Stochastic Optimization Methods for Biomolecular Structure Prediction

Thomas Herges and Wolfgang Wenzel

Research Center Karlsruhe
Institute for Nanotechnology
P.O. Box 3640, D-76021 Karlsruhe, Germany

ABSTRACT

Protein structure prediction (PSP) remains one of the main outstanding challenges of theoretical biophysical chemistry. Its goal is to predict the fully three-dimensional tertiary structure of the protein on the basis of its amino acid sequence alone. Here we report on the development of a *specific biomolecular forcefield* with an implicit solvent model and *novel simulation methods* that enhance the simulation speed by several orders of magnitude. We report on the progress of two folding studies for the 36 amino-acid avian pancreatic polypeptide (1PPT) and the 42 amino-acid headgroup of the HIV-1 accessory protein (1F4I).

Keywords: protein structure prediction, protein folding, stochastic optimization

1 INTRODUCTION

While protein sequencing techniques have made enormous progress in the last decade, experimental methods for protein structure determination are orders of magnitude more involved and more expensive than sequencing techniques. Unfortunately, sequence information alone is often insufficient to elucidate the biological function or mechanism of a protein [1]. Information regarding the unique three-dimensional 'native' structure which most proteins spontaneously assume, is a prerequisite for their proper function. Although their number is steadily growing, the protein database (PDB), presently contains only about 13,000 spatially resolved structures[2]. The large pool of available, but unsequenced proteins is likely to contain a wealth of important biological and biomedical information [3].

There are many questions regarding the details of the function of proteins, in particular those of dynamical nature, that are difficult to address experimentally at the present time. Many of these problems, as well as questions regarding protein-protein association or protein-ligand interactions, would benefit from accurate theoretical methods for PSP. In particular simulation techniques addressing protein-ligand interactions would contribute significantly to the development of new pharmaceutical agents for a variety of diseases[4].

2 METHODOLOGY

Heuristic methods based on sequence homology, as well as traditional simulation strategies based on standard forcefields have so far proven insufficient to generically predict the structure of many naturally occurring proteins. The most rigorous approach, the simulation of the folding process of the protein in the solvent, requires enormous computational resources [5] and has not proven successful for all but small peptides[6].

One may therefore ask whether it is possible to devise alternate, more efficient simulation strategies. An obvious starting point is the elimination of the explicit treatment of the solvent molecules [7], which often consumes the majority of the numerical effort associated with the simulation of the overall system. Upon closer inspection of this approximation, we find that the introduction of an implicit solvent model has far deeper implications on PSP than the obvious reduction of the computational effort resulting from the reduction of the degrees of freedom of the simulation. We note that the overwhelming majority of the entropic contribution to the folding process are solvent contributions, mediated by the hydrophobic and hydrophilic effects of the different amino acid side chains[1]. Incorporating these terms into an implicit solvent model we obtain in conjuction with the internal energy of the protein a good model for the total *free energy* of the system[8]. As indicated above most proteins attain a unique stable native structure. If the protein is in thermodynamic equilibrium with its environment, this structure must therefore correspond to the global minimum of its free energy surface. As is well known from the simulation of many physical systems with complex dynamics, it is possible to locate the thermodynamically stable state of the system using *stochastic optimization methods* without recourse to its dynamics *orders of magnitude faster* than in a traditional MD simulation [9].

To implement this approach to PSP, one must develop a suitable forcefield to describe the internal energy of a protein with an adequate implicit solvent model. In addition one needs efficient global optimization methods that are able to reliably locate the global minimum of the resulting free energy landscape of the protein. In the past several years we have implemented such a strategy,

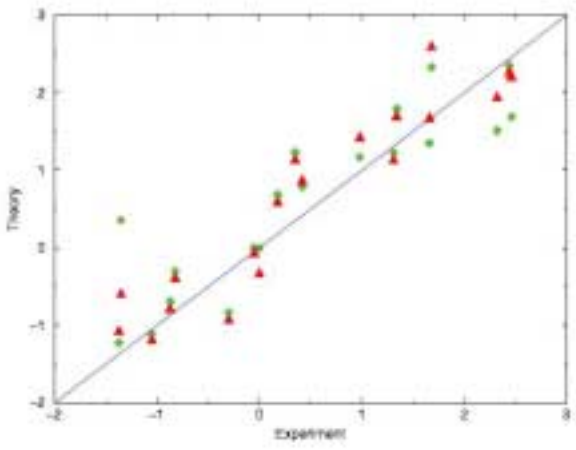

Figure 1: Correlation between the free energies of solvation between experimental data fir Gly-X-Gly and two solvent accessible surface area based models (in units of kcal/mol) that differ in the number of atom groups used in the fit. The INT forcefield uses the fit indicated by the red triangles with an RMS error of less than 0.5 kcal/mol.

developing both forcefields and stochastic optimization methods suitable to this task. In the following sections we describe the ingredients of this approach and give an overview of our results.

2.1 Biomolecular Forcefield

Over the last decades many forcefields [10]–[13] have been developed to investigate numerous phenomena in physical, organic and inorganic chemistry. The difficulties encountered in PSP justify the development of specific forcefields for the following reasons: By exploiting the fact that only a limited number of building blocks occur, their ingredients may be specifically adapted to provide a more accurate description of the system. Secondly, we are interested only in the low-energy conformations of the model. As a result, many degrees of freedom that are associated with covalent interactions, e.g. bond stretching, may be neglected. The degrees of freedom considered are only rotations about the dihedral angles of the backbone and of freely rotatable single bonds of the sidechains. This reduction of the number of degrees of freedom leads to a dramatic increase in the efficiency of the simulation.

The INT forcefield was specifically adapted to proteins and peptides by: (i) fitting the van-der-Waals parameters extracted from the PDB database, (ii) the use of group-specific dielectric constants [14] to account for the nontrivial electrostatic interactions in the interior of proteins, (iii) environment dependent partial charges that model changes in the acid-base equilibria encountered in comparing folded and unfolded protein conformations, (iv) novel models for backbone-backbone hydrogen bonding that better reproduce secondary structure elements of a selected set of proteins in the PDB and (v) implicit area based solvent model (see Figure 1) to describe the interactions of the protein with its environment [7]. Each of these ingredients represents an attempt to model the complex underlying physics of the protein with the simplest and computationally most efficient approximation. In combination with efficient optimization techniques they open a perspective to predict the structure of proteins and peptides at the all-atom level with present-day computational resources.

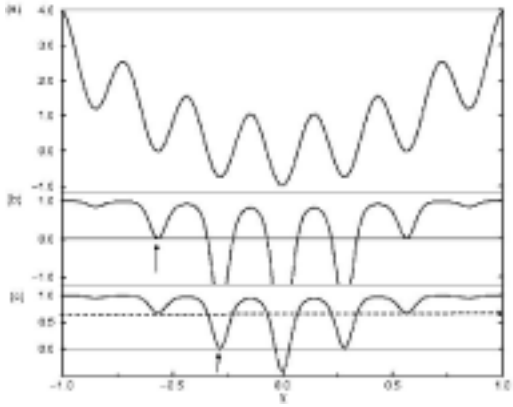

Figure 2: Schematic one dimensional potential energy surface and its transformations under the STUN procedure, provided that the local minima indicated by the arrows have been found. Part (a) shows the original potential energy surface, parts (b) and (c) the transformed PES under the assumption that the minima indicated by the arrow are the best configurations found so far in the simulation, respectively.

2.2 Stochastic Optimization Methods

Stochastic optimization methods are now being used in a multitude of applications, ranging from circuit design on silicon wafers to airline flight schedules. Their objective is to minimize a given cost function that depends on a large number of discrete or continuous variables[15], [16]. The degree of difficulty in stochastic optimization depends strongly on the number of degrees of freedom and the complexity of the PES.

In this study we have used the simulated annealing method (SA) [17], the parallel tempering method (PT) [18], [19] and the stochastic tunneling method (STUN)[20]. The latter incorporates the ability to escape metastable states by letting the particle in the minimization process "tunnel" forbidden regions of the PES. We retain the idea of a biased random walk, but apply a non-linear transformation to the potential energy surface:

$$E_{\mathrm{STUN}}(x) = 1 - \exp\left[-\gamma(E(x) - E_0)\right] \qquad (1)$$

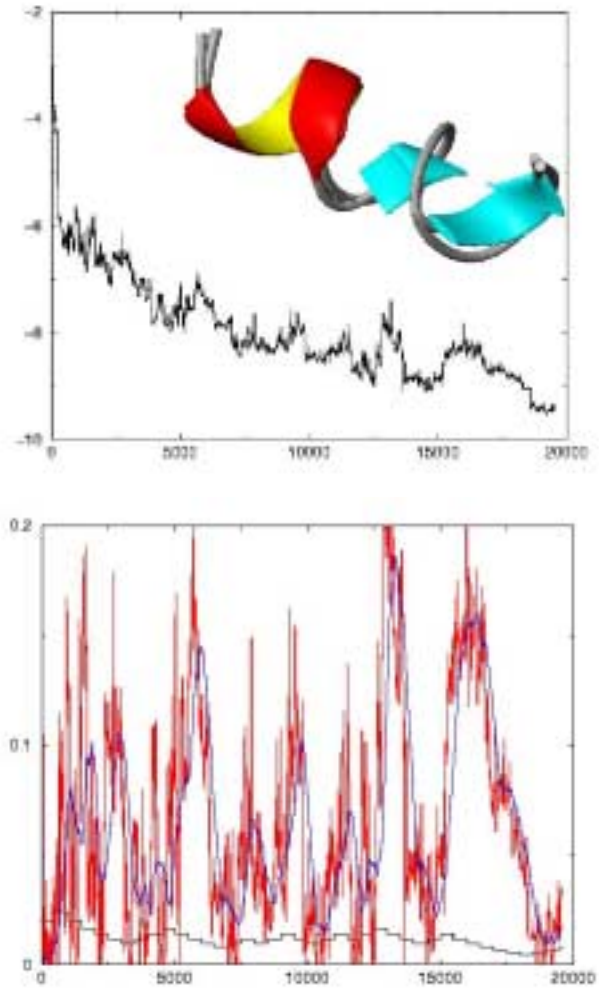

Figure 3: Folding of a 13 amino acid helix fragment of 1HRC (Residues: 92-105) with STUN. The top of the figure shows the total energy of the system as a function of the number of simulation steps and an overlay of the resulting structure and the NMR structure (inset) . The lower part shows the effective energy (red), its moving average (dashed) and the effective inverse temperature of the STUN procedure.

where E_0 is the lowest minimum encountered by the dynamical process so far (see Figure 2). This effective potential preserves the locations of all minima, but maps the entire energy space from E_0 to the maximum of the potential onto the interval $[0, 1]$. At a given finite temperature of O(1), the dynamical process can therefore pass through energy barriers of arbitrary height, while the low energy-region is resolved even better than in the original potential. The degree of steepness of the cutoff is controlled by the tunneling parameter γ.

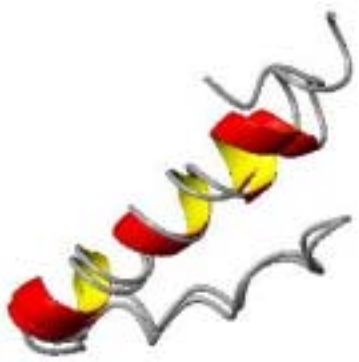

Figure 4: Overlay of the simulated and NMR structure of 1PPT

3 RESULTS

We have first investigated the folding of small peptide fragments that are believed to assume a unique three dimensional structure even when removed from their environment in the protein. The inset of figure 3 shows the overlay of the crystal structure of a helical 13 amino-acid residue fragment of the 1HRC protein with the structure we have obtained in STUN simulations. Encouragingly, the backbone configurations of these two structures are identical to better than experimental resolution. Figure 3 (a) shows the evolution of the total energy of the structure from an unfolded configuration to the folded configuration as a function of the number of energy evaluations. Figure 3(b) shows the effective energy and the effective temperature. Several heating an cooling cycles were required to fold the helix fragment and "tunneling phases" that occur when the effective energy is relatively high significantly aided the search process. In these phases the original energy of the system undergoes significant fluctuations that are much larger in magnitude than the difference in energy of two successive metastable states. Circumnavigating these energy barriers in a traditional simulation would significantly slow the optimization process. We conducted several dozen STUN runs for this, as well as for other fragments that were investigated to verify that the structure we had obtained corresponds to the global optimum of the system. For 1HRC we found no competing structures with either PT or SA. We noted that in SA the helix could not be folded even with a tenfold increase of the computational effort. Hence STUN appears to present a viable and efficient optimization strategy to optimize peptide fragments of this length.

Using PT simulations we attempted to fold the autonomously folding 36 amino-acid avian pancreatic peptide (1PPT) [21] by varying the relative strength of the solvent contributions in the forcefield. Depending on the

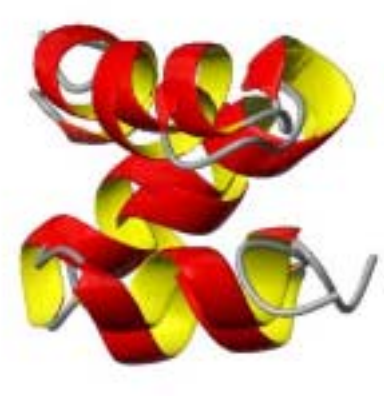

Figure 5: Overlay of the simulated and NMR structure of residues 1-42 of the three-helix HIV-1 accessory protein (1F4I).

value of this parameter, we can stabilize completely helical (no solvent) to completely collapsed configurations (unphysically strong solvent interactions). For an appropriate choice, a very good overlap between the simulated and the NMR structure of this peptide can be obtained (see Figure 4). Finally we have attempted to fold the three-helix HIV-1 accessory protein (1F4I) [22]. We conducted several runs and found the structure depicted in Figure 5 among the most stable, but not as the optimal structure in our simulations. This suggests a rational decoy strategy to systematically improve the forcefield the we presently implement. We generate a large set of "good" candidates that compete with the NMR structure. As long as one of these decoys has a better energy than the native configuration, the forcefield must be modified to stabilize the native configuration in comparison to all other decoys. When this is achieved we generate new decoys by refolding the peptide, generating either new configurations that are yet again better in energy than the NMR structure or ultimately folding the peptide. This strategy is presently implemented in ongoing work.

4 Summary and Conclusions

We have motivated the use of stochastic optimization methods as a technique to predict the structure of complicated biomolecules. To implement this approach, a forcefield that parameterizes the free energy of the underlying model must be developed, such a forcefield must contain an implicit parameterization of the interactions of the biomolecule with the solvent. We have argued that there is a rational, decoy-based strategy to develop a biomolecular forcefield that can be used to predict the structure of short peptide fragments using stochastic optimization techniques such as the stochastic tunneling method. We have illustrated the success of this approach in the folding of short peptide fragments and autonomously folding peptides and proteins. Stochastic optimization methods permit an analysis of this problem and a systematic strategy for the improvement of the forcefield several orders of magnitude faster than competing simulation techniques.

Acknowledgments: We acknowledge the support and many useful discussions with Susan Gregurick and John Moult in the course of this work. This work was funded by the Deutsche Forschungsgemeinschaft (We 1863/11-1), the BMBF and the Bode foundation.

REFERENCES

[1] C. Branden and J. Tooze. *Introduction to Protein Folding.* Garland, 2nd edition, 1999.
[2] RCBS data bank: http://www.rcsb.org/pdb, 2001.
[3] D. Baker and A. Sali. *Science*, 294:93, 2001.
[4] K. Gubernator, editor. *Structure Based Ligand Design.* Wiley, 1998.
[5] Y. Duan and P. A. Kollman. *Science*, 23:740, 1998.
[6] X. Daura, et. al. *J. Molc. Biologyd*, JMB:925, 1998.
[7] D. Eisenberg and A.D. McLachlan. *Nature*, 319:199, 1986.
[8] M. Daune. *Molecular Biophysics: Structures in Motion.* Oxfird Scientific, 1999.
[9] U Hansmann and Y. Okamoto. *Curr. Op. Struct. Biol.*, 9:177, 1999.
[10] W.F. van Gunsteren and H.J.C. Berendsen. The Groningen Molecular Simulation Manual Groningen University, 1987.
[11] MacKerell Jr. et. al. *J. Phys. Chem.*, B102:3586, 1998.
[12] W. L. Jorgensen and N. A. McDonald. *J. Mol. Struct.*, 424:145, 1998.
[13] Y. Duan, L. Wang, and P.A. Kollman. *Proc. Nat. Acad. Science USA*, 95:9897, 95.
[14] F. Avbelj and J. Moult. *Biochemistry*, 34:755, 1995.
[15] C. L. Brooks, J. N. Onuchic, and D. J. Wales. *Science*, 293:612, 2001.
[16] D.J. Wales and H.A. Scheraga. *Science*, 285:1368, 1999.
[17] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. *Science*, 220:671–680, 1983.
[18] A. P. Lyubartsev et. al. *J. Chem. Phys.*, 96:1776, 1992.
[19] E. Marinari and G. Parisi. *Europhysics Letters*, 451:1992, 19.
[20] W. Wenzel and K. Hamacher. *Phys. Rev. Lett.*, 82:3003, 1999.
[21] T.L. Blundell et. al. *Proc. Natl. Acad. Sciences (USA)*, 7:4175, 1981.
[22] E. S. Withers-Ward, T.D. Mueller, I.S. Chen, and J. Feigon. *Biochemistry*, 39:14103, 2000.