

***Ab initio* Protein Structure Prediction Using Pathway Models**

Xin Yuan, Yu Shao, and Christopher Bystroff

Department of Biology, Rensselaer Polytechnic Institute, Troy, NY12180

Abstract

Ab initio prediction is the challenging attempt to predict protein structures based only on sequence information and without using templates. It is often divided into two distinct sub-problems: (1) the scoring function that can distinguish between native or native-like structures from non-native ones, and (2) the method of searching the conformational space. Currently there does not exist a reliable scoring function that can always drive a search to the native fold, and there is no general search method that can guarantee a significant sampling of near-natives. Pathway models combine the scoring function and the search. In this short review, we explore some of the ways pathway models are used in folding, in published works since 2001, and present a new pathway model HMMSTR-CM, that uses a fragment library and a set of nucleation/propagation-based rules. The new method was used for ab initio predictions as part of CASP5. This work was presented at the Winter School in Bioinformatics, Bologna, Italy, Feb 10-14, 2003.

Introduction

Protein structure prediction methods have implicit underlying principles that fall into two categories: evolution and folding. Evolution-based methods seek to find conserved sequence patterns, while folding method simulate the physical process of folding. A folding pathway is a time series of protein folding events. Most molecular simulation methods, including molecular dynamics (MD) and Monte Carlo (MC), create a pathway implicitly. Other methods enforce

certain characteristics of the folding events during the simulation, including some genetic algorithms, neural nets, and a new rule-based approach.

Detailed molecular representations

The MD approach to folding draws its strength from the fundamental nature of its physics-based energy function. Unfortunately, unless simplified models can be used, long simulations are still far too costly to be practical. Head-Gordon and Crivelli have developed the global optimization methods called Stochastic Perturbation with Soft Constraints. Their MD approach s atom-based energy function and novel hydrophobic solvation function is able to discriminate against misfolds. However, the method is still computationally expensive, and it needs improvement in β -strand and loop matching (Crivelli *et al.*, 2002).

In Beveridge s protocol, they combined an AMBER united atom empirical energy functions, a GBSA (Generalized Born/Solvent Accessibility) for solvent dielectric polarization, van der waals and cavitation effects, and a multiple copy MCSA (Monte Carlo simulated annealing) searching scheme, which is able to escape to some extent from meta-stable local minimum. The results show that the method is able to recover the structures of test cases within 6.0 RMS (Liu & Beveridge, 2002).

Simplified models and lattice simulations

In Gibbs s *ab initio* method, the protein conformation is represented using backbone torsion angles and fixed side chains. An evolutionary Monte Carlo algorithm is developed to search through this restricted conformational space. The simple physiochemical force field based on hydrophilic, hydrophobic, steric, and hydrodron-bonding potentials is used to assess the energies.

The 3-D structures of polypeptide chains up to 38 residues have been accurately predicted (Gibbs *et al.*, 2001).

Scheraga's group used the hierarchical approach for global optimization of an off-lattice simplified chain, with a modified united-residue (UNRES) force field and their Conformational Space Annealing (CSA) Global Optimization procedure. Good results have been obtained for both a four- and a three-helix protein (Saunders *et al.*, 2002).

LINUS developed by Rose's group is a Monte Carlo program that emphasizes the role of steric interactions and conformational entropy. Simple scoring functions represent the hydrogen bonds and hydrophobic interactions (Srinivasan & Rose, 2002).

Lattice-based studies represent proteins on a cubic or tetrahedral lattice, and this reduces the conformational space enormously, making even exhaustive simulations possible for short chains (e.g. 27 residues).

The recent face-centered cubic lattice model in Skolnick's group includes the interactions between hydrophobic residues, repulsive interactions between hydrophobic and polar residues, and orientation-dependent polar-polar interactions. Their Replica Exchange Monte Carlo method is able to reproduce a cooperative all-or-none folding transition and the cooperative formation of secondary structure upon the folding transition (Pokarowski *et al.*, 2003).

Skolnick also proposed a lattice-based parallel hyperbolic sampling (PHS) Monte Carlo algorithm through the logarithmic flattening of the local high energy barriers by an inverse hyperbolic sine function, which can overcome the local minima trapping and speed up the

thermalization of the protein folding process (meaning the time spent to reach equilibrium). They applied the method to side chain only protein model and able to identify much lower energy structures and explore a larger conformational space than the replica sampling MC method. They also pointed out that the minimum relative RMSD (mrRMSD) is more favorable than lowest-energy for prediction quality. The drawback of PHS is that for a relatively smooth energy landscape system it might be less efficient than other methods (Zhang *et al.*, 2002).

Fragment Libraries

The hierarchical condensation of a polypeptide may be roughly modeled by simulations that draw from a fragment library. Each fragment is a preferred conformer for a segment of the chain, usually defined by sequence statistics or motif patterns. Fragment library simulations leapfrog the earliest steps in folding, that being the formation of local structure.

Levitt's group constructed proteins from different-sized fragment libraries (four to seven residues) using simulated-annealing k-means clustering method. Their discrete approximation model is able to achieve 1 accuracy with lower complexity for four- and five-residue fragments. However, the complexity for longer fragments still needs to be improved (Kolodny *et al.*, 2002). Their study demonstrates that it is sufficient to use fragments for protein structure simulations. This has the relevance to the work of several other groups, including Karplus (Karplus *et al.*, 2003), Baker (Bonneau *et al.*, 2002), Jones (Jones, 2001), and Bystroff (Shao & Bystroff, 2003), all of whom use the fragment libraries for simulations.

In general, fragment library simulations use knowledge-based potentials and a simplified sidechain representation while swapping fragments drawn from a library. The first such program was Baker's ROSETTA algorithm (Bonneau *et al.*, 2002), automated in Bystroff's I-sites/Rosetta

server (Bystroff & Shao, 2002), which explores conformation space using MC simulated annealing and a Bayesian knowledge-based potential. Jones's FRAGFOLD starts with a library of common supersecondary units and also applies simulated annealing (Jones, 2001). Karplus's Undertaker uses HMMs (and other sources) to build a fragment library and optimizes the "cost of burial" (Karplus *et al.*, 2003). Fragment library simulations have had perhaps the broadest success in ab initio prediction in the last three CASP meetings.

HMMSTR-CM: Rule-based folding in 2D.

HMMSTR-CM is a new algorithm based on HMMSTR (Bystroff *et al.*, 2000) that was used for predicting contact maps in CASP5. The approach is not a simulation but a set of knowledge-based potentials and rules for building a protein contact maps. Nucleation/condensation-type folding pathways are encoded in the rule set. A contact map is a low resolution, 2-D representation of a protein's 3-D structure. Contact maps have been used as a tool for the protein structure prediction (Fariselli *et al.*, 2001; Olmea *et al.*, 1999; Pollastri & Baldi, 2002), especially as an easily data-minable representation of 3D structure (Hu *et al.*, 2002; Zaki & Bystroff, 2001; Zhang & Kim, 2000). Contact maps may be projected into 3D using existing algorithms (Brunger *et al.*, 1986; Taylor & Aszodi, 1994; Vendruscolo *et al.*, 1997).

For a given protein sequence, its contact potential map (Figure 1(a)) is calculated using HMMSTR, a hidden Markov model for local sequence structure correlations. A contact potential is the negative log-likelihood of a contact between a pair of Markov states, one at each of two positions in the sequence. The Markov states for each position in the sequence are assigned using the forward/backward algorithm (Rabiner, 1989). In Figure 1(a), a low contact potential is colored red and a high contact potentials are blue. Secondary structures, which can be predicted

directly by HMMSTR, can also be identified in the contact potential map. For example, strong i to $i+4$ contacts indicate predicted helix, and predicted beta strands tend to have low contact potentials with other strands. In Figure 1 (a), 3 helices and 4 (or 5) beta strands can be identified.

HMMSTR-CM initially over-predicts contacts, with few false negatives. Thus, the accuracy of the *ab initio* approach depends on the accuracy of pruning false positives. A nucleation propagation folding pathway scheme is used to find the true contacts. Its success depends strongly on the choice of the initial nucleation site. The strategy of the prediction is: (1) predict the secondary structure; (2) choose a folding nucleation site by assigning local contacts; (3) propagate from the nucleation site by assigning or removing contacts based on physicality and propagation rules (Table 1). The prediction is finished when all pairs are assigned either a contact or non-contact and none of the rules are violated.

Results from CASP5

Here, we will discuss one example of a HMMSTR-CM prediction of a CASP5 target.

Summaries of prediction methods, including this one, can be found in a special edition of the journal *Proteins, Structure, Function and Genetics* this year, to be dedicated to the CASP5 prediction experiment.

Target T0130 has 116 residues arranged in a three-layer α/β sandwich. The contact potential map is shown in Figure 1(a). By choosing different nucleation sites, we found more than one way to derive a physically possible topology. In this case, we selected to start the pathway with $\beta_2\alpha_2\beta_3$. The following is the sequence of operations that built the prediction. This sequence of events is the predicted folding pathway.

(1) Parallel β contacts were assigned between β_2 and β_3 ;

(2) Anti-parallel contacts were assigned to β_1 and β_2 . All other β contacts to β_2 were pruned.

(3) There were two ways to make a right-handed crossover from β_3 to β_4 (Fig. 1(c-d)). Since β_1 is more hydrophobic than β_3 , we paired β_1 with β_4 . All other β contacts to β_1 were pruned, and contacts between α_2 and α_3 were pruned since they are now on opposite sides of the sheet.

(4) α_1 must be on the opposite side of the sheet from α_3 , since α_3 extends across the sheet. Contacts were assigned between α_1 and α_2 .

The completed TOPS diagram and contact map accurately match the true structure (Figure 1(b)). The prediction has 42% contact coverage and 29% accuracy. However, if we count near misses (-1 residue), the coverage is 75% and the accuracy is 57%. Note that the long range contacts between the β_1 and β_4 were correctly predicted. Long range contacts are difficult to predict using purely statistical methods.

Identification of the folding nucleation site is the critical step in this approach. Once the nucleation site is chosen, the subsequent contact assignments are often unambiguous. The choice of the nucleation site in T0130 was relatively easy. Only one of the three parallel $\beta\alpha\beta$ units had a high score. The hairpin between β_1 and β_2 would also be a correct choice, but the selection of $\beta_2\alpha_2\beta_3$ eliminated more of the potential incorrect folding pathways. This prediction turned out to be topologically correct.. In other cases, the wrong structure was chosen for the nucleation site, and the algorithm failed. When the correct nucleation site was assigned retrospectively, the correct topology could be identified, but this has not yet been cross-validated.

Summary

Simulating the physical process of protein folding has taken many algorithmic forms, distinguished by the differing levels of detail in the representation of the model. Detailed all-atom representations continue to be popular, while simplified models have proven successful in blind predictions. Pathways have recently been defined for a 2D contact map representation, and this approach shows potential for modeling the folding process without simulations.

Table 1. Rules for folding in contact map space. Contacts are assigned if they have the lowest contact free energy and satisfy the following rules.

Physicality and Propagation Rules
1. <i>Propagation rule</i> : residues i and j are separated by no more than 5 sequence positions or are both within 5 residues of the same block of previously defined contacts.
2. <i>Maximum neighbor rule</i> : one residue can have at the most 12 contacts.
3. <i>Maximum mutual contact rule</i> : if residue i and j are in contact, there are at the most 6 residues in contact with both i and j .
4. <i>β pairing rule</i> : a β strand can be in contact with at the most 2 other β strands.
5. <i>β sheet rule</i> : any two pairing strands are either parallel or antiparallel.
6. <i>Helix mutual contact rule</i> : a residue can not be in contact at the same time with the residues on the opposite sides of a helix.
7. <i>Helix rule</i> : within a helix, only the contact between residue i and $i+4$ is allowed.
8. <i>β rule</i> : no contact is allowed within any strand
9. <i>Right-hand crossover rule</i> : Crossovers between parallel strands of the same sheet (paired or not) are right-handed (especially if the crossover contains a helix)
10. <i>Helix crowding rule</i> : If a helix can go to either side of a sheet, it picks the side with fewer crossovers.
11. <i>Strand burial rule</i> : if a strand can pair with either of two other strands, it chooses the one that is more non-polar.

REFERENCES

- Bonneau, R., Strauss, C. E., Rohl, C. A., Chivian, D., Bradley, P., Malmstrom, L., Robertson, T. & Baker, D. (2002). De novo prediction of three-dimensional structures for major protein families. *J Mol Biol* 322, 65-78.
- Brunger, A. T., Clore, G. M., Gronenborn, A. M. & Karplus, M. (1986). Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: application to crambin. *Proc Natl Acad Sci U S A* 83, 3801-5.
- Bystroff, C. & Shao, Y. (2002). Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. *Bioinformatics* 18 Suppl 1, S54-61.
- Bystroff, C., Thorsson, V. & Baker, D. (2000). HMMSTR: A hidden markov model for local sequence-structure correlations in proteins. *Journal of Molecular Biology* 301, 173-90.
- Crivelli, S., Eskow, E., Bader, B., Lamberti, V., Byrd, R., Schnabel, R. & Head-Gordon, T. (2002). A physical approach to protein structure prediction. *Biophys J* 82, 36-49.
- Fariselli, P., Olmea, O., Valencia, A. & Casadio, R. (2001). Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins Suppl* 5, 157-62.
- Gibbs, N., Clarke, A. R. & Sessions, R. B. (2001). Ab initio protein structure prediction using physicochemical potentials and a simplified off-lattice model. *Proteins* 43, 186-202.
- Hu, J., Shen, X., Shao, Y., Bystroff, C. & Zaki, M. J. (2002). *BIOKDD 2002, Edmonton, Canada*.
- Jones, D. T. (2001). Predicting novel protein folds by using FRAGFOLD. *Proteins Suppl* 5, 127-32.
- Karplus, K., Karchin, R. & Hughey, R. (2003). *Unifying secondary- structure, fold- recognition, and new- fold methods for protein structure prediction*. <http://www.cse.ucsc.edu/~karplus/papers/sam-undertaker.pdf>. Accessed May, 12, 2003
- Kolodny, R., Koehl, P., Guibas, L. & Levitt, M. (2002). Small libraries of protein fragments model native protein structures accurately. *J Mol Biol* 323, 297.
- Liu, Y. & Beveridge, D. L. (2002). Exploratory studies of ab initio protein structure prediction: Multiple copy simulated annealing, AMBER energy functions, and a generalized born/solvent accessibility solvation model. *Proteins* 46, 128-46.
- Olmea, O., Rost, B. & Valencia, A. (1999). Effective use of sequence correlation and conservation in fold recognition. *J Mol Biol* 293, 1221-39.
- Pokarowski, P., Kolinski, A. & Skolnick, J. (2003). A Minimal Physically Realistic Protein-Like Lattice Model: Designing an Energy Landscape that Ensures All-Or-None Folding to a Unique Native State. *Biophys J* 84, 1518-26.
- Pollastri, G. & Baldi, P. (2002). Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics* 18 Suppl 1, S62-S70.
- Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc IEEE* 77, 257-286.
- Saunders, J. A., Gibson, K. D. & Scheraga, H. A. (2002). Ab initio folding of multiple-chain proteins. *Pac Symp Biocomput*, 601-12.
- Shao, Y. & Bystroff, C. (2003). Predicting inter-residue contacts using templates and pathways. *Proteins, Structure, Function and Genetics* in press.

- Srinivasan, R. & Rose, G. D. (2002). Ab initio prediction of protein structure using LINUS. *Proteins* 47, 489-95.
- Taylor, W. R. & Aszodi, A. (1994). Building protein folds using distance geometry; towards a modeling and prediction method. In *The Protein Folding Problem and Tertiary Structure Prediction* (Grand, K. M. & Le, S., eds.). Birkhauser, Boston.
- Vendruscolo, M., Kussell, E. & Domany, E. (1997). Recovery of protein structure from contact maps. *Fold Des* 2, 295-306.
- Zaki, M. J. & Bystroff, C. (2001). Mining Residue Contacts in Proteins. In *Data Mining for Scientific and Engineering Applications* (R. Grossman, C. Kamath, P. Kegelmeyer, Kumar, V. & Namburu, R., eds.), pp. 141-164. Kluwer Academic Publishers, Boston, MA.
- Zhang, C. & Kim, S. H. (2000). Environment-dependent residue contact energies for proteins. *Proc Natl Acad Sci U S A* 97, 2550-5.
- Zhang, Y., Kihara, D. & Skolnick, J. (2002). Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins* 48, 192-201.