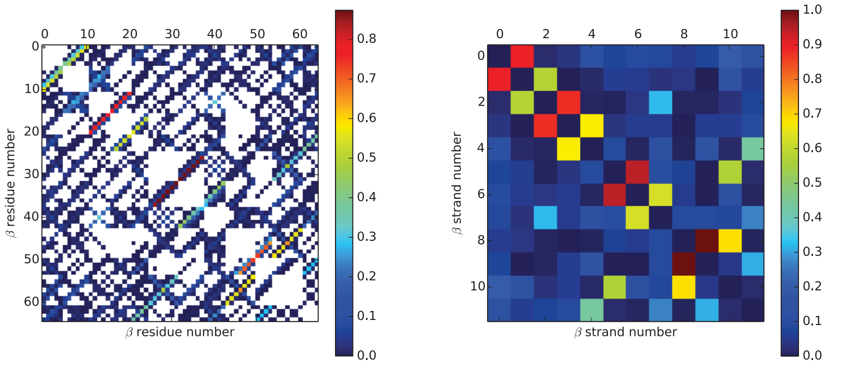


Figure 5: β residue contact probability (left) and β strand probability (right) for 1A45 from CMM [9]. Only residues for which Psipred [6] predicted a β contact are included in the β residue-contact probability plot and residues with a β probability of 0 are shown in white. All strands predicted by CMM [9] are included in the β strand pseudoenergy matrix.

of the target. Because this is not the case, we only show $P(I|\mathbf{R})$ on the strand level in Figure 5. It shows that not only the contact probability, but also the direction of the β strands can be seen in the contact prediction. The β strand probability matrix can be used to calculate the formation of β sheets. Unfortunately, Burkoff et al. [9] did not deliver the calculation of the β strand formation from the β strand probability map in their software package. Therefor we do not use the predictions of the software package from Burkoff et al. [9] in our simulations of section 4.

3.3.3 BCov

BCov [8] determines the β sheet topology in three steps:

1. Calculation of the residue contact propensity with PSICOV
2. Calculation of the β strand score for every possible pairing
3. Calculation of the β sheet topology with integer programming optimization

PSICOV calculation of the residue contact propensity PSICOV [13] uses the MSA information to create a covariance matrix $\mathbf{M}$ of size $21m \times 21m$ where m is the number of sequences in the MSA, according to their occurrence:

$$M_{ij}^{ab} = f_{ij}(a,b) - f_i(a)f_j(b)$$

Here, a and b are the amino acids (1-21 for the 20 naturally occurring amino acids and 1 for a gap in the MSA alignment), i and j are residue positions in the MSA sequence, and f is the frequency of occurrence in the MSA. By minimizing

$$\sum_{i,j=1}^{d} M_{ij}^{ab}(M^*)_{ij}^{ab} - \log\left(\det\left((M^*)^{ab}\right)\right) + \rho\sum_{i,j=1}^{d}|(M^*)_{ij}^{ab}|$$

for every a and b with ρ as a sparsity parameter, one obtains as a result of the minimization a sparse inverse matrix $\mathbf{M}^*$. Using $\mathbf{M}^*$ one can calculate the corrected contact score B_{ij}^c via

$$\begin{aligned} B_{ij}^c &= B_{ij} - \frac{\overline{B}_{i,-}\overline{B}_{-,j}}{\overline{B}} \\ B_{ij} &= \sum_{a,b}|(M^*)_{ij}^{ab}|. \end{aligned}$$

Here, $\overline{B}_{i,-}$ is the mean of the contact predictions between residue position i and all other residues, and $\overline{B}$ is the mean of the full contact score matrix. As a result, one obtains a contact map from PSICOV [13].

As before, the MSA of our example protein 1A45 does not fulfill the conditions that PSICOV [13] needs to determine a contact map. Thus, we cannot use the MSA information to calculate the β sheet topology of 1A45 with BCov [8].

Calculation of β strand scores and β sheet topology If possible, BCov [8] uses the contact scores from PSICOV [13] to calculate interaction scores. First it reduces the PSICOV [13] dimension which is a general matrix for all possible contacts to a matrix where only β contacts are involved. Then it calculates the submatrix $\mathbf{S}_{ij}$ which contains the interaction scores.

$$S_{ij}(s_i, s_j) = \begin{cases} s_{\|}(s_i, s_j) & if \quad i < j \\ 0 & if \quad i = j. \\ s_{\perp}(s_i, s_j) & if \quad i > j \end{cases}$$

The scores $s_{\|}(s_i, s_j)$ and $s_{\perp}(s_i, s_j)$ for parallel and antiparallel β strands are a result of the β contacts involved in the strands s_i and s_j. Linear integer programming is then used to maximize

$$\sum_{i,j=1}^{n} S_{ij}X_{ij}$$

where the matrix $\mathbf{X}$ is the solution matrix. Several constraints are set to make sure that the solution is binary, compatible with parallel or antiparallel assignment and the maximum number of β strand partners for every strand is 2.

BCov [8] gives the residues which form β strands as an output in table form. The result for 1A45 is given in Table 2.

residue 1	residue 2
2	16
3	15
4	14
5	13
6	12
34	167
35	166
36	165
37	164
42	56
43	55
44	54
45	53
75	92
76	91
77	90
78	89
79	88
80	87
100	124
101	123
102	122
103	121
104	120
129	141
130	140
131	139
132	138

Table 2: Two residues in contact which form β strand pairs as predicted by BCov [8] for protein with PDB code 1A45.

4 β constraints

In section 3 we described how to obtain a prediction of β contacts which form β sheets. In this section we will show how we include this information as constraints in our force field. A very general formulation for the energy from a force field is

$$\begin{aligned} E &= E_{bonded} + E_{nonbonded} \\ E_{bonded} &= E_{bond} + E_{angle} + E_{dihedral} + E_{improper} \\ E_{nonbonded} &= E_{electrostatic} + E_{LJ} \end{aligned}$$

We apply additional constraints between the residues predicted to form a β contact using one of the β predictors introduced in section 3. We use the C_α atoms of the residues as they are a good approximation for the center of the residues. Furthermore, we apply distance constraints, as we obtain some kind of distance constraints from the contact prediction. The combination with other constraints, such as angle constraints, would also be possible because atoms in β residues follow a specific conformation pattern in β sheets due to sterical requirements.
The constraints are nonbonded constraints as there is no covalent bond between the atoms forming a β sheet. On the other hand, they are short-range interactions as their task is to help forming β sheets during simulations. For small distances one can do a Taylor expansion around the minimum. If the gradient of the functional term is not too big, one can stop the Taylor expansion after the second order which gives a harmonic approach for the C_α distance constraints:

$$E_{cons} = \sum_{i=1}^{N_{\beta,con}} K \left(||\mathbf{r}_{i,1} - \mathbf{r}_{i,2}|| - d \right)^2 .$$

Here, E_{cons} is the potential energy of the additional distance constraint, $||.||$ the Euclidean distance, $\mathbf{r}_1$ and $\mathbf{r}_2$ the positions of the C_α atoms of the residues predicted to be in β contact, d the typical distance between C_α atoms in β sheets, $N_{\beta,con}$ the number of predicted contacts and K the force constant, for which the optimal value has to be determined. The average distance between C_α atoms of pairing residues in β strands is 4.7 Å. However, the exact interstrand distance depends on the pairing residues, neighbouring residues and on the question if the involved strands have an alignment only with one or with two stands. We set $d = 5$ Å for our initial simulations. K should not be too high because the C_α atoms of the involved pairs need some flexibility to find the perfect distance to form a β strand. On the other hand, K should not be too low either because the pairing partners would not find each other if the constraint is not high enough. We tested different values for K to find the optimal value enabling for β strand formation.

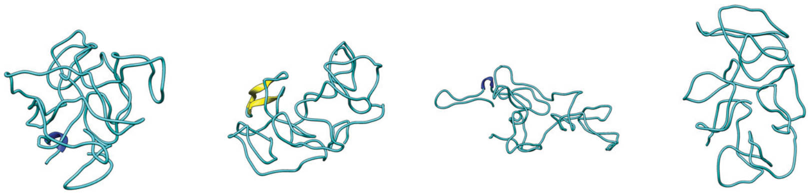


Figure 6: Lowest energy structures of 1A45 after 300 BH steps using RM dihedral angle moves. The values for the force constant from left to right are 0.01 $\mathrm{kcal\,mol^{-1}\mathring{A}^{-2}}$, 0.1 $\mathrm{kcal\,mol^{-1}\mathring{A}^{-2}}$, 1 $\mathrm{kcal\,mol^{-1}\mathring{A}^{-2}}$ and 10 $\mathrm{kcal\,mol^{-1}\mathring{A}^{-2}}$.

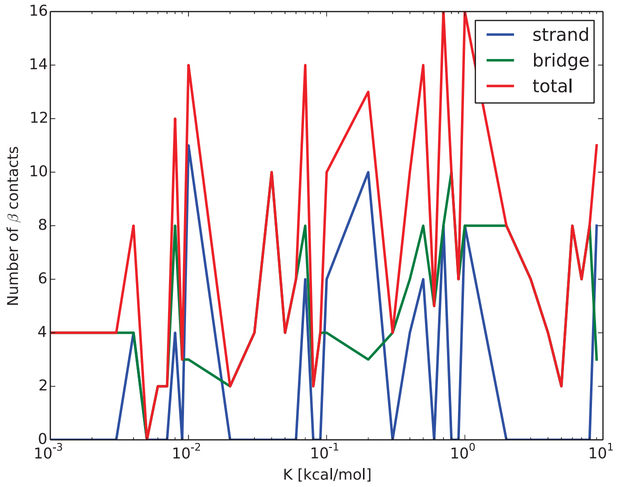
4.1 Determination of the force constant K and strand distance d for β constraint

We performed basin-hopping (BH) simulations with Ramachandran (RM) dihedral angle moves. The methods are described in section 5. We performed simulations with 300 BH steps from the fully extended structure of our example protein 1A45 with four different values for the force constant K. We used the prediction from BCov [8] as input for the constraints. We show the structures with lowest energy obtained in each of the four BH runs in Figure 6. One can see that β strands are only formed for $K = 0.1$ $\mathrm{kcal\,mol^{-1}\mathring{A}^{-2}}$, but not for $K = 0.01$ $\mathrm{kcal\,mol^{-1}\mathring{A}^{-2}}$, $K = 1$ $\mathrm{kcal\,mol^{-1}\mathring{A}^{-2}}$ and $K = 10$ $\mathrm{kcal\,mol^{-1}\mathring{A}^{-2}}$. The number of β contacts for different values of K are plotted in Figure 7. Even though there is no specific value of K for which β strand contacts are preferably created, it can be seen that simulations with $K < 0.01$ $\mathrm{kcal\,mol^{-1}\mathring{A}^{-2}}$ or $K > 1$ $\mathrm{kcal\,mol^{-1}\mathring{A}^{-2}}$ are not useful to enforce β strand contacts. While there are more β residues involved for $K = 1$ $\mathrm{kcal\,mol^{-1}\mathring{A}^{-2}}$, more residues form β strands for $K = 0.1$ $\mathrm{kcal\,mol^{-1}\mathring{A}^{-2}}$.

We do not expect to have a significant influence of the force constant K on the optimal distance restraint d between the C_α atoms of pairing residues in β strands. To test this, we set $K = 0.1$ $\mathrm{kcal\,mol^{-1}\mathring{A}^{-2}}$ as it seems to be a good value for β strand formation and repeated the simulations with different values for d in the interval $d \in [4$ Å, 6 Å$]$. The number of β contacts as a function of d are shown in Figure 7. Most contacts involved in β strands were detected for distance constraints of 4.8 Å and 5.0 Å.

We repeated the simulations, but using the constraints from the BetaPro prediction. The results are shown in Figure 8. Also in case of BetaPro predictions, $K < 0.01$ $\mathrm{kcal\,mol^{-1}\mathring{A}^{-2}}$ or $K > 1$ $\mathrm{kcal\,mol^{-1}\mathring{A}^{-2}}$ are not suitable for enforcing β contacts. There is no evidence for the best distance d for the simulations with BetaPro constraints with protein 1A45.

Based on these findings we decided to use $K = 0.1$ $\mathrm{kcal\,mol^{-1}\mathring{A}^{-2}}$ and $d = 5.0$



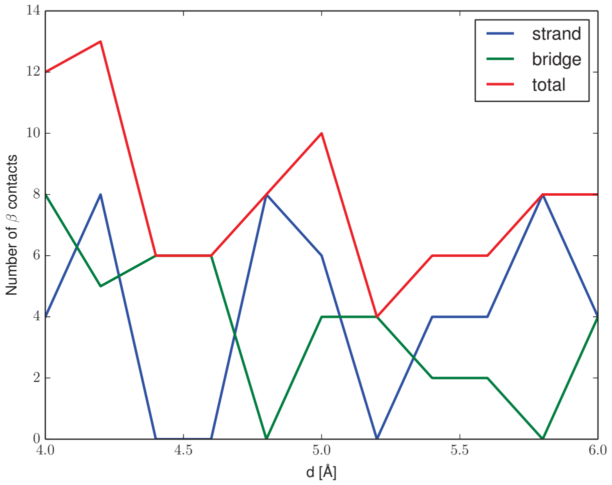


Figure 7: Number of β contacts of the lowest energy structure after 300 steps of BH simulations for different values of K and constant distance constraint d = 5.0 Å (top) and for different values of d and constant value of K = 0.1 kcal mol^{-1}Å^{-2} (bottom). Blue represents the number of residues involved in all β contacts, green the number of residues involved in β strand contacts and red the number of residues involved in β bridges according to DSSP [7]. BCov [8] constraints were used.

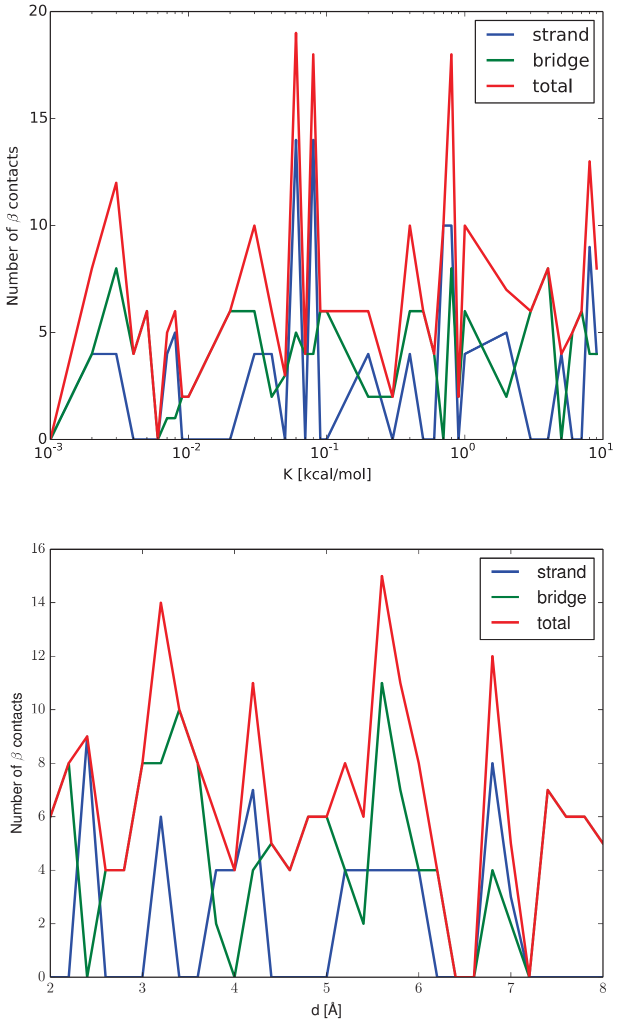


Figure 8: Number of β contacts of the lowest energy structure after 300 steps of BH simulations for different values of K and constant distance constraint $d = 5.0$ Å (top) and for different values of d and constant value of $K = 0.1$ kcal mol^{-1}Å^{-2} (bottom). Blue represents the number of residues involved in all β contacts, green the number of residues involved in β strand contacts and red the number of residues involved in β bridges according to DSSP [7]. BetaPro [2] constraints were used.

Å for all subsequent simulations.

4.2 Comparison of β constraints from BCov and BetaPro

In order to compare the effectiveness of the integration of β constraints in our simulations we performed a simulation with the previously determined values $K = 0.1$ kcal mol^{-1}Å^{-2} and $d = 5.0$ Å for all β constraints predicted by BetaPro [2] and BCov [8] for six different proteins with the PDB [3] codes 1G43, 1IS1, 1L5B, 1NEP, 1OK0 and 1UMI taken from the BetaSheet916 dataset [2]. The structures of the proteins were downloaded from the RCSB protein data bank [3] and used as target structures. All structures have a significant β content. We use the CHARMM22 force field [25, 26] to model the peptides, and the generalized Born model FACTS [27] to describe the aqueous solvent. For the calculation of the nonbonded interactions, the cutoff scheme suggested in the FACTS documentation is employed, i.e., truncation of both long-range electrostatics at 12Å using a shift function and the van der Waals energy with a polynomial switching function applied between 10 and 12Å. We performed 6 ns molecular dynamics (MD) simulations at an elevated temperature of $T = 500$ K using a Langevin thermostat with frictional coefficient 5 ps^{-1} to produce 6 unfolded starting structures per peptide for the subsequent folding simulations. The root mean square deviations (RMSDs) of the C$_\alpha$ atoms between the starting structures and the corresponding target structure are 44.2–88.2 Å for 1G43, 79.2–129.1 Å for 1IS1, 14.4–51.0 Å for 1L5B, 68.0–84.3 Å for 1NEP, 13.2–40.8 Å for 1OK0 and 85.9–102.1 Å for 1UMI. Most of the initial structures are in an extended formation. In Figure 9 the target structure and one representative starting structure are shown for each peptide. We performed three independent simulations with different random start values for all starting structures. The length of the simulation was 100 BH steps. The predictions from BCov and BetaPro were used as β constraints. In this simulations, we stopped each energy minimization after max. 200 minimization steps as our aim is to find the constrained distances fast. We calculated the C_α distance d_{CA} between all residues pairs which are predicted to be in β contact after every BH step. We averaged these distances over all residue pairs, all runs and all initial structures. In order to know how fast the simulation finds a distance d_{CA} close to the constrained distance $d = 5.0$ Å between the residue pairs, we subtracted this value from the averaged distance $\langle d_{CA} \rangle$. A normalization to the value after the first BH step (averaged over all runs) $\langle d_{CA} \rangle_0$ was performed due to different separations in residue pair distances in the predictions of BetaPro [2] and BCov [8]. The results for the six proteins are shown in Figure 10. All simulations except for the BCov predictions of 1L5B and 1UMI find the β contacts within 100 BH steps. The results show that it is not needed to run very long simulations and minimize the structures to very low RMS forces to fulfill the constraints. The comparison of the results for BCov [8] and BetaPro [2] shows that for 1IS1, 1NEP and 1OK0 the β contacts are found at a similar speed while for the other 3 proteins BetaPro outperforms BCov. This difference can be explained with the separation between residues which are predicted to have a β contact. The

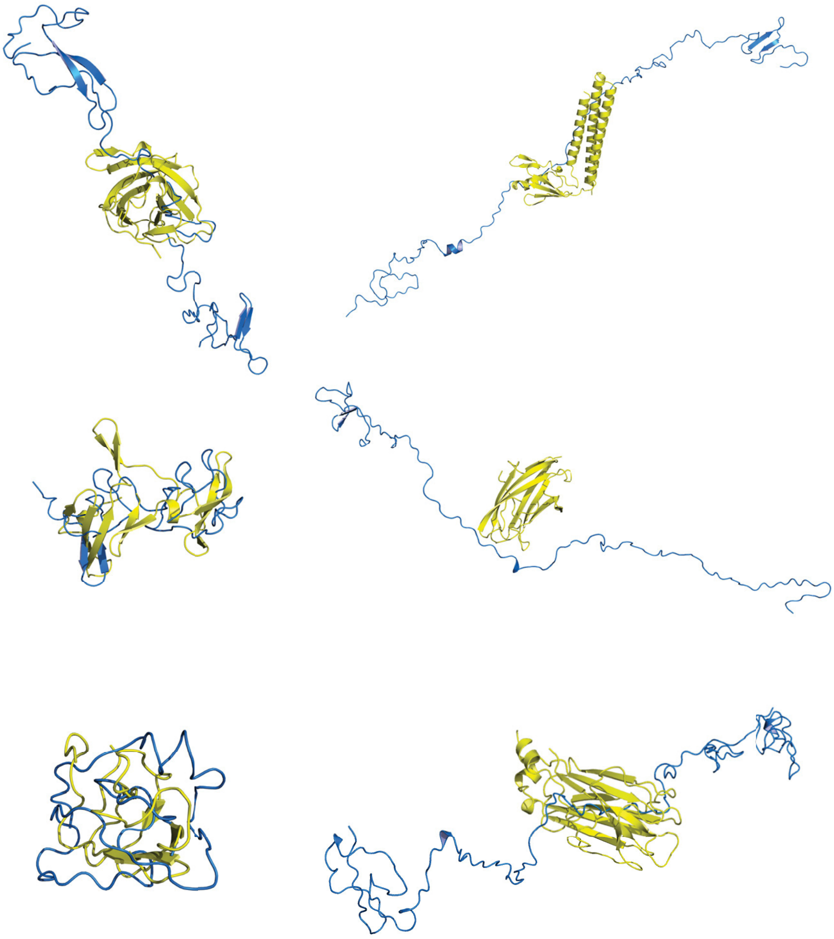


Figure 9: Representative initial structure (blue) and target structure (yellow) for 1G43 (top, left), 1IS1 (top, right), 1L5B (center, left), 1NEP (center, right), 1OK0 (bottom, left) and 1UMI (bottom, right).

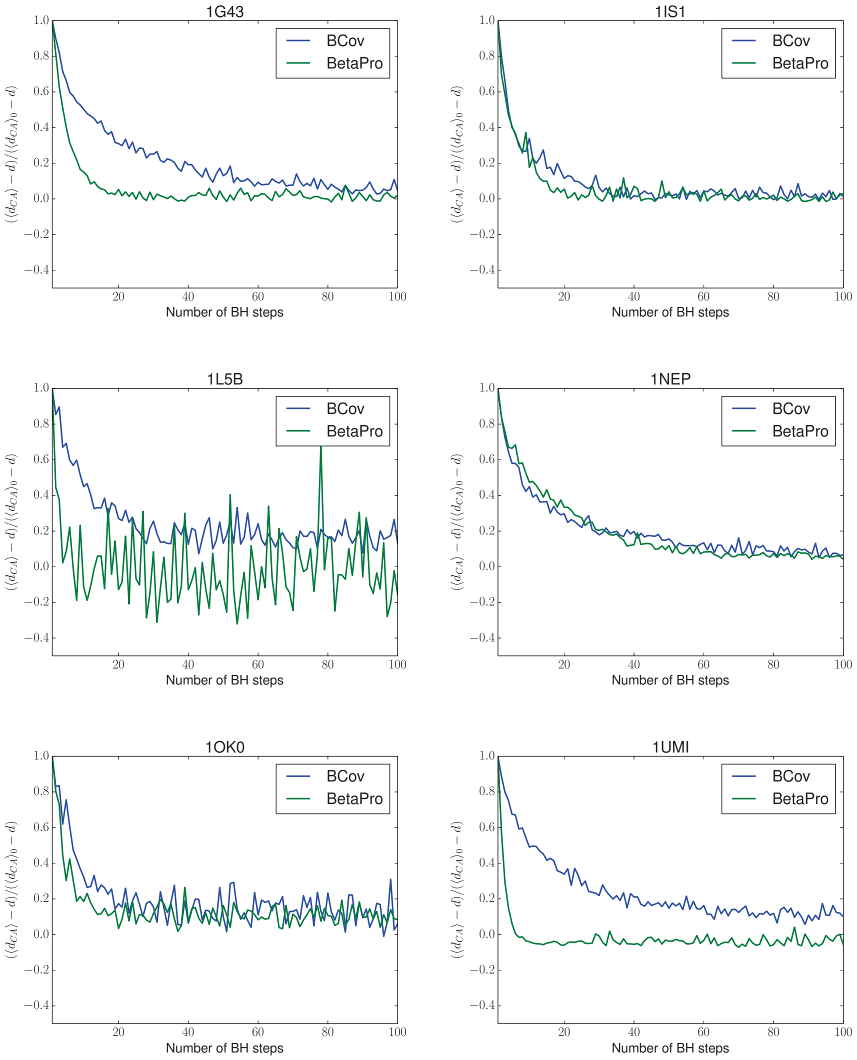


Figure 10: Normalized C_α distance of residue pairs predicted to have a β contact as a function of the number of BH steps for the proteins 1G43 (top, left), 1IS1 (top, right), 1L5B (center, left), 1NEP (center, right), 1OK0 (bottom, left) and 1UMI (bottom, right). The simulations for the predictions from BetaPro [2] and BCov [8] are shown in green and blue, respectively.

Protein	BCov	BetaPro
1G43	26.1	14.2
1IS1	40.5	33.0
1L5B	23.7	8.3
1NEP	26.8	24.2
1OK0	17.3	11.2
1UMI	35.9	10.7

Table 3: Averaged residue number separation between residues in a β contact for the predictions from BCov and BetaPro for the proteins 1G43, 1IS1, 1L5B, 1NEP, 1OK0 and 1UMI.

average separation in terms of residue numbers in the polypeptide chain for the six proteins is shown in table 3. In all cases, BetaPro predicts residue pairs to be in β contact whose residues are closer along the polypeptide chain compared to the prediction by BCov. We also find that the difference in the number of BH steps needed to find the constrained distance is largest for the proteins with the largest difference for the average separations as predicted by BetaPro and BCov (1UMI, 1L5B, 1G43). We conclude that it is easier to establish the β contacts if they are separated by fewer residues. Furthermore, independent of the accuracy of the predictors, BetaPro predicts more short- and medium-range contacts than BCov while BCov predicts more long-range contacts than BetaPro.

5 Methods

5.1 Basin-hopping

In the basin-hopping (BH) approach to global optimization [28–30] moves are proposed by perturbing the current geometry, and are accepted or rejected based upon the energy difference between the local minimum obtained by minimization from the instantaneous configuration and the previous minimum in the chain. In effect the potential energy surface is transformed into the basins of attraction [31, 32] of all the local minima, so that the energy for configuration $\mathbf{r}$ is

$$\widetilde{E}(\mathbf{r}) = \min\{E(\mathbf{r})\},$$

where min denotes minimization. Large steps can be taken to sample this transformed landscape, since the objective is to step between local minima. Furthermore, there is no need to maintain detailed balance when taking steps, because the BH approach attempts to locate the global potential energy minimum and is not intended to sample thermodynamic properties. The BH algorithm has been implemented in the GMIN program [33] and has already been employed to find the global minimum of peptides and peptide complexes in previous work [34–44].

To perturb the current geometry we have the option of taking steps in dihedral angle space for the backbones and side chains of the peptides [36], where we consider dihedral angles defining planar structures, such as rings, as non-twistable in order to maintain the planar geometry [45]. In earlier work, we selected a certain number of the rotatable dihedral angles for the backbone and side chains with different twisting probabilities depending on the position along the peptide chain [36] and twisted them up to a maximum angle, which can be initially set by the user and is normally in the range of 20° to 60°. Recently, we also employed dihedral trial moves to contiguous residues along the chain with the help of secondary structure information and generalized rotation moves [46]. Here we describe a new dihedral angle move set based on Ramachandran plots of amino acids.

5.2 Ramachandran moves

Previously, we employed random dihedral angle moves in basin-hopping simulations. Random moves for the backbone dihedral angles Φ and Ψ have the advantage, that one samples the full range of dihedral angle pair combinations in a relatively small number of MC moves. However, it is well known that specific dihedral angle regions of the Ramachandran plot of proteins are more populated than other regions. We want to use this fact for the Monte Carlo moves and started by creating the Ramachandran plots for every amino acid. We then converted these Ramachandran plots into probability functions $P(\Phi, \Psi)$ for every amino acid, which in turn are used to create dihedral angle moves.

Determination of Ramachandran plots We created Ramachandran plots in a similar way like Chen et al. [47]. We downloaded the structures of the 488

PDB codes used by Chen et al. [47] from the PDB data bank [3]. We recorded the dihedral angle values of every (Φ, Ψ) pair together with their amino acid A. The results for all 20 amino acids are shown in figure 11.

We created 324 quadratic clusters of size 20° × 20° for every amino acid A where the centers of these clusters are located at $(-170° + n \times 20°, -170° + m \times 20°)$ in the Ramachandran plot of the amino acid, and n and m are integers between 0 and 17. For each amino acid, we associated every recorded (Φ, Ψ) dihedral angle pair with its corresponding cluster C_{nm}. We counted the number of associations $AS(C_{nm})$ for every cluster C_{nm}. The probability $P(C_{nm})$ for every cluster $AS(C_{nm})$ was calculated via

$$P(C_{nm}) = \frac{AS(C_{nm})}{\sum_{n=1}^{17} \sum_{m=1}^{17} AS(C_{nm})}$$

to ensure normalization,

$$\int_{\mathbb{R}^2} \mathrm{d}\Phi \mathrm{d}\Psi P(\Phi, \Psi) = 1$$

The probability plots for all 20 amino acids are shown in figure 12.

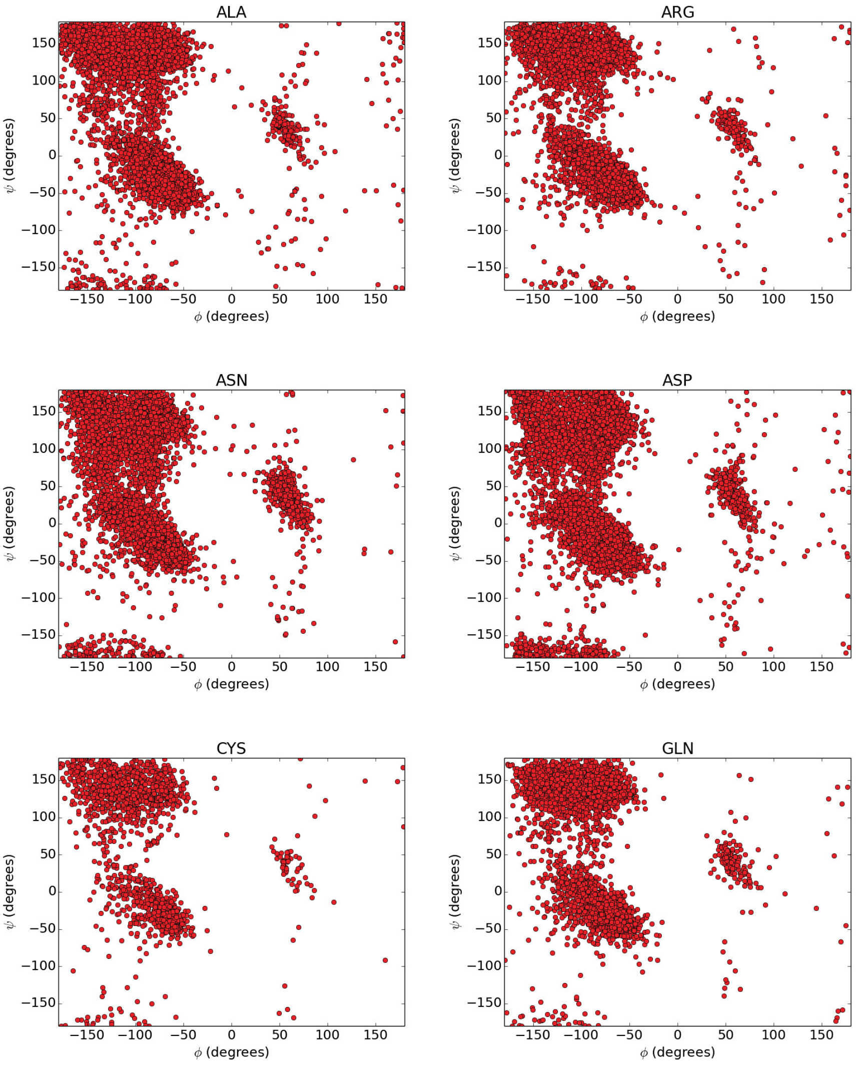
Ramachandran moves In GMIN, dihedral angle moves are performed in two steps:

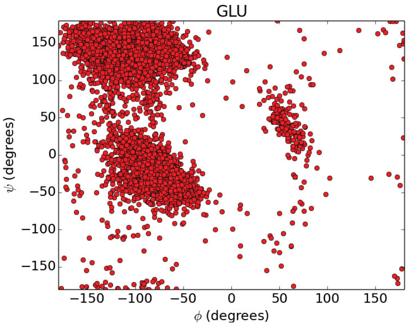
1. Residue choice
2. Dihedral angle choice

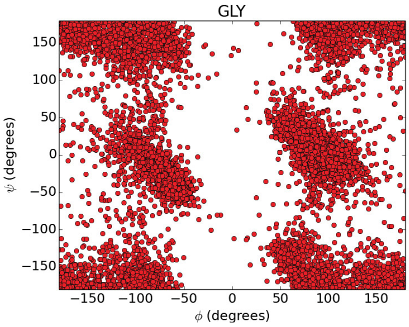
In the original implementation, residues were chosen randomly from the protein sequence [36]. Here, the minimum and maximum number of residues to be chosen depends on the total number of residues in the protein chain. The dihedral angles of a chosen residue are changed with a probability which depends on the position of this residue in the polypetide chain: The maximum probability p_{max} is set to the residues at the termini of the chain and the minimum probability p_{min} for the residues in the center of the chain. A linear increase from the center to the termini was used. Furthermore, the dihedral angles are changed randomly with a user-defined maximum dihedral angle change from the previous value.

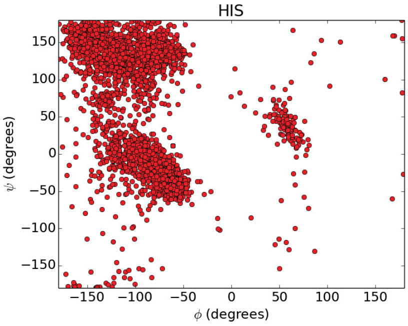
However, the value of the new dihedral angles determines the effectiveness of the dihedral angle moves. In this study, we use the information of the Ramachandran probability plots for changing the dihedral angles to improve the efficiency of basin-hopping. After we have chosen randomly a residue from the polypeptide chain, we randomly select a (Φ_R, Ψ_R) pair with $-180° \leq \Phi_R < 180°$ and $-180° \leq \Psi_R < 180°$. We then multiply the probability $P(\Phi_R, \Psi_R)$ corresponding to the amino-acid specific cluster to which (Φ_R, Ψ_R) belongs to (see above) with a constant and amino acid-independent factor f,

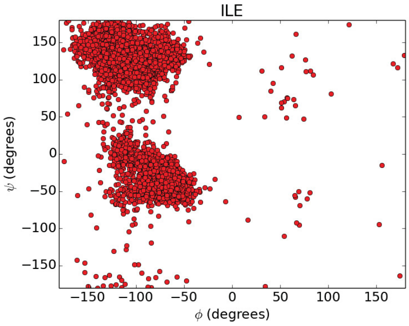
$$P_R = f * P(\Phi_R, \Psi_R)$$

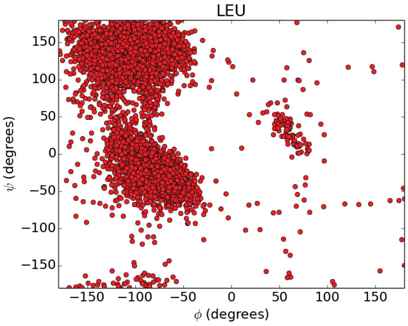

ALA
ARG
ASN
ASP
CYS
GLN
ψ (degrees)
ϕ (degrees)
−150
−100
−50
0
50
100
150

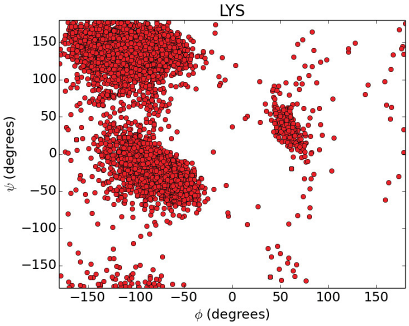

GLU
ψ (degrees)
ϕ (degrees)
150
100
50
0
−50
−100
−150
−150 −100 −50 0 50 100 150

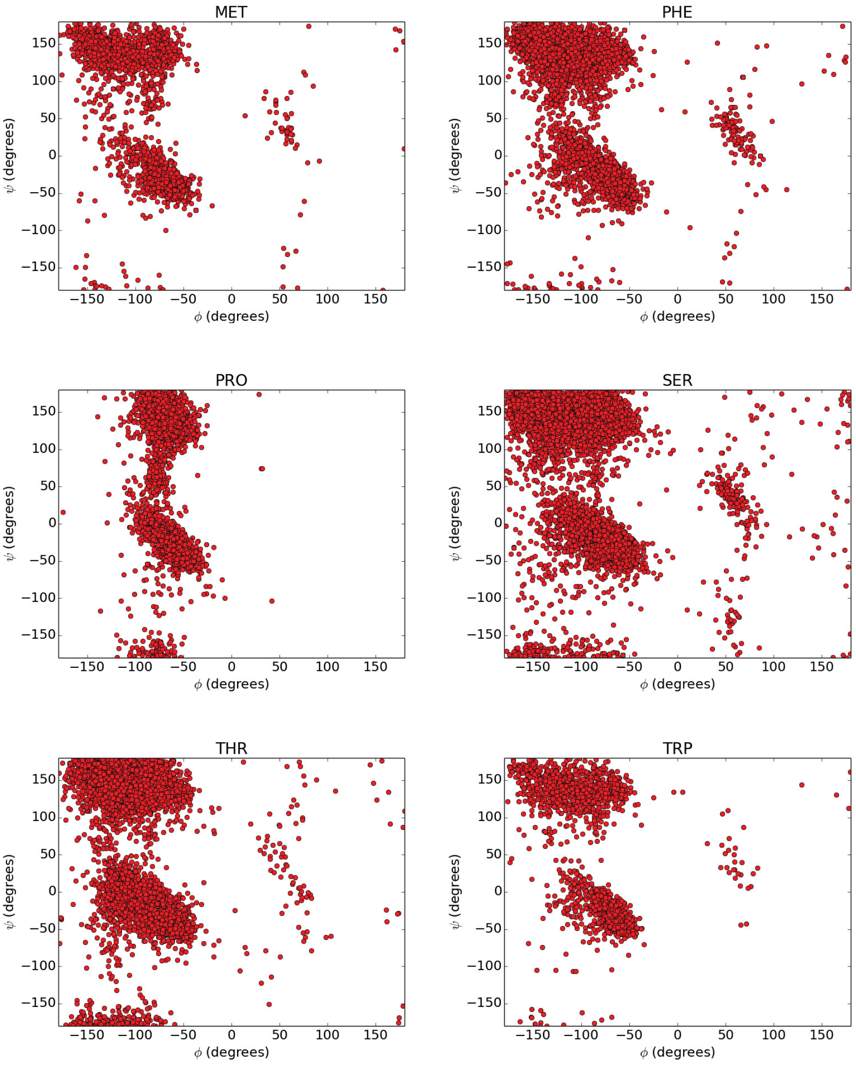

GLY
ψ (degrees)
ϕ (degrees)
150
100
50
0
−50
−100
−150
−150 −100 −50 0 50 100 150


HIS
ψ (degrees)
ϕ (degrees)
150
100
50
0
−50
−100
−150
−150 −100 −50 0 50 100 150


ILE
ψ (degrees)
ϕ (degrees)
150
100
50
0
−50
−100
−150
−150 −100 −50 0 50 100 150


LEU
ψ (degrees)
ϕ (degrees)
150
100
50
0
−50
−100
−150
−150 −100 −50 0 50 100 150


LYS
ψ (degrees)
ϕ (degrees)
150
100
50
0
−50
−100
−150
−150 −100 −50 0 50 100 150


MET
PHE
PRO
SER
THR
TRP
ψ (degrees)
ϕ (degrees)
150
100
50
0
−50
−100
−150
−150
−100
−50
0
50
100
150

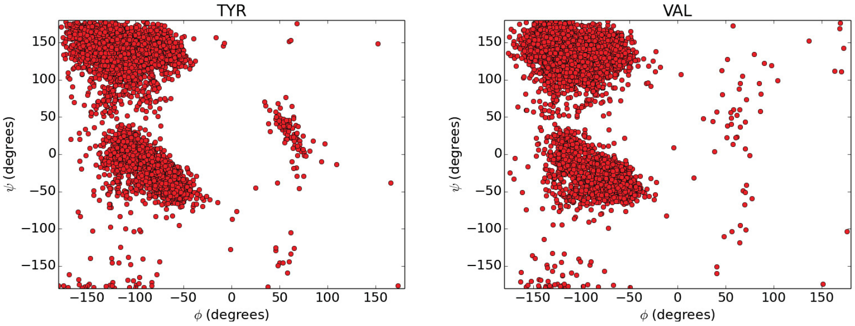
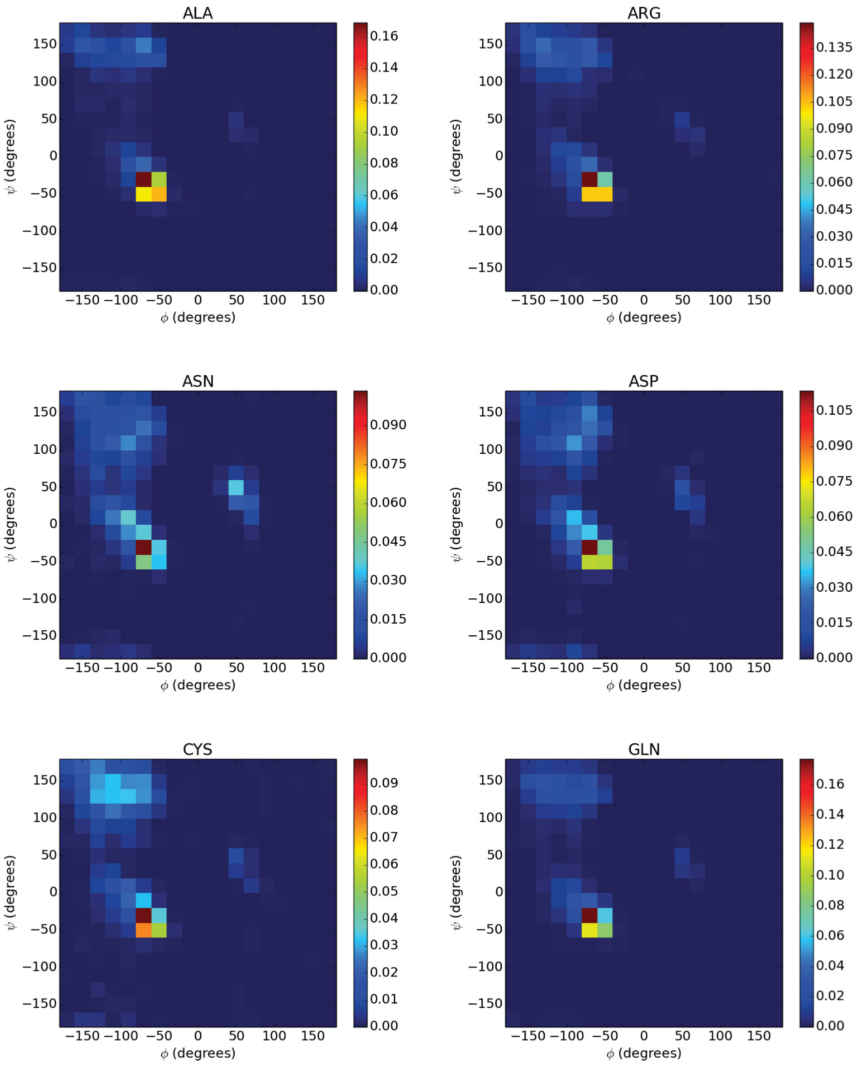


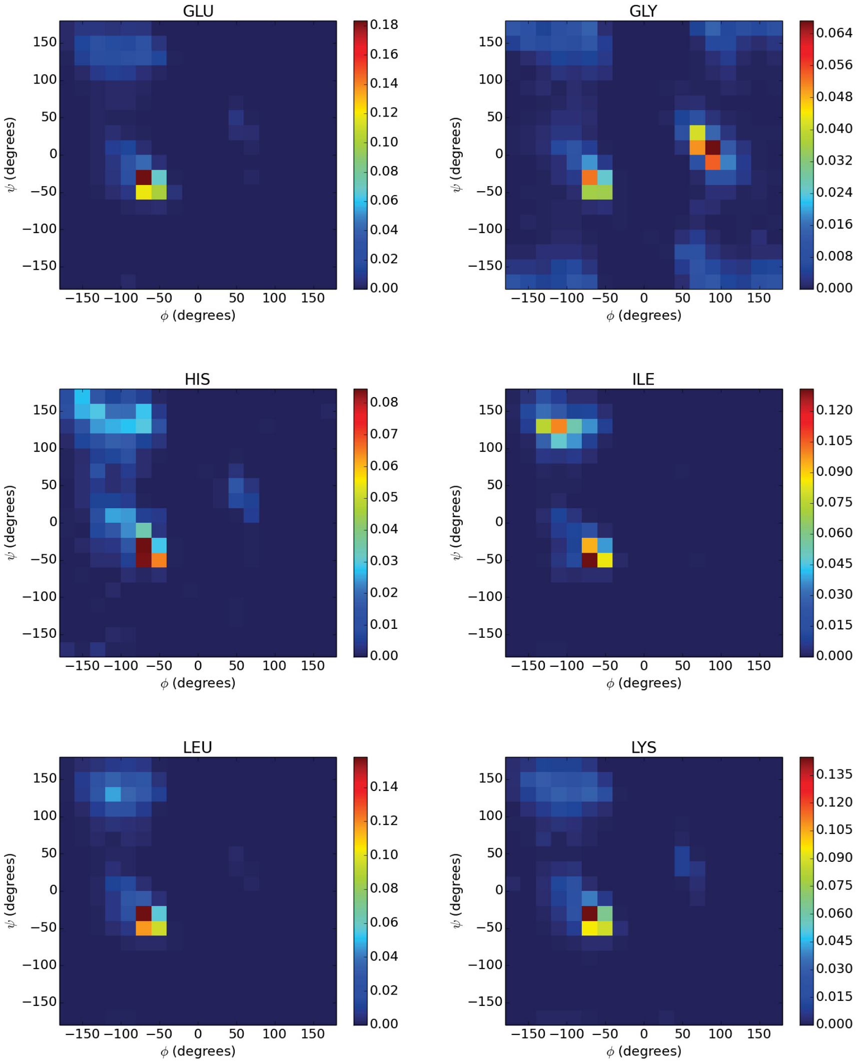
Figure 11: Ramachandran plots for all 20 amino acids. Every dot represents one dihedral angle pair obtained from the analysis of 488 proteins.

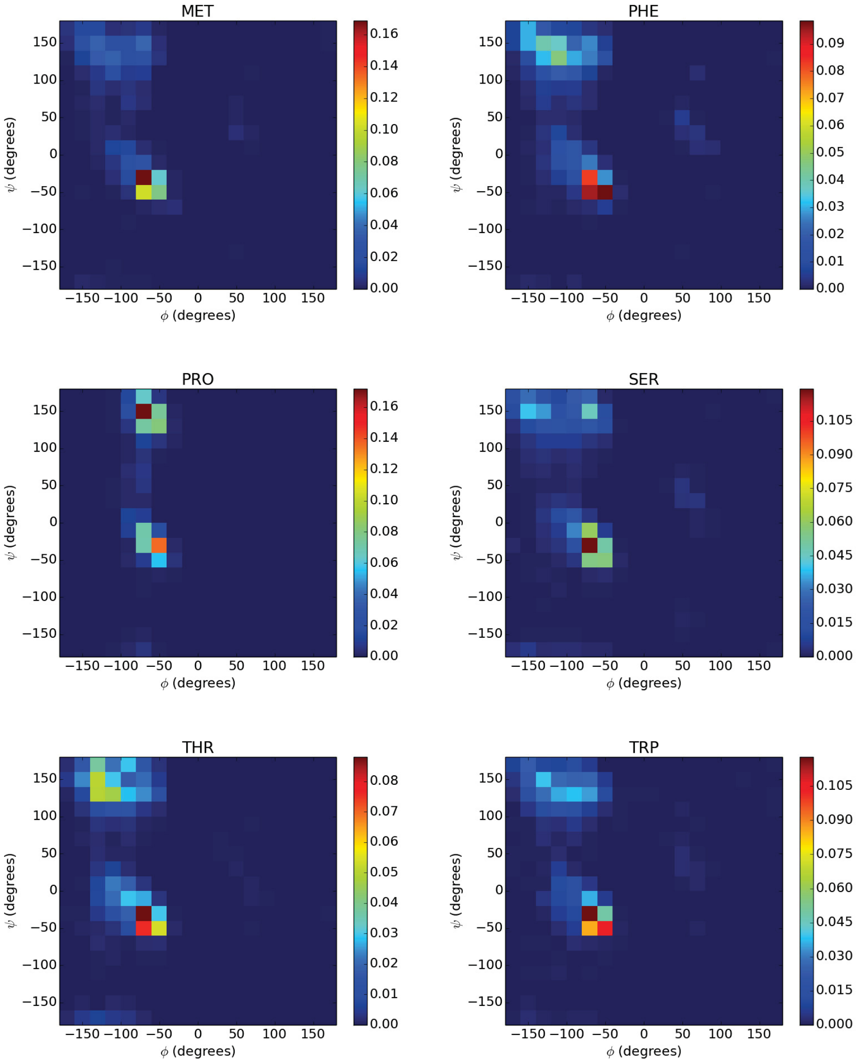
and compare P_R with a fixed acceptance threshold P_A. P_A is a randomly chosen value in the interval $[0, P_{max})$ where P_{max} is a constant value specified at the beginning of the simulation. It reflects the maximum possible probability, which P_R has to exceed in order to accept the dihedral angle move: if $P_R \geq P_A$, then a dihedral angle move is performed. If $P_R < P_A$, a new dihedral angle pair (Φ_R, Ψ_R) is chosen and its weighted probability P_R is compared with another $P_A \in [0, P_{max})$. The procedure is repeated until a pair (Φ_R, Ψ_R) is found with $P_R \geq P_A$.

Once such a pair is found, we compare the weighted probability P_R with the sequence threshold P_S. If $P_R \geq P_S$ we perform a sequence change while in case of $P_R < P_S$ a point change is performed. A point change means that only the dihedral angles of the chosen residues are changed to their new value pair (Φ_R, Ψ_R). If we perform a sequence change, we randomly choose two integers i and j between 0 and 2 and change all residues with indices in the range of $[R - i, R + j]$ where R is the index of the chosen residue in the primary sequence of the protein. The lower and upper bound of this sequence is constrained by the termini of the protein, e.g. if $R < i + 1$, then we change all residues with indices $[1, R + j]$. Similarly, if $R > N - j$ with N as the number of amino acids in the sequence, then we change all residues with indices $[R - i, N]$. Because i and j are randomly chosen, there is a probability of 1/9 that $i = j = 0$ is selected. In this case, the sequence change is identical to a point change. We introduce sequence changes because our first goal is to find secondary structure elements. In secondary structure elements such as α helices or β sheets, the dihedral angles (Φ, Ψ) of consecutive residues adopt only values from rather populated regions of the Ramachandran plot (see figures 11 and 12). Their probabilities are higher than the probabilities of other regions. We use this property by applying sequence changes to the affected dihedral angle pairs (Φ, Ψ) for residues $[R - i, R + j]$ if $P_R \geq P_S$.

The three user-supplied parameters f, P_{max} and P_S are not independent from


ALA
ARG
ASN
ASP
CYS
GLN
ψ (degrees)
ϕ (degrees)
150
100
50
0
−50
−100
−150
−150 −100 −50 0 50 100 150
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.135
0.120
0.105
0.090
0.075
0.060
0.045
0.030
0.015
0.000
0.090
0.075
0.060
0.045
0.030
0.015
0.000
0.105
0.090
0.075
0.060
0.045
0.030
0.015
0.000
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00


GLU
GLY
HIS
ILE
LEU
LYS
ψ (degrees)
ϕ (degrees)
−150 −100 −50 0 50 100 150
150 100 50 0 −50 −100 −150
0.18 0.16 0.14 0.12 0.10 0.08 0.06 0.04 0.02 0.00
0.064 0.056 0.048 0.040 0.032 0.024 0.016 0.008 0.000
0.08 0.07 0.06 0.05 0.04 0.03 0.02 0.01 0.00
0.120 0.105 0.090 0.075 0.060 0.045 0.030 0.015 0.000
0.14 0.12 0.10 0.08 0.06 0.04 0.02 0.00
0.135 0.120 0.105 0.090 0.075 0.060 0.045 0.030 0.015 0.000


MET
PHE
PRO
SER
THR
TRP
ψ (degrees)
ϕ (degrees)
150
100
50
0
−50
−100
−150
−150 −100 −50 0 50 100 150
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0.105
0.090
0.075
0.060
0.045
0.030
0.015
0.000

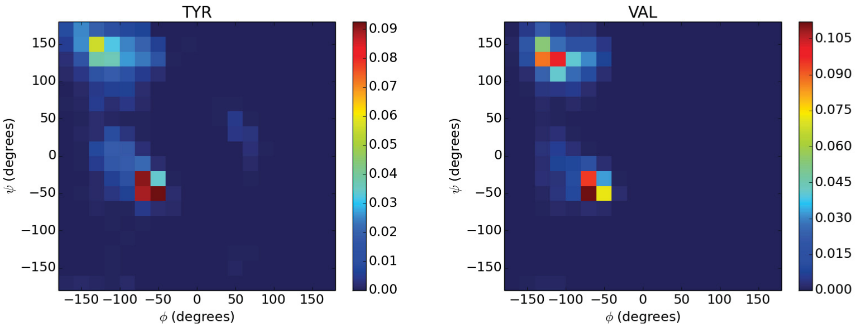


Figure 12: Ramachandran probability plots for all 20 amino acids. The plots are created by clustering the plots of Figure 11 in grids of size 20° × 20° and normalization. The color code represents the probability from 0 (blue) to the maximum value of each amino acid (red).

each other. The two independent parameters are P_{max}/f and P_S/f. The inverse of the first parameter is proportional to the acceptance of the move while the inverse of the second parameter is proportional to the probability of a sequence change. We performed a simulation with 1000 BH steps for the peptide with PDB code 1ERP starting from the fully extended structure with $P_{max} = 1.0$ and $P_S = 0.05$ for different values of f. These values for P_{max} and P_S were chosen because with them secondary structure elements for small proteins could be found with short BH runs. Note that $P_A \in [0, P_{max})$ can be lower than P_S. For different values of f, we counted the number of dihedral angle pair tries (Φ_R, Ψ_R) needed before a dihedral angle move was accepted, i.e. $P_R \geq P_A$. The results are shown in figure 13. One can see that too low values of f increase substantially the number of attempts and thus the computing time needed to find an acceptable dihedral angle pair. We chose to use a lower limit for the value of f which ensures that the average number of dihedral angle attempts is lower than 20. In this case, we get the condition $f \geq 5$.

The choice of the parameter f for constant values of P_{max} and P_S does not only have an influence on the computing time. For $f = 0$ no dihedral angle pair (Φ_R, Ψ_R) will fulfill the condition for a dihedral angle move. As we increase f, different scenarios are possible. To illustrate them, we saved the dihedral angles of the structures after every BH step of the simulation of 1ERP for different values of f. The results are plotted in figure 14. For small values of f, most of the dihedral angle pairs are concentrated in a very small area. The by far most populated area corresponds to the dihedral angles of α helices. If one wants to have MC moves which create preferably α helices, one would use values of f lower than 1 for the given choice of P_{max} and P_S.

An increase in f to values higher than 1 creates more dihedral angle moves from the β sheet region. The corresponding region around $(\Phi, \Psi) = (-105°,$

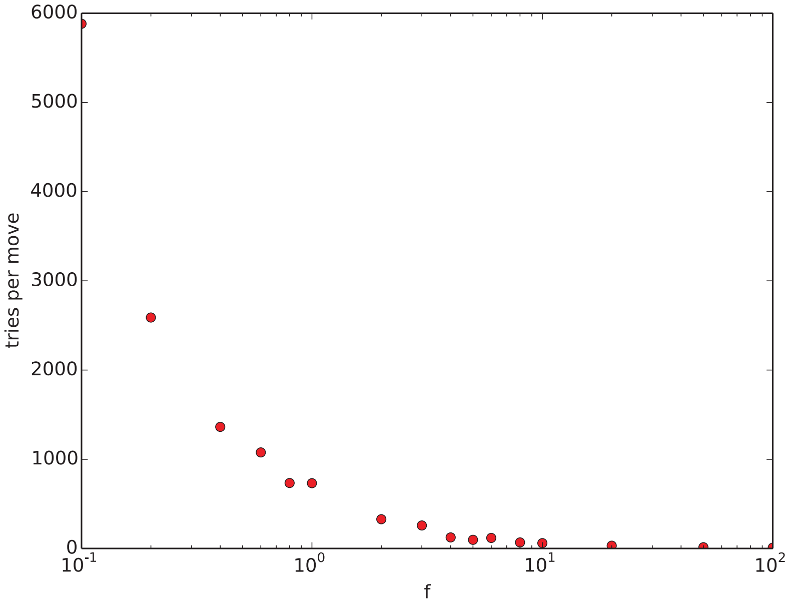
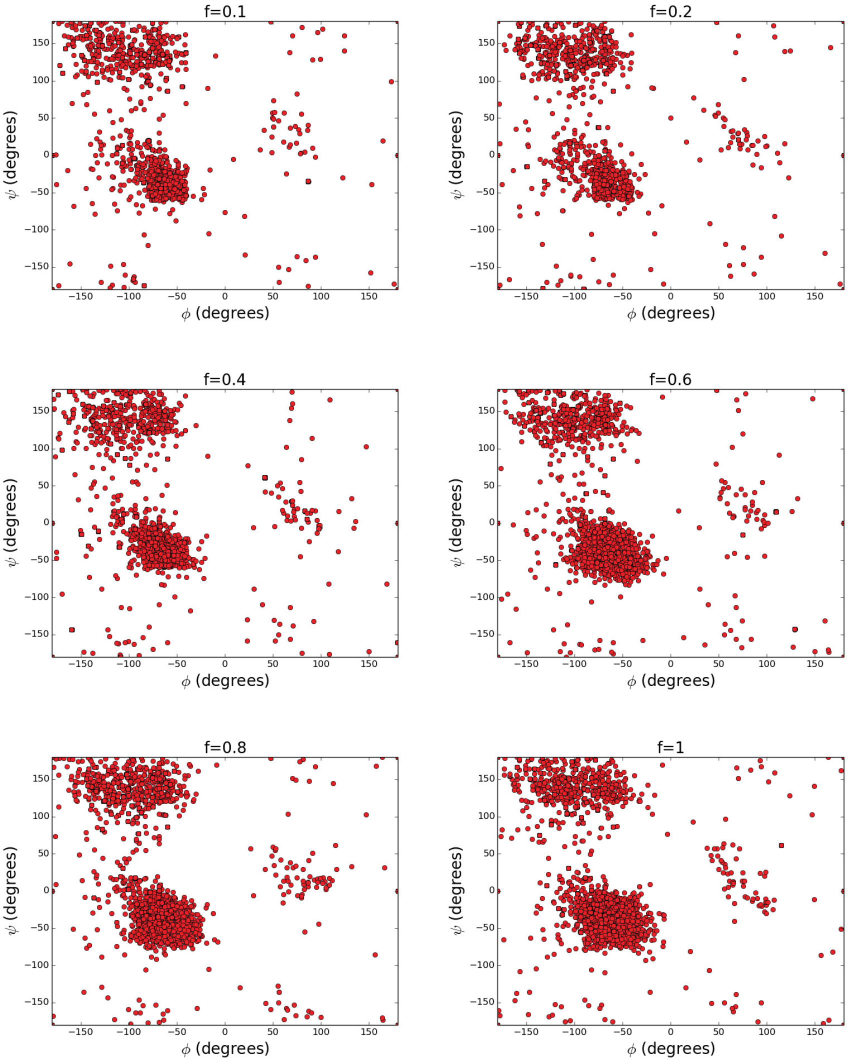
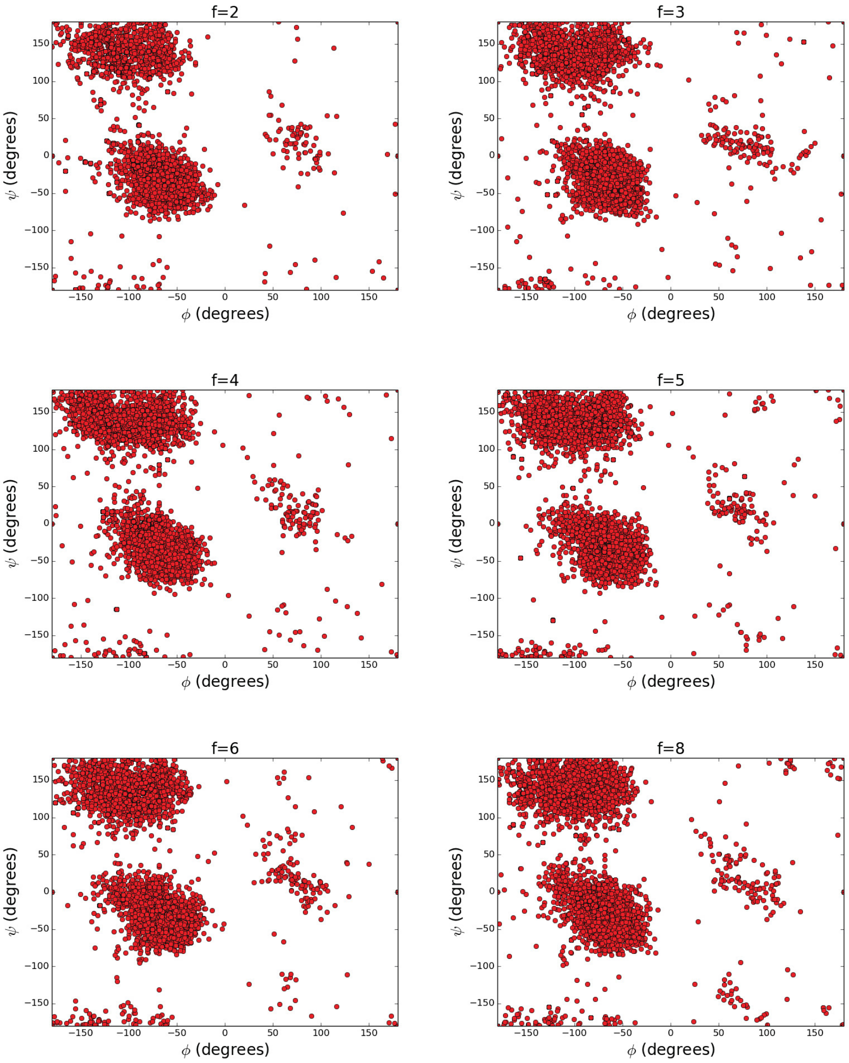


Figure 13: Average number of tries per dihedral angle move which were needed to find a dihedral angle pair whose weighted probability P_R is higher than or equal to the acceptance threshold P_A. Results are shown for different values of f.


f=0.1
f=0.2
f=0.4
f=0.6
f=0.8
f=1
ψ (degrees)
ϕ (degrees)
150
100
50
0
−50
−100
−150
−150
−100
−50
0
50
100
150


f=2
f=3
f=4
f=5
f=6
f=8
ψ (degrees)
ϕ (degrees)
150
100
50
0
−50
−100
−150
−150
−100
−50
0
50
100
150

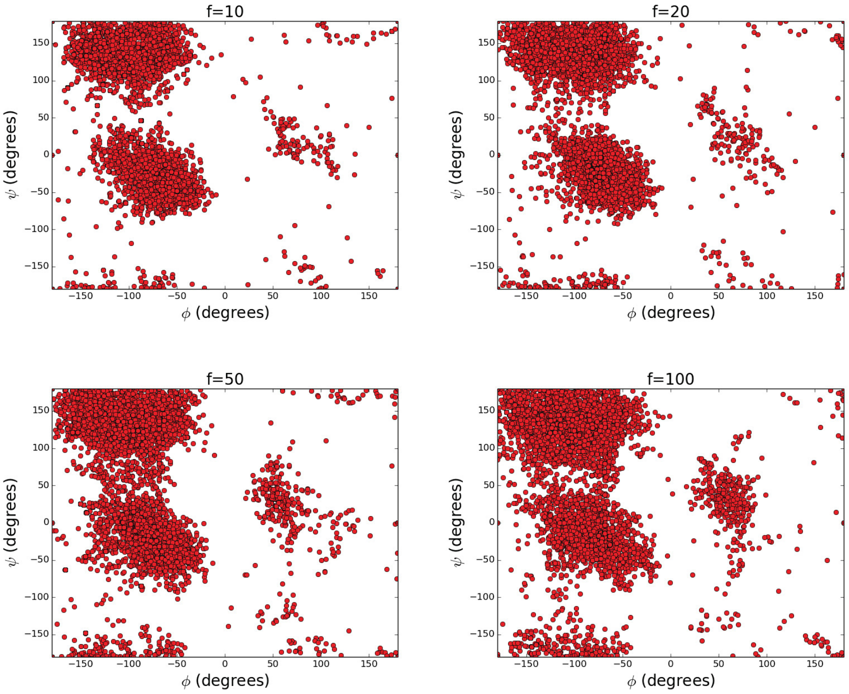


Figure 14: Ramachandran plots obtained from the BH simulations of 1ERP with $P_{max} = 1.0$, $P_S = 0.05$ and different values of f. Every plot shows all dihedral angle pairs of the protein of all 1000 structures, which were saved after each of the 1000 BH steps.

130°) is increasingly populated and becomes wider for increasing values of f. A value of $f = 20$ for the given choice of P_{max} and P_S would be a good choice to have moves for both α helices and β sheets. A further increase of f leads to a population increase of the region of left-handed helices around $(\Phi, \Psi) = (50°, 30°)$. An even further increase of f causes that less likely (Φ, Ψ) pairs get accepted for a dihedral angle move, i.e. more regions of the Ramachandran space become populated (see $f = 100$ in figure 14). It should be noted, that values in figure 14 are after minimization, i.e. these Ramachandran plot will never be totally random because the dihedral angles of the configuration prefer stable regions after minimization.

If the value of P_{max}/f is too high, one would spend too much computing time with unsuccessful attempts to find a dihedral angle pair (Φ_R, Ψ_R) whose probability is high enough to perform a dihedral angle move. On the other hand, if P_{max}/f is too low, one would allow too many dihedral angles from the normally unpopulated regions of the Ramachandran plot which is against our intention to focus on highly populated regions for a fast detection of secondary structure elements. We performed the same simulations like above, but now with constant values $f = 1$ and $P_S = 0.05$ and different values of P_{max}. The results are shown in figure 15. The acceptance values P_A increase with increasing maximum value P_{max}. As the acceptance rate is proportional to f/P_A, more attempts are needed before a dihedral angle change is accepted. With our previous condition that the average number of attempts should not be higher than 20, we get the condition $P_{max} \leq 0.1$. Previously, we found $f \geq 5$ which would give us the condition $\frac{f}{P_{max}} \geq 50$. On the other hand, we want to fix P_{max} and P_S and change the behavior of the moves only with f. Therefore, we need a little bit more flexibility for our choice of parameters and also accept lower values of f/P_{max}. Our general condition for good acceptance is

$$\frac{f}{P_{max}} \geq 10. \tag{1}$$

If this condition is fulfilled one can change the moves to more or fewer sequence moves by changing P_S with constant f and P_{max}, or changing f with constant P_{max} and P_S and condition 1. If P_S/f is too high in comparison to P_A/f (which means $P_S \gg P_A$), we would create just very few or no sequence changes which we need to form secondary structure elements. If $P_S \approx P_A$, we only perform sequence changes which ignores the fact that we also need point changes for the amino acids which are not involved in α helices or β sheets, and are likely to have (Φ, Ψ) values from less populated Ramachandran regions.

In summary, the chosen values for P_{max} and P_S for the test simulations of 1ERP lead to all kinds of dihedral angle moves for $f \in [0.1, 100]$ (figure 14). Condition 1 gives us the requirement that we should focus on the interval $f \in [0.5, 100]$. We are especially interested in dihedral angle moves leading to (Φ, Ψ) values from the α helical or β sheet regions. The corresponding values $f = 0.8$ for dihedral angles from the α helical and $f = 20$ for dihedral angle moves from the β sheet region fulfill this condition. For the presented simulations in section 4, we therefore set $P_{max} = 1.0$ and $P_S = 0.05$. We further chose $f = 20$ because

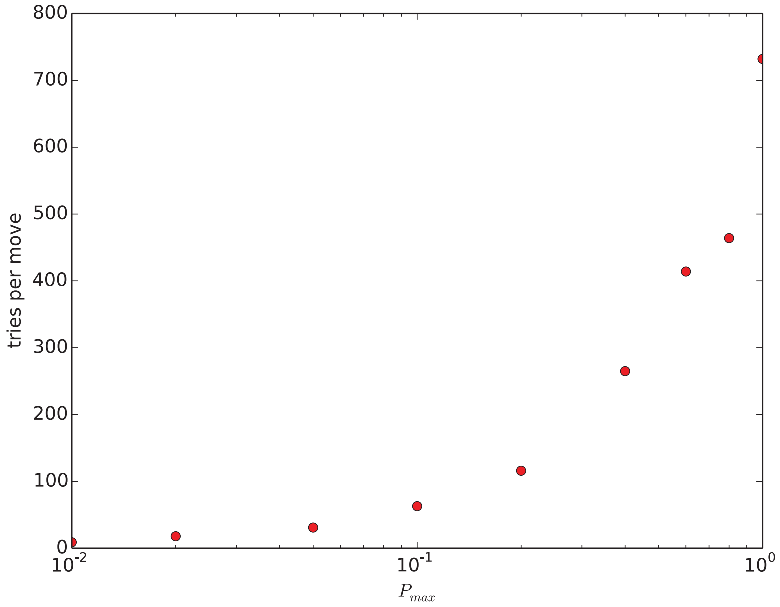


Figure 15: Average number of tries per dihedral angle move which were needed to find a dihedral angle pair whose weighted probability P_R is higher than the acceptance threshold P_A.

our aim is to find the correct alignment of previously constrained β strands.

Comparison of random and Ramachandran moves In order to evaluate our new Ramachandran moves, we performed BH simulations with our new move set and with the previous random dihedral angle move set (see above). We expect to find a native-like structure in both cases for long simulations. However, our focus is on the fast detection of secondary structure elements and structures close to the native structure. We therefore performed short BH simulations with a small protein with PDB [3] code 1YRF. The experimental structure of the protein contains three helices. We generated a totally extended structure of the amino acid sequence of the protein with VMD [48] and started simulations using the Ramachandran move set with previously determined P_{max} and P_S (see above) and three values for f: $f = 0.1$ (less influence of the Ramachandran move set), $f = 0.8$ (α moves) and $f = 100$ (strong preference to regions from the left- and right-handed helical regions and the β sheet region). We executed the same simulation for random moves. We saved the structures after every BH step. Figure 16 shows the RMSD of the structures of the first 15 BH steps to the target structure for all simulations. The simulations with random moves and Ramachandran moves using $f = 100$ even lead to an initial increase of the RMSD in comparison to the extended structure. After this initial BH step, the RMSD decreases in both cases. However, the decrease for the simulation with Ramachandran moves is much larger. For the simulations with the other two values of f, the RMSD decreases from the beginning using the effect of a smart MC move. In effect, the RMSD of the simulation with $f = 0.8$ decreases to 5 Å after 15 BH steps. Two helices are already formed after 7 BH steps for this simulation while the helices are completely missing in the lowest RMSD structure of the simulation with random moves after 1000 BH steps as figure 17 shows. Simulations with other proteins containing α helices show the same advantage in terms of BH steps needed to find the correct helices. However, β strands could not be found significantly faster than with random moves leading to the necessity of the usage of β sheet predictors, which we will further analyze in the future.

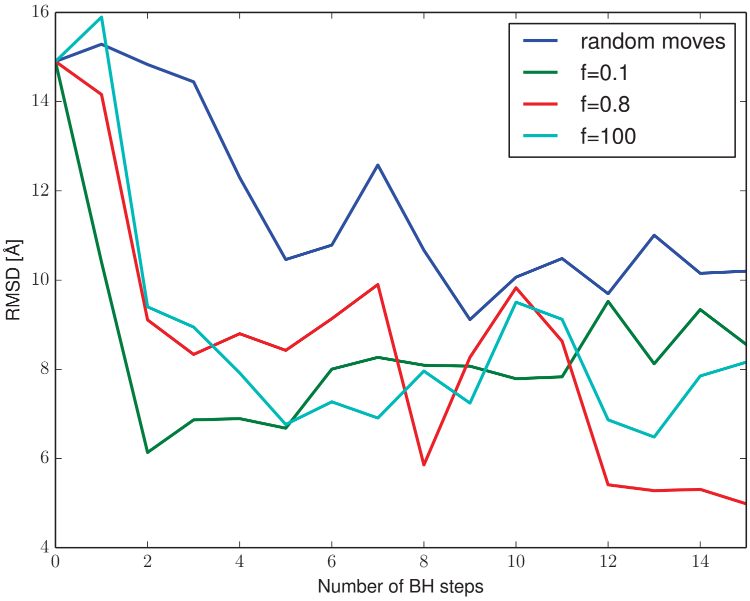
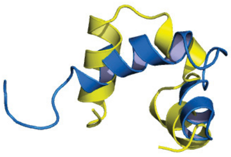


Figure 16: RMSD to the target structure for the structures after every BH step of the simulations of protein with PDB [3] code 1YRF with Ramachandran moves with $f = 0.1$ (green), $f = 0.8$ (red) and $f = 100$ (turquoise) and with random moves (blue). The structure for BH step 0 is the structure after an initial minimization of the protein.



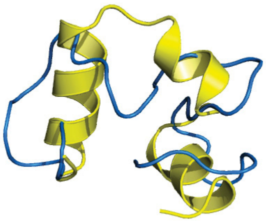


Figure 17: Structure after 7 BH steps for the simulation with Ramachandran moves and $f = 0.8$ (left) and lowest RMSD structure for the simulation with random moves (right) in blue together with the experimental structure in yellow.

6 Conclusion

It is well known, than the choice of Monte Carlo moves has a significant influence on the accuracy and speed of Monte Carlo simulations. The typical MC moves for proteins are dihedral angle moves as dihedral angles describe the structure between neighbouring residues and the side chains. We developed a new dihedral angle move set which is based on the Ramachandran plots of the 20 naturally occurring amino acids. To this end, we counted all dihedral angles of 488 proteins. Based on the statistical distribution of their dihedral angles in the Ramachandran plots of an amino acid, we created a probability function for every amino acid. We combined these probability functions with the basin-hopping approach to global optimization via a new move set. In this move set, we use point and sequence changes to change the dihedral angles of individual and neighbouring residues, respectively. Sequence changes were introduced to enable folding of secondary structure elements like α helices or β sheets in a cooperative manner. Our simulations for the protein 1YRF show that this approach is superior in the ability of forming α helices in few BH steps in comparison to previously used random dihedral angle moves. In order to improve the formation of β sheets, we implemented β contact constraints as obtained from a β sheet predictor. To this end, we compared three different state-of-the-art β sheet predictors. We chose two of them and used their predictions to include structural constraints of β contacts in the basin-hopping approach. We performed short simulations on 6 proteins and found that contacts predicted by BetaPro [2] can be established in fewer BH steps than contacts predicted by BCov [8].

References

[1] John Moult, Jan T. Pedersen, Richard Judson, and Krzysztof Fidelis. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 23(3):ii–iv, 1995.

[2] Jianlin Cheng and Pierre Baldi. Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. *Bioinformatics*, 21(suppl 1):i75–i84, 2005.

[3] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[4] Sven Mika and Burkhard Rost. Uniqueprot: creating representative protein sequence sets. *Nucleic Acids Research*, 31(13):3789–3791, 2003.

[5] B.V. Norledge, R.E. Hay, O.A. Bateman, C. Slingsby, and H.P.C. Driessen. Towards a molecular understanding of phase separation in the lens: a comparison of the x-ray structures of two hightcγ-crystallins, γe and γf, with two lowtcγ-crystallins, γb and γd. *Experimental Eye Research*, 65(5):609 – 630, 1997.

[6] David T Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2):195 – 202, 1999.

[7] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[8] Castrense Savojardo, Piero Fariselli, Pier Luigi Martelli, and Rita Casadio. Bcov: a method for predicting β-sheet topology using sparse inverse covariance estimation and integer programming. *Bioinformatics*, 29(24):3151–3157, 2013.

[9] Nikolas S. Burkoff, Csilla Várnai, and David L. Wild. Predicting protein β-sheet contacts using a maximum entropy-based correlated mutation measure. *Bioinformatics*, 29(5):580–587, 2013.

[10] Luca Miceli, Luigi Palopoli, SimonaE. Rombo, Giorgio Terracina, Giuseppe Tradigo, and Pierangelo Veltri. Experimental evaluation of protein secondary structure predictors. In Gabrielle Allen, Jarosław Nabrzyski, Edward Seidel, GeertDick Albada, Jack Dongarra, and PeterM.A. Sloot, editors, *Computational Science ICCS 2009*, volume 5544 of *Lecture Notes in Computer Science*, pages 848–857. Springer Berlin Heidelberg, 2009.

[11] Falk Hoffmann and Birgit Strodel. Protein structure prediction: assembly of secondary structure elements by basin-hopping. *ChemPhysChem*, 2014. DOI: 10.1002/cphc.201402247.

[12] Claudio Mirabello and Gianluca Pollastri. Porter, paleale 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. *Bioinformatics*, 29(16):2056–2058, 2013.

[13] David T. Jones, Daniel W. A. Buchan, Domenico Cozzetto, and Massimiliano Pontil. Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190, 2012.

[14] Baris E. Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H. Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

[15] Baker laboratory. Hmmer, version 3.1b1. `http://hmmer.org/`, 2013.

[16] T.J.P. Hubbard. Use of β-strand interaction pseudo-potentials in protein structure prediction and modelling. In *System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on*, volume 5, pages 336–344, Jan 1994.

[17] Minoru Asogawa. Beta-sheet prediction using inter-strand residue pairs and refinement with hopfield neural network. In *Ismb,5*, pages 48–51, 1997.

[18] Hongyao Zhu and Werner Braun. Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of β-sheet formation in proteins. *Protein Science*, 8(2):326–342, 1999.

[19] Pierre Baldi, Gianluca Pollastri, Claus A. F. Andersen, and Søren Brunak. Matching protein b-sheet partners by feedforward and recurrent neural networks. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 25–36. AAAI Press, 2000.

[20] Robert E. Steward and Janet M. Thornton. Prediction of strand pairing in antiparallel and parallel β-sheets using information theory. *Proteins: Structure, Function, and Bioinformatics*, 48(2):178–191, 2002.

[21] Marco Lippi and Paolo Frasconi. Prediction of protein β-residue contacts by markov logic networks with grounding-specific weights. *Bioinformatics*, 25(18):2326–2333, 2009.

[22] R. Rajgaria, Y. Wei, and C. A. Floudas. Contact prediction for β and α-β proteins using integer linear optimization and its impact on the first principles 3d structure prediction method astro-fold. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1825–1846, 2010.

[23] Zafer Aydin, Y. Altunbasak, and Hakan Erdogan. Bayesian models and algorithms for protein β-sheet prediction. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 8(2):395–409, March 2011.

[24] Mr. Bayes and Mr. Price. An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s. *Philosophical Transactions*, 53:370–418, 1763.

[25] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, 4:187–217, 1983.

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

[27] U. Haberthür and A. Caflisch. Facts: Fast analytical continuum treatment of solvation. *J. Comput. Chem.*, 29:701–715, 2008.

[28] Z Li and H A Scheraga. Monte carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Sciences*, 84(19):6611–6615, 1987.

[29] David J. Wales and Jonathan P. K. Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116, 1997.

[30] David J. Wales and Harold A. Scheraga. Global optimization of clusters, crystals, and biomolecules. *Science*, 285(5432):1368–1372, 1999.

[31] P. G. Mezey. *Potential Energy Hypersurfaces*. Elsevier, Amsterdam, 1987.

[32] David J. Wales. Basins of attraction for stationary points on a potential-energy surface. *J. Chem. Soc., Faraday Trans.*, 88:653–657, 1992.

[33] David J. Wales. Gmin: A program for basin-hopping global optimization. `http://www-wales.ch.cam.ac.uk/software.html`, 2014.

[34] Philippe Derreumaux. A diffusion process-controlled monte carlo method for finding the global energy minimum of a polypeptide chain. i. formulation and test on a hexadecapeptide. *The Journal of Chemical Physics*, 106(12), 1997.

[35] Philippe Derreumaux. Folding a 20 amino acid $\alpha\beta$ peptide with the diffusion process-controlled monte carlo method. *The Journal of Chemical Physics*, 107(6), 1997.

[36] Paul N. Mortenson and David J. Wales. Energy landscapes, global optimization and dynamics of the polyalanine ac(ala)8nhme. *The Journal of Chemical Physics*, 114(14), 2001.

[37] Paul N. Mortenson, David A. Evans, and David J. Wales. Energy landscapes of model polyalanines. *The Journal of Chemical Physics*, 117(3), 2002.

[38] Joanne M. Carr and David J. Wales. Global optimization and folding pathways of selected α-helical proteins. *The Journal of Chemical Physics*, 123(23):–, 2005.

[39] A. Verma, A. Schug, K. H. Lee, and W. Wenzel. Basin hopping simulations for all-atom protein folding. *The Journal of Chemical Physics*, 124(4):–, 2006.

[40] Birgit Strodel and David J. Wales. Implicit solvent models and the energy landscape for aggregation of the amyloidogenic kffe peptide. *Journal of Chemical Theory and Computation*, 4(4):657–672, 2008.

[41] Birgit Strodel, Jason W. L. Lee, Christopher S. Whittleston, and David J. Wales. Transmembrane structures for alzheimer's aβ1-42 oligomers. *Journal of the American Chemical Society*, 132(38):13300–13312, 2010.

[42] Olujide O. Olubiyi and Birgit Strodel. Structures of the amyloid β-peptides aβ1-40 and aβ1-42 as influenced by ph and a d-peptide. *The Journal of Physical Chemistry B*, 116(10):3280–3291, 2012.

[43] Falk Hoffmann and Birgit Strodel. Protein structure prediction using global optimization by basin-hopping with nmr shift restraints. *The Journal of Chemical Physics*, 138(2):–, 2013.

[44] Mark T. Oakley and Roy L. Johnston. Energy landscapes and global optimization of self-assembling cyclic peptides. *Journal of Chemical Theory and Computation*, 0(0):null, 0.

[45] Marianne S. Bauer, Birgit Strodel, Szilard N. Fejer, Elena F. Koslover, and David J. Wales. Interpolation schemes for peptide rearrangements. *The Journal of Chemical Physics*, 132(5):–, 2010.

[46] Kenji Mochizuki, Chris S. Whittleston, Sandeep Somani, Halim Kusumaatmaja, and David J. Wales. A conformational factorisation approach for estimating the binding free energies of macromolecules. *Phys. Chem. Chem. Phys.*, 16:2842–2853, 2014.

[47] William W. Chen, Jae Shick Yang, and Eugene I. Shakhnovich. A knowledge-based move set for protein folding. *Proteins: Structure, Function, and Bioinformatics*, 66(3):682–688, 2007.

[48] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

Chapter 5

Conclusions

Different methods have been used to predict the structures of proteins. Experimentally, more than 90,000 protein structures have been solved and deposited in the RCSB protein data bank [28]. With increasing computer resources the search of protein structures moves increasingly towards computer simulations. Molecular dynamics and Monte Carlo simulations are the most used simulation methods for modelling biomolecules. Molecular dynamics simulations can, in principle, find the full folding pathway of a protein. However, molecular dynamics simulations of big systems require a lot of computing resources and time. Monte Carlo simulations can speed up the simulations. Here, trial moves are used and accepted or rejected based on a certain criterion. Under the assumption that the lowest energy in the energy landscape of a protein can be associated with the native structure of this protein, the energy of a protein can be used as the criterion. The search for the lowest energy in a high-dimensional energy landscape can be performed by global optimization methods. The basin-hopping approach to global optimization [55–57] transforms the energy landscape of the protein into basins of attraction which allows an effective search, overcoming high energy barriers. Basin-hopping methods have been applied to find the global minimum of peptides and peptide complexes in previous work [49, 61–69]. In this thesis, we presented three extensions to the basin-hopping approach to global optimization, making it more efficient for its application to proteins.

In our first study, we included chemical shifts as structural restraints in the basin-hopping method. Chemical shifts are calculated at every minimization step with the chemical shift predictor CAMSHIFT [44]. The chemical shift difference between the calculated and the target chemical shift is converted into an energy via a penalty function, which forces the calculated chemical shifts in the direction of their target values. Simulations for the peptides 1LE0, 1L2Y and 1YRF have been performed. There is a clear improvement in terms of computational time for the prediction of all three protein structures for the simulation with chemical shift restraints in comparison with simulations without chemical shift restraints. We determined the optimal values for the two adjustable parameters combining CAMSHIFT and the CHARMM force field with the FACTS implicit solvent model. We could detect the structures of all three proteins with an RMSD of lower than 3 Å if we exclude the flexible termini whose chemical shifts are not calculated by CAMSHIFT. We performed simulations with incomplete chemical shift assignments and proved that just one type of chemical shift assignment is needed to detect the correct secondary structure elements. We can conclude that chemical shifts help significantly to detect the structure of a protein. In the future, the approach can be extended to use secondary structure information with chemical shift restraints as it is known that some types of chemical shifts show a specific upward or downward deviation from their random coil values depending on their presence in an α helix or β sheet. The simple approach allows an easy extension to other structural restraints and for investigations with experimental NMR studies.

The efficiency of Monte Carlo simulations depends on the trial moves for the generation of new structures. The standard Monte Carlo moves for proteins are dihedral angle moves. In our second study, we tested different kinds of Monte Carlo moves. We developed a secondary-to-tertiary basin-hopping approach which successfully and reliably predicts the three-dimensional structure of proteins. The approach starts with the prediction of the secondary structure of the protein. To this end, we compared three secondary structure predictors on their accuracy to predict α helices and β sheets correctly. We have demonstrated, that PORTER [20] outperforms the other secondary structure predictors by far. We used the predictions of PORTER [20] to fix the dihedral angles of the amino acids which were predicted to be involved in α helices and β sheets. Furthermore, MC moves were only applied to the intervening residues. We performed simulations for the proteins 1LE0, 1L2Y and 1ERP and random dihedral angle moves which were limited to a user-defined maximum dihedral angle change. For all proteins, structures with an RMSD of lower than 2 Å could be found. The result shows that the secondary-to-tertiary approach works on timescales lower than comparable MD simulations and with less Monte Carlo steps than in comparable studies. We also have shown that a maximum dihedral angle change of 30° needs more time to fold a protein than simulations with a maximum dihedral angle change of 60° or 90°. We refined the tertiary contacts in a second run by releasing the secondary structure constraints. Here, we applied different frequency schemes of backbone to side chain dihedral angle moves. We could show that this second Monte Carlo move set improved the structures of all tested proteins. For example, the Cα RMSD of the lowest-energy structure of 1LE0 could be decreased by a factor 2. The comparison of the different frequency schemes revealed that both backbone and side chain dihedral angle moves are needed to refine the structure. We could also fold larger proteins with α helices while further developments are needed for proteins with β sheets.

Knowledge-based moves can improve the Monte Carlo search. To this end, we developed new dihedral angle moves for proteins. The moves are based on the statistical distribution of the dihedral angles in the two dimensional dihedral angle space. We analyzed all dihedral angles of 488 proteins from the RCSB protein data bank [28] and created a Ramachandran plot for each amino acid. We clustered these Ramachandran plots and converted the results into an amino acid specific probability function. We created a new move set which selects dihedral angles based on this probability functions. Furthermore, we introduced new sequence moves which enable to establish secondary structure elements faster than with previous methods. We demonstrated that with this move set, proteins with α helices can be folded within less than 100 BH steps. However, for proteins with β sheets further method improvements are necessary. Therefor, we included the information from β sheet predictors as structural constraints. First, we compared the state-of-the-art β sheet predictors BetaPro [72] and BCov [73] and tested how efficient β contacts get established as a result of the constraints from these β sheet predictions. Simulations of small peptides shew that constraints from BetaPro [72] are more efficient than constraints from BCov [73]. Our goal is to fold larger proteins with the help of the implemented Ramachandran dihedral angle moves and β constraints. The approach is expected to improve all basin-hopping simulations of proteins using dihedral angle Monte Carlo moves. In summary, in this thesis it was demonstrated that structural restraints and knowledge-based Monte Carlo moves can remarkably improve the efficiency of the basin-hopping approach to global optimization. We plan to combine the three methods introduced in this work to a protein folding package which is then expected to become an important competitor in the Critical Assessment of Techniques for Protein Structure Prediction [8], the state-of-the-art test for protein structure prediction.

Bibliography

[1] Jean Bréfort. Gnome chemistry utils. `http://gchemutils.nongnu.org/`, 2010.

[2] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, 2004.

[3] Robert Murray. *Harper's illustrated biochemistry*. Lange Medical Books/McGraw-Hill, New York, 2006.

[4] Guido Hartmann. The structure and action of proteins. von r. e. dickerson und i. geis. harper and row, publishers, new york-evanston-london 1969. 1. aufl., viii, 120 s., zahlr. abb., paperback dm 20.50. *Angewandte Chemie*, 82(18):780–780, 1970.

[5] Donald Voet. *Biochemistry*. J. Wiley and Sons, Hoboken, NJ, 2004.

[6] Valerie Daggett. α sheet the toxic conformer in amyloid diseases. *Accounts of Chemical Research*, 39(9):594–602, 2006. PMID: 16981675.

[7] Roger S. Armen, Mari L. DeMarco, Darwin O. V. Alonso, and Valerie Daggett. Pauling and corey's α-pleated sheet structure may define the prefibrillar amyloidogenic intermediate in amyloid disease. *Proceedings of the National Academy of Sciences of the United States of America*, 101(32):11622–11627, 2004.

[8] John Moult, Jan T. Pedersen, Richard Judson, and Krzysztof Fidelis. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 23(3):ii–iv, 1995.

[9] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[10] Dmitrij Frishman and Patrick Argos. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, 23(4):566–579, 1995.

[11] Frederic M. Richards and Craig E. Kundrot. Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure*. *Proteins: Structure, Function, and Bioinformatics*, 3(2):71–84, 1988.

[12] Anthony V. Guzzo. The influence of amino acid sequence on protein structure. *Biophysical Journal*, 5(6):809 – 822, 1965.

[13] J.W. Prothero. Correlation between the distribution of amino acids and α helices. *Biophysical Journal*, 6(3):367 – 370, 1966.

[14] Marianne Schiffer and Allen B. Edmundson. Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophysical Journal*, 7(2):121 – 135, 1967.

[15] D. Kotelchuck and H. A. Scheraga. The influence of short-range interactions on protein conformation, ii. a model for predicting the α-helical regions of proteins. *Proceedings of the National Academy of Sciences*, 62(1):14–21, 1969.

[16] P. N. Lewis, N. G[unk]o, M. G[unk]o, D. Kotelchuck, and H. A. Scheraga. Helix probability profiles of denatured proteins and their correlation with native structures. *Proceedings of the National Academy of Sciences*, 65(4):810–815, 1970.

[17] Peter Y. Chou and Gerald D. Fasman. Prediction of protein conformation. *Biochemistry*, 13(2):222–245, 1974.

[18] J. Garnier, D.J. Osguthorpe, and B. Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*, 120(1):97 – 120, 1978.

[19] David T Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2):195 – 202, 1999.

[20] Gianluca Pollastri and Aoife McLysaght. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, 21(8):1719–1720, 2005.

[21] Tho Hoan Pham, Kenji Satou, and Tu Bao Ho. Support vector machines for prediction and analysis of β and γ-turns in proteins. *Journal of Bioinformatics and Computational Biology*, 03(02):343–358, 2005.

[22] Qidong Zhang, Sukjoon Yoon, and William J. Welsh. Improved method for predicting β-turn using support vector machine. *Bioinformatics*, 21(10):2370–2374, 2005.

[23] Ofer Dor and Yaoqi Zhou. Achieving 80structure prediction by large-scale training. *Proteins: Structure, Function, and Bioinformatics*, 66(4):838–845, 2007.

[24] Claudio Mirabello and Gianluca Pollastri. Porter, paleale 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. *Bioinformatics*, 29(16):2056–2058, 2013.

[25] Debora S. Marks, Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3d structure computed from evolutionary sequence variation. *PLoS ONE*, 6(12):e28766, 12 2011.

[26] Pande laboratory. Folding@home. `http://folding.stanford.edu/`, 2013.

[27] Baker laboratory. Rosetta@home. `http://boinc.bakerlab.org/`, 2013.

[28] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[29] K Wuthrich. Protein structure determination in solution by nuclear magnetic resonance spectroscopy. *Science*, 243(4887):45–50, 1989.

[30] F. Castellani, B. van Rossum, A. Diehl, M. Schubert, K. Rehbein, and H. Oschkinat. Structure of a protein determined by solid-state magic-angle-spinning nmr spectroscopy. *Nature*, 420(1):98–102, 2002.

[31] Gordon S. Rule and T. Kevin Hitchens. *Fundamentals of Protein NMR Spectroscopy*. Springer, Dordrecht, 2006.

[32] D.S. Wishart, B.D. Sykes, and F.M. Richards. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *Journal of Molecular Biology*, 222(2):311 – 333, 1991.

[33] Annalisa Pastore and Vladimir Saudek. The relationship between chemical shift and secondary structure in proteins. *Journal of Magnetic Resonance (1969)*, 90(1):165 – 176, 1990.

[34] Gerhard Wagner, Arthur Pardi, and Kurt Wuethrich. Hydrogen bond length and proton nmr chemical shifts in proteins. *Journal of the American Chemical Society*, 105(18):5948–5949, 1983.

[35] DavidA. Case. Calibration of ring-current effects in proteins and nucleic acids. *Journal of Biomolecular NMR*, 6(4):341–346, 1995.

[36] Silvia Spera and Ad Bax. Empirical correlation between protein backbone conformation and cα and cβ 13c nuclear magnetic resonance chemical shifts. *Journal of the American Chemical Society*, 113(14):5490–5492, 1991.

[37] Xiao-Ping Xu and David A. Case. Probing multiple effects on 15n, 13cα, 13cβ, and 13c chemical shifts in peptides using density functional theory. *Biopolymers*, 65(6):408–423, 2002.

[38] Yang Shen, Oliver Lange, Frank Delaglio, Paolo Rossi, James M. Aramini, Gaohua Liu, Alexander Eletsky, Yibing Wu, Kiran K. Singarapu, Alexander Lemak, Alexandr Ignatchenko, Cheryl H. Arrowsmith, Thomas Szyperski, Gaetano T. Montelione, David Baker, and Ad Bax. Consistent blind protein structure generation from nmr chemical shift data. *Proceedings of the National Academy of Sciences*, 105(12):4685–4690, 2008.

[39] Andrea Frank, Ionut Onila, Heiko M. Möller, and Thomas E. Exner. Toward the quantum chemical calculation of nuclear magnetic resonance chemical shifts of proteins. *Proteins: Structure, Function, and Bioinformatics*, 79(7):2189–2202, 2011.

[40] Maria Giovanna Chini, Catharine R. Jones, Angela Zampella, Maria Valeria D'Auria, Barbara Renga, Stefano Fiorucci, Craig P. Butts, and Giuseppe Bifulco. Quantitative nmr-derived interproton distances combined with quantum mechanical calculations of 13c chemical shifts in the stereochemical determination of conicasterol f, a nuclear receptor ligand from theonella swinhoei. *The Journal of Organic Chemistry*, 77(3):1489–1496, 2012.

[41] Andrea Cavalli, Xavier Salvatella, Christopher M. Dobson, and Michele Vendruscolo. Protein structure determination from nmr chemical shifts. *Proceedings of the National Academy of Sciences*, 104(23):9615–9620, 2007.

[42] Yang Shen and Ad Bax. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *Journal of Biomolecular NMR*, 38(4):289–302, 2007.

[43] Stephen Neal, AlexM. Nip, Haiyan Zhang, and DavidS. Wishart. Rapid and accurate calculation of protein 1h, 13c and 15n chemical shifts. *Journal of Biomolecular NMR*, 26(3):215–240, 2003.

[44] Kai J. Kohlhoff, Paul Robustelli, Andrea Cavalli, Xavier Salvatella, and Michele Vendruscolo. Fast and accurate predictions of protein nmr chemical shifts from interatomic distances. *Journal of the American Chemical Society*, 131(39):13894–13895, 2009. PMID: 19739624.

[45] Yang Shen, Frank Delaglio, Gabriel Cornilescu, and Ad Bax. Talos+: a hybrid method for predicting protein backbone torsion angles from nmr chemical shifts. *Journal of Biomolecular NMR*, 44(4):213–223, 2009.

[46] Yang Shen and Ad Bax. Protein backbone and sidechain torsion angles predicted from nmr chemical shifts using artificial neural networks. *Journal of Biomolecular NMR*, 56(3):227–241, 2013.

[47] Beomsoo Han, Yifeng Liu, SimonW. Ginzinger, and DavidS. Wishart. Shiftx2: significantly improved protein chemical shift prediction. *Journal of Biomolecular NMR*, 50(1):43–57, 2011.

[48] Paul Robustelli, Kai Kohlhoff, Andrea Cavalli, and Michele Vendruscolo. Using {NMR} chemical shifts as structural restraints in molecular dynamics simulations of proteins. *Structure*, 18(8):923 – 933, 2010.

[49] Falk Hoffmann and Birgit Strodel. Protein structure prediction using global optimization by basin-hopping with nmr shift restraints. *The Journal of Chemical Physics*, 138(2):–, 2013.

[50] Christian B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[51] David Applegate. *The traveling salesman problem : a computational study.* Princeton University Press, Princeton, 2006.

[52] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[53] E. Marinari and G. Parisi. Simulated tempering: A new monte carlo scheme. *EPL (Europhysics Letters)*, 19(6):451, 1992.

[54] W. Wenzel and K. Hamacher. Stochastic tunneling approach for global minimization of complex potential energy landscapes. *Phys. Rev. Lett.*, 82:3003–3007, Apr 1999.

[55] Z Li and H A Scheraga. Monte carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Sciences*, 84(19):6611–6615, 1987.

[56] David J. Wales and Jonathan P. K. Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116, 1997.

[57] David J. Wales and Harold A. Scheraga. Global optimization of clusters, crystals, and biomolecules. *Science*, 285(5432):1368–1372, 1999.

[58] P. G. Mezey. *Potential Energy Hypersurfaces.* Elsevier, Amsterdam, 1987.

[59] David J. Wales. Basins of attraction for stationary points on a potential-energy surface. *J. Chem. Soc., Faraday Trans.*, 88:653–657, 1992.

[60] David J. Wales. Gmin: A program for basin-hopping global optimization. `http://www-wales.ch.cam.ac.uk/software.html`, 2014.

[61] Philippe Derreumaux. A diffusion process-controlled monte carlo method for finding the global energy minimum of a polypeptide chain. i. formulation and test on a hexadecapeptide. *The Journal of Chemical Physics*, 106(12), 1997.

[62] Philippe Derreumaux. Folding a 20 amino acid $\alpha\beta$ peptide with the diffusion process-controlled monte carlo method. *The Journal of Chemical Physics*, 107(6), 1997.

[63] Paul N. Mortenson and David J. Wales. Energy landscapes, global optimization and dynamics of the polyalanine ac(ala)8nhme. *The Journal of Chemical Physics*, 114(14), 2001.

[64] Paul N. Mortenson, David A. Evans, and David J. Wales. Energy landscapes of model polyalanines. *The Journal of Chemical Physics*, 117(3), 2002.

[65] Joanne M. Carr and David J. Wales. Global optimization and folding pathways of selected α-helical proteins. *The Journal of Chemical Physics*, 123(23):–, 2005.

[66] A. Verma, A. Schug, K. H. Lee, and W. Wenzel. Basin hopping simulations for all-atom protein folding. *The Journal of Chemical Physics*, 124(4):–, 2006.

[67] Birgit Strodel and David J. Wales. Implicit solvent models and the energy landscape for aggregation of the amyloidogenic kffe peptide. *Journal of Chemical Theory and Computation*, 4(4):657–672, 2008.

[68] Birgit Strodel, Jason W. L. Lee, Christopher S. Whittleston, and David J. Wales. Transmembrane structures for alzheimer's aβ1-42 oligomers. *Journal of the American Chemical Society*, 132(38):13300–13312, 2010.

[69] Olujide O. Olubiyi and Birgit Strodel. Structures of the amyloid β-peptides aβ1–40 and aβ1–42 as influenced by ph and a d-peptide. *The Journal of Physical Chemistry B*, 116(10):3280–3291, 2012.

[70] Tilo Strutz. *Data fitting and uncertainty : a practical introduction to weighted least squares and beyond ; with 23 tables and 71 test questions and examples ; [with online-service.* Vieweg + Teubner, Wiesbaden, 2011.

[71] Kenji Mochizuki, Chris S. Whittleston, Sandeep Somani, Halim Kusumaatmaja, and David J. Wales. A conformational factorisation approach for estimating the binding free energies of macromolecules. *Phys. Chem. Chem. Phys.*, 16:2842–2853, 2014.

[72] Jianlin Cheng and Pierre Baldi. Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. *Bioinformatics*, 21(suppl 1):i75–i84, 2005.

[73] Castrense Savojardo, Piero Fariselli, Pier Luigi Martelli, and Rita Casadio. Bcov: a method for predicting β-sheet topology using sparse inverse covariance estimation and integer programming. *Bioinformatics*, 29(24):3151–3157, 2013.