Protein structure prediction based on statistical potential

Shaojian Sun, Ning Luo, Rick L. Ornstein,* and Robert Rein
Department of Biophysics, Roswell Park Cancer Institute, Buffalo, New York 14263; Department of Biophysical Science, State University of New York at Buffalo, Buffalo, New York 14214; and *Molecular Science Research Center, Pacific Northwest Laboratory, Richland, Washington 99352

INTRODUCTION

Great computational difficulties have been encountered in protein structure prediction and protein folding from primary sequence, which mainly arise from two related aspects of the problem: First, the system is highly heterogeneous and has many degrees of freedom; second, the number of minima in the free energy landscape of the system depends exponentially on the total number of degrees of freedom of the system (1). Methods such as molecular mechanics and molecular dynamics with full atomic representation are precluded from protein folding studies due to their computational complexity. To circumvent these two major difficulties, an alternative approach has been developed in which the essential points are to reduce the degrees of freedom in amino acids by a simplified geometric representation and to smooth the potential hypersurface by an average potential function over the geometric reduction. The assumption of our reduced representation model (RRM) is that the overall folded structure of a protein is insensitive to the fine details of atomic interactions. This RRM is developed with the aim of (*a*) obtaining a better understanding of the basic physics of the protein folding problem, and (*b*) achieving a higher efficiency in the large scale conformational search of protein so that meaningful folded protein structures can be generated from the information of the primary sequence only.

The geometric representation of protein used in the current RRM has been set up in following way (2) (*a*) all backbone bond lengths and bond angles are kept at their ideal values; (*b*) peptide bond dihedral angles are fixed in the *trans* conformation; and (*c*) a single virtual atom is used to represent each side chain at the position that is determined by averaging the coordinates of the heavy atoms in the side chain. The geometric variables which determine protein conformation in this representation are ϕ and ψ defined by the bonds on either side of the C^α atom of a given residue. This geometric representation can fit most of the known crystal structures very well.

A statistical potential function has been adopted in the current RRM, which is derived from the statistical distribution of the conformations of the native proteins in the Brookhaven Protein Data Bank (3, 4). It is assumed that this statistical distribution reflects a mean field approximation of the interactions as well as the statistics of geometric constraints. The interaction potential function consists of two parts: The nonlocal part, which has been used in the pioneering work by Crippen et al. (5) and Wilson and Doniach (2), and the local part. The local part, which is unique to the present study, characterizes the interactions of singlet residues in a mean field potential and the interactions between the residues which are neighbors (doublets) along the primary sequence, and is a function of the backbone dihedral angles (ϕ, ψ). The specific form of the potential function used here is as follows:

$$H = \sum_i E^{(1)}_{k_i}(\phi_i, \psi_i) + \sum_i E^{(2)}_{k_{i-1},k_i}(\phi_i, \psi_i) + \sum_{i<j-2} (E^{(3)}_{k_i,k_j} r^{C^\alpha}_{ij}) + \sum_{i<j-2} E^{(4)}_{k_i,k_j}(r^{SC}_{ij}), \quad (1)$$

where i or j is the position of a residue in the primary sequence, k_i is the amino acid type of the residues i, $r^{C^\alpha}_{ij}$ is the distance between the C^α's of the residue i and j, and r^{SC}_{ij} is the distance between the centroids of the side-chains of the residues i and j. The model potential function is different from those that have been used in previous studies by including explicitly the local interaction energy terms to represent the effects of the steric constraints, which are manifested in the chirality of the backbone structures and the preferred conformations within the regular secondary structures of proteins.

To effectively search a larger region of the conformation space, we have employed the simulated annealing (2, 6) as energy minimization algorithm in the current RRM. A monotonic temperature decrease and multiple sites changes of (ϕ, ψ) of protein conformation search have been implemented in our simulation. Secondary structure assignments from the corresponding crystal structures have been used to bias the conformational search (for α-helix, $\phi \in [-140°, -30°]$, $\psi \in [-80°, 30°]$, for β-sheet $\phi \in [-180°, -30°]$, $\psi \in [30°, -160°]$).

0006-3495/92/04/104/03 $2.00

RESULTS AND DISCUSSION

An obvious advantage of the RRM is that the conformation space has been reduced greatly so that the number of the local minima in the potential function is also reduced exponentially. The reduction makes it possible to calculate the secondary structure and the approximate tertiary structure of a protein based on its primary sequence only from random starting conformations.

Table 1 depicts the current RRM simulation results for melittin (26 residues). Each of the numbered structures is computed by a complete simulation started from a random conformation and then the structure is minimized by a simulated annealing conformational search till no further optimized conformation could be found. More data on this protein and several other proteins are being compiled and will be published elsewhere.

Simulations with the biased conformational search for melittin and six other proteins (avian pancreatic polypeptide inhibitor, crambin, ferredoxin, bovine pancreatic trypsin inhibitor, neurotoxin and ubiquitin) indicate that the current RRM in most cases reconstructs the secondary structure regions (α, β and random) in the computed tertiary structures with significant fidelity. This can be explained as following: (*a*) the mean field statistical potential used in the RRM contains enough statistical information on the formation of the secondary structures, and the local and long range interactions favor the secondary structures formation according to the protein data bank statistics; (*b*) the current RRM has taken into consideration of both the local interactions and the overall energetics of the protein during the process of conformational search so that the assessment of the secondary structures by the current RRM, for the proteins we studied, is much improved in comparison with the case where only nonlocal potential is used. This indicates the importance of the local interaction terms in the potential function.

The general topology of the computed structures of the tested proteins is similar to the corresponding crystal structures. While some of the computed structures of melittin have fairly large RMS and DME to the crystal structures, there are several minimized structures which have low RMS and DME to the crystal structures (Table

TABLE 1 Simulations for Melittin, a membrane protein of 26 residues

No.	E_{start}	E_{end}	DME	RMS	Total Contact	Gyration Radius
1	4815.4	1072.5	7.22	7.64	372	7.6
2	5041.4	1087.3	6.54	8.15	286	8.9
3	1912.6	1152.9	5.75	7.40	260	9.0
4	5417.8	1035.6	3.60	5.12	264	10.9
5	5117.5	1078.7	6.25	7.49	308	8.4
6	5392.5	1153.8	5.40	6.34	348	8.5
7	4019.0	1086.2	6.50	8.15	268	9.2
8	4850.9	1066.5	6.11	7.21	304	8.5
9	3391.5	1105.4	5.29	6.61	312	8.8
10	3704.4	1020.9	6.55	7.40	308	9.1
11	7633.3	1047.0	2.31	3.26	266	10.8
12	5147.3	991.4	5.08	6.05	258	10.4
13	7270.2	1113.9	3.39	4.35	292	9.9
14	5748.9	990.0	5.13	5.75	342	8.6
15	5132.4	1041.3	6.47	7.34	350	8.1
16	4046.6	1058.7	6.31	7.28	310	8.5
17	6130.8	1054.6	5.81	6.83	308	8.5
18	3447.6	1063.1	6.54	7.76	316	8.7
19	4450.6	1106.0	6.34	7.30	330	8.6
20	6838.1	1128.5	2.75	4.61	256	11.9
21	5329.3	1103.6	3.34	5.04	332	8.5

21 structures have been computed, both of the energy of starting conformation and optimized conformation are listed here. The starting conformations are randomly chosen. The average total contact and radius of gyration for 21 structures are respectively 304 and 9.1 Å. The crystal structure has the total contact 300 and radius of gyration 11.1 Å. The contacts were defined for each pair of residues whose C^{α} were separated <10.0 Å. RMS and DME are defined respectively as RMS $= \{1/N \sum_i^N (r_i - r_i^c)^2\}^{1/2}$, DME $= \{[2/N(N-1)] \sum_{ij}^N (r_{ij} - r_{ij}^c)^2\}^{1/2}$, where superscript ^{c} indicates the conformation to which the comparison is made, it is usually the crystallography conformation of the protein, which in most case is not far from the native conformation. Distance matrix error, in some cases, werves as a better parameter to compare the overall similarity of two conformations.

1). The lowest DME among the computed structure is 2.31 Å. Despite the difference in the structures computed from the different random starting conformations, the overall average of the computed structures for melittin is similar to that of the crystal structure. This indicates that the current RRM contains certain essential features of folded proteins. We have noted also that the computed structures tend to be over folded by about 0.5 Å. The over folding of melittin is more serious because melittin is a membrane protein. The total contact obtained by averaging 21 computed structures is 305, which is fairly close to the crystal structure contact of 300, however the average radius of gyration is almost 2.0 Å less than that of the crystal structure. This over folding phenomenon is due to the following two facts: (*a*) the geometric reduction used in the RRM has reduced certain geometric hindrance in the protein structure; (*b*) the mean field statistical potential is too soft to maintain the van der Waals excluded volume.

This research is funded in part by the Laboratory Directed Research and Development (LDRD) Program of Pacific Northwest Laboratory. Partial support from NASA grant (NAGW-1546) and from the National Foundation of Cancer Research is also acknowledged.

Pacific Northwest Laboratory is operated for the United States Department of Energy by Battelle Memorial Institute under contract DE-AC06-76RLO 1830.

REFERENCES

1. Gibson, K. D., and H. A. Scheraga. 1988. The multi-minimum problem in protein folding. *In* Structure and Expression Vol. 1: From Structure to Ribosomes. Academic Press, Inc., San Diego. 67.
2. Wilson, C., and S. Doniach. 1989. A computer model to dynamically simulate protein folding: studies with crambin. *Proteins.* 6:193.
3. Bernstein, F. C., 1977. Protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535.
4. Abola, E. E., F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng. 1987. Protein data bank. *In* Crystallographic Databases: Information Content, Software Systems, Scientific Application. F. H. Allen, and R. Sievers, editors. Data Commission of the International Union of Crystallography, Bonn/Cambridge/ Chester. 107–132.
5. Crippen, G. M., and V. N. Viswanadhan. 1984. Sidechain and backbone potential function for conformational analysis of proteins. *Int. J. Pept. Protein Res.* 24:279.
6. Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science (Wash. DC).* 220:671.