Institute for Advanced Simulation

De Novo Protein Folding with Distributed Computational Resources

Timo Strunk, Abhinav Verma, Srinivasa Murthy Gopal,
Alexander Schug, Konstantin Klenin,
and Wolfgang Wenzel

published in

Multiscale Simulation Methods in Molecular Sciences,
J. Grotendorst, N. Attig, S. Blügel, D. Marx (Eds.),
Institute for Advanced Simulation, Forschungszentrum Jülich,
NIC Series, Vol. **42**, ISBN 978-3-9810843-8-2, pp. 397-420, 2009.

http://www.fz-juelich.de/nic-series/volume42

De Novo Protein Folding with Distributed Computational Resources

Timo Strunk[1], Abhinav Verma[2], Srinivasa Murthy Gopal[3], Alexander Schug[4], Konstantin Klenin[1], and Wolfgang Wenzel[1,5]

[1] Forschungszentrum Karlsruhe
Institute for Nanotechnology , P.O. Box 3640, 76021 Karlsruhe, Germany
E-mail: wenzel@int.fzk.de

[2] Centro de Investigaciones Biologicas
Ramiro de Maeztu 9, 28040 Madrid, Spain

[3] Michigan State University
Department of Biochemistry & Molecular Biology
East Lansing, MI 48824-1319, USA

[4] Center for Theoretical Biological Physics (CTBP)
UCSD, La Jolla, CA 92093, USA

[5] DFG Center for Functional Nanotechnology
Karlsruhe Institute for Technology
76131 Karlsruhe, Germany

Proteins constitute a major part of the machinery of all cellular life. While sequence information of many proteins is readily available, the determination of protein three-dimensional structure is much more involved. Computational methods increasingly contribute to elucidate protein structure, conformational change and biological function. Simulations also help us understand, why naturally occurring proteins fold with high precision into a unique three-dimensional structure, in which they can perform their biological function. Here we summarize recent results of a free-energy approach to simulate protein large-scale conformational change and folding with atomic precision. In the free-energy approach, which is based on Anfinsen's thermodynamic hypothesis, the conformational ensemble can be sampled with non-equilibrium methods, which accelerates the search of the high-dimensional protein landscape and permits the study of larger proteins at the all-atom level.

1 Introduction

Proteins are the workhorses of all cellular life. They constitute the building blocks and the machinery of all cells. Proteins perform a variety of roles in the cell: structural proteins constitute the building blocks for cells and tissues, enzymes, like pepsin, catalyze complex reactions, signaling proteins, like insulin, transfer signals between or within the cells. Transport proteins, like hemoglobin, carry small molecules or ions, while receptor proteins like rhodopsin generate response to stimuli. The mechanisms of all these biophysical processes depend on the precise folding of their respective polypeptide chains[1].

From the work of C.B. Anfinsen and co-workers in the 1960s we know that the amino acid sequence of a polypeptide chain in the appropriate physiological environment can fully determine its folding into a so-called native conformation[2]. Unlike man-made polymers of similar length, functional proteins assume unique three-dimensional structures under physiological conditions and there must be rules governing this sequence-to-structure

transition. Protein structures can be determined experimentally, by X-ray crystallography[3] or NMR methods[4], but these experiments are still challenging and do not work for all proteins. From the theoretical standpoint it is still not possible to reliably predict the native three-dimensional conformation of most proteins given their amino acid sequence alone[5–8].

The triplet genetic code by which the DNA sequence determines the amino acid sequence of polypeptide chains is well understood. However, unfolded polypeptide chains lack most of the properties needed for their biological function. The chain must fold into its native three dimensional conformation in order to perform its function[9]. Despite much research in this direction and the emergence of novel folding paradigms during the last decade, much of the mechanism by which the protein performs this auto-induced folding reaction is still unclear[6].

Therefore it would be very helpful to develop methods for protein structure prediction on the basis of the amino acid sequence alone. Even if this goal it is not fully realized, methods that can complete partially resolved experimental protein structures would be very helpful to determine the structure of proteins where neither theoretical methods nor experimental techniques alone can succeed[10]. For the trans-membrane family of proteins, present day experimental methods fail, which is responsible for the entire communication of the cell with its environment[11]. Theoretical methods would be very helpful to investigate these proteins. There are large number of related questions, for instance regarding the interactions of a given protein with a large variety of other proteins, where theoretical methods could also contribute to our understanding of biological function.

Related to the question of protein structure prediction is the question of how the proteins attain their final conformation - the so called protein folding problem. It remains one of the astonishing mysteries responsible for the evolution of life how these complex molecules can attain a unique native conformation with such precision. No man-made polymer of similar size is able to assemble into a predetermined structure with the precision encountered in the proteins that have evolved in nature.

Given its complexity it is not surprising that the protein folding process occasionally fails, and many of such failures are related to cellular disfunction or disease[12,13]. Therefore it is important not only to be able to predict the final structure of proteins but also very desirable to understand the mechanisms by which proteins fold.

Many theories and computational methods have been developed to understand the folding process. Simplified models have been applied to understand its physical principles[14]. Lattice based methods were among the first models that allowed efficient sampling of conformational space[15–17]. The lattice models, either 2D square or 3D cubic, were used to study protein folding and unfolding, but they were too simplified for protein structure prediction. Subsequently "Gō-Models" were developed, where only native contacts interact favorably[18], and were useful to characterize some aspects of the folding of small proteins. Further development led to statistically obtained knowledge based potentials[19–21]. These potentials were obtained and parameterized on the structures available from the Protein Data Bank. The knowledge based potentials are mostly used for fold recognition or protein structure prediction.

With the increase in computational resources and speed, all-atom molecular dynamics simulations of protein folding have been undertaken. For most proteins, it is still not feasible to determine the protein structure from extended conformations using a single molecular dynamics simulation. This is due to the fact that at the all-atom level, the typical

time step in a molecular dynamics simulation is about 1-2 femtoseconds while the protein folding occurs at millisecond timescale. A single such simulation would need years to complete. Replica exchange MD simulations have been successful in folding proteins from extended conformations, but are still limited to the size of 20-30 amino acids[22–26].

In this review we explore an alternative approach for protein structure prediction and folding that is based on the Anfinsen's hypothesis[2] that most proteins are in thermodynamic equilibrium with their environment in their native state. For proteins of this class the native conformation corresponds to the global optimum of the free energy of the protein. We know from many problems in physics and chemistry that the global optimum of a complex energy landscape can be obtained with high efficiency using stochastic optimization methods[27–29]. These methods map the folding process found in nature onto a fictitious dynamical process that explores the free-energy surface of the protein. By construction these fictitious dynamical processes not only find the conformation of lowest energy, but typically characterize the entire low-energy ensemble of competing metastable states.

This review is structured as follows: The second section introduces the protein the protein free-energy forcefield PFF02 and methods to efficiently explore the protein free-energy surface with stochastic simulation methods. In the next section, we review all-atom folding simulations for various proteins with the free-energy approach. The key results of these investigations and opportunities for further work are outlined in the last section.

2 Free-Energy Forcefields and Simulation Methods

2.1 The free-energy forcefield PFF02

We have recently developed an all-atom (with the exception of apolar CH_n groups) free-energy protein forcefield (PFF01) that models the low-energy conformations of proteins with minimal computational demand.[9,14] The forcefield parameterizes the internal free energy of a particular protein backbone conformation, excluding backbone entropy and thus makes different discrete conformational states directly comparable with regard to their stability. The effect of backbone entropy of a particular state can be assessed with Monte Carlo simulations at a finite temperature.

PFF02 contains the following non-bonded interactions:

$$V\left(\{\vec{r_i}\}\right) = \sum_{ij} V_{ij}\left[\left(\frac{R_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{ij}\frac{q_i q_j}{\varepsilon_{g(i)g(j)} r_{ij}}$$

$$+ \sum_i \sigma_i A_i + \sum_{hbonds} V_{hb} + V_{tor}$$

Here r_{ij} denotes the distance between atoms i and j and g(i) the type of the amino acid i. The Lennard Jones parameters (V_{ij},R_{ij}) for potential depths and equilibrium distance) depend on the type of the atom pair and were adjusted to satisfy constraints derived from a set of 138 proteins of the PDB database.[18−20] The non-trivial electrostatic interactions in proteins are represented via group-specific and position dependent dielectric constants $\varepsilon_{g(i)g(j)}$, depending on the amino-acids to which the atoms i and j belong. Interactions with the solvent were first fit in a minimal solvent accessible surface model[21] parameterized by free energies per unit area σ_j to reproduce the enthalpies of solvation of the

Gly-X-Gly family of peptides[22]. A_j corresponds to the area of atom i that is in contact with a fictitious solvent.

Hydrogen bonds are described via dipole-dipole interactions included in the electrostatic terms and an additional short range term for backbone-backbone hydrogen bonding (CO to NH) which depends on the OH distance, the angle between N, H and O along the bond and the angle between the CO and NH axis.[9] In comparison to PFF01, the forcefield PFF02 contains an additional term that differentiates between the backbone dipole alignments found in different secondary structure elements (included in the electrostatic potential between atoms i and j belonging to the backbone NH or CO groups via the dielectric constants $\varepsilon_{g(i)g(j)}$)[23] and a torsional potential for backbone dihedral angles V_{tor}, which gives a small contribution (about 0.3 kcal/mol) to stabilize conformations with dihedral angles in the beta sheet region of the Ramachandran plot.[14,24]

2.2 Stochastic Simulation Methods

Proteins assume unique three dimensional structures after being synthesized into a linear chain of amino acids. In the free-energy approach this native conformation corresponds to the global optimum of the free-energy forcefield. In order to fold proteins with free-energy methods, we need to use efficient sampling methods to reliably locate the associated global minima of the free-energy surface. The low-energy region of the free-energy landscape of proteins is extremely rugged due to the close packing of the atoms in the native conformation. Sampling this surface efficiently is therefore the central cogmputational bottleneck of this approach.

2.2.1 Monte Carlo

Most stochastic methods originate from the Monte Carlo method that explores the energy landscape by random changes in the geometry of the molecule. In this way large regions of the configurational space can be searched in finite time, without regard of the kinetics of the process. A Monte Carlo simulation is composed of the following steps:

1. Specify the initial coordinates (R_0).
2. Generate new coordinates by random change to initial coordinates ($R^{'}$).
3. Compute transition probability $T(R_0, R^{'})$.
4. Generate a uniform random number RAN in range [0,1].
5. If $T(R_0, R^{'}) < RAN$, then discard the new coordinates and goto step 2.
6. Otherwise accept the new conformation and goto step 2.

The most popular realization of the Monte Carlo method for molecular systems is the Metropolis method (see flowchart in Figure 1), which uses $T(R_0, R^{'}) = e^{-\Delta V/kT}$ if $\Delta V > 0$, and unit probability otherwise.

In Monte Carlo simulations, the system has no "memory" between two steps, *i.e.*, the probability that the system might revert to its previous state is as probable as choosing any

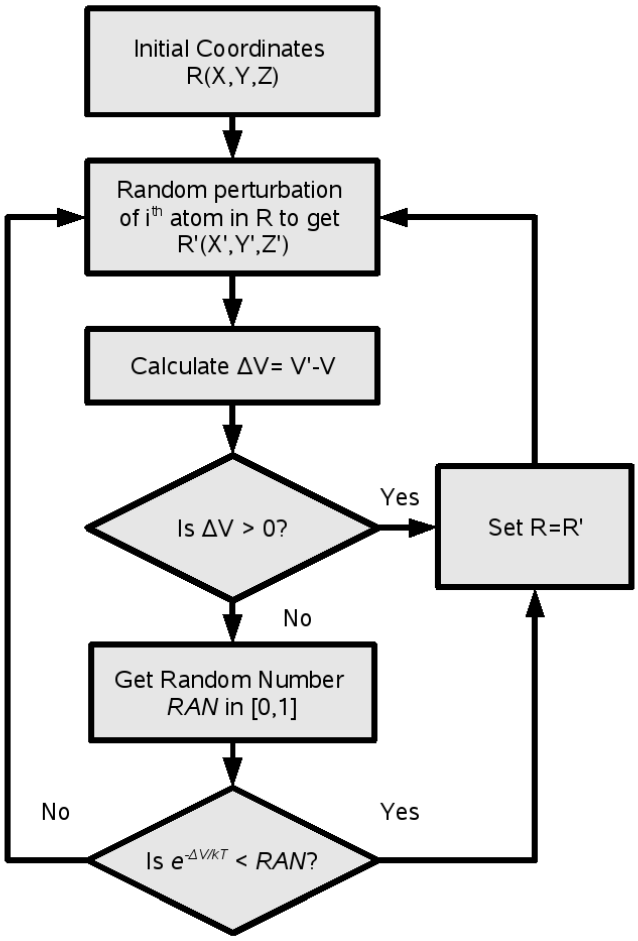


Figure 1. Schematic representation of Metropolis method.

other state. As a result of the stochastic simulation a large number of configurations is accumulated, which can be used to calculate thermodynamic properties of the system. Monte Carlo is not a deterministic method (as molecular dynamics), but gives rapid convergence of the thermodynamic properties[30].

2.2.2 Improved Sampling Techniques

Due to its popularity a large number of modifications and improvements of the Monte Carlo technique have been suggested and many of them have been used in the context of protein simulations:

- Simulated annealing: In this approach[31] barriers in the simulation are avoided by

starting the simulation at some high temperature and slowly lower the temperature of the simulation until the target temperature is reached. At high temperature the exploration of the phase space is very rapid, while near the end of the simulation the true thermodynamic probabilities of the system are sampled.

- Stochastic tunneling: Here a potential energy surface is transformed by using a non-linear transformation to suppress the barriers which are significantly above the present best energy estimate[32]. The transformed energy surface which is used for exploration of global minimum is given by

$$E_{STUN} = ln(x + \sqrt{x^2 + 1})$$

with $x = \gamma(E - E_0)$, where E is the present energy, E_0 is best estimation so far and γ the transformation parameter, which controls the rate of rise for the transformation.

- Parallel tempering: This method is Monte Carlo implementation of the replica exchange molecular dynamics method described. A modified version of this method, which uses an adaptive temperature control and replication step, has been employed for exploration of protein energy surfaces[33].

- Basin hopping technique (BHT): In this scheme the original potential energy surface is simplified by replacing the energy of each conformation with the energy of a nearby local minimum[34]. The minimization is carried out on the simplified potential (see section 2.2.3).

- Evolutionary strategy: This scheme is a multi-process extention of the BHT. Several concurrent simulations are carried out in parallel on a population. The population is evolved towards a global optimum of energy with a set of rules which enforce energy improvement and population diversity (see section 2.2.4).

2.2.3 Basin Hopping Technique

BHT[35] employs a relatively straightforward approach to eliminate high-energy transition states of the free-energy surface: The original free-energy surface is simplified by replacing the energy of each conformation with the energy of a nearby local minimum. In many applications the additional effort for the minimization step is more than compensated by the improved efficiency of the stochastic search. This process leads to a simplified potential on which the simulations search for the global minimum. This replacement eliminates high-energy barriers in the stochastic search that are responsible for the freezing problem in simulated annealing. A one dimensional schematic representation of BHT is shown in Figure 2. Every basin hopping cycle (minimization step) tries to locate a local minima and thus it simplifies the original potential energy surface (PES) (black curve) into an effective PES (blue curve) which is then searched for the global minima.

The basin hopping technique and its derivatives have been used previously to study the potential-energy surface of model proteins and polyalanines using all-atom models[36–39]. Here we replace the gradient-based minimization step used in many prior studies with a simulated annealing run[31], because local minimization generates only very small steps on the free energy surface of proteins. In addition, the computation of gradients for the SASA

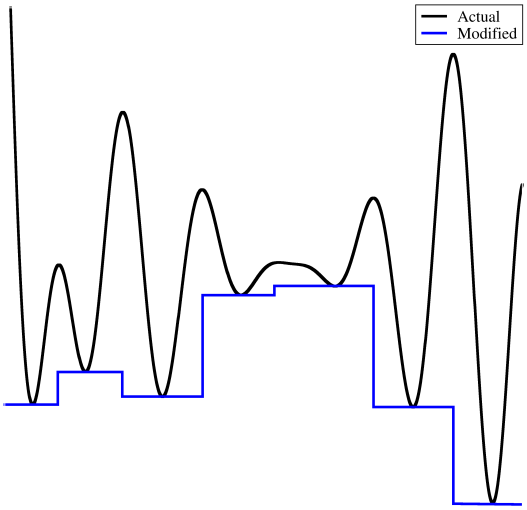


Figure 2. Schematic representation of Basin Hopping technique. The modified potential is obtained by replacing every point on the curve to its neared local minimum.

(Solvent Accessible Surface Area) is computationally prohibitive. Within each simulated annealing simulation, new configurations are accepted according to the Metropolis criterion, while the temperature is decreased geometrically from its starting to the final value.

The starting temperature and cycle length determine how far the annealing step can deviate from its starting conformation. The final temperature must be chosen small compared to typical energy differences between competing metastable conformations, to ensure convergence to a local minimum. The annealing protocol is thus parameterized by the starting temperature T_S, the final temperature T_F, and the number of steps. We investigated various choices for the numerical parameters of the method but have always used a geometric cooling schedule. At the end of one annealing cycle the new conformation is accepted if its energy difference to the current configuration was no higher than a given threshold energy ϵ_T, an approach recently proven optimal for certain optimization problems[40]. We typically used a threshold acceptance criteria of 1-3 kcal/mol.

2.2.4 Evolutionary Algorithms

The popular BHT method [41,34] for global optimization eliminates high-energy potential-energy surface (PES) by replacing the energy of each conformation with the energy of a nearby local minimum. For protein folding we have replaced the original local minimization by simulated annealing(SA). In the course of our folding studies, we find that independent BHT simulations often find identical structures corresponding to same local(global) minimum. As a result, each independent simulation reconstructs the full folding path independently. It would be very desirable to develop methods, where several concurrent simulations exchange information to *learn* from each other. For a PES having many local

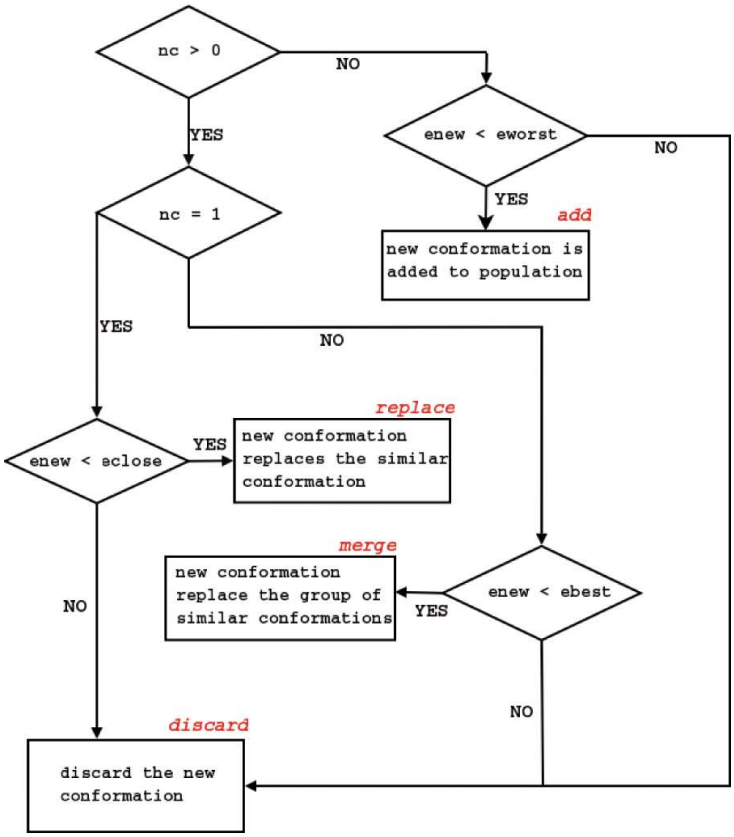


Figure 3. A flowchart illustrating the population update. See the text for an explanation

minima, independent simulations limit the efficient exploration of the PES. Also, occasionally BHT simulations go astray, ending the search in a wrong energy basin of the PES. We have developed a *greedy* version of BHT[42] which overcome these problems to a certain extent.

We have therefore generalized the BHT approach to a population of size N which is iteratively improved by P concurrent dynamical processes[33]. The population is evolved towards a optimum of the free energy surface with a ES that balances the energy improvement with population diversity. In the ES, conformations are drawn from the *active* population and subjected to an annealing cycle. At the end of each cycle the resulting conformation is either integrated into the active population or discarded. The algorithm was implemented as a master-client model in which idle clients request a task from the master. The master maintains the *active* conformation of the population and distributes the work to the clients. Each step in the algorithm has three phases:

1. Selection: A conformation is drawn randomly from the *active* population. We have used a uniform probability distribution with population of 20 conformers.

2. Annealing cycle: We use a simulated annealing schedule with T_{start} drawn from an exponential distribution and T_{end} fixed at 2K. The number of steps per cycle is increased as $10^5 \times \sqrt{cycle}$.

3. Population update: We have adjusted the acceptance criterion for newly generated conformations to balance the population diversity and energy enrichment. We define the two structures as *similar* if they have bRMSD less than 3 Å to each other. We define an *active* population as the pool containing mutually different lowest energy conformers. The master finds number of similar structures(*nc*) and then performs one of the following operations on complete population.

   (a) Add: If the new conformation is not *similar* to any structure(nc=0) in the population, we add it to the population, provided its energy is less than the energy of conformation with highest energy(E_{worst})

   (b) Replace: If the new conformation (with energy E_{new}) is *similar to one* existing structure in the population (with energy E_{old}), it replaces that structure provided $E_{\mathrm{new}} < E_{\mathrm{old}} + \Delta$ (see below).

   (c) Merge: If the new conformation has *several similar* structures, it replaces this group of structures provided its energy is less than the best one of the group E_{best} plus an acceptance threshold Δ.

   A flowchart illustrating the population update tasks of the master is shown in Fig. 3. In our first BHT/ES simulations we have used a fixed energy threshold (Δ) acceptance criterion. Here we have implemented a *variable* energy threshold which we define as $\Delta = A \times tanhD$, where

$$D = \frac{E_{\mathrm{new}} - E_{\mathrm{best}}}{A},$$

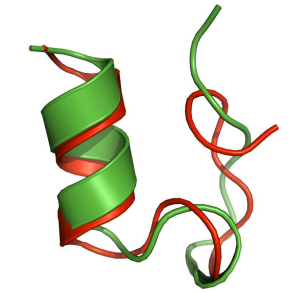
   where A is the energy threshold (3kcal/mol), Enew is energy of the new structure, E_{best} is the lowest energy structure in the population. This choice of the energy criterion ensures that the conformation with the best energy is never replaced, while conformations higher in energy are more easily replaced in the secure knowledge that they are far from optimal. The rules for the *replace* and *merge* operations ensure the structural diversity of the population and its continued energetic improvement (on average).

3 Folding Simulations

3.1 Helical Proteins

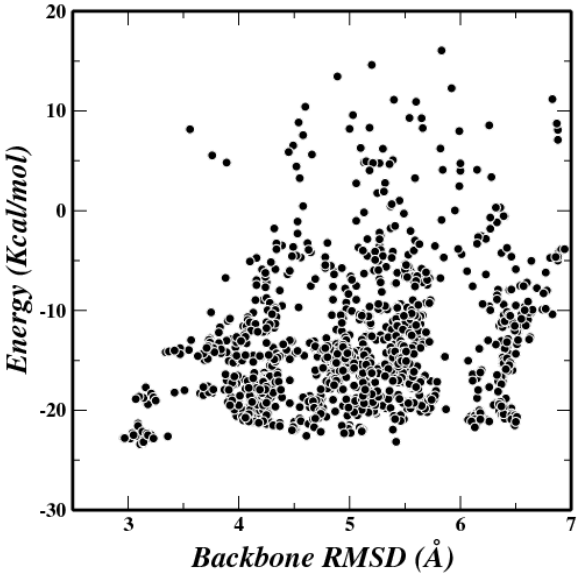
3.1.1 The tryptophan cage miniprotein

Tryptophan cage or trp-cage protein[43] has been the subject of various theoretical studies and it has been of great scientific interest. It had been reported to fold using replica



(a) Overlay

(b) C_β-C_β matrix



(c)

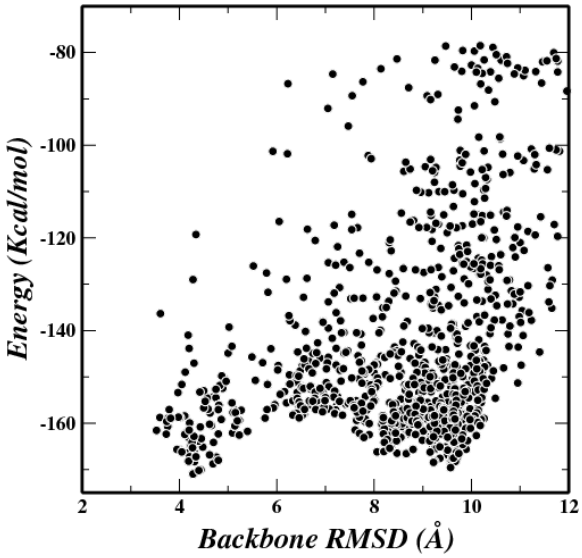
Figure 4. 1L2Y: Overlay of predicted (red) structure to experimental (green) structure. C_β-C_β distance overlay matrix and Energy vs. RMSD plot.

exchange MD and a variety of other simulations[44,27,45–47,29,48]. We performed 20 independent basin hopping simulations starting with the completely extended conformations in PFF02 with 100 cycles. The starting conformation had a RMSD of 12.94 Å to the native conformation and was completely extended manually (by setting all backbone dihedral angles except proline to 180°). The starting temperatures were chosen from a distribution of exponentially distributed temperatures and the number of steps increased with the BHT cooling cycle by $10^4\sqrt{n_m}$ where n_m is the number of minimization cycles.

The lowest energy structure converges to a native like conformation with RMSD of 3.11 Å to the native conformation. For the sake of uniformity in case of NMR resolved experimental structures, we compare the RMSD to the first model in the protein data bank file. The lowest energy structure had an energy of -23.4 Kcal/mol. Figure 4(c) shows the scatter plot of the conformations visited by the basin hopping simulations on the free energy surface. The overlay of native conformation (green) with the lowest energy conformation (red) is shown in Figure 4(a) and the corresponding C_β-C_β overlay matrix is shown in Figure 4(b). The C_β-C_β overlay matrix quantifies the tertiary alignment along with secondary structure formation by taking the difference between all C_β distances of predicted and native conformation. Black regions indicate excellent agreement in the formation of native contacts while white regions indicate larger deviations.

(a) Overlay

(b) C_β-C_β matrix



(c)

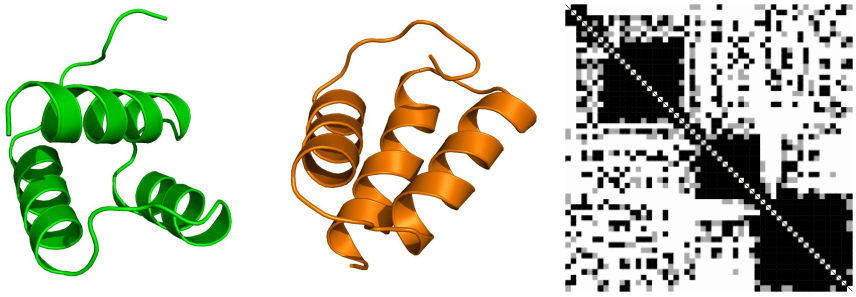
Figure 5. 1ENH: Overlay of predicted (red) structure to experimental (green) structure. C_β-C_β distance overlay matrix and Energy vs. RMSD plot.

3.1.2 The engrailed Homeodomain - 1ENH

The 54 amino acid engrailed homeodomain protein[49] is a three helical orthogonal bundle protein which has been subjected to detailed molecular dynamics simulations[50,51]. It was not possible to fold this protein using basin hopping technique due to the previously described freezing problem in the basin hopping simulations.

Here we studied the folding of engrailed homeodomain in PFF02 using the evolutionary algorithm with a maximum population of 64 conformations and 512 processors[52]. The lowest energy structure converges to 4.28 Å to the native conformation with the energy of -170.95 Kcal/mol. 1ENH has a unstructured tail at the N-terminus; after excluding this seven amino acid region, the RMSD reduces to only 3.4Å.

The scatter plot of conformations visited during the simulation are shown in Figure 5(c). Seven out of the total population of 64 structures are less than 4.5 Å RMSD to the native conformation. The overlay of the lowest energy conformation (red) with the native conformation (green) is shown in Figure 5(a) and the corresponding C_β-C_β overlay matrix is shown in Figure 5(b). There are also competing conformations (within 2 Kcal/-mol) with large RMS deviations encountered in the simulations. One such conformation is shown in Figure 6). These conformations have the same secondary structure, but a different tertiary structure alignment. The C_β-C_β overlay matrix for the misfolded conformation also confirms that all the three helices are properly predicted but their tertiary arrangement



(a) (b) (c)

Figure 6. 1ENH: Overlay of misfolded (orange) structure to experimental (green) structure and C_β-C_β distance overlay matrix.

is completely different. This indicates that various conformations exist in the low energy region of the 1ENH which are similar in secondary structure content.

No two helices in the misfolded conformation are in agreement with the respective helices in the native state. Independently, helix-1 (E8-E20), helix-2 (E26-L36) and helix-3 (A40-K43) are nearly perfectly predicted and have RMS of only 0.56, 0.42 and 0.47 Å respectively.

As about 10% of the population is native-like and the misfolded conformations we can conclude that the folding is reproducible.

3.2 Hairpins

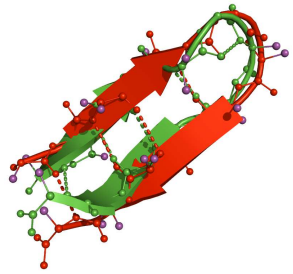
Hairpins are the simplest beta sheet structures with only two strands in antiparallel directions that are connected together with a turn. Hydrogen bonding and the packing of the protein itself plays a crucial role here in the folding of such small polypeptides. There are not many hairpin proteins that are not stabilized by external interaction with ions or with the formation of disulphide bridges.

3.2.1 trp-zippers

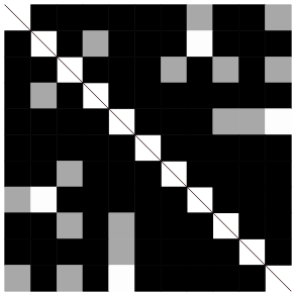
The tryptophan zippers are small monomeric stable β-hairpins that adopt an unique tertiary fold without requiring metal binding, unusual amino acids, or disulfide crosslinks[53]. We were able to fold various tryptophan zippers using PFF02 and basin hopping technique (not shown here).

We studied the folding of 1LE0 with EA using 128 processors on Marenostrum cluster at the Barcelona supercomputer center starting from completely extended conformations. We performed twenty cycles of evolutionary algorithm. The lowest energy conformation reached in the simulation had a RMSD of only 1.5 Å to the native conformation with the energy of -29.97 Kcal/mol.

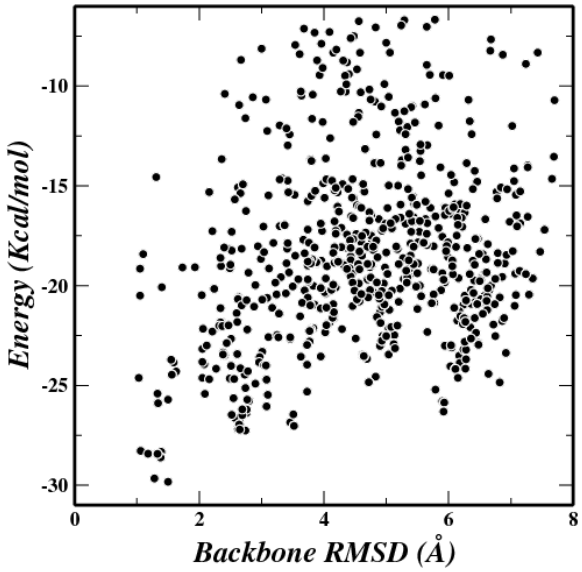
The scatter plot of the conformations visited during the simulations is shown in Figure 7(c). The scatter plot shows that the native-like conformations lie significantly below any other conformation. Twelve out of the 64 conformations from the final population



(a) Overlay



(b) C_β-C_β matrix



(c)

Figure 7. 1LE0: Overlay of predicted (red) structure to experimental (green) structure. C_β-C_β distance overlay matrix and Energy vs RMSD plot.

are less than 3.0 Å to the native conformation. The protein folds in less than 90 minutes using 128 processors in parallel by means of the twenty cycles of evolutionary algorithm amounting to 77×10^6 function evaluations or about 9 CPU days.

The overlay of the predicted conformation (red) with the native conformation (green) is shown in Figure 7(a) and the corresponding C_β-C_β overlay matrix is shown in Figure 7(c). Large black regions in the C_β-C_β overlay matrix indicates the agreement of native contacts between the two conformations.

As hydrogen bonding plays an important role in the formation and topology of β-sheet structures, it is important to compare the hydrogen bonding pattern in the lowest energy conformations as two β-sheet conformations might look very similar to the eye, but they might have completely different topology resulting from shifting of backbone hydrogen bonds.

The pattern of backbone hydrogen bonds is shown in Table 1 for the native and the predicted conformation. These were calculated with MOLMOL using the standard definitions (Distance=2.4Å and angle=35°). Four out of the five backbone hydrogen bonds of the native structure are predicted correctly in the lowest energy structure found in the simulations.

As about 20% of the population converged to native-like conformations with much lower energies, we conclude the folding of tryptophan zipper as reproducible and predictive.

Hydrogen bond	Native	Predicted
03 THR HN ⟶ 10 THR O	X	X
05 GLU HN ⟶ 08 LYS O	X	X
07 ASN HN ⟶ 05 GLU O	X	
10 THR HN ⟶ 03 THR O	X	X
12 LYS HN ⟶ 01 SER O	X	X

	Secondary Structure	RMSD (Å)
Native	CEEECSSSEEEC	-
Predicted	CEEEETTEEEEC	1.52

Table 1. 1LE0: Backbone hydrogen bond pattern of the native and predicted conformations and secondary structure information.

3.2.2 HIV-1 V3 loops

We studied the folding of 14 amino acid HIV-1 $V3_{MN}$ loop 1NIZ[54] in PFF02 using a greedy version of the basin hopping technique[55].

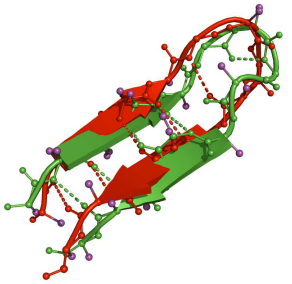
In basin hopping simulations there is a threshold energy acceptance criterion at the end of every basin hopping cycle. In our previous simulations, we have used this threshold acceptance criterion of 1-3 Kcal/mol depending upon this size of the protein. In the greedy version of basin hopping the threshold energy is varied depending upon the best energy found so far in the simulation. Here we calculated the threshold as $(\epsilon_S - \epsilon_B)/4$, where ϵ_S is the starting energy and ϵ_B is the best energy found so far in the simulation. This choice implies that the conformation with the best energy is never replaced with a conformation that is higher in energy and thus introduces a "memory effect" in the simulation. For the simulations that are higher in energy, the increased threshold value implies a higher acceptance probability of conformations with higher energy.

We did 200 cycles of greedy basin hopping simulations in PFF02. The simulations were started with completely extended conformation that had the RMSD of 12 Å to the native state. The lowest energy structure found in the simulation had the RMSD of only 2.04 Å to the native state.

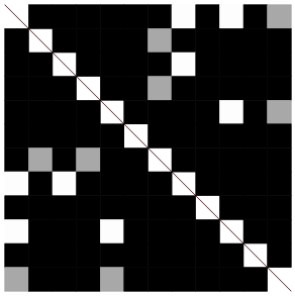
Hydrogen bond	Native	Predicted
02 ARG HN ⟶ 13 THR O	X	X
04 HIS HN ⟶ 11 PHE O	X	X
06 GLY HN ⟶ 09 ARG O		X
08 GLY HN ⟶ 06 GLY O	X	
11 PHE HN ⟶ 03 HIS O	X	X
13 THR HN ⟶ 01 ARG O	X	X

	Secondary Structure	RMSD (Å)
native	CEEEECSSCEEEEC	-
predicted	CEEEECSSCEEEEC	2.04

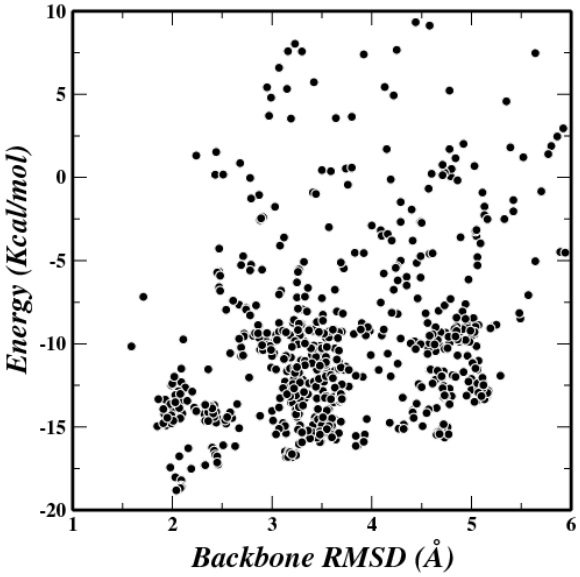
Table 2. 1NIZ: Backbone hydrogen bond pattern between native and predicted conformations and secondary structure information.



(a) Overlay



(b) C_β-C_β matrix



(c)

Figure 8. 1NIZ: Overlay of predicted (red) structure to experimental (green) structure. C_β-C_β distance overlay matrix and Energy vs. RMSD plot.

The scatter plot of the conformations visited during the simulations is shown in Figure 8(c). The scatter plot shows a single downhill folding funnel for this hairpin. Eight out of the ten independent simulations converged to less than 3.5 Å RMSD to the native conformation.

The overlay of the lowest energy conformation (red) with the native conformation (green) is shown in Figure 8(a) and the corresponding C_β-C_β distance matrix is shown in Figure 8(c). Large black regions in the C_β-C_β overlay matrix indicates the agreement of native contacts between the two conformations.

Again, we did the backbone hydrogen bond analysis. Four out of the five backbone hydrogen bonds of the native structure were correctly predicted in the lowest energy structure found in the simulations. The pattern of the backbone hydrogen bonds is shown in Table 2. The secondary structure of the predicted and native conformation is also shown in Table 2. The letters in the secondary structure correspond to DSSP definitions.

As eight of the ten simulations converged to the native-like conformation without any competing metastable conformations, the folding is concluded as reproducible and predictive.

3.3 A mixed secondary structure protein

Zinc fingers are among the most abundant proteins in eukaryotic genomes and occur in many DNA binding domains and transcription factors[56]. They participate in DNA recognition, RNA packaging, transcriptional activation protein folding and assembly and apoptosis. Many zinc fingers contain a $\mathrm{Cys_2His_2}$ binding motif that coordinates the Zn-ion in $\alpha\beta\beta$ -framework[57–59] and much effort is towards the engineering of novel zinc fingers[60]. A classical zinc finger motif binding DNA is illustrated in Fig. 9.

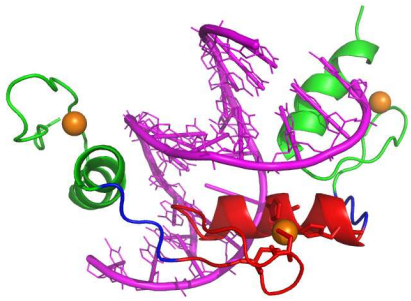


Figure 9. A classical $\mathrm{Cys_2His_2}$ zinc finger motif with Zn-ion(orange) and DNA (magenta).

The reproducible folding of such proteins with mixed secondary structure, however, remains a significant challenge to the accuracy of the all-atom forcefield and the simulation method[61]. We use the all-atom free-energy forcefield PFF02 to predictively fold the 23-51 amino-acid segment of the N-terminal sub-domain of ATF-2 (PDBID 1BHI)[62], a 29 amino acid peptide that contains the basic leucine zipper motif . 1BHI folds into the classical TFIIIa conformation found in many zinc-finger like sub-domains. The fragment contains all the conserved hydrophobic residues (PHE25, PHE36, LEU42) of the classical zinc finger motif and the CYS27, CYS32, HIS45, HIS49 zinc binding pattern.

Starting from a completely unfolded conformation with no secondary structure (16 Å backbone RMSD (bRMSD) to native) we performed 200 cycles of the evolutionary algorithm. The distribution of bRMSD versus energy of all accepted conformations during the simulation (Fig. 10) demonstrates that the simulation explores a wide variety of conformations, with regard to their free-energy and their deviation from the native conformation.

Among the ten energetically lowest conformations (see Table 3) six fold into near-native conformations with bRMSDs of 3.68-4.28 Å, while four fold to conformations with a larger bRMSD. The three energetically best conformations are all near-native in character. An overlay with the experimental conformation (left panel of Fig. 11) illustrates that the helix, beta-sheet and both turns are correctly formed. The hydrophobic residues, which determine the packing of the beta-sheet against the helix, are illustrated in blue in the figure. The helical section (GLU39-GLU50) and the beta-sheet (PHE25-LEU26 and ARG35-PHE36) deviate individually by 1.6 Å and 2.4 Å bRMSD from their experimental counterparts, respectively.

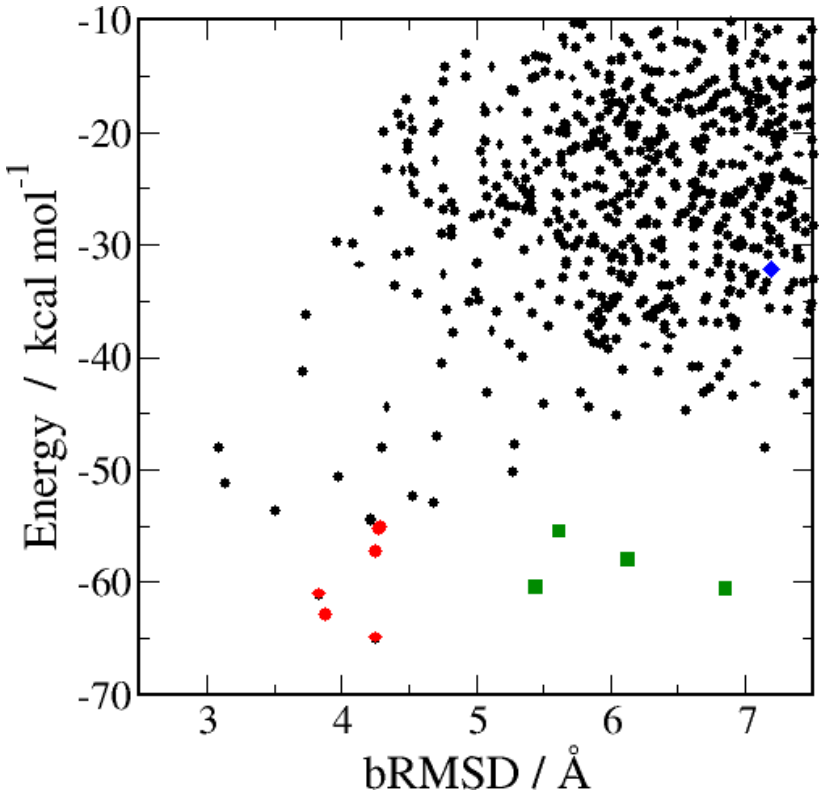


Figure 10. Free energy versus bRMSD of all accepted conformations in the simulation. The best 10 structures are highlighted as: red circles(native-like), green squares(non-native). The folding intermediate is denoted by blue diamond

The overall difference between the experimental and the folded conformations stems from the relative arrangement of the beta-sheet with respect to the helix, which is dominated by unspecific hydrophobic interactions. All conserved hydrophobic sidechains are also buried in the folded structure. The zinc-coordinating cysteine residues (CYS27,CYS32) are within 2 Å of their native positions and available association with the Zn-ion.

Fig. 12 shows the convergence of the energy. After about 120 attempted updates per population member (3.5×10^8 function evaluations) the population converged to the native ensemble. According to the funnel paradigm for protein folding[63], tertiary structure forms as the protein slides downhill on the free-energy surface from the unfolded ensemble towards the native conformation. Each annealing cycle generates a small perturbation on the existing conformation, which averages to a 0.5 Å bRMSD change (max 3 Å initially). As new low-energy conformations replace old conformations, the population slides as a whole down the funnel of the free energy landscape.

Ensemble averages as a function of time over the moving population are thus associated with different stages of the structure formation process. In the lower panels of Fig. 12, we

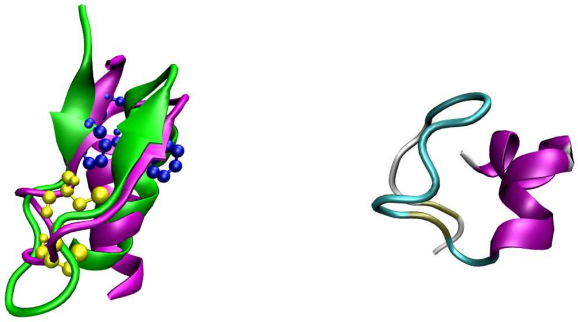


Figure 11. Left: Overlay of the native (green) and folded (magenta) conformations. The conserved hydrophobic residues are shown in blue and Zn binding cysteines are shown in yellow. Right: The intermediate conformation with partially formed helix and β sheet.

plot the average helical content and the number of beta-sheet H-bonds as a function of the cycle number. Following a rapid collapse to a compact conformation, the helix forms first, followed by the formation of the beta sheet. The analysis of the folding funnel upwards in energy illustrates that the lowest energy metastable conformations correspond to a partial unzipping of amino acids PHE25-ARG35, while the conserved cysteine residues are still buried. Even much higher on the free energy funnel (blue diamond in Fig. 10), we find many structures that have much residual structure, but essentially not long-range native contacts.

The preformed sheet-region is stabilized by the hydrogen bonds (LEU26-CYS27, ARG35) and packs at the right angle to the helix, the hydrophobic residues are only partially buried. This conformational freedom may be relevant in DNA binding, where the helical part of the zinc finger packs into the major groove of the DNA.

De novo folding of the zinc finger domain permits a direct sampling of the relevant low-energy portion of the free-energy surface of the molecule as the first step towards the elucidation of the structural mechanisms involved in DNA binding[64]. We find that much of the structure of the zinc finger is formed even in the absence of the metal ion that is ultimately required for the stabilization of the native conformation. Because the algorithm tracks the development of the population it is possible to reconstruct a folding pathway by reconstructing the sequence of events starting with converged conformation and moving backwards to the completely unfolded conformation.

We have thus demonstrated predictive all-atom folding of the DNA binding zinc-finger motif in the free-energy forcefield PFF02. This investigation offers the first unbiased characterization of the low-energy part of the free-energy surface of the zinc finger motif, which is unattainable in coarse grained, knowledge-based models. We find that the helix forms first along the folding path and acts as a template against which a variety of near-native beta-sheet backbone arrangements can pack. There are many zinc fingers with bRMSD

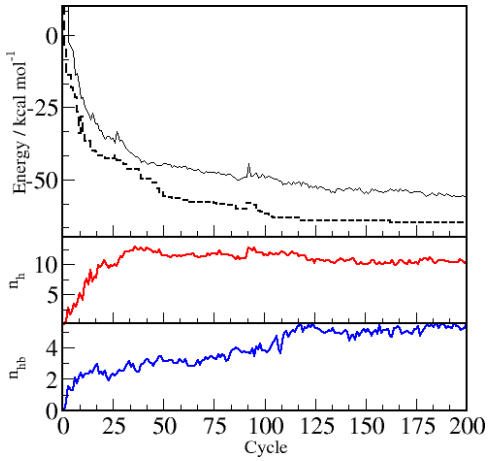


Figure 12. Top: Average (solid line) and best (dashed line) energies as functions of the number of the simulation cycle for the Zinc Finger, Middle: number of amino acids (n_h) in a helical conformation (as computed by DSSP) and Bottom: number of hydrogen bones (n_{hb}) as function of the ES cycle number

#	Energy kcal/mol	bRMSD Å	Secondary Structure
E01	-64.94	4.25	CC**EE**CTTTTSCC**EE**SSC**HHHHHHHHHHHH**C
E02	-62.84	3.88	CC**EE**CTTTTSCC**EE**SSC**HHHHHHHHH**STTC
E03	-61.05	3.83	CC**EE**CTTTTCCC**EE**SSC**HHHHHHHHH**STTC
E04	-60.51	6.85	CC**EE**CTTTTSCC**EE**CSC**HHHHHHH**SCCCCC
E05	-60.40	5.44	CCBBCTTTTCCCBCCSC**HHHHHHHH**CCCBC
E06	-57.93	6.12	CC**EE**CTTTTSCC**EE**CSC**HHHHHHH**SCCCCC
E07	-56.21	4.25	CC**EEEE**CSSSSC**EEEE**SC**HHHHHHHHHHH**C
E08	-55.44	5.61	CCSSSCSSCCSSCCCSC**HHHHHHHHH**TTTC
E09	-55.18	4.27	CCCC**EE**CTTSSC**EE**CS**HHHHHHHHHH**CSCC
E10	-55.02	-4.29	CCCCBTTTTBTTCCCSS**HHHHHHHHHHHH**C

Table 3. Energy, bRMSD and secondary structures of the 10 lowest energy structures

of less than 2 Å to 1BHI[62]. Thus, this investigation provides one important step in the theoretical understanding of zinc-finger formation and function.

4 Summary

These investigations demonstrate that the free-energy approach is able to predict the native state of a wide range of proteins at the global minimum of their free energy surface[27, 65–71]. Protein folding with free energy methods is much faster than the direct simulation of the folding pathway by kinetic methods such as molecular dynamics. Using just standard PCs we can fold a simple hairpin with fifteen to twenty amino acids in a matter of hours, at most in a day[69]. Unfortunately even for free energy methods the computational cost rises steeply with the system size.

The second ingredient in protein folding studies, aside from the force field, are therefore the simulation protocols, which ultimately determine whether the global optimum of the forcefield is determined accurately and reliably. We have reviewed key aspects of such methods, e.g. the stochastic tunneling or the basin hopping technique, which had proven successful in folding studies for small proteins. One of the key limitations of these methods is that they map the global optimization problem onto a single fictitious dynamical process, while in principle, many concurrent processes can be used[28, 29, 65].

We have therefore also discussed an evolutionary algorithm[72] for massively parallel architectures, such as the BlueGene architecture, which keeps a diverse population on the master, while the clients sample the protein landscape simultaneously. This algorithm scales very well with the number of processors used (up to 4096 tested on the IBM BlueGene). Using this algorithm we folded various proteins such as 40 amino acid HIV accessory protein (1F4I) and 54 amino acid engrailed homeodomain protein (1ENH) in a single day. The folding of the engrailed homeodomain protein was carried out in a single day using 512 processors on the Barcelona Mare Nostrum Supercomputer, the current largest supercomputer in Europe. Folding of the tryptophan zipper protein (1LE0) was possible in only 14 minutes using 128 processors[69].

To date we have succeeded to develop methods to find the native state of various proteins by locating the global minimum of the free energy surface[28]. There are, however, a large number of questions that remain to be addressed. Fortunately there are complementary methods, which in combination with the free-energy methodology developed here, can address these problems. For example, we have neglected the details of the kinetics of protein folding in our approach. As stated earlier, its important to study kinetics of folding to understand protein folding mechanism and to predict folding rates. Because free-energy methods sample exhaustively the low-energy conformations of the protein that are accessible under physiological conditions it may be possible to reconstruct the folding kinetics on the basis of that ensemble of conformations. This can be achieved by a dynamical analysis of the low energy region by using master equations assuming diffusive processes between similar conformations.

With the development of the all-atom protein forcefield (PFF02) we have made a significant step towards a universal free-energy approach to protein folding and structure prediction[68]. The massively parallel simulation methods developed in the last few years now permit the protein folding of medium-size proteins from random initial conformations. This work thus lays the foundations to further explore the mechanism of protein folding, to understand protein stability and ultimately develop methods for *de novo* protein structure prediction.

References

1. C. Branden and J. Tooze. *Introduction to protein structure.* Routledge, 1999.
2. C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.
3. L. Stryer. Implications of x-ray crystallographic studies of protein structure. *Annu. Rev. Biochem.*, 37:25–50, 1968.
4. G. Wagner, S. G. Hyberts, and T. F. Havel. Nmr structure determination in solution: A critique and comparison with x-ray crystallography. *Annu. Rev. Biophys. Biomol. Struct.*, 21:167–98, 1992.
5. Richard Bonneau and David Baker. Ab initio protein structure prediction: Progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.*, 30:173–89, 2001.
6. D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294:93–96, 2001.
7. J. Moult, K. Fidelis, A. Zemia, and T. Hubbard. Critical assessment of methods of protein structure (casp): round iv. *PROTEINS:Structure, Function, and Bioinformatics*, 45:2–7, 2001.
8. C. Hardin, M.P. Eastwood, M. Prentiss, Z. Luthey-Schulten, and P. Wolynes. Folding funnels: The key to robust protein structure prediction. *J. Comp. Chem.*, 23:138–146, 2003.
9. Jeremy M. Berg, John L. Tymoczky, and Lubert Stryer. *Biochemistry, fifth edition*. Michelle Julet, 2002.
10. B. Rost. Protein secondary structure prediction continues to rise. *J. Struct. Biol*, 134:204–18, 2001.
11. W. Khlbrandt and E. Gouaux. Membrane proteins. *Current Opinion in Structural Biology*, 9:445–7, 1999.
12. C. M. Dobson. The structural basis of protein folding and its links with human disease. *Phil. Trans. R. Soc. Lond. B*, 356:133–145, 2001.
13. M.B. Pepys. In J.G. Ledingham D.J. Weatherall and D.A. Warrel, editors, *The Oxford Textbook of Medicine (third ed.)*. Oxford University Press, Oxford, 1995.
14. H. S. Chan and K. A. Dill. Protein folding in the landscape perspective: Chevron plots and non-arrhenius kinetics. *Proteins: Struc. Func. and Gen.*, 30:2–33, 1998.
15. K. F. Lau and K. A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22:3986–97, 1989.
16. Ken A. Dill, Sarina Bromberg, Kaizhi Yue, Klaus M. Fiebig, David P. Yee, Paul D. Thomas, and Hue Sun Chan. Principles of protein folding- a perspective from simple exact models. *Protein Science*, 4:561–602, 1995.
17. Eugene Shakhnovich, G. Farztdinov, A.M. Gutin, and Martin Karplus. Protein folding bottlenecks: A lattice monte carlo simulation. *Phys. Rev. Lett.*, 67(12):1665–1668, 1991.
18. N. Go and H. A. Scheraga. Analysis of the contribution of internal vibrations to the statistical weight of equilibirium conformations of macromolecules. *J. Chem. Phys.*, 51:4751–4767, 1969.
19. M. J. Sippl, G. Nemethy, and H. A. Scheraga. Intermolecular potentials from crstal data. 6. determination ofemperical potentials for o-h$\cdots$o=c hydrogen bonds from packing configurations. *J. Phys. Chem.*, 88:6231–6233, 1984.

20. G. Casari and M. J. Sippl. Structure derived hydrophobic potentials. a hydrophobic potential derived from x ray structures of globular proteins is able to identify native folds. *J. Molec. Biol.*, 224:725–732, 1992.
21. M. J. Sippl. Knowledge-based potentials for proteins. *Current Opinion in Structural Biology*, 5:229–35, 1995.
22. F. Rao and A. Caflisch. Replica exchange molecular dynamics simulations of reversible folding. *J. Chem. Phys.*, 119:4035–4042, 2003.
23. P. H. Nguyen, G. Stock E. Mittag, C. K. Hu, and M. S. Li. Free energy landscape and folding mechanism of a β-hairpin in explicit water: A replica exchange molecular dynamics study. *Proteins: Struc. Func. and Gen.*, 61:705–808, 2005.
24. Y.M. Rhee and V.S. Pande. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys. J*, 84:775–786, 2003.
25. S.-Y. Kim, J. Lee, and J. Lee. Folding of small proteins using a single continous potential. 120:8271–8276, 2004.
26. J.-E. Shea and C. L. Brooks III. From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem.*, 52:499–535, 2001.
27. A. Schug, T. Herges, and W. Wenzel. Reproducible protein folding with the stochastic tunneling method. *Phys. Rev. Lett.*, 91:1581021–4, 2003.
28. A. Schug, B. Fischer, A. Verma, H. Merlitz, W. Wenzel, and G. Schoen. Biomolecular structure prediction stochastic optimization methods. *Advanced Engineering Materials*, 7(11):1005–1009, 2005.
29. A. Schug, T. Herges, A. Verma, K. H. Lee, and W. Wenzel. Comparison of stochastic optimization methods for all-atom folding of the trp-cage protein. *Chemphyschem*, 6:2640–6, 2006.
30. A. R. Leach. *Molecular Modelling: Principles and Applications*. Pearson Education Ltd., 2001.
31. S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–80, 1983.
32. K. Hamacher and W. Wenzel. A stochastic tunnelling approach for global minimization. *Phys. Rev. E*, 59:938, 1999.
33. A. Schug, T. Herges, A. Verma, and W. Wenzel. Investigation of the parallel tempering method for protein folding. *Phys. Cond. Matter, special issue: Structure and Function of Biomolecules (in press)*, 2005.
34. D. M. Leitner, C. Chakravarty, R. J. Hinde, and D. J. Wales. Global optimization by basin-hopping and the lowest energy structures of lennard jones clusters containing upto 110 atoms. *Phys. Rev E*, 56:363, 1997.
35. A. Nayeem, J. Vila, and H. A. Scheraga. A comparative study of the simulated-annealing and monte carlo-with-minimization approaches to the minimum-energy structures of polypeptides: [Met]-enkephalin. *J. Comp. Chem.*, 12:594–605, 1991.
36. R. A. Abagyan and M. Totrov. Biased probability monte carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.*, 235:983–1002, 1994.
37. D. J. Wales and P. E. J. Dewsbury. Effect of salt bridges on the energy landscape of a model protein. *J. Chem. Phys.*, 121:10284–90, 2004.
38. P. N. Mortenson and D. J. Wales. Energy landscapes, global optimisation and dynam-

ics of the polyalanine Ac(ala)8 NHMe. *J. Chem. Phys.*, 114:6443–54, 2001.
39. P. N. Mortenson, D. A. Evans, and D. J. Wales. Energy landscapes of model polyalanines. *J. Chem. Phys.*, 117:1363–76, 2002.
40. J. Schneider, I. Morgenstern, and J. M. Singer. Bouncing towards the optimum: Improving the results of monte carlo optimization algorithms. *Phys. Rev. E*, 58:5085–95, 1998.
41. A. Nayeem, J. Vila, and H.A. Scheraga. A comparative study of the simulated-annealing and monte carlo-with-minimization approaches to the minimum-energy structures of polypeptides: [met]-enkephalin. *J. Comp. Chem.*, 12(5):594–605, 1991.
42. W. Wenzel. Predictive folding of a β hairpin in an all-atom free-energy model. *Europhys. Letters*, 76:156, 2006.
43. J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen. Designing a 20-residue protein. *Nat. Struct. Biol.*, 9:425–30, 2002.
44. C. D. Snow, B. Zagrovic, and V. S. Pande. Folding kinetics and unfolded state topology via molecular dynamics simulations. *J. Am. Chem. Soc.*, 124:14548–14549, 2002.
45. F. Ding, S. V. Buldyrev, and N. V. Dokholyan. Folding trp-cage to nmr resolution native structure using a coarse-grained protein model. *Biophys. J.*, 88:147–55, 2005.
46. A. Linhananta, J. Boer, and I. MacKay. The equilibrium properties and folding kinetics of an all-atom go- model of the trp-cage. *J. Chem. Phys*, 122:1–15, 2005.
47. A. Schug, W. Wenzel, and U. H. E. Hansmann. nergy landscape paving simulations of the trp-cage protein. *J. Chem Phys.*, 122:1–7, 2005.
48. J. Juraszek and P. G. Bolhuis. Sampling the multiple folding mechanisms of trp-cage in explicit solvent. *Proc. Natl. Acad. Sci. USA*, 103:15859–64, 2006.
49. N. D. Clarke, C. R. Kissinger, J. Desjarlais, G. L. Gilliland, and C. O. Pabo. Structural studies of the engrailed homeodomain. *Protein Sci.*, 3:1779–87, 1994.
50. U. Mayor, N. R. Guydosh, C. M. Johnson, J. G. Grossmann, S. Sato S, G. S. Jas, S. M. Freund, D. O. Alonso, V. Daggett, and A. R. Fersht. The complete folding pathway of a protein from nanoseconds to microseconds. *Nature*, 421:863–7, 2003.
51. V. Daggett and A. Fersht. The present view of the mechanism of protein folding. *Nat. Rev. Mol. Cell. Biol.*, 4:497–502, 2003.
52. A. Verma and W. Wenzel. All-atom protein folding in a single day. submitted, 2006.
53. A. G. Cochran, N. J. Skelton, and M. A. Starovasnik. Tryptophan zippers: stable, monomeric β-hairpins. *Proc. Natl. Acad. Sci. USA*, 98:5578–83, 2001.
54. M. Sharon, N. Kessler, R. Levy, S. Zolla-Pazner, M. Gorlach, and J. Anglister. Alternative conformations of HIV-1 V3 loops mimic β-hairpins in chemokines, suggesting a mechanism for coreceptor selectivity. *Structure*, 11:225–236, 2003.
55. A. Verma and W. Wenzel. De-novo all atom folding of a HIV-1 V3 hairpin loop in an imporved free energy forcefield. Submitted, 2006.
56. J. H. Laity, B. M. Lee, and P. E. Wright. Zinc finger proteins: new insights into structural and functional diversity. *Curr. Opin. Struct. Biol*, 11:39–46, 2001.
57. MS Lee, GP Gippert, KV Soman, DA Case, and PE Wright. Three-dimensional solution structure of a single zinc finger DNA-binding domain. *Science*, 245(4918):635–637, 1989.
58. NP Pavletich and CO Pabo. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science*, 252(5007):809–817, 1991.

59. Scot A. Wolfe, Lena Nekludova, and Carl O. Pabo. Dna recognition by cys2his2 zinc finger proteins. *Annual Review of Biophysics and Biomolecular Structure*, 29(1):183–212, 2000.
60. Fyodor D. Urnov, Jeffrey C. Miller, Ya-Li Lee, Christian M. Beausejour, Jeremy M. Rock, Sheldon Augustus, Andrew C. Jamieson, Matthew H. Porteus, Philip D. Gregory, and Michael C. Holmes. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature*, 435(7042):646–651, June 2005.
61. A. Abagyan and M. Totrov. Ab initio folding of peptides by the optimal-bias monte carlo minimization procedure. *J Comput Phys*, 402-412:151, 1999.
62. A. Nagadoi, K. Nakazawa, H. Uda, K. Okuno, T. Maekawa, S. Ishii, and Y. Nishimura. Solution structure of the transactivation domain of atf-2 comprising a zinc finger-like subdomain and a flexible subdomain. *J. Mol. Biol.*, 287:593–607, 1999.
63. J. N. Onuchic, Z. Luthey-Schulten, and P.G. Wolynes. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.*, 48:545–600, 1997.
64. J. H. Laity, H. J. Dyson, and P. E. Wright. Dna-induced alpha-helix capping in conserved linker sequences is a determinant of binding affinity in $cys_2 - his_2$ zinc fingers. *J. Mol. Biol*, 295:719–727, 2000.
65. A. Verma, A. Schug, K. H. Lee, and W. Wenzel. Basin hopping simulations for all-atom protein folding. *J. Chem. Phys.*, 124:044515, 2006.
66. A. Quintilla, E. Starikov, and W. Wenzel. De novo folding of two-helix potassium channel blockers. *J. Chem. Theory and Computation.*, 3:1183–92, 2007.
67. A. Schug, T. Herges, and W. Wenzel. All atom folding of the three helix hiv accessory protein with an adaptive parallel tempering method. *Proteins*, 57(4):792–798, 2004.
68. A. Verma and W. Wenzel. Towards an all-atom free-energy forcefield for protein folding. in preparation, 2006.
69. Abhinav Verma, Srinivasa M. Gopal, Jung S. Oh, Kyu H. Lee, and Wolfgang Wenzel. All-atom de novo protein folding with a scalable evolutionary algorithm. *J. Comp. Chem*, 28:2552–2558, 2007.
70. Srinivasa M. Gopal and Wolfgang Wenzel. De novo folding of the dna-binding atf-2 zinc finger motif in an all-atom free-energy forcefield. *Angewandte Chemie International Edition*, 45(46):7726–7728, 2006.
71. A. Verma and W. Wenzel. Protein structure prediction by all-atom free-energy refinement. *BMC Structural Biology*, 7:12, 2007.
72. A. Schug and W. Wenzel. Reproducible folding of a four helix protein in an all-atom forcefield. *J. Am. Chem. Soc.*, 126(51):16736–16737, 2004.