

Comparative Modeling of Mainly- β Proteins by Profile Wrapping

Andrew V. McDonnell^{1,2}, Matthew Menke^{1,2}, Nathan Palmer^{1,2}, Jonathan King⁴, Lenore Cowen^{5,6}, Bonnie Berger^{2,3,6}

Keywords: protein folding, protein motif recognition, comparative modeling, statistical prediction, beta-strand interaction

1 Introduction

This work addresses the problem of predicting, from sequence alone, three-dimensional atomic coordinates for mainly- β protein families of low sequence homology (less than 15%). A method is presented that uses sequence profiles along with empirically-derived pairwise β -strand interaction probabilities to boost detection of β -sheet propensity signal in protein sequences. We show that this new profile-based method provides adequate β -signal amplification to facilitate an accurate alignment of a query sequence onto a super-secondary structural template. From these alignments, we are able to produce putative structures for the aligned regions of the β -helix and β -trefoil motifs to an average C_α RMSD of 2.0Å and 4.5Å, respectively, in leave-family-out cross-validation. Side-chain positions are also predicted for the structures.

2 Motivation

The *comparative modeling* problem is: given only the target amino acid sequence for a protein, and a superfamily or fold class, predict whether the protein folds into a three-dimensional structure which is a member of that superfamily, or fold class (i.e., the structural motif recognition problem); if so, give an accurate residue-by-residue alignment of the portions of the query sequence onto a super-secondary structural template, and finally, produce a prediction of the structure's atomic coordinates based on this alignment. This work studies the comparative modeling of two motifs where producing the correct sequence-target alignment has been considered to be an extremely difficult problem.

3 Algorithm

In this work, we extend our methodology for structural motif recognition of mainly- β structures via the statistical capture of long-distance pairwise β -strand interactions[2, 4] to attack the more difficult problem of comparative modeling for protein domains with low sequence similarity across families. First, we modify the algorithms to produce an alignment of a target sequence onto an abstract motif template. We then map the predicted sequence-template

¹These authors contributed equally to this work.

²Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139

³Department of Mathematics, MIT, Cambridge, MA 02139

⁴Department of Biology, MIT, Cambridge, MA 02139

⁵Department of Computer Science, Tufts University, Medford, MA 02155

⁶Address correspondence to Dr. Bonnie Berger, bab@mit.edu, or Dr. Lenore Cowen, cowen@eecs.tufts.edu.

alignment to known three-dimensional structures and model the sidechains in order to predict the full atomic coordinates.

In order to increase the accuracy of this alignment, we generalize our algorithm to operate on a *sequence profile* rather than a single sequence. Sequence profiles present information about residue conservation at each position, and what *types* of substitutions are allowed at each location[3]. Our method takes advantage of the fact that β -strand interactions act as a stabilizing mechanism for mainly- β structures by considering the potential mutations suggested by the target's profile when evaluating pairwise β -strand alignments.

These techniques, in conjunction with abstract structural templates derived from structural alignments, enables accurate alignment of sequences to these templates. Combining these improvements with a backbone-dependant rotamer library sidechain modeling program [1], we are able to accurately model these motifs. The result is the first algorithm that can produce predicted three dimensional coordinates for mainly- β protein families of low (less than 15%) sequence similarity from sequence data alone.

4 Results

A program that implements this new algorithm, **BetaWrapPro**, was developed, and is available at <http://betawrappro.csail.mit.edu>. On the 14 known β -helices, this program produces accurate sequence-structure alignments for 76% of the predicted residues. For the β -trefoils, 86% of the alignments are accurate. Given these high-quality sequence-structure alignments, **BetaWrapPro** is able to generate accurate three-dimensional structure predictions for the target motifs. The accurately aligned regions of the β -helix template average less than 2.0Å C α RMSD, while those of the β -trefoils average 4.5Å RMSD.

We have found that **BetaWrapPro** is also highly successful at detecting the motifs across these families despite the lack of sequence identity, achieving 100% sensitivity at 99.5% specificity in cross-validation on the β -helices in our data set, and 100% sensitivity at 92.5% specificity in cross-validation for the β -trefoils. This is an improvement over the results for **BetaWrap** (99.0% sensitivity, 95.0% specificity) and **Wrap-and-Pack** (88.9%, 96.3%).

References

- [1] M.J. Bower, F.E. Cohen, and R.L. Dunbrack Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *J. Mol. Biol.*, 267:1268–1282, 1997.
- [2] P. Bradley, L. Cowen, M. Menke, J. King, and B. Berger. BETAWRAP: Successful prediction of parallel β -helices from primary sequence reveals an association with many microbial pathogens. *Proc Natl Acad Sci, USA*, 98:14819–14824, 2001.
- [3] M. Gribskov, R. Lüthy, and D. Eisenberg. Profile analysis. *Methods in Enzymology*, 183:146–159, 1990.
- [4] M. Menke, E. Scanlon, J. King, B. Berger, and L. Cowen. Wrap-and-pack: a new paradigm for beta structural motif recognition with application to recognizing beta trefoils. In *Proceedings of the eighth annual international conference on Computational molecular biology*, pages 298–307. ACM Press, 2004.