Protein Structure Prediction – An *Ab Initio* Approach

Rajgopal Srinivasan
Jenkins Department of Biophysics,
Johns Hopkins University,
3400 N. Charles Street,
Baltimore, MD 21218.
U.S.A
E-mail: raj@roselab.jhu.edu

George D. Rose
Jenkins Department of Biophysics,
Johns Hopkins University,
3400 N. Charles Street,
Baltimore, MD 21218.
U.S.A
E-mail: rose@roselab.jhu.edu

Keywords: protein folding, LINUS, secondary structure

Abstract

We describe LINUS (Local Independently Nucleated Units of Structure), a computer program for simulating the protein folding process. At its core, LINUS is a Metropolis Monte Carlo procedure with a 'smart' move set for efficient exploration of conformational space and a simple energy function to rank conformations. It is shown that LINUS can successfully anticipate a large fraction of the native state secondary and super-secondary structure.

Introduction

Ever since Anfinsen[1] showed that a knowledge of the primary sequence of a protein should be sufficient to predict the tertiary structure of the protein there have been several attempts, with varying degrees of success, to predict the tertiary structure of a protein using sequence information alone.

Central to protein folding is the existence of a close packed core[2] enriched in hydrophobic residues[3] wherein sequentially distant residues are brought into spatial proximity. How does a protein screen the enormous number of conformations available to select the native conformation reliably in a biological time-scale remains to be fully elucidated.[4] While a folding protein achieves this spontaneously this has been vexing for the protein folder. In other words, proteins do not have a folding problem, humans do.

Current approaches to solving the protein folding problem can be classified into direct and template based methods. In template based methods the sequence of the protein is compared against a library of known structures using a suitable scoring function and the template that scores best, subject to a minimum threshold is chosen as the most likely fold.[5] Direct methods, on the other hand, take only the sequence as input and combine it with an algorithm for efficient exploration of conformation space to find a conformation that is lowest in energy for a suitable energy function.

In this paper we describe LINUS, a direct procedure for studying protein folding that has been implemented in a computer program. This procedure is an implementation of a hierarchic model of protein folding.[6] In this model it is hypothesized that folding begins with the formation of structural elements that are local in sequence and of marginal stability. These structures can combine to form intermediates of increasing complexity leading finally to the formation of the native structure of the protein. The salient feature of this model is that folding is local at all stages of the folding process.

Methods

LINUS operates on a protein sequence, starting in an extended conformation (ϕ=-120° and ψ=120°). The protein molecule is represented by all its heavy atoms using idealized geometry.[7]

Energy Function

The energy function used in LINUS is a simple one containing only three terms: a repulsive term to make sure that no two atoms overlap and two attractive terms that favor the formation of hydrogen bonds and promote contacts between hydrophobic atoms.

The repulsive component is the only accurately represented term in the energy function. It is implemented by rejecting conformations in which the separation between two atoms is less than the sum of their hard sphere radius. The hard sphere radii of the atoms are available in a previous publication.[8] All pairs of atoms are evaluated except those which are separated by 3 bonds or less.

A hydrophobic contact is assigned between side chain carbon atoms i and j of two residues when

distance(i, j) < radius(i) + radius(j) + 1.4 Å

where radius(x) is the atom' s contact radius. The maximal value is realized when the two atoms are in contact, and it scales linearly to zero as the separation distance increases to 1.4 Å. The maximal value is 0.5 units when both residues are hydrophobic (Cys, Ile, Leu, Met, Phe, Trp, Val), 0.25 units when one residue is hydrophobic and the other is amphipathic (Ala, His, Thr, Tyr), and 0.0 units for all other combinations. A salt bridge is assigned to contacts between oppositely charged groups (namely, Arg or Lys with Glu or Asp), with a maximal strength of 0.5 units that scales linearly to 0.0 over a separation interval of 1.4Å.

An attractive hydrogen bond energy of 0.5 units is assigned to residues *i* and *j* when the distance between the amide nitrogen of *i* and the carbonyl oxygen of *j* is <3.5Å, and the out-of-plane dihedral O(*j*)- N(*i*) - CA(*i*) - C(*i* + 1) > 140°. This score scales linearly to 0.0 as the distance between donor and acceptor increases from 3.5 to 5.0 Å. All backbone amide nitrogens (except proline) are considered H-bond donors, and all backbone carbonyl oxygens are considered H-bond acceptors. Additionally, the side chains of Ser, Thr, Asn, Asp, Gln, and Glu are also considered H-bond acceptors, with a maximal score of 1.0 unit. Two additional restrictions also apply: (i) a donor and acceptor must be at least three residues apart in sequence, and (ii) no donor can participate in more than one H-bond.

Finally a main chain torsion term is used to chase away residues from the right hand side of the ϕ, ψ map. Specifically a residue with $\phi > 0°$ is penalized 1 unit, unless it is glycine or asparagine in which case it is rewarded 1 unit.

Move Set

We employ a ' smart' move set in which the conformation of 3 contiguous residues is changed at a time. The move set is comprised of 4 different moves. These move sets have been chosen to reflect the observed conformations in experimentally solved protein structures. Specifically the move set incorporates the known tendency of proteins to populate the α-helical, β-strand and β-turn conformations predominantly. In detail, the move set is comprised of:

- α-helix move in which the three residues are assigned conformations with $\phi = -64\pm15^o$ and $\psi = -43\pm15^o$.
- β-strand move in which the three residues are assigned conformations with $\phi = -120\pm15^o$ and $\psi = 135\pm15^o$.
- β-turn move in which either the first two or last two residues are assigned one of four β-turn conformations. The remaining residue is assigned a random conformation from the allowed region of the Ramachandran map. The four possible β-turn moves are as follows.

- Type I β-turn $\phi = -60\pm15^o$ and $\psi = -30\pm15^o$ for the first residue in the turn and $\phi = -90\pm15^o$ and $\psi = 0\pm15^o$ for the second residue.
- Type II β-turn $\phi = -60\pm15^o$ and $\psi = 120\pm15^o$ for the first residue in the turn and $\phi = -80\pm15^o$ and $\psi = 0\pm15^o$ for the second residue.
- Type I' β-turn $\phi = 60\pm15^o$ and $\psi = 30\pm15^o$ for the first residue in the turn and $\phi = 90\pm15^o$ and $\psi = 0\pm15^o$ for the second residue.
- Type II' β-turn $\phi = 60\pm15^o$ and $\psi = -120\pm15^o$ for the first residue in the turn and $\phi = -90\pm15^o$ and $\psi = 0\pm15^o$ for the second residue.

- A coil move in which the conformation of each of the three residues is chosen randomly from the allowed region of the Ramachandran map.

  Sidechain torsions are sampled randomly in the interval -180° to 180°.

Simulation Procedure

Protein conformational space is explored using a standard Metropolis Monte Carlo procedure.[9] Starting with an extended conformation, C, a three residue segment is chosen at random, and its conformation is perturbed using a randomly chosen move from the previously described move set, to generate a new conformation, C*. If the new conformation is free of hard sphere overlaps and its energy (E_{C*})is lower than the energy of C (E_C) or if the Metropolis criterion ($e^{(Ec - Ec*)/T} > x$, where x is a random number in the interval (0, 1] and T is 0.5) is true set C to C*, otherwise C* is rejected and C is retained. In choosing a move all move types in the move set are equi-probable. This procedure is repeated 10000 x (N-2) times, where N is the number of residues in the protein. In evaluating the energy of the protein conformation only attractive interactions between pairs of residues that are no farther than 6 from one another are considered. This constraint serves as a mechanism for enforcing that only local interactions are considered. Sampled structures are stored after every N-2 moves have been attempted.

After the completion of the simulation, saved structures are analyzed to determine the fraction of helix, strand, turn and coil secondary structure populated by each residue. The assignment of secondary structure is done using a backbone torsion based procedure described previously.[8] We now repeat the simulation, allowing attractive interactions between residues that are separated by up to 18 residues. In this second stage of the simulation we use the observed secondary structure distribution for each residue at the end of the first stage as the sampling probabilities. For e.g., if a residue had a secondary structure distribution of 0.6 for α-helix, 0.2 for β-strand, 0.1 for β-turn and 0.1 for coil, then the α-helix move will be chosen with a probability of 0.6, a β-strand move will be chosen with a probability of 0.2 and so on.

Results

The LINUS procedure has been applied to several proteins. It is observed that by the end of the first stage that a substantial fraction of the residues in every protein populates

the native secondary structure. This suggests that local interactions play a significant role in the folding process. We also observe super secondary structure formation at the second stage of the simulation. The significance of these results and their implication for protein folding form the rest of the paper.

While the LINUS procedure has been applied to several proteins we discuss 3 examples, fragment of Protein G (pdb code 1pga), plastocyanin (pdb code 2pcy) and ribonuclease H (pdb code 2rn2). In all three cases, in addition to the solved crystal structure, experimental studies on the folding of the protein are also available, facilitating detailed comparison to the simulation results.

Protein G

The data in Table 1 shows that the simulation uncovers a significant bias to the native structure even when attractive interactions are allowed only between residues that are separated by no more than 5 intervening residues. Fragment studies of Blanco and Serrano[10] also show native conformational biases for both hairpins and the central helix. Thus, the simulation recaptures the known experimental tendencies.

Plastocyanin

The simulation is largely in agreement (Table 2) with the known structure of plastocyanin, except for the region around residue 60, wherein the simulation shows a bias for turn/helical structures while the solved structure shows a strand in this region. Interestingly, the fragment studies of Dyson and co-workers[11] shows that this region has a turn like conformation. This suggests that LINUS is capturing events in the folding process.

Ribonuclease H

Summarizing multiple kinetic and equilibrium experiments, Chamberlain and Marqusee[12] find a self-consistent hierarchic folding pathway for the molecule in which helices A and D fold first and are then augmented by helix B and β-strand 4. Each of these regions has pronounced, native-like biases. In fact, the only discrepant region between the native structure and the simulated biases is around residues 78-82, corresponding to an irregular kink between helices B and C.

Discussion

The results presented and other unpublished results show that LINUS has a considerable amount of success in anticipating the folded conformation of several proteins. This is a surprising result considering the simplifications employed in the simulation procedure. The LINUS procedure emphasizes the organizing role of hard sphere repulsions with attractive interactions restricted to sequentially local residues. While considerable debate has raged on whether secondary structure is a consequence of local or non-local interactions, and the degree of sophistication required in the

construction of the energy functions to simulate protein folding the results here suggest that simple models can do a surprisingly good job in capturing the essential features of protein folding. For sure, the LINUS procedure is imperfect, but it provides a zeroth order model which can be elaborated to include other more sophisticated energy functions that may improve it' s prediction prowess. In fact, CASP4, showed that the LINUS procedure was quite successful in predicting structures of new folds. Further research, currently underway, involving improved sampling schemes and better energy functions could improve the prediction accuracy of LINUS.

References

1. Anfinsen, C.B. Principles that govern the folding of protein chains. Science, 181:223-230, 1973.
2. Richards, F.M. Areas, volumes, packing and protein structure. Annu. Rev. Biophys. Bioeng. 6:151-176, 1977.
3. Kauzmann, W. Some factors in the interpretation of protein denaturation. Adv. Prot. Chem. 14:1-64, 1959.
4. Levinthal, C. Are there pathways for protein folding? J.Chim.Phys. 65:44-45, 1968.
5. Luthy, R., Bowie, J., Eisenberg, D. Assessment of protein models with three-dimensional profiles. Nature(London) 356:83-85, 1992.
6. Rose, G.D. Hierarchic organization of domains in globular proteins. J. Mol. Biol. 134:447-470, 1979.
7. Engh, R. A., Huber, R. Accurate bond angle and angle parameters for X-ray protein structure refinement. Acta. Cryst. . 47:392-400, 1991.
8. Srinivasan, R., Rose, G. D. A physical basis for protein secondary structure. Proc. Natl. Acad. Sci. USA . 96:14258-14263, 1999.
9. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. Equation of state calculations by fast computing machines. J. Chem. Phys. . 21:1087-1092, 1953.
10. Blanco, F. J., Serrano, L. Folding of protein G B1 domain studied by the conformational characterization of fragments comprising its secondary structure elements. Eur J Biochem . 230:634-649, 1995.
11. Dyson, H. J., Sayre, J. R., Merutka, G., Shin, H. C., Lerner, R. A., Wright, P. E. Folding of peptide fragments comprising the complete sequence of proteins. Models for initiation of protein folding II. Plastocyanin. J. Mol. Biol. 226:819-835, 1992.
12. Llinas, M.; Marqusee, S. Subdomain interactions as a determinant in the folding and stability of T4 lysozyme. Protein Science . 7:96-104, 1998.
13. Srinivasan, R., Rose, G.D., *Ab initio* prediction of protein structure using LINUS. Proteins: 47:489-495, 2002.

Table 1. Secondary Structure Distribution for Protein G

Struture	*Residues*	*Helix*	*Strand*	*Turn*	*Coil*
Strand	2 to 7	4	75	10	11
Strand	12 to 20	22	43	17	18
Strand	42 to 45	22	49	14	15
Strand	51 to 55	4	69	15	12
Helix	23 to 36	55	24	13	8

Table 2. Secondary Structure Distribution for Plastocyanin

Struture	*Residues*	*Helix*	*Strand*	*Turn*	*Coil*
Strand	2 to 5	12	43	19	26
Strand	18 to 22	2	73	13	13
Strand	25 to 31	9	70	12	9
Strand	36 to 42	3	73	14	10
Strand	45 to 47	3	69	13	15
Strand	56 to 58	26	44	18	11
Strand	61 to 63	38	35	14	13
Strand	68 to 74	40	37	14	9
Strand	79 to 84	7	74	10	9
Strand	93 to 99	3	55	14	28

Table 3. Secondary Structure Distribution for Ribonuclease H

Struture	*Residues*	*Helix*	*Strand*	*Turn*	*Coil*
Helix	44 to 58	40	31	18	12
Helix	72 to 78	31	39	14	16
Helix	101 to 112	53	22	13	12
Helix	128 to 142	46	31	14	10
Strand	4 to 13	3	55	19	23
Strand	17 to 28	13	28	19	39
Strand	31 to 39	26	39	15	19
Strand	41 to 43	21	54	13	12
Strand	61 to 69	19	60	13	7
Strand	96 to 98	1	83	6	10
Strand	114 to 122	19	58	13	10