Stochastic Optimization Methods for Protein Folding

ALEXANDER SCHUG, ABHINAV VERMA, KYU HWAN LEE, WOLFGANG WENZEL

Forschungszentrum Karlsruhe,
Institut für Nanotechnologie,
P.O. Box 3640, 76021 Karlsruhe,
GERMANY
Forschungszentrum Karlsruhe,
Institut für Nanotechnologie,
P.O. Box 3640, 76021 Karlsruhe,
GERMANY
Supercomputing Materials Laboratory
Korean Institute for Science and Technology
Seoul
KOREA
http://www.fzk.de/biostruct

Abstract: - We recently developed an all-atom free energy forcefield (PFF01) for protein structure prediction with stochastic optimization methods. We demonstrated that PFF01 correctly predicts the native conformation of several proteins as the global optimum of the free energy surface. Here we review recent folding studies, which permitted the reproducible all-atom folding of the 20 amino-acid trp-cage protein, the 40-amino acid three-helix HIV accessory protein and the sixty amino acid bacterial ribosomal protein L20 with a variety of stochastic optimization methods. These results demonstrate that all-atom protein folding can be achieved with present day computational resources for proteins of moderate size.

Keywords: protein folding, stochastic optimization

1. Introduction

Ab-initio protein tertiary structure prediction (PSP) and the elucidation of the mechanism of the folding process are among the most important outstanding problems of biophysical chemistry [1,2]. The many complementary proposals for PSP span a wide range of representations of the protein conformation, ranging from coarse grained models to atomic resolution. The choice of representation often correlates with the methodology employed in structure prediction, ranging from empirical potentials for coarse grained models [3,4] to complex atom-based potentials that directly approximate the physical interactions in the system. The latter offer insights into the mechanism of protein structure formation and promise better transferability, but their use incurs large computational costs that has confined all-atom protein structure prediction to all but the smallest peptides [5,6].

It has been one of the central paradigms of protein folding that proteins in their native conformation are in thermodynamic equilibrium with their environment [7]. Exploiting this characteristic the structure of the protein can be predicted by locating the global minimum of its free energy surface without recourse to the folding dynamics, a process which is potentially much more efficient than the direct simulation of the folding process. PSP based on global optimization of the free energy may offer a viable alternative approach, provided that suitable parameterization of the free energy of the protein in its environment exists and that global optimum of this free energy surface can be found with sufficient accuracy [8].

We have recently demonstrated a feasible strategy for all-atom protein structure prediction [9,10,11] in a minimal thermodynamic approach. We developed an all-atom free-energy forcefield for proteins (PFF01), which is primarily based on physical interactions with important empirical, though sequence independent, corrections [11]. We already demonstrated the reproducible and predictive folding of four proteins, the 20 amino acid trp-cage protein (1L2Y) [9,12], the structurally conserved headpiece of the 40 amino acid HIV accessory protein (1F4I) [10,13] and the sixty amino acid bacterial ribosomal protein L20 [14]. In addition we showed that PFF01 stabilizes the native conformations of other proteins, e.g. the 52 amino-acid protein A [5,15], and the engrailed homeodomain (1ENH) from Drosophilia melangaster [16].

2. Forcefield

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 [17,10,11]. In the folding process at physiological conditions the degrees of freedom of a peptide are confined to rotations about single bonds. The forcefield is parameterized with the following non-bonded interactions:

$$V(\{\vec{r}_i\}) = \sum_{ij} V_{ij} \left[\left(\frac{R_{ij}}{r_{ij}} \right)^{12} - \left(\frac{2R_{ij}}{r_{ij}} \right)^{6} \right] \quad (1)$$

$$+ \sum_{ij} \frac{q_i q_j}{\varepsilon_{g(i)g(j)} r_{ij}} + \sum_i \sigma_i A_i + \sum_{\text{hbonds}} V_{hb}.$$

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 as a set of 138 proteins of the PDB database [18,17,19]. The non-trivial electrostatic interactions in proteins are represented via group-specific dielectric constants ($\varepsilon_{g(i),g(j)}$ depending on the amino-acid to which atom i belongs). The partial charges q_i and the dielectric constants were derived in a potential-of-mean-force approach [20]. Interactions with the solvent were first fit in a minimal solvent accessible surface model [21] parameterized by free energies per unit area σ_i to reproduce the enthalpies of solvation of the Gly-X-Gly family of peptides [22]. A_i is the area of atom i that is in contact with a ficticious 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 [11].

3. Optimization Methods

The low-energy free energy landscape of proteins is extremely rugged due to the comparatively close packing of the atoms in the native structure. Suitable optimization methods must therefore be able speed the simulation by avoiding high energy transition states, adapt large scale move or accept unphysical intermediates. Here we report on four different optimization methods, the stochastic tunneling method [23], the basin hopping technique [24,25], the parallel tempering method [26,27] and a recently employed evolutionary technique. The stochastic tunneling method and the basin hopping approach are an inherently sequential algorithms, which evolve a single configuration according to a given stochastic process. In contrast, parallel tempering and evolutionary techniques are inherently parallel optimization strategies that are well suited to presently available multiprocessor architectures with low bandwidth connections. Since all-atom protein structure prediction remains a computationally challenging problem it is important to search for suitable optimization methods that are capable to exploit such architectures, i.e. a high degree of parallelism with little communication is desirable.

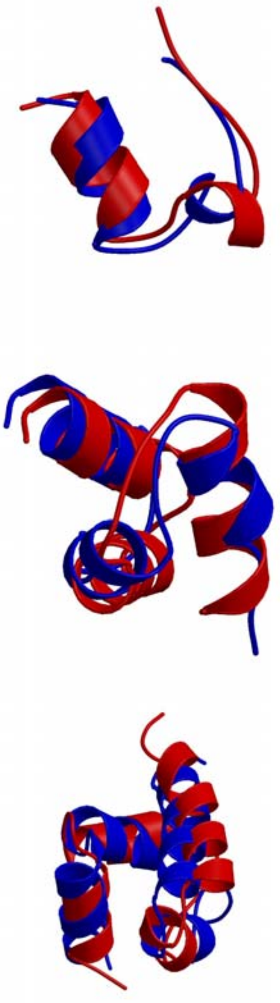

Figure 1. Overlay of the native(red) and folded (blue) structures of trp-cage protein [28], the HIV accessory protein [13] and the bacterial ribosomal protein L20 [14].

4. Results

4.1. The trp-cage protein

Using the PFF01 forcefield we simulated 20 independent replicas of the 20 amino acid trp-cage protein [29,6] (pdb code 1L2Y) with a modified versions of the stochastic tunneling method [23,9]. Six of 25 simulations reached an energy within 1 kcal/mol of the best energy, all of which correctly predicted the native experimental structure of the protein (see Fig 1 (top)). We find a strong correlation between energy and RMSD deviation to the native structure for all simulations. The conformation with the lowest energy had a backbone root mean square deviation of 2.83 Å.

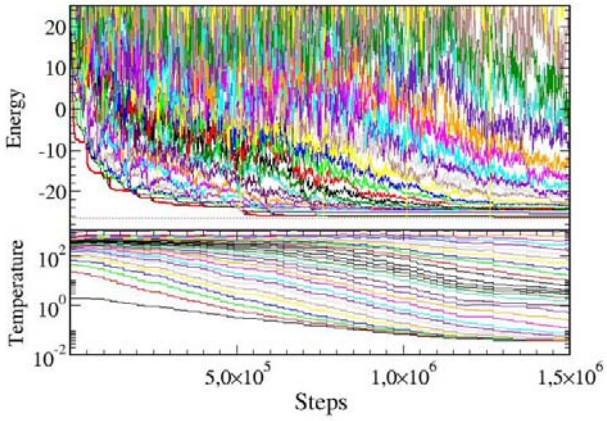

Figure 2. Energies (upper panel) and temperatures (lower panel) of the 30 replica modified parallel tempering simulation of the trp-cage protein reported in the text. The dotted line in the upper panel corresponds to the estimate of the global optimum of the free energy (obtained independently). The lower panel demonstrates a rapid equilibration of the temperatures during the simulation..

Table 1: Distance to native structure and energies of the best decoys for the HIV accessory protein

RMSB	ENERGY
0.00	
2.34	-119.54
2.41	-117.52
2.76	-116.25
2.40	-115.85
2.43	-114.67
6.48	-114.06
2.57	-113.65
4.61	-107.72
4.14	-106.29
5.92	-103.88

The good agreement between the folded and the experimental structure is also evident from Figure (1)(center), which shows the secondary structure

We also folded this protein with the parallel tempering method [12]. Figure (2) shows the energies and corresponding temperatures for a simulation using thirty replicas. The temperature adjustment scheme results in a temperature distribution that permits frequent exchange of replicas and significantly speeds convergence. The best final structure associated with the lowest temperature in the simulation with 30 replicas had a RMSB deviation of 2.01 Å.

Finally we have folded the trp-cage protein protein with the basin hopping technique. Very high starting temperatures above 600 K are required to permit a sufficient exploration of the free energy surface. the length increased with the square root of the cycle number, we found much lower energies for the latter after investing the same total number of function evaluations in each run. Using the basin hopping method with a starting temperature of $T_s = 800K$ and a final temperature of $T_f = 3K$ the lowest six of 20 simulations converged to A total of 12 of these simulations approached the native conformation as its estimate of the optimum. While all methods correctly identify the folding funnel, the basin hopping approach results in the lowest energies. Note that the second best simulation has an RMSB of only 1.8Å to the native conformation and loses in energy with less the 0.5 kcal/mol.

4.2 The HIV accessory protein

We also applied a the modified basin hopping or Monte-Carlo with minimization (MCM) strategy [8,25] to fold the structurally conserved 40-amino acid headpiece of the HIV accessory protein [10]. We performed twenty independent simulations and found the lowest five to converge to the native structure (see Table 1) [14]. The first non-native decoy appears in position six, with an energy deviation of 5 kcal/mol and a significant RMSB deviation. The table demonstrates that all low-energy structures have essentially the same secondary structure, i.e. position and length of the helices are always correctly predicted, even if the protein did not fold correctly.

alignment of the native and the folded conformations. The good physical alignment of the helices illustrates the importance of hydrophobic contacts to correctly fold this protein. An independent measure to assess the quality of these contacts is to compare the C_β-C_β distances (which correspond to the NOE constraints of the NMR experiments that determine tertiary structure) in the folded structure to those of the native structure. We found that 66 % (80 %) of the C_β-C_β distance distances agree to within one (1.5) standard deviations of the experimental resolution. We also performed a simulation of the HIV accessory protein using the adapted parallel tempering method [13] on 20 processors of an INTEL XEON PC cluster All simulations were started with random conformations at high temperatures to allow for rapid, unbiased relaxation of the structures The final conformation with the lowest energy/temperature had converged to within 1.23 / 2.46 Å backbone root mean square (RMSB) deviation to the best known decoy / NMR structure of the HIV accessory protein. The overlay of the experimental and the converged structure (see Figure (1)) demonstrates the good agreement between the conformations, the difference in NOE constraints demonstrates that not only short range, but also long range distances are correctly predicted.

Table 2: Energies and root-mean square deviations for the 10 best decoys of the bacterial ribosomal protein L20.

Energy	RMSB
-167.87	4.64
-166.15	8.25
-165.91	4.41
-164.11	5.54
-163.99	3.79
-163.93	4.04
-163.45	8.52
-163.20	4.37
-162.67	5.55
-162.52	3.78

4.3 Bacterial Ribosomal Protein L20

For the 60 amino acid bacterial ribosomal protein L20 (pdb-code 1GYZ) we experimented with an evolutionary technique. Starting from a seed population of random structures we performed the folding simulation in three phases: (1) generation of starting structures of the population, (2) evolutionary improvement of the population and (3) refinement of the best resulting structures to ensure convergence.

The energies and structural details of the best ten resulting conformations are summarized in Table 2. Again the best conformation had approached the native conformation to about 4.6 Å RMSB deviation. In total six of the lowest ten conformations approach the native structure, while four others misfolded. The final population contains in excess of 20% of near native conformations, its native content increased sixty-fold during the simulation.

5. Summary

Since the native structure dominates the low-energy conformations arising in all of these simulation, our results demonstrate the feasibility of all-atom protein tertiary structure prediction for three different proteins ranging from 20-60 amino acids in length with a variety of different optimization methods. The free energy approach thus emerges as viable trade-off between predictivity and computational feasibility. While sacrificing the folding dynamics, a reliable prediction of its terminus, the native conformation — which is central to most biological questions — can be achieved.

The computational advantage of the optimization approach stems from the possibility to visit unphysical intermediate conformations with high energy during the search. This goal is realized with different mechanism in all of the employed stochastic optimization methods. In the stochastic tunnelling method the nonlinear transformation of the PES permits the dynamical process to traverse arbitrarily high energy barriers at low temperatures, in basin hopping and parallel tempering, simulation phases at very high temperatures accomplish the same objective.

This review indicates that all-atom protein structure prediction with stochastic optimization methods becomes feasible with present-day computational resources. The fact that three proteins were reproducibly folded with different optimization methods to near-native conformation increases the confidence in the parameterization of our all-atom protein forcefield PFF01. The presently available evidence indicates that the comparatively straightforward basin hopping routine is a good work horse to evolve individual conformations. The resolution of several independent basin hopping simulations may be enhanced by the use of evolutionary algorithms such as the one used for the bacterial ribosomal protein L20 While the present results demonstrate proof of principle, much work, remains to be done to arrive at an optimal strategy.

References

[1] D. Baker and A. Sali. Protein structure prediction and structural genomics. Science, 294:93–96, 2001.

[2] J. Schonbrunn, W. J. Wedemeyer, and D. Baker. Protein structure prediction in 2002. Curr. Op. Struc. Biol., 12:348–352, 2002.

[3] N. Go and H. A. Scheraga. On the use of classical statistical mechanics in the treatment of polymer chain conformation. Macromolecules, 9:535–542, 1976.

[4] P. Ulrich, W. Scott, W.F. W. F. van Gunsteren, and

A. E. Torda. Protein structure prediction forcefields: Paramterization with quasi newtonian dynamics. Proteins, SF&G, 27:367–384, 1997.

[5] C. D. Snow, H. Nguyen, V. S. Pande, and M. Gruebele. Absolute comparison of simulated and experimental protein folding dynamics. Nature, 420:102–106, 2002.

[6] C. Simmerling, B. Strockbine, and A. Roitberg. All-atom strucutre prediction and folding simulations of a stable protein. J. Am. Chem. Soc., 124:11258–11259, 2002.

[7] C. B. Anfinsen. Principles that govern the folding of protein chains. Science, 181:223–230, 1973.

[8] Z. Li and H.A. Scheraga. Monte carlo minimization approach to the multiple minima problem in protein folding. Proc. Nat. Acad. Sci. U.S.A., 84:6611–6615, 1987.

[9] A. Schug, T. Herges, and W. Wenzel. Reproducible protein folding with the stochastisc tunneling method. Phys. Rev. Letters, 91:158102, 2003.

[10] T. Herges and W. Wenzel. Reproducible in-silico folding of a three-helix protein in a transferable all-atom forcefield. Phys. Rev. Letters, 94:018101, 2004.

[11] T. Herges and W. Wenzel. An All-Atom Force Field for Tertiary Structure Prediction of Helical Proteins. Biophys. J., 87(5):3100–3109, 2004.

[12] A. Schug, T. Herges, and W. Wenzel. All-atom folding of the trp-cage protein in an all-atom forcefield. Europhyics Lett., 67:307–313, 2004.

[13] A. Schug, T. Herges, and W. Wenzel. All-atom folding of the three-helix hiv accessory protein with an adaptive parallel tempering method. Proteins, 57:792–798, 2004.

[14] A. Schug, T. Herges, and W. Wenzel. Predictive in-silico all-atom folding of a four helix protein with a free-energy model. J. Am. Chem. Soc., 126:16736–7, 2004.

[15] H. Gouda, H. Torigoe, A. Saito, M. Sato, Y. Arata, and I. Shimanda. Three-dimensional solution structure of the b domain of staphylococcal protein a: comparisons of the solution and crystal structures. Biochemistry, 40:9665–9672, 1992.

[16] U. Mayor, N. R. Guydosh, C. M. Johnson, J. G. Grossmann, S. Sato, G. S. Jas, S. M. V. Freund, D. O. V. Alonso, V. Daggett, and A. R. Fersht. The complete folding pathway of a protein from nanoseconds to microseconds. Nature, 421:863–867, 2003.

[17] T. Herges, H. Merlitz, and W. Wenzel. Stochastic optimisation methods for biomolecular structure prediction. J. Ass. Lab. Autom., 7:98–104, 2002.

[18] R. Abagyan and M. Totrov. Biased probability monte carlo conformation searches and electrostatic calculations for peptides and proteins. J. Molec. Biol., 235:983–1002, 1994.

[19]T. Herges, A. Schug, B. Burghardt, and W. Wenzel. Exploration of the free energy surface of a three helix peptide with stochastic optimization methods. Intl. J. Quant. Chem., 99:854–893, 2004.

[20] F. Avbelj and J. Moult. Role of electrostatic screening in determining protein main chain conformational preferences. Biochemistry, 34:755–764, 1995.

[21] D. Eisenberg and A. D. McLachlan. Solvation energy in protein folding and binding. Nature, 319:199–203, 1986.

[22] K. A. Sharp, A. Nicholls, R. Friedman, and B. Honig. Extracting hydrophobic free energies from experimental data:relationship to protein folding and theoretical models. Biochemistry, 30:9686–9697, 1991.

[23] W. Wenzel and K. Hamacher. Stochastic tunneling approach for global optimization of complex potential energy landscapes. Phys. Rev. Lett., 82:3003, 1999.

[24] 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.

[25] J. P.K. Doye and D. Wales. On potential energy surfaces and relaxation to the global minimum. J. Chem. Phys., 105:8428, 1996.

[26] G. J. Geyer. Stat. Sci., 7:437, 1992.

[27] K. Hukushima and K. Nemoto. Exchange monte carlo method and application to spin glass simulations. Journal of the Physical Society of Japan, 65:1604–1608, 1996.

[28]A. Schug, A. Verma, T. Herges, and W. Wenzel. Comparison of stochastic optimization methods for all-atom folding of the trp-cage protein. submitted to Proteins, 2005.

[29] J. W. Neidigh, R. M. Fesinmeyer, and N. H. Anderson. Designing a 20-residue protein. Nature Struct. Biol., 9:425–430, 2002.