De Novo Prediction of Protein Tertiary Structure

by

Jeffrey Skolnick[a] and Andrzej Kolinski[a,b]
[a]Department of Molecular Biology
The Scripps Research Institute
10666 N. Torrey Pines Road
La Jolla, CA 92037

[b]Department of Chemistry
University of Warsaw
Pasteura 1, 02-093
Warsaw, Poland

I. Introduction

The acquisition of the ability to predict the tertiary structure of a globular protein from its amino acid sequence is a long sought objective of molecular biology [1]. The solution to the folding problem is not only of fundamental interest but is also of practical importance. The understanding of the rules of protein folding would allow us to produce proteins having desired properties such as enzymes with enhanced specificity, thermal stability or designed activity. The knowledge of the structure of an enzyme of interest is a necessary prerequisite for rational drug design; however, the time consuming experimental techniques for structural determination are a major bottleneck in the development of drugs. Ultimately, it might even prove possible to design proteins having a particular novel function. Over the shorter term, the solution of the protein folding problem would allow for the wealth of sequence information provided by the human genome project to be exploited; often sequence information is difficult if not impossible to translate into structure and ultimately function.

In what follows, we describe an application of a recently developed hierarchical approach to protein folding which only requires sequence information and does not require any information about the folded state[2,3]. Using the identical parameter set, this approach has been applied to the prediction of the native conformation of a number of small proteins having a relatively simple native topology [3]. For the one of the sequences designed by DeGrado and coworkers [4], in agreement with experiment [5], we predict a molten globule state in which left and right turning, four helix bundles are isoenergetic [2]. The prediction that both right and left turning bundles are isoenergetic was made without knowledge of the experimental result [6]. For a redesigned sequence in which 14 of the sites are replaced by Val, Ile and aromatics, we predicted independent experiment, a low temperature, native state that should be a right turning, four helix bundle with well defined tertiary contacts [7]. The existence of well defined native contacts was also confirmed by experiment. Whether the right turning bundle is the correct fold remains to be established. These simulations illustrate how by sequence modification, a system can pass from a molten globule state having substantial secondary structure but poorly defined tertiary contacts to one with well-defined tertiary contacts. However, it is not yet possible to fold an arbitrary protein from sequence information alone. As a step in this direction, we describe here the folding of some naturally occurring proteins.

II. Overview of the Hierarchical Approach to Protein Folding

For a given protein sequence, a series of independent annealing lattice simulations are run to identify the thermal transition region. Then, a series of at least ten independent simulations are run for 10^6 Monte Carlo time steps. The lowest energy structures obtained in each run are stored and compared on the basis of rms and contact map overlap. Structures that are at least 30 k_BT higher in energy than the average of the lowest energy structures are dismissed. The remaining structures are subject to further refinement at low temperature. If more than one topology is recovered, the lowest energy structures of the various topologies are compared. The family of structures with the lowest energy is defined as "native", and this family is subjected to the full atom rebuilding protocol. In addition, to confirm the consistency of the results, the lowest energy structures from each family are compared using our inverse folding protocol (this uses a different energy scale) [8]. An overview is presented in Figure 1.

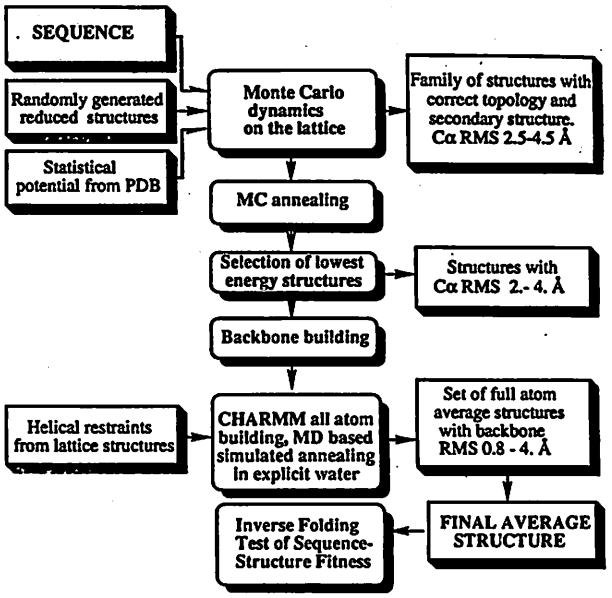

Figure 1. Flow chart depicting the hierarchical approach to protein folding.

In the lattice model of the protein, adjacent Cαs are chosen from 1.22*{(2,2,0),(2,2,1),(3,0,0),(3,1,0),(3,1,1)}. Here, the underlying cubic lattice mesh size is 1.22 Å. This lattice reproduces protein Cα coordinates at the level of 0.7 Å rms deviation from the corresponding crystal structures [9]. Side chains are treated as single spheres located at the center of mass. Moreover, the side chain positions are off-lattice, i.e., they are not restricted to cubic lattice points.

The interaction scheme is logically divided into three parts: There are short range interactions reflecting sequence specific, local conformational propensities. Then, there are hydrogen bond terms designed to enhance and regularize the secondary structural elements of the model proteins. Finally, there are long range or tertiary interactions that have several contributions. These are a one body centrosymmetric potential, a pairwise potential and a multibody potential. Thus, the interaction energy of a model protein can be schematically written as follows:

$$E = E_{local}+E_{rotamer}+ E_{h\text{-}bond}+E_{pair}+E_{multibody} \quad (1)$$

where E_{local} accounts for the intrinsic amino acid pair specific preferences for local secondary structure. $E_{rotamer}$ is the side chain rotamer energy that depends on the amino acid and the local backbone conformation. $E_{h\text{-}bond}$ is a cooperative potential that simulates the hydrogen bond network in proteins. E_{one} is the amino acid specific centrosymmetric potential that reflects the tendency of the given amino acid to be buried inside the globule or exposed to solvent. E_{pair} is the pairwise interaction free energy of the side groups. $E_{multibody}$ accounts for the higher order packing patterns between side chains found in globular proteins. We point out that for all statistical contributions to the potential which based on a structural database, none of the proteins we attempt to fold are in the database nor are they homologous to any of the proteins in the database. A further description of the interaction scheme may be found elsewhere [2].

III. Applications

The first naturally occurring protein we attempted to fold is the 60 residue fragment comprising the B domain of staphylococcal protein A, whose solution structure was recently determined by NMR [10]. It adopts a three helix bundle in solution. This stands in contrast to its crystal structure in an immunoglobulin complex where the C-terminal helix is apparently disordered [11]. In solution, helix I runs from residues 10-19, helix II runs from residues 25-37 and helix III runs from residues 42-55, with the remainder of the chain partially disordered. Helices II and III form an antiparallel hairpin, and helix I crosses helix III at an angle of about 30°.

Using the full interaction scheme described above, the resulting family of folded structures is a three helix bundle with well-defined helices running from residues 13-19, 25-37 and 42-55; slight fraying or extension of the helices beyond these regions is sometimes observed, but

in many cases residues 10-12 are also helical. The average Cα rms deviation for residues 13-55 of the fine hybrid lattice structures from the NMR solution structure is 2.25 Å.

We next summarize the results of the folding of Rop monomer containing 120 amino acids, which is a redesigned version of the dimeric protein (1rop) found in E. coli [12]. Experimentally, Rop dimer adopts a four helix bundle native state, similar to a bundle formed by two supercoiled, coiled coils. The sequence of the monomer, designed to be a left turning bundle, is MTKQEKTALNMARFIRSQTLTLLEKLNELDADEQADICESLHDHADELYR CSLASFKKPGQIDEQADICESLHDHADELYRSCLARFGGSKQEKTALNMARFIRSQT LTLLELNELAKG. Topology assembly consists of formation of a helical hairpin (mostly the central one), followed by formation of a three helix bundle intermediate which is much longer lived. In successful assembly, this is followed by formation of the four helix bundle. In 11 of 12 independent folding runs, the left turning bundle is adopted, whereas the apparently misfolded right turning bundle is formed only once. The left turning bundle is more stable by about 40 k_BT. Subsequent refinement produced a family of structures whose rms deviation from the dimer based native structure is in the range of 3.6-4.2 Å for the Cα coordinates. The rms deviation of individual helices ranges from 1.0-2.4 Å rms, with the central helical hairpin having an rms from the equivalent residues in the dimer in many cases of 2 Å. Moreover, these structures clearly exhibit excellent side chain fixation. It is unclear at this point, how close the actual monomer structure is to the dimer, as it has not yet been solved. With respect to the dimer, the predicted structures have a smaller supertwist.

An obvious concern is whether the model is biased towards highly helical structures. To address this, we attempted to fold the 46 residue minimal α/β protein crambin whose native conformation is comprised of a small helical hairpin and three extended chains arranged in a minimal antiparallel β–sheet. Three pairs of crosslinked cystines make the native conformation of this molecule very stable. In our simulations, we assumed that in the native state, the cystines are crosslinked, but the identity of the partners is unspecified. Our statistical pair potential indicates that Cyx-Cyx interactions are favored by -5.2 k_BT and Cys-Cyx interactions are repulsive by 1.6 k_BT. Thus, the folding simulation can also simulate the equilibrium between the oxidation states of cysteines and cystines. However, because the Cyx-Cyx interaction is very strong, the overall folding efficiency is quite low, with only 1/10 of the simulations leading to the native state. The other conformations have the topology of crambin but their secondary structure is highly distorted; they range from 4.3-5.5 Å rms deviation from the Cαs of the crystal structure.

The problem is that an S-S bond is so strong that once it forms, it greatly slows down the rate of local conformational rearrangement leading to correct secondary structure. To surmount this problem, we ran the system at a constant higher temperature where the S-S bond dissociation rate is sufficiently large. We then collected statistics about the secondary structural preferences to adopt helix/turn or extended/loop states. The helix probability profiles clearly predict the existence of two helices. The first runs from residues 7-17, two residues less than in the crystal structure, whereas the second helix from 23-30 is correctly predicted. The short helical fragment from residues 42-44 is not correctly predicted. It is worthwhile noting that many standard secondary structure prediction schemes [13-15] do not predict any helices in crambin; rather, crambin is predicted to be an all β–protein. Only the Chou-Fasman [16] method predicts a portion of the central helix. Thus, these calculations represent a preliminary application of the models as a secondary structure predictor.

The folding algorithm is supplemented by a small energetic bias towards the predicted helical regions. This strategy increases the folding efficiency to about 50% and yields structures whose backbone rms is about 4.0 Å from native. The remainder of the folds have the correct topology but with distorted, mostly helical, secondary structure. Since these distorted folds have energies which are about 20% higher than in the putative native conformation, they are dismissed. Three independently obtained lowest energy lattice structures were refined in long runs to yield an average coordinate backbone rms for residues 3-42 (1-46) of 3.61 (3.76) Å and a distance rms of 2.6 Å. The distances between the α-carbons of three pairs of cystine residues are reproduced within 1.1 ± 0.2 of the crystal structure.

Thus, while more work is required to solve the full folding problem, these simulations demonstrate that our hierarchical approach provides one promising path towards its solution.

References

(1) Creighton, T. E. *Biochem. J.* **1990**, *270*, 1-16.
(2) Kolinski, A.; Godzik, A.; Skolnick, J. *J. Chem. Phys.* **1993**, *98*, 7420-7433.
(3) Skolnick, J.; Kolinski, A.; Brooks III, C. L.; Godzik, A.; Rey, A. *Current Biology* **1993**, *3*, 414-423.
(4) DeGrado, W. F.; Wasserman, Z. R.; Lear, J. D. *Science* **1989**, *243*, 622-628.
(5) Raleigh, D. P.; DeGrado, W. F. *J. Am. Chem. Soc.* **1992**, *114*, 10079-10081.
(6) Handel, T. M.; Williams, S. A.; DeGrado, W. F. *Science* **1993**, *261*, 879-885.
(7) Handel, T.; DeGrado, W. F. *Biophysical J.* **1992**, *61*, A265.
(8) Godzik, A.; Skolnick, J.; Kolinski, A. *J. Mol. Biol* **1992**, *227*, 227-238.
(9) Godzik, A.; Kolinski, A.; Skolnick, J. *J. Compt. Chem.* **1993**, in press.
(10) Gouda, H.; Torigoe, H.; Saito, A.; Sato, M.; Arata, Y.; Schimada, I. *Biochemistry* **1992**, *40*, 9665-9672.
(11) Deisenhofer, J. *Biochemistry* **1981**, *20*, 2361.
(12) *Protein design exercises 86*; Sander, C., Ed.; EMBL: Heidelberg, 1986; Vol. 1.
(13) Holley, L. H.; Karplus, M. *Proc. Natl. Acad. Sci., U.S.A* **1989**, *86*, 152-156.
(14) Garnier, J.; Osguthorpe, D. J.; Robson, B. *J. Mol. Biol.* **1978**, *120*, 97-120.
(15) Zhang, X.; Mesirov, J. P.; Waltz, D. L. *J. Mol. Biol.* **1992**, *225*, 1049-1063.
(16) Chou, P. Y.; Fasman, G. D. *Adv. in Enzymology* **1978**, *47*, 45-148.