Predicting Protein Inter-Residue Contacts Using Templates and Pathways

Yu Shao and Chris Bystroff

Department of Biology, Rensselaer Polytechnic Institute

1 Introduction

We developed a rule-based algorithm, called HMMSTR-CM, for protein folding pathways using contact maps. Contact maps are square symmetrical Boolean matrices that represent protein tertiary structures in a two-dimensional format. Our earlier work (Zaki et al (2000)) 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 (Nolting and Andert (2000), Baldwin (1995), Fersht (1995)). 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.

2 Methods

Single sequences were submitted to PSI-BLAST (Altschul et al (1997)). The resulting multiple sequence alignment was converted to a sequence profile. The target sequence profile was used to generate contact maps using HMMSTR-CM. HMMSTR is a hidden Markov model for local sequence-structure correlations in proteins (Bystroff et al (2000)). Each HMMSTR state is a position in an I-sites Library motif (Bystroff and Baker (1998)). These motifs are short sequence patterns that may fold independently. The position-specific Markov state probability matrix ($\gamma$, as described in Rabiner’s tutorial (Rabiner (1989))) was calculated for the target sequence, and was pre-calculated for each of the 1239 templates. Each value $\gamma(i, q)$ is the confidence for HMMSTR state $q$ at sequence position $i$, calculated using the forward/backward algorithm (Baum and Petrie (1966)). The pairwise contact potential between any two HMMSTR states $p$ and $q$ ($G(p, q, s)$), where $s$ is the sequence separation, was calculated as the log of the mutual probability of these two states in contacting residues ($C_{\alpha} - C_{\alpha}$ distance < 8Å) for proteins in the PDBselect database. 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 $G$. In general, $E$ readily identifies possible contacts between $\beta$ strands, and also finds super-secondary structure motifs such as the right-handed parallel $\beta\alpha\beta$ motif and the $\alpha$ -corner. The contact free energy ($CFE$) was calculated by summing the elements of $E$ that are present in $C$, which is the target contact map candidate derived from the target-template alignment generated by the
BayesAligner (Zhu (1998)). Contacts with sequence separations less than 4 were ignored. 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 relative lengths of the sequences and the number of gaps in the alignment, were also considered. 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. By combining the top scoring predictions, we could "grow" the incomplete pattern into a complete one. 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, $E$, 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 strands and sparse contacts to helices. 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 contacts around the nucleation site, subject to a set of common-sense rules, which were compiled in order to enforce physical reality and protein-like characteristics. TOPS diagrams (Sternberg and Thornton (1976)) were drawn for the growing structure as a visual aid. The prediction was complete when all of the remaining contacts were rejected.

3 Results

HMMSTR-CM has been tested in CASP5. CASP5 Target T0157 is an example of a successful prediction using a single template. All visible secondary structure units are correctly predicted (18 residues are missing in the crystal structure), and all of the true contacts have 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 pathway prediction. Nucleating the pathway at $\beta_3\alpha_2\beta_5$ and propagating produced a TOPS diagram that agreed with one of the top templates, 1HJR, and this template was therefore chosen to prune the consensus contact map. Target T0147 is an example of a successful prediction using the consensus method. The threading method found 4 templates that had top $CFE$ scores and also shared common structural components. By combining the results from those top scoring templates, the final prediction is better than any of the contact maps from the single templates. Target 130 is an example of a successful ab initio prediction using folding pathway. It has 116 residues arranged in a three-layer sandwich (Fig. 1(b)). The contact potential map is shown in Fig. 1(a). Identification of the folding nucleation site is the critical step, since 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 by first assigning the antiparallel contacts to $\beta_1$ and $\beta_2$. There were two ways to make a right-handed crossover from $\beta_3$ to $\beta_4$ (Fig. 1(c-d)). Since $\beta_1$ is more hydrophobic than $\beta_3$, we paired $\beta_1$ with $\beta_4$. $\alpha_1$ must be on the opposite side of the sheet from $\alpha_3$, since $\alpha_3$ extends across the sheet. 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 (±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 using purely statistical methods.
Figure 1: 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 lower right triangle is the contact map of the true structure of T0130. (b) The true 3-D structure of T0130 (c) The correct TOPS diagram. The circles represent helices and triangles represent strands. The dotted line indicates the non-polar strand and the solid line indicates the amphipathic strand. (d) The wrong TOPS diagram.
4 Conclusions

Results of the HMMSTR-CM method on CASP5 targets reveal that the folding pathways for some $\alpha/\beta$ proteins are sometimes unambiguous given the correct choice of the folding nucleation site. Pathway predictions improved the selection of a remote homolog for one threading target. Consensus contact maps are more accurate than maps from single templates. The contact map format is a useful intermediate-level of representation that facilitates rule-based algorithm development.

References


