Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 5027–5029, May 1994
Biophysics

An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure

ALESSANDRO MONGE*†, RICHARD A. FRIESNER*†, AND BARRY HONIG†‡

*Department of Chemistry and †Center for Biomolecular Simulation, Columbia University, New York, NY 10027; and ‡Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032

Communicated by William A. Goddard, October 11, 1993

ABSTRACT **An algorithm is described to assemble the three-dimensional fold of a protein starting from its secondary structure. A reduced representation of the polypeptide chain is used together with a crude potential based on pair hydrophobicities. The method is shown to be successful in locating the native topology for two 4-α-helix bundles, myohemerythrin and cytochrome *b*-562.**

The native fold of protein molecules is the result of a delicate balance of forces involving intramolecular as well as protein–solvent interactions. Over the last 15 years, empirical force fields have been developed to describe relative molecular energies as a function of atomic positions. The prediction of the three-dimensional structure of a protein by using such force fields would require the global optimization of an extremely large number of interdependent variables and is at present an unattainable task. In recent years several protein models based on simplified chain geometries and energetics have been explored, leaving unresolved though the question whether a simplified description can be effective in predicting protein structure. Here we show that a very simple model can identify the native conformation at low resolution when protein secondary structure is specified. We describe our methodology for two 4-α-helix bundles, myohemerythrin and cytochrome *b*-562, and discuss the implications for protein folding prediction.

The search for the native structure in protein models employing empiric force fields is hindered by the multiple-minima problem (1), i.e., the existence of an astronomically large number of local minima that reduces tremendously the effectiveness of any of the searching algorithms available today. Scheraga and coworkers (ref. 2 and references therein) have considered several strategies to attack the multiple-minima problem. One possibility to get around this problem is to try to reduce its complexity by designing simplified models where coarse representations of the protein chain are used together with primitive interaction potentials encoding information about known protein structures [such as profile Hamiltonians (3), contact potentials (4), or associative-memory Hamiltonians (5)]. Obviously, simplified models can be expected only to produce low-resolution structures, which could then be refined using more detailed descriptions.

In the spirit of such a hierarchical description of protein folding, we investigated the question whether it is possible to locate the native structure by employing a reduced representation and by specifying secondary structure. Our working hypothesis is that if one assigns secondary structure, a simple interaction energy can be chosen such that the potential surface of the simplified representation of the protein with its secondary structure frozen in place has relatively few minima, one of which corresponds to the native structure. Computer simulations (employing minimization or simulated annealing) of a model of this type should then be relieved from the multiple-minima problem and be able to find the native fold.

Several previous workers have investigated the idea that protein structure can be understood by packing secondary-structure elements (6–8). Reasonable (although generally rather qualitative) results have been obtained in most of these studies, which have encouraged us to pursue the research described here. Our work differs from the papers cited above, however, in the details of the model and the computational algorithms; as is shown below, we have been able to obtain relatively accurate structures with an algorithm that is entirely automated, i.e., would function in the same way in the absence of any prior knowledge of the tertiary structure.

We use backbone atoms N, C_α, and C to represent the polypeptide chain and internal dihedral angles to describe protein conformations. Side chains are identified by the C_β atom position. In secondary-structure regions, the dihedral angles are kept fixed to their x-ray structure values. A discrete set of six pairs of ϕ, ψ angles with $\omega = 180°$ is used for residues located in loop regions (9). The interaction between residues is described by a simple hydrophobic pair potential, derived from the analysis of the protein data base by Casari and Sippl (10). Each pair interaction depends linearly on the C_β–C_β distance between residues, with a prefactor given by the sum of the structure-derived hydrophobicities for the two residues. The van der Waals core repulsion is modeled by requiring that the C_α–C_α distance between residues be larger than r_{min} (set equal to 3.8 Å in our calculations).

An algorithm to assemble a three-dimensional folded structure was optimized by implementing the following scheme. Given a protein conformation, a new one is obtained by randomly choosing one of the loops and ϕ, ψ angles for each of its residues from the set of six dihedral angle states. An approximate radius of gyration (based on the centers of the secondary structure and loop regions) is calculated and the new conformation is rejected if it is more than 25% less compact than the present one. If the new configuration is accepted, first its energy is evaluated and then, if the Metropolis test is satisfied, the chain is checked to be self-avoiding.

We applied our methodology to myohemerythrin (11) and cytochrome *b*-562 (12), two 4-α-helix bundles consisting of 118 and 106 residues, respectively. In a typical run, we start with each of the interhelix loops in an extended conformation and sample from 300,000 to 1,000,000 configurations. Quenching or simulated annealing is used to find low-energy structures. Our results indicate that in a batch of 20 runs typically four or five final structures are very close in energy to the native one, and their C_α rms deviation is between 4 and 5 Å (Table 1). The low-energy-generated structures are similar to each other, with rms deviations among them at 4–5 Å. Figs. 1 and 2 show an example of such structures for myohemerythrin and for cytochrome *b*-562, respectively. As can be observed, the basic topology of the 4-α-helix bundle is

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Table 1. Results of a batch of 20 runs for myohemerythrin

IR	Final energy, arbitrary units	Final RG, Å	RMS, Å
01 I	−121.88	15.43	14.97
02 I	−95.64	15.76	9.18
03 I	−110.54	15.16	14.59
04 I	−119.13	15.06	9.46
05 I	−117.52	14.41	8.04
06 I	−121.98	14.57	8.20
07 I	−146.60	14.25	4.71
08 I	−109.16	14.99	11.16
09 I	−130.82	14.68	9.04
10 I	−141.14	13.53	8.50
11 I	−106.38	14.88	13.06
12 I	−149.46	14.12	4.83
13 I	−110.80	14.69	9.92
14 I	−146.94	13.68	4.12
15 I	−119.82	15.06	13.78
01 C	−152.65	13.83	5.01
02 C	−95.51	14.80	10.38
03 C	−149.20	13.83	4.48
04 C	−109.25	15.39	8.89
05 C	−96.25	15.42	12.44

IR is a run identifier (I and C correspond to runs on an IBM 550 and on a Convex C210, respectively), RG is the radius of gyration, and RMS is the C_α rms deviation from the native structure. The native energy (in arbitrary units) is −157.02 and the radius of gyration is 13.77 Å. Similar results were obtained for cytochrome *b*-562.

recovered, with deviations mainly due to a tilted helix or to a twisted pair of helices (as in the case shown here). These deviations are reflected in C_α distance plots, which for the most part though show similar features to the native one.

Our ability to recover a low-resolution native structure for the two 4-α-helix bundles considered here is related to the fact that the potential surface of our model does not have a large number of local minima. In fact, the results shown in Table 1 were obtained by pure Monte Carlo quenching and did not require simulated annealing. Indirect evidence that the potential surface has relatively few minima is given by the results obtained using a biased potential, with favorable

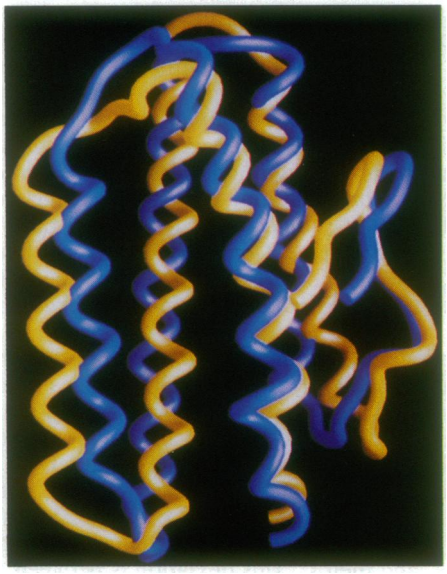

FIG. 1. Overlap of the x-ray structure of myohemerythrin, in blue, and one of the low-energy structures produced by our algorithm, in yellow. The C_α rms deviation between the two structures is 4.12 Å (if only the helical regions are considered, the C_α rms deviation is 3.87 Å).

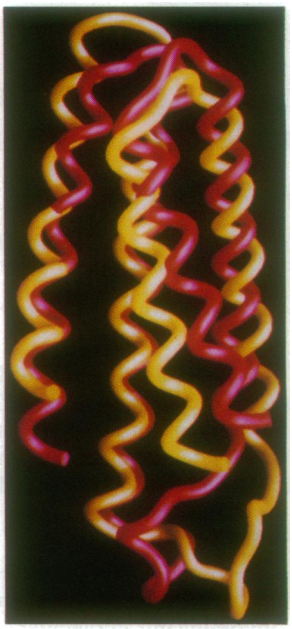

FIG. 2. Overlap of the x-ray structure of cytochrome *b*-562, in red, and one of the low-energy structures produced by our algorithm, in yellow. The C_α rms deviation between the two structures is 4.09 Å (if only the helical regions are considered, the C_α rms deviation is 3.37 Å).

contributions for pair interactions corresponding to native interhelix contacts. In this case the occurrence of native contacts in nonnative conformations is responsible for the appearance of deeper local minima, and the native structure is found only if a simulated annealing scheme is used.

An obvious limitation of the methodology we have described is that only low-resolution structures can be generated since the description of the protein chain is very coarse and specific interactions (such as hydrogen bonds) are not modeled by the simple potential used. Nonetheless, our results indicate that a potential based on relative hydrophobicities can successfully distinguish the native fold. The essential role of hydrophobic interactions in our folding procedure is demonstrated by the results of simple experiments where we randomly shuffled the sequence of myohemerythrin, maintaining the same amino acid composition. In this case, the native state is no longer the lowest-energy one, and our minimization procedure identifies structures with no resemblance to the native one that are significantly less compact. We also turned all the residues hydrophobic and observed that many local minima appear, corresponding to compact structures with a large number of hydrophobic contacts. As far as the potential is concerned, our present assumption is that the effect of the solvent can be described implicitly and that protein interactions can be expressed by an effective residue–residue force field. Such interactions can also be represented by a potential that is a sum of terms depending on the identity of each residue and its environment (3).

We recognize that the 4-α-helix bundle structure is one of the easier ones to recover for an approach such as ours and that a demonstration of generality of our algorithm will require obtaining results of similar quality for more complex topologies (e.g., including different arrangements of helices and β-sheets). However, it should be noted that the two helix bundles considered here are quite different (the rms deviation obtained from the comparison of the two native structures is 7.45 Å) and there is a large number of alternative structures that one could separate with regard to the details of helix

packing. Our methodology is able to reproduce these details quite reliably, as Figs. 1 and 2 indicate.

Simplified models for protein folding have been presented in the past by several authors. These models fall into a number of different categories. For example, Covell and Jernigan (13), Hinds and Levitt (14), Chan and Dill (15, 16), and Shakhnovich and coworkers (17) have investigated lattice models of small proteins (both actual and hypothetical) by using exact enumeration techniques and energy screening via various potential functions (e.g., contact potentials). They have demonstrated that simple potentials are capable of producing high rankings for native-like structures. While such methods have provided substantial insight into the energetic and topological basis of protein folding (although there are a number of technical difficulties, e.g., mapping lattice models onto detailed protein conformations), the computational scaling of enumeration algorithms with protein size eventually becomes exponential, rendering the study of larger proteins (including the 4-α-helix bundles considered here) impractical.

Sun (18) has investigated a reduced representation model similar to the one described in this paper, employing an explicit representation of the backbone atoms with a simplified side-chain description; the potential is also based on data-base statistics but differs in its details from the one presented herein. He uses a genetic algorithm to determine minimized structures for three small proteins, ranging from 18 to 36 residues. With the additional constraint of fixing the radius of gyration to the native value, reasonable agreement with the experimental structures is obtained in all cases. The radius of gyration constraint is certainly quite efficacious for the small proteins considered, particularly when the number of conformations satisfying the constraint is limited. However, a comparison between this approach and our strategy of fixing secondary structure needs to be made, at least from the standpoint of our objectives, in the context of larger proteins.

Skolnick and coworkers have carried out a significant number of studies of folding of small and medium size proteins ($\simeq$100 residues) using both lattice (19, 20) and off-lattice (21) models (although the backbone contains only C_α atoms explicitly) via dynamical Monte Carlo methods and simulated annealing. These studies are probably closest to our own in spirit. However, the details of the models are very different. In particular these workers use biasing potentials in the turn regions, taken from the native structure, to drive the system into the correct folded state. They also adjust numerous parameters in their potential to render the simulations tractable computationally and to generate reasonable final structures. Our view is that fixing secondary structure is a much more straightforward methodology conceptually, is substantially easier to connect with other sources of information like NMR data and secondary-structure prediction algorithms, and is computationally more efficient. However, this will have to be determined by the test of time and it may turn out that both approaches are valuable.

Despite the obvious similarities to the diffusion–collision model (22, 23), the simulations reported here have no direct implications for protein folding kinetics, as the simulation algorithm is not intended to be representative of real-time dynamics. Moreover, the locking in of secondary structure in any case precludes consideration of the time scale for helix formation as compared to that for assembling the hydrophobic core.

In conclusion, we have shown that it is possible to locate the native structure of a protein at low resolution by using a model in which secondary structure is assigned and a very crude potential is used. We have demonstrated this for myohemerythrin and cytochrome *b*-562, by using a simple and efficient algorithm (the central processing unit time for sampling 300,000 configurations was $\approx$12 min on a Convex C210, $\approx$30 min on an IBM 550, and $\approx$35 min on an SGI Indigo R4000). Concerning our approach of fixing secondary structure, it should be pointed out that although a significant reduction of the configuration space is obtained in our algorithm by fixing the helical regions, the study of self-avoiding random walks with preset helices by Cohen and Sternberg (24) has shown that the knowledge of the location of helical regions alone does not significantly facilitate the generation of the tertiary structure. What is most striking about the results in the present paper, as compared to the work of other authors described above (13–21), is the ease with which two fairly large proteins were folded into relatively high-quality structures. The computation time is minimal and the algorithm used to carry out minimization is quite primitive. All of this suggests that with more effort, folding of larger and more complex proteins will be possible using our methodology. Ultimately, the procedure described here could be applied to proteins of unknown structure when secondary-structure assignments and other distance constraints are imposed using data obtained by CD, NMR, and/or other experimental techniques.

We acknowledge stimulating discussions with An-Suei Yang and John Gunn. This work was supported in part by National Institutes of Health Grant P41 RR06892 (to R.A.F.) and National Institutes of Health Grant GM30518 (to B.H.). The figures were produced with the program GRASP developed by Anthony Nicholls and B.H. (Columbia University).

1. Gibson, K. D. & Scheraga, H. A. (1988) in *Structure and Expression: From Proteins to Ribosomes*, eds. Sarma, M. H. & Sarma, R. H. (Adenine, Guilderland, NY), Vol. I, p. 67.
2. Scheraga, H. A. (1989) *Chem. Scr.* **29A**, 3.
3. Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991) *Science* **253**, 164–170.
4. Miyazawa, S. & Jernigan, R. L. (1985) *Macromolecules* **18**, 534–552.
5. Friedrichs, M. S. & Wolynes, P. G. (1989) *Science* **246**, 371–373.
6. Cohen, F. E., Richmond, T. J. & Richards, F. M. (1979) *J. Mol. Biol.* **132**, 275–288.
7. Chou, K.-C., Maggiora, G. M., Némethy, G. & Scheraga, H. A. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 4295–4299.
8. Carlacci, L. & Chou, K.-C. (1990) *Protein Eng.* **3**, 509–514.
9. Rooman, M. J., Kocher, J.-P. A. & Wodak, S. J. (1991) *J. Mol. Biol.* **221**, 961–979.
10. Casari, G. & Sippl, M. J. (1992) *J. Mol. Biol.* **224**, 725–732.
11. Sheriff, S., Hendrickson, W. A. & Smith, J. L. (1987) *J. Mol. Biol.* **197**, 273–296.
12. Lederer, F., Glatigny, A., Bethge, P. H., Bellamy, H. D. & Mathews, F. S. (1981) *J. Mol. Biol.* **148**, 427–448.
13. Covell, D. G. & Jernigan, R. L. (1990) *Biochemistry* **29**, 3287–3294.
14. Hinds, D. A. & Levitt, M. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 2536–2540.
15. Chan, H. S. & Dill, K. A. (1989) *J. Chem. Phys.* **92**, 3118–3135.
16. Chan, H. S. & Dill, K. A. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 6388–6392.
17. Shakhnovich, E., Farztdinov, G., Gutin, A. M. & Karplus, M. (1991) *Phys. Rev. Lett.* **67**, 1665–1668.
18. Sun, S. (1993) *Protein Sci.* **2**, 762–785.
19. Skolnick, J. & Kolinski, A. (1990) *Science* **250**, 1121–1125.
20. Kolinski, A. & Skolnick, J. (1993) *J. Chem. Phys.* **98**, 7420–7433.
21. Rey, A. & Skolnick, J. (1993) *Proteins Struct. Funct. Genet.* **16**, 8–28.
22. Karplus, M. & Weaver, D. L. (1976) *Nature (London)* **260**, 404–406.
23. Bashford, D., Cohen, F. E., Karplus, M., Kuntz, I. D. & Weaver, D. L. (1988) *Proteins Struct. Funct. Genet.* **4**, 211–227.
24. Cohen, F. E. & Sternberg, M. J. E. (1980) *J. Mol. Biol.* **138**, 321–333.