Predictive in Silico All-Atom Folding of a Four-Helix Protein with a Free-Energy Model

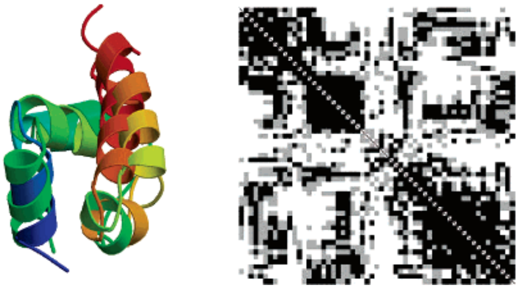
Alexander Schug and Wolfgang Wenzel*

Forschungszentrum Karlsruhe, Institute for Nanotechnology, P.O. Box 3640, 76021 Karlsruhe, Germany

Available genomic and sequence information for proteins contains a wealth of biomedical information that becomes accessible when translated into three-dimensional structure.[1] The development of predictive first principles all-atom folding methods would significantly benefit the understanding of protein function, association, and dynamics but is complicated by the associated large computational costs, which have been argued to grow exponentially with the number of amino acids.

Here we report the predictive folding of the 60 amino acid four-helix bacterial ribosomal protein L20 (BRPL20),[2] the largest protein predictively folded to date, in an all-atom free-energy force field. Our free-energy approach is based on the thermodynamic hypothesis[3] that for many proteins the thermodynamic equilibrium and native conformations coincide. The native state thus corresponds to the global minimum of the free-energy landscape,[4] which can be found using stochastic optimization methods orders of magnitude faster than direct simulation of the folding pathway. We used the free-energy protein force field (PFF01),[5] which was demonstrated to stabilize the native structure of several helical proteins against independently generated decoys. PFF01 represents all atoms individually (with the exception of hydrogen in CH$_n$ groups) and models the complex electrostatic interactions in proteins with group-specific dielectric constants and contains a SASA based implicit solvent model. During the folding process, we consider only variations of the dihedral angles of the backbone and the side chains, while keeping all other angles and bond lengths fixed. With this force field, we were previously able to predict the tertiary structure of the 20-amino acid trp-cage protein,[6,7] the 36-amino acid villin headpiece,[8] and the 40-amino acid headgroup of the HIV accessory protein,[9,10] and investigations for β-sheet proteins are presently under way.

The simulations were performed with a simple evolutionary strategy using a distributed master−client model in which idle clients request a task from the master. The master maintains a list of open tasks comprising the active conformations of the population. The client then performs either a Monte Carlo (MC) or a simulated annealing (SA)[11] simulation of specified length on the conformation. Conformations are drawn randomly from the active population. When the client returns a new conformation after completing its task, the result is stored. The new conformation replaces the energetically worst conformation in the active population, provided its energy is lower than the highest energy of the population and that it differs by at least 3 Å backbone root-mean-square deviation (RMSB) from all members of the active population. If the new conformation has an RMSB of less than 3 Å to some conformation of the population, it replaces this conformation if it is lower in energy. The population was initially seeded with 100 random conformations with an average (RMSB) deviation of 12.2 Å to the NMR conformation.



The folding simulation has three phases. In the first phase, we performed high-temperature (500 K) MC simulations of 50 000 steps each with a 20% reduced solvent strength to facilitate the rapid formation of secondary structure. It has been argued that hydrophobic collapse competes with secondary structure formation in protein folding.[12] In the collapsed conformational ensemble large-scale conformational changes, such as those required for secondary structure formation, occur only rarely. The goal of this simulation phase was the generation of a wide variety of competitive starting conformations for further refinement.

At the end of this simulation, we had gathered more than 17 000 distinct decoys that were ranked according to their total energy as well as according to the individual energy terms. For each criterion, we selected the best 50 conformations and eliminated duplicates to arrive at a population of 266 starting structures for the second stage of the relaxation procedure. This population was relaxed in 14 000 SA simulations as described above. We then selected the 50 conformations best in total energy for further refinement and performed 5500 SA simulations on this subpopulation. The length of the individual relaxation simulations was gradually increased from 10^5 steps to 2.3×10^6 steps per simulation.

In the final population the energetically lowest conformation approached the native state to 4.6 Å RMSB. The good alignment of the helices illustrates the importance of hydrophobic contacts to correctly fold this protein (see Figure 1a), while the $C^\beta{-}C^\beta$ distance map (Figure 1b) shows the presence of important long-range native contacts. The approximations incurred by the use of computationally efficient electrostatic and solvation models are likely to limit the accuracy of the predicted structure, which might be further refined in all-atom explicit water simulations.[13] Separate refinement simulations of the NMR structure in PFF01 found an RMSB deviation of 1.6 Å for the best decoy,[5] but at significantly higher energy (−151.4 kcal/mol).

In total, six of the lowest 10 conformations had approached the native state (see Figure 2 and Supporting Information) from different directions in the folding funnel. Because the selection criterion for the active population precluded occurrence of the same configuration to within 3 Å RMSB, this dominance of near-native conformations of the total ensemble is particularly encouraging.

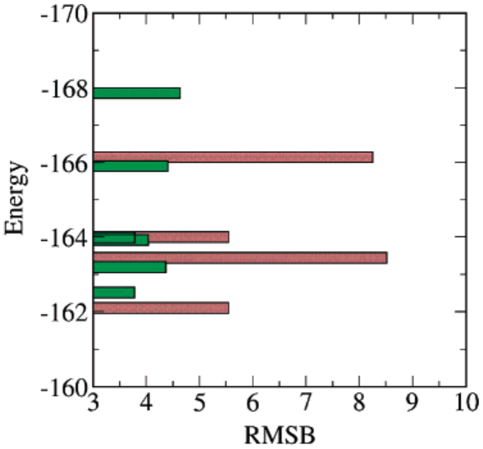


Figure 2. Energy vs RMSB deviation of the best 10 conformations of the final simulated population (see Supporting Information). Green bars indicate near-native conformations, and red bars indicate competing metastable decoys. Note that selection procedure of the evolutionary algorithm precludes the native conformation to occur more than once. All metastable conformations have nearly identical secondary structure as the native conformation, but the helices are arranged differently.

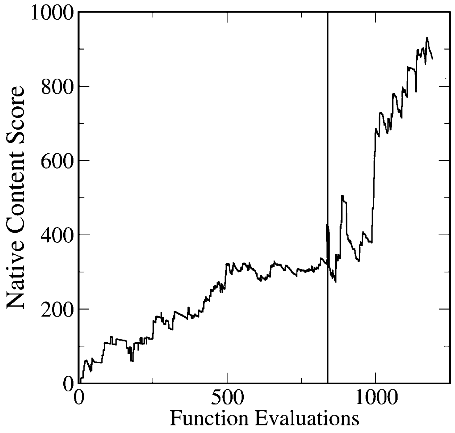


Figure 3. Native content (in arb units) as a function of the number of function evaluations (in millions). The vertical line indicates the pruning step of the population from 266 to 50 members, and the native score of the latter population was scaled to agree with the previous population before pruning.

Next we calculated the "native content" of the simulated ensemble as a weighted average of the structural deviations of the population and the native conformation (Figure 3). For a population of size N we add $100(N - \mathrm{R} + 1)/N$ for each near-native decoy ranked at position R by energy to the total native score of this population. A score of 100 thus corresponds to a native decoy placed at the top position, while a near-native decoy at the very bottom contributes just unity. Non-native conformations contribute nothing. The final population contains in excess 20% of near-native conformations, its native score exceeds 800, and the score increased more than 60-fold during the simulation.

Since the native structure dominates the low-energy conformations arising in the simulation, these results demonstrate the feasibility of all-atom protein tertiary structure prediction for BRPL20 with our free-energy optimization approach using the PFF01 force field. In total we invested 1.25×10^9 function evaluations, roughly corresponding to a 1-ms MD simulation time. Note, however, that in each move only a small fraction of the atoms move, so that the energy evaluation is much less expensive than an MD step in which all atoms move.

The free-energy approach emerges as a viable tradeoff between predictivity and computational feasibility. The computational efficiency of the optimization approach stems from the possibility to visit unphysical intermediate high-energy conformations during the search. Its application necessitates the use of approximations for the electrostatic and solvent interactions, which are known to limit the accuracy of the predicted conformations.

The methodology presented here may thus be used to generate coarse grained models of the protein tertiary structure, which can be refined in more elaborate but short explicit water simulations.[13] While sacrificing the analysis of the folding process,[14] the strength of our approach lies in the reliable prediction of the native conformation with present day computational resources from random initial conditions. The approach used here is predictive and particularly well suited to worldwide distributed computational architectures, because the optimization problem, in contrast to the simulation of the folding path, is nonsequential in nature.

Acknowledgment. We are grateful to S. Gregurick and J. Pedersen for discussions and portions of the code used in these simulations. This work was supported by the DFG and BMWF and partially performed at the KIST teraflop cluster.

Supporting Information Available: Energies and RMSB deviations and secondary structure content of the 10 lowest conformations. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Baker, D ; Sali, A *Science* **2001**, *294*, 93
(2) Raibaud, S ; Lebars, I ; Guillier, M ; Chiaruttini, C ; Bontems, F ; Rak, A ; Garber, M ; Allemand, F ; Springer, M ; Dardel, F *J. Mol. Biol* **2002**, *323*, 143
(3) Anfinsen, C B *Science* **1973**, *181*, 223
(4) Onuchic, J N ; Luthey-Schulten, Z ; Wolynes, P G *Annu. Rev. Phys. Chem.* **1997**, *48*, 545
(5) Herges, T ; Wenzel, W *Biophys. J.* **2004**, *87*, 3100
(6) Schug, A ; Herges, T ; Wenzel, W *Phys. Rev. Lett.* **2003**, *91*, 158102
(7) Schug, A ; Herges, T ; Wenzel, W *Europhys. Lett.* **2004**, *67*, 30
(8) Herges, T ; Wenzel, W Folding and Misfolding the Villin Headpiece, submitted for publication
(9) Herges, T ; Wenzel, W *Phys. Rev. Lett.*, in press; http://www arxiv org/pdf/physics/0310146
(10) Schug, A ; Herges, T ; Wenzel, W *Proteins: Struct., Funct., Genet.* **2004**, *57*, 792
(11) Kirkpatrick, S ; Gelatt, C D ; Vecchi, M P *Science* **1983**, *220*, 671
(12) Zhou, Y ; Karplus, M *J. Mol. Biol.* **1999**, *293*, 917
(13) Simmerling, C ; Strockbine, B ; Roitberg, A *J. Am. Chem. Soc* **2002**, *124*, 11258
(14) Garcia, A ; Onuchic, J *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13898

JA0453681