Learning Biophysically-Motivated Parameters for Alpha Helix Prediction

Blaise Gassend, Charles W. O'Donnell, William Thies, Andrew Lee,
Marten van Dijk, and Srinivas Devadas

MIT Computer Science and Artificial Intelligence Laboratory,
Cambridge MA 02139, USA
{gassend, cwo, thies, andy_lee, marten, devadas}@mit.edu

Keywords: protein structure prediction, protein secondary structure, all-alpha proteins, support vector machines, dynamic programming

1 Introduction

Our goal is to develop a state-of-the-art protein secondary structure predictor, with an intuitive and biophysically-motivated energy model. The lack of experimentally determined free energy values makes it difficult to design accurate cost functions that can be optimized by predictors. Our technique uses a cost function comprised of unknown parameters, and applies Support Vector Machines (SVMs) to learn parameters that correctly predict known protein structures. So far, we have focused on the prediction of all-alpha proteins and have shown that a model with 302 parameters can achieve a Q_α value (percent of correctly predicted residues) of 77.6% and a $SOV99_\alpha$ (see [3]) value of 73.4%. As detailed in an accompanying technical report [1], these performance numbers are among the best for techniques that do not rely on multiple sequence alignments.

2 Method

Our method assumes that a protein's secondary structure can be found by minimizing a free-energy function G that is computed as a sum of elementary free-energies. For example, an elementary free-energy might represent the energetic cost of a given residue appearing in an alpha helix. The predicted structure minimizes the sum of these costs. For fixed values of the free-energy parameters, there are well-known algorithms to perform such a minimization (for example, dynamic programming). Thus, our main task is to find the unknown elementary free-energies by using a database of known protein structures. A general Support Vector Machine algorithm has been proposed that can be applied to this task [2].

The key ideas of the algorithm are illustrated in Figure 1. First, the problem is converted into an exponentially large system of inequalities that the elementary free-energies must satisfy: for each sequence x_i, the correct secondary structure y_i must have a lower free-energy $G'(x_i, y_i)$ than for any of the incorrect secondary structures y^j. Next, a tractable subset of these inequalities is selected. This subset may not have any solutions, because there might not exist a set of free-energies that is compatible with the whole database of training structures. Alternately, if the problem does have solutions, it will probably have many. The SVM techniques of margin maximization and slack variables are used to translate the reduced problem into a quadratic program that has a unique solution.

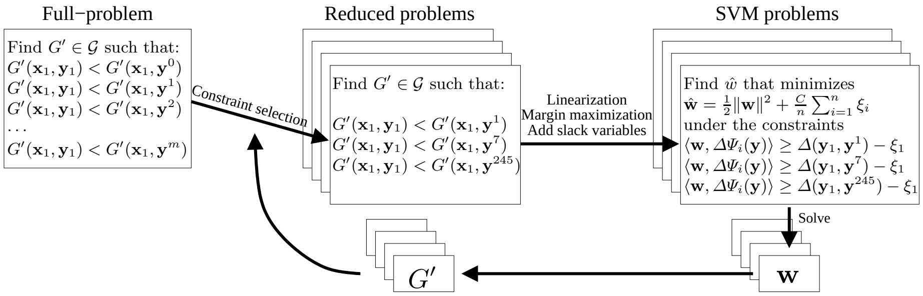
Full−problem
Find $G' \in \mathcal{G}$ such that:
$G'(\mathbf{x}_1, \mathbf{y}_1) < G'(\mathbf{x}_1, \mathbf{y}^0)$
$G'(\mathbf{x}_1, \mathbf{y}_1) < G'(\mathbf{x}_1, \mathbf{y}^1)$
$G'(\mathbf{x}_1, \mathbf{y}_1) < G'(\mathbf{x}_1, \mathbf{y}^2)$
...
$G'(\mathbf{x}_1, \mathbf{y}_1) < G'(\mathbf{x}_1, \mathbf{y}^m)$
Constraint selection
Reduced problems
Find $G' \in \mathcal{G}$ such that:
$G'(\mathbf{x}_1, \mathbf{y}_1) < G'(\mathbf{x}_1, \mathbf{y}^1)$
$G'(\mathbf{x}_1, \mathbf{y}_1) < G'(\mathbf{x}_1, \mathbf{y}^7)$
$G'(\mathbf{x}_1, \mathbf{y}_1) < G'(\mathbf{x}_1, \mathbf{y}^{245})$
Linearization
Margin maximization
Add slack variables
SVM problems
Find $\hat{w}$ that minimizes
$\hat{\mathbf{w}} = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n} \xi_i$
under the constraints
$\langle \mathbf{w}, \Delta\Psi_i(\mathbf{y}) \rangle \geq \Delta(\mathbf{y}_1, \mathbf{y}^1) - \xi_1$
$\langle \mathbf{w}, \Delta\Psi_i(\mathbf{y}) \rangle \geq \Delta(\mathbf{y}_1, \mathbf{y}^7) - \xi_1$
$\langle \mathbf{w}, \Delta\Psi_i(\mathbf{y}) \rangle \geq \Delta(\mathbf{y}_1, \mathbf{y}^{245}) - \xi_1$
Solve
$\mathbf{w}$
G'

Fig. 1. Overview of the learning algorithm.

The quadratic program is solved to produce a candidate set of elementary free-energies, which is used by a structure predictor (not shown) to find a new structure for each sequence x_i. If any of these structures are incorrect, they are added to the subset of incorrect structures y^j, a new quadratic program is created, and a new candidate solution is found. We repeat these steps until all generated structures are correct (or a suitable termination condition is satisfied).

3 Results

We applied this method to the case of all-alpha protein secondary structure prediction. We worked with a set of 300 non-homologous all-alpha proteins taken from EVA's sequence-unique subset of the PDB, July 2005.

For each run of our algorithm, we randomly selected a 150 protein training set and an independent 150 protein test set. The training set is used to learn elementary free-energies, and the test set is used to evaluate the result. Our predictor minimizes the free-energy function G using the Viterbi algorithm on a simple 7-state Finite State Machine. Table 1 summarizes our results. The prediction accuracy is competitive with other state-of-the-art predictors that do not rely on sequence alignment data. Further, while some techniques require upwards of 10,000 parameters, our predictor uses only 302 parameters in the form of elementary free-energies [1].

Description	SOV99$_\alpha$ (%) (train)	SOV99$_\alpha$ (%) (test)	Q$_\alpha$ (%) (train)	Q$_\alpha$ (%) (test)	Training time (s)
Best run for SOV99$_\alpha$	76.4	75.1	79.6	78.6	123
Average of 20 runs	75.1	73.4	79.1	77.6	162

Table 1. Performance of our algorithm on all-alpha protein structure prediction.

4 Conclusion

This work is a promising first pass at using SVM techniques to find the elementary free-energies needed to predict protein secondary structure. The method we use is general and can be extended beyond the all-alpha case described here. In future work, we plan to extend this method to super-secondary structure prediction, generating contact maps of individual hydrogen bonds in beta sheets.

References and Bibliography

1. Blaise Gassend et al. Secondary Structure Prediction of All-Helical Proteins Using Hidden Markov Support Vector Machines. Technical Report MIT-CSAIL-TR-2005-060, MIT, October 2005.
2. I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support Vector Machine Learning for Interdependent and Structured Output Spaces. In *ICML*, 2004.
3. A. Zemla, Ceslovas Venclovas, K. Fidelis, and B. Rost. A Modified Definition of Sov, a Segment-Based Measure for Protein Secondary Structure Prediction Assessment. *Proteins*, 34(2), 1999.