

# Predicting Interresidue Contacts Using Templates and Pathways

Yu Shao and Christopher Bystroff\*

Department of Biology, Rensselaer Polytechnic Institute, Troy, New York

**ABSTRACT** We present a novel method, HMMSTR-CM, for protein contact map predictions. Contact potentials were calculated by using HMMSTR, a hidden Markov model for local sequence structure correlations. Targets were aligned against protein templates using a Bayesian method, and contact maps were generated by using these alignments. Contact potentials then were used to evaluate these templates. An *ab initio* method based on the target contact potentials using a rule-based strategy to model the protein-folding pathway was developed. Fold recognition and *ab initio* methods were combined to produce accurate, protein-like contact maps. Pathways sometimes led to an unambiguous prediction of topology, even without using templates. The results on CASP5 targets are discussed. Also included is a brief update on the quality of fully automated *ab initio* predictions using the I-sites server. *Proteins* 2003;53:497–502.

© 2003 Wiley-Liss, Inc.

**Key words:** predictions; contact maps; HMMSTR; rule-based; protein folding; I-sites; Rosetta; hidden Markov models

## INTRODUCTION

Traditional structure prediction methods represent proteins either as three-dimensional structures or linear strings of secondary structure symbols. Contact maps are square symmetrical Boolean matrices that represent protein tertiary structures in a two-dimensional (2D) format. The 2D format has simplified the process of developing a rule-based algorithm for protein-folding pathways. The new algorithm, called HMMSTR-CM, has been tested on CASP5 targets.

Two-dimensional flat images are more readily discernable to the eye and more memorable than complex, rotating three-dimensional (3D) images. With only a little training, a student can learn to quickly distinguish a contact map for an  $\alpha/\beta$  barrel from a three-layer  $\alpha/\beta$  fold, different topologies which are very similar in their secondary structures. Similarities between distant homologues or analogs of  $\alpha/\beta$  and all  $\beta$  folds can be seen easily in contact maps, even when the 3D structures superimpose poorly. It makes sense that if our eyes can recognize protein folds from 2D patterns, that we may be able to program a computer to do so and thereby create a new tool for learning the rules of folding.

Contact maps may be projected into three-dimensions if they satisfy the conditions of a sphere intersection graph of

a self-avoiding chain,<sup>1</sup> which all protein contact maps do but not all predictions. Methods that reconstruct the protein structure from its contact map have been developed.<sup>2–5</sup>

Previous contact map prediction methods have used neural nets,<sup>6,7</sup> correlated mutations,<sup>8–11</sup> and association rules.<sup>12,13</sup> Neural net-based predictions had an average accuracy of about 21% overall,<sup>14</sup> whereas higher accuracies were reported for local contacts,<sup>7</sup> but the accuracy is lower for all- $\alpha$  proteins.

Our earlier work<sup>13</sup> led us to believe that two important factors were missing in contact map predictions. First, typical predicted contact maps were ambiguous or physically impossible in 3D. Second, the order of appearance of contacts was not considered, even though much is known about folding pathways.<sup>15–18</sup> In the new approach, we tried to incorporate “physicality” and protein-like characteristics by using protein templates and simple rules. The rules consist of common sense facts for packing of secondary structures. Rules for the order of appearance were derived from the general assumptions of a nucleation/propagation pathway.<sup>15</sup>

## MATERIALS AND METHODS

The results of two methods are discussed here: the I-sites server, which is fully automated, and HMMSTR-CM, which was only partially automated. The two methods consist of suites of programs having a common origin in the I-sites Library or its hidden Markov model incarnation HMMSTR. The I-sites server uses the folding simulation program ROSETTA,<sup>19</sup> originally developed in the Baker laboratory. The strategy used by the I-sites server has been previously documented,<sup>20</sup> and no significant changes were made to it before predicting the CASP5 targets.

The initial steps in processing the query sequence, database search, and building of a sequence profile are common to the two methods. Single sequences were submitted to PSI-BLAST,<sup>21</sup> searching *nr* with an E value cutoff of 0.001. The resulting multiple-sequence alignment was converted to a sequence profile, as previously described. The target sequence profile was used to generate 3D coordinate (TS) using the I-sites server or contact maps

\*Correspondence to: Christopher Bystroff, Department of Biology, Rensselaer Polytechnic Institute, 110 8th St., Troy, NY 12180. E-mail: bystrc@rpi.edu

Received 13 February 2003; Accepted 19 May 2003

(RR) using HMMSTR-CM. I-sites generates a fragment library, and with it, Rosetta searches conformational space using the Monte Carlo fragment insertion method.<sup>19,22,23</sup>

### HMMSTR-CM: Threading and Ab Initio Contact Map Predictions

HMMSTR is a hidden Markov model for local sequence structure correlations in proteins.<sup>24</sup> Each HMMSTR state is a position in an I-sites Library motif.<sup>25</sup> These motifs are short sequence patterns that may fold independently. The position-specific Markov state probability matrix ( $\gamma$ , as described in Rabiner's tutorial<sup>26</sup>) was calculated for the target sequence and was precalculated for each of the 1239 templates. Each value  $\gamma(i,q)$  is the confidence for HMMSTR state  $q$  at sequence position  $i$ , calculated by using the forward/backward algorithm.<sup>27</sup>

The target  $\gamma$ -matrix was aligned against each template  $\gamma$ -matrix by using Bayesian adaptive sequence alignment method<sup>28</sup> (BayesAligner), where the alignment matrix was composed of sums over the joint probabilities of HMMSTR states (i.e.,  $A_{i,j} = \sum_q \gamma_{(i,q)}^{\text{target}} \gamma_{(j,q)}^{\text{template}}$ ). The BayesAligner produced a single score and any number of alignments. Templates with low alignment scores were rejected. Otherwise, 100 alignments were selected from each template at random for further evaluation.

Alignments were removed from this set if they were highly fragmented. A "compactness score" was calculated as the length of the longest contiguously aligned region, ignoring small gaps ( $\leq 3$  residues). The template distances at the ends of the aligned blocks were enforced to be physically possible values (i.e.,  $D_{i,j} \leq 3.8 \times |i - j|$ ) by trimming the aligned blocks if necessary. Candidate contact maps ( $C$ ) were generated by using the alignments and the contact maps of the templates with top compactness scores. Each  $C$  was scored by using the "contact free energy" (CFE).

The pairwise contact potential between any two HMMSTR states  $p$  and  $q$  ( $G(p,q,s)$ ) was calculated as the log of the mutual probability of these two states in contacting residues ( $C_\alpha$ - $C_\alpha$  distance  $< 8 \text{ \AA}$ ), for proteins in the PDBselect database (Eq. 1).

$$G(p,q,s) = -\log \frac{\sum_{\text{PDBselect}} \sum_{i \ni D_{i,i+s} < 8\text{\AA}} \gamma(i,p)\gamma(i+s,q)}{\sum_{\text{PDBselect}} \sum_i \gamma(i,p)\gamma(i+s,q)} \quad (1)$$

The sensitivity of discriminating contacts from non-contacts improved greatly by calculating  $G$  as a function of the sequence separation  $s$  ( $4 \leq s \leq 20$ ). For sequence separations  $> 20$ ,  $s = 20$  was used. The total number of potential functions  $G$  was 1037153, one for every pair of 247 Markov states in HMMSTR and every separation distance from 4 to 20.  $G$  may be viewed as the knowledge-based energy of contacts between local structure motifs.

The contact potential between residues  $i$  and  $j$  ( $E(i,j)$ ) in the target was calculated as the probability-weighted sum of the pairwise potential functions  $G$  (Eq. 2).

$$E(i,j) = \sum_p \sum_q \gamma(i,p)\gamma(j,q)G(p,q,s) \quad (2)$$

where  $s = |i - j|$ .  $E$  is called the target contact potential map. In general,  $E$  readily identifies possible contacts between  $\beta$  strands and also finds supersecondary structure motifs such as the right-handed parallel  $\beta\alpha\beta$  motif and the  $\alpha\alpha$ -corner.

The contact free energy (CFE) was calculated by summing the elements of  $E$  that are present in  $C$ . Contacts with sequence separations  $< 4$  were ignored (Eq. 3).

$$\text{CFE} = \sum_{i,j \ni C_{ij} = 1 \cap (j > (i+3))} E(i,j) - \langle E \rangle \quad (3)$$

where  $\langle E \rangle$  is the mean contact potential for the target. For each target, we calculated the CFE for all templates and all alignments and chose one or more template/alignment with the best CFE. Other factors, such as the compactness score, were also considered.

The automated selection of templates was sometimes overruled by our ab initio analysis, described below. If the propagation rules favored one topology over another and a template of the favored topology was present in our list of top scorers, we would select that template over a higher scoring one.

### Consensus and Composite Contact Map Predictions

Often several of the top-scoring templates contained the same fold or substructure. Consensus was considered a strong indicator, especially if the fold was uncommon. Multiple candidates were sometimes used to construct a single composite map. In practice, consensus similarity between many structures is difficult to see in a 3D multiple superposition but is easy to see in superimposed contact maps. By combining the top scoring predictions, we could "grow" the incomplete pattern into a complete one.

However, simply overlapping and combining the contact maps may introduce "noise"—contacts that make the prediction physically impossible. (An impossible contact map is the one that cannot be projected into 3 dimensions.) Manual postprocessing, including rule-based manual editing (discussed below) were needed to enforce the physical reality of the final contact map.

### Ab Initio Rule-Based Pathway Predictions

A rule-based structure propagation model was used either in conjunction with templates, consensus templates, or ab initio (without templates). Given a contact potential map,  $E$  (see Eq. 2), we kept the contacts that were better than a cutoff value to create the initial contact map. The initial map was often characterized by dense blocks of contacts between  $\beta$ -strands and sparse contacts to helices.

If we kept all of these contacts, the map would be physically impossible. A set of common sense rules (Table I) were compiled to enforce physical reality and protein-like characteristics. These rules were enforced as the map was pruned by using a pathway scheme.

To start the folding pathway, we selected one or more triangular local regions with many high-confidence contacts as the nucleation site(s). We propagated the prediction in both directions by assigning or erasing blocks of

**TABLE I. Physicality and Propagation Rules Used in Ab Initio Predictions<sup>†</sup>**

1. Maximum neighbor rule: One residue can have at the most 12 contacts.
2. Maximum mutual contact rule: If residue  $i$  and  $j$  are in contact, there are at the most six residues in contact with both  $i$  and  $j$ .
3.  $\beta$ -pairing rule: A  $\beta$ -strand can be in contact with at the most two other  $\beta$ -strands.
4.  $\beta$ -sheet rule: Any two pairing strands are either parallel or antiparallel.
5. Helix mutual contact rule: A residue cannot be in contact at the same time with the residues on the opposite sides of a helix.
6. Helix rule: Within a helix, only the contact between residue  $i$  and  $i + 4$  is allowed.
7.  $\beta$ -rule: No contact is allowed within any strand.
8. Right-hand crossover rule: Crossovers between parallel strands of the same sheet (paired or not) are right-handed (especially if the crossover contains a helix)
9. Helix crowding rule: If a helix can go to either side of a sheet, it picks the side with fewer crossovers.
10. Strand burial rule: If a strand can pair with either of two other strands, it chooses the one that is more nonpolar.

<sup>†</sup>We only consider contacts separated by at least three residues.

contacts around the nucleation site, subject to the rules. TOPS diagrams<sup>29</sup> were drawn for the growing structure as a visual aid. The prediction was complete when all of the remaining contacts were rejected.

## RESULTS AND DISCUSSION

### I-Sites/Rosetta Server Automated Predictions

The automated I-sites Rosetta server (<http://www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php>) predicted the tertiary structures for all CASP5 targets in the NF and FR categories, as well as several of the FR/CM targets that were not identified as homologues of known structures. The successes and failures of the server may be summarized in a few broad statements.

For the 44 submitted predictions, 67% of the residues were found in “topologically correct” large fragments, defined as fragments of  $\geq 30$  residues with  $\text{RMSD} < 6 \text{ \AA}$ . This is a slight improvement over the results from CASP4, but because the method was not significantly modified during that period, the improvement may be just a statistical fluctuation. The average contact order of server predictions was significantly lower than that of the average target, as was reported previously. The server inserted approximately the right number ( $\sim 8\%$ ) of left-handed ( $\phi > 0$ ) residues overall, but too few of them (14%) are located on glycines.

Topologically correct large fragments were often incorrect locally, probably because Rosetta occasionally needs to insert a very unlikely fragment to get the non-local contacts to come together. The I-sites server correctly identified the overall antiparallel  $\beta$  topology of target 162, a new fold.

### HMMSTR-CM: Contact Map Predictions

We predicted contact maps for 42 CASP5 targets in the FR and NF categories using HMMSTR-CM; 12 were

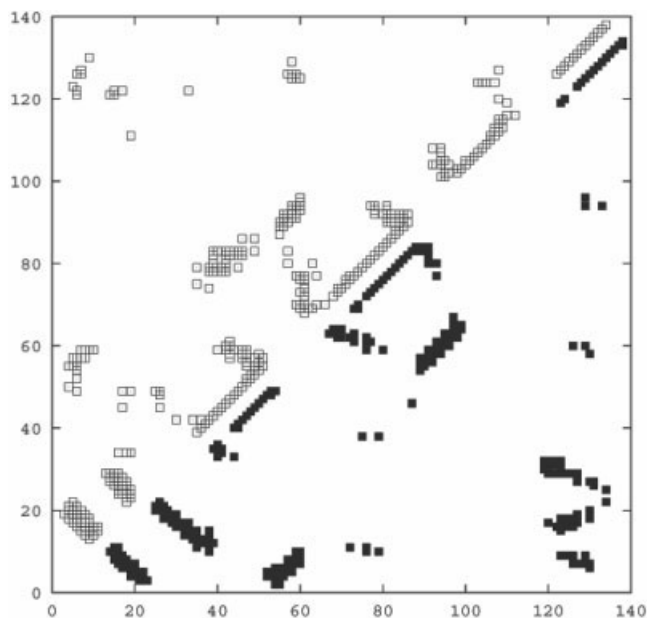


Fig. 1. Comparison between the predicted and the true contact maps of target 157. **Upper left:** Predicted contact map. **Lower right:** True contact map.

predicted with the consensus method, 22 were predicted with single templates, and 6 were ab initio predictions. Two targets were predicted with partial templates with remaining parts predicted ab initio. For all but three targets, we submitted single predictions. Here we discuss a few of the successes and failures.

### Using a Single Template

Target T0157 is an example of a successful prediction using a single template (Fig. 1). All visible secondary structure units are correctly predicted (18 residues are missing in the crystal structure), and all of the true contacts have a better-than-average  $E(i,j)$  score. A consensus map of the top scoring six templates was plotted, and this map, along with the  $E(i,j)$  map, was used to do an ab initio prediction. Nucleating the pathway at  $\beta_4\alpha_2\beta_5$  and propagating produced a TOPS diagram that agreed with one of the toptemplates, 1HJR, and this template was, therefore, chosen to prune the consensus contact map. The two N-terminal hairpins are slightly underpredicted, and a contact between helices 1 and 2 is overpredicted (Fig. 2).

This method fails if the secondary structure prediction is inaccurate. In targets T0129, T0134, and T0174, HMMSTR significantly underpredicted the helices.

### Using a Consensus Template

Target T0147 is an example of a successful prediction using multiple templates. The threading method found four templates that had top CFE scores and also shared common structural components. Three of those templates are eight-stranded  $\alpha/\beta$  barrels, and the other consists of two parallel  $\alpha/\beta$  domains. T0147 is an  $\alpha/\beta$  barrel with seven parallel  $\beta$ -strands. Templates with good CFE scores existed, but none of them predicted all of the first five

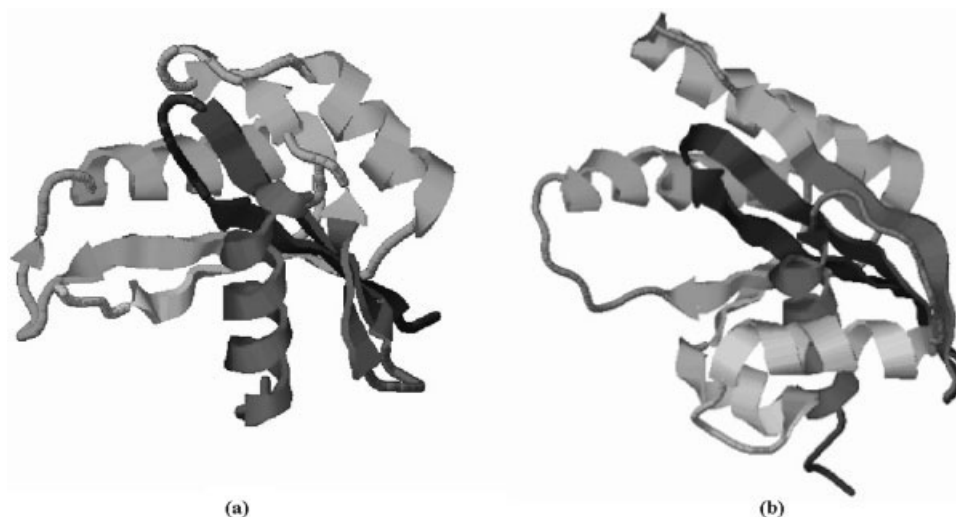


Fig. 2. Comparison of the topologies of the target T0157 and its threading template: **a**: The true structure of T0157. **b**: The structure of the threading template 1HJR.

helices and parallel  $\beta$  strand contacts correctly. By combining the results from those top scoring templates, the final prediction is better than any of the contact maps from the single templates. In particular, we found parallel contacts between the first six  $\beta$ -strands correctly. Visual inspection of the templates confirmed that they share the same topology, but we believe that finding a structural similarity and combining structures is easier and more easily automated in the contact map format rather than in 3D coordinate space.

Target T0147 also reveals a weakness of the method. HMMSTR, which is trained to recognize recurrent supersecondary motifs, does not recognize the unusual substructure at the C-terminus, three short helices instead of the usual  $\beta\alpha\beta$  motif. The consensus method tends to bias the prediction toward the more common folds.

### Using No Template

When the results of threading were not clear-cut, we tried to construct a contact map directly from the contact potential map,  $E$ , using physicality and propagation rules (Table I). Given a reasonably accurate secondary structure prediction of HMMSTR, the coverage of true contacts in  $E$  was generally high, but the map also included many false contacts. The accuracy of the *ab initio* approach depended on the accuracy of contact map pruning. Accurate pruning was found to depend strongly on the choice of the initial nucleation site.

Target 130 is an example of a successful *ab initio* prediction. It has 116 residues arranged in a three-layer  $\alpha/\beta$  sandwich. The contact potential map is shown in Figure 3(a). By choosing different nucleation sites, there was more than one way to derive a physically possible and high scoring topology. In this case, we selected to start the pathway with  $\beta_2\alpha_2\beta_3$ . The pathway was propagated as follows:

1. Parallel  $\beta$  contacts were assigned between  $\beta_2$  and  $\beta_3$ .

2. Antiparallel contacts were assigned to  $\beta_1$  and  $\beta_2$ . All other  $\beta$  contacts to  $\beta_2$  were pruned.
3. There were two ways to make a right-handed crossover from  $\beta_3$  to  $\beta_4$  [Fig. 3(c)–(d)]. Because  $\beta_1$  is more hydrophobic than  $\beta_3$ , we paired  $\beta_1$  with  $\beta_4$ . All other  $\beta$  contacts to  $\beta_1$  were pruned, and contacts between  $\alpha_2$  and  $\alpha_3$  were pruned because they are now on opposite sides of the sheet.
4.  $\alpha_1$  must be on the opposite side of the sheet from  $\alpha_3$ , because  $\alpha_3$  extends across the sheet. Contacts were assigned between  $\alpha_1$  and  $\alpha_2$ .

The completed TOPS diagram and contact map accurately match the true structure. The prediction has 42% contact coverage and 29% accuracy. However, if we count near misses ( $\pm 1$  residue), then the coverage is 75% and the accuracy is 57%. Note that the long-range contacts between the  $\beta_1$  and  $\beta_4$  were correctly predicted. Long-range contacts are difficult to predict with 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  $\beta_1$  and  $\beta_2$  would also be a correct choice, but the selection of  $\beta_2\alpha_2\beta_3$  eliminated more of the potential incorrect folding pathways.

Target 138 is an example where we chose the wrong nucleation site (data not shown). T0138 is 135 residues long and has multiple  $\beta\alpha\beta$  motifs. From its contact potential, two possible  $\beta\alpha\beta$  nucleation sites could be identified. We chose  $\beta_2\alpha_1\beta_3$ . If we had chosen the correct nucleation site,  $\beta_3\alpha_2\beta_4$ , there would be an unambiguous propagation all the way to the N-terminus, giving the correct 2134 strand order. Our erroneous choice of the nucleation site led to the incorrect strand order 2314.

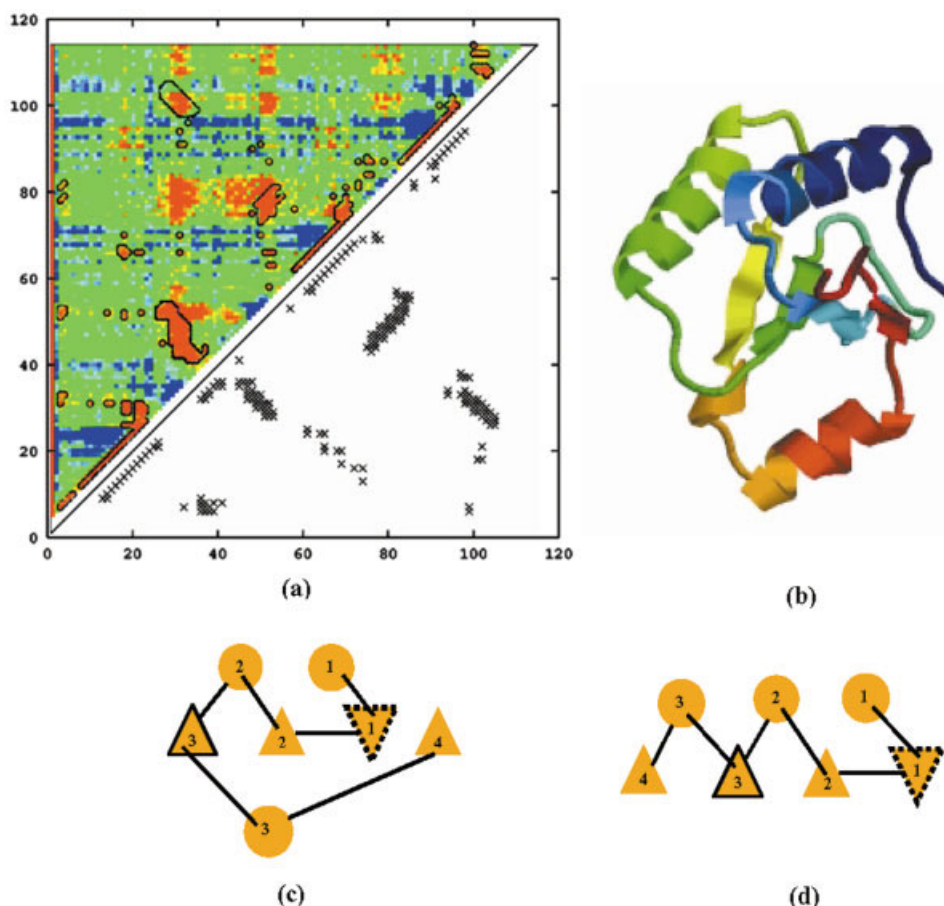


Fig. 3. T0130. **a**: The upper left triangle is the superposition of the predicted contact maps of T0130 on top of its contact potential map. The predicted contacts are represented by the black outlines. The color of the contact potential map ranges from red to blue, indicating the potential from low (favorable) to high (unfavorable). The lower right triangle is the contact map of the true structure of T0130. **b**: The true 3D structure of T0130. **c**: The correct TOPS diagram. The circles represent helices and triangles represent strands. The dotted line indicates the nonpolar strand and the solid line indicates the amphipathic strand. **d**: The wrong TOPS diagram.

### Potential Improvements

By gaining insight about how different parts of the protein pack together, we can improve the accuracy of the *ab initio* method. This will be necessary to make the whole prediction process automatic. The rule-based pathway approach depends on the correct assignment of the fold class of the target (all- $\alpha$ ,  $\alpha/\beta$ ,  $\alpha + \beta$ , or all  $\beta^{30}$ ), because the rules of propagation depend on choices of the final topology. Generally, this assignment is not difficult. So far, pathways have been applied only to the  $\alpha/\beta$  class, but a different set of rules may be envisioned for the packing of helices and all  $\beta$  proteins.

The difficulty of choosing the correct nucleation site increases with protein size, because there are more to choose from. For larger proteins, more than one correct nucleation site may be required. Because the propagation pathways have only a few steps, one approach could be an exhaustive, recursive search of all allowed topologies starting with each potential nucleation site. This could be done automatically.

Our predictions have many false contacts adjacent to true contacts (e.g., a “fat”  $\beta$ -hairpin prediction) even though it is predicted at the right position. Rules to prune this type of false contacts—in other words, to beautify the predicted contact blocks—would increase the accuracy of our prediction.

### CONCLUSIONS

We have developed methods for calculating an interresidue contact potential map for a protein sequence, for aligning that map to templates, and for pruning that map by using a folding pathway model. Results on CASP5 targets reveal that the folding pathways for some  $\alpha/\beta$  proteins are unambiguous given the correct choice of the folding nucleation site. Pathway predictions improved the selection of a remote homologue for one threading target. Consensus contact maps are more accurate than maps from single templates. The contact map format (RR) is a useful intermediate level of representation that facilitates rule-based algorithm development.

## REFERENCES

1. Michael TS, Quint T. Sphere of influence graphs in general metric spaces. *Math Comput Model* 1999;29:45–53.
2. Crippen GM, Havel, T. F. Distance geometry and molecular conformation. *Chemometrics Series*, 15. New York: John Wiley & Sons; 1988.
3. Aszodi A, Munro RE, Taylor WR. Distance geometry based comparative modelling. *Fold Des* 1997;2:S3–S6.
4. Brunger AT, Clore GM, Gronenborn AM, Karplus M. Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: application to crambin. *Proc Natl Acad Sci USA* 1986;83:3801–3805.
5. Vendruscolo M, Kussell E, Domany E. Recovery of protein structure from contact maps. *Fold Des* 1997;2:295–306.
6. Fariselli P, Casadio R. A neural network based predictor of residue contacts in proteins. *Protein Eng* 1999;12:15–21.
7. Pollastri G, Baldi P. Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics* 2002;18 Suppl 1:S62–S70.
8. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des* 1997;2:S25–S32.
9. Singer MS, Vriend G, Bywater RP. Prediction of protein residue contacts with a PDB-derived likelihood matrix. *Protein Eng* 2002;15:721–725.
10. Ortiz AR, Kolinski A, Skolnick J. Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments. *J Mol Biol* 1998;277:419–448.
11. Zhang C, Kim SH. Environment-dependent residue contact energies for proteins. *Proc Natl Acad Sci USA* 2000;97:2550–2555.
12. Hu J, Shen X, Shao Y, Bystrhoff C, Zaki MJ. *BIOKDD 2002*, Edmonton, Canada; 2002.
13. Zaki MJ, Shan J, Bystrhoff C. *Proceedings IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, Arlington, VA, USA; 2000.
14. Fariselli P, Olmea O, Valencia A, Casadio R. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins* 2001;Suppl 5:157–162.
15. Nolting B, Andert K. Mechanism of protein folding. *Proteins* 2000;41:288–298.
16. Baldwin RL. The nature of protein folding pathways: the classical versus the view. *J Biomol NMR* 1995;5:103–109.
17. Fersht AR. Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc Natl Acad Sci USA* 1995;92:10869–10873.
18. Galzitskaya OV, Ivankov DN, Finkelstein AV. Folding nuclei in proteins. *FEBS Lett* 2001;489:113–118.
19. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 1997;268:209–225.
20. Bystrhoff C, Shao Y. Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. *Bioinformatics* 2002;18 Suppl 1:S54–S61.
21. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
22. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystrhoff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* 1999;34:82–95.
23. Bonneau R, Strauss CE, Baker D. Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. *Proteins* 2001;43:1–11.
24. Bystrhoff C, Thorsson V, Baker D. HMMSTR: a hidden markov model for local sequence-structure correlations in proteins. *J Mol Biol* 2000;301:173–190.
25. Bystrhoff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol* 1998;281:565–577.
26. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE* 1989;77:257–286.
27. Baum LE, Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. *Ann Math Stat* 1966;37:1554–1563.
28. Zhu J, Liu JS, Lawrence CE. Bayesian adaptive sequence alignment algorithms. *Bioinformatics* 1998;14:25–39.
29. Sternberg MJ, Thornton JM. On the conformation of proteins: the handedness of the beta-strand-alpha-helix-beta-strand unit. *J Mol Biol* 1976;105:367–382.
30. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;247:536–540.