Predicting Protein-Protein Interactions with an Application to β-Strand Ordering

Shawn Martin[1] (smartin@sandia.gov), W. Michael Brown[1], Charlie Strauss[2], **Mark. D. Rintoul**[1], and Jean-Loup Faulon[3]

[1]Sandia National Laboratories, Albuquerque, NM; [2]Los Alamos National Laboratories, Los Alamos, NM; and [3]Sandia National Laboratories, Livermore, CA.

Abstract

As a part of the project entitled "Carbon Sequestration in *Synechococcus sp.*: From Molecular Machines to Hierarchical Modeling," we have developed a computational method for predicting protein-protein interactions from amino acid sequence and experimental data (Martin, Roe et al. 2005). We describe our method, show benchmarking results, and apply the method to ordering β-strands in proteins.

Predicting Protein-Protein Interactions

Both computational and experimental methods are available to infer protein-protein interactions. Most of the computational methods are based on genome sequence analysis and comparison across species. Experimental methods include two-hybrid systems, mass spectrometry, and protein chips. We propose a method which uses *both* sequence analysis and experimental data.

Signature Product

Our method is based on the use of the *signature* molecular descriptor, originally developed for use in cheminformatics (Faulon *et al*., 2003). The signature descriptor enables us to encode variable length amino acid sequences using neighborhood information.

$$\text{LTRPKYRNP} \xrightarrow{\Phi} \begin{pmatrix} \vdots \\ 1 \quad \text{K(PY)} \\ 0 \quad \text{R(LP)} \\ \vdots \end{pmatrix}$$

In the case of protein-protein interactions, we adapt the signature descriptor by using *signature products*, or tensor products of signature vectors.

$$\Phi(\text{LTRPKYRNP}) \otimes \Phi(\text{EYVEALYDFE AQQ} \ldots \text{NKIGIFPANY VKPA}) =$$

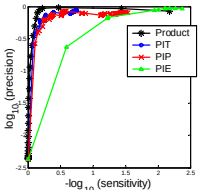
$$\begin{pmatrix} \vdots \\ 1 \quad \text{K(PY)A(EL)} \\ 0 \quad \text{N(PL)A(EL)} \\ \vdots \end{pmatrix}$$

We incorporate the signature product into a SVM framework by replacing the standard inner product with an inner product of signature products (a signature product kernel).

Benchmarks

We have implemented an compared our algorithm to similar existing algorithms (Sprinzak and Margalit, 2001; Bock and Gough, 2001; Jansen *et al*., 2003) using 10-fold cross validation accuracy, precision TP/(TP+FP), and sensitivity TP/(TP+FN).

Yeast SH3	*Accuracy*	*Precision*	*Sensitivity*
Ligand-Only	73.7	75.5	63.1
Product	80.7	81.4	75.2
PSSM	75.4	68.8	81.3
Full Yeast	*Accuracy*	*Precision*	*Sensitivity*
Product	69.0	71.5	63.2
InterPro	70.8	86.5	49.2
Sprinzak	68.8	79.8	50.0
H. Pylori	*Accuracy*	*Precision*	*Sensitivity*
Product	83.4	85.7	79.9
Bock&Gough	75.8	80.2	69.8

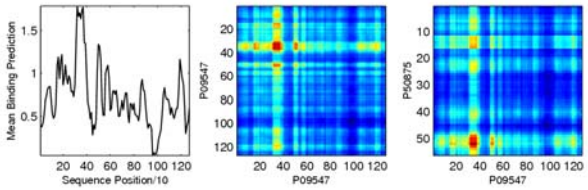


In addition, we have recently obtained 91% accuracy using the COG *Synechocystis Sp.* network and 69% accuracy using the COG *Nostoc Sp.* network.

Domain Identification

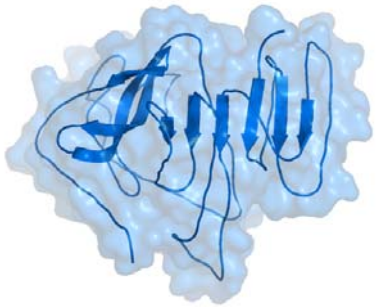
We also examined the ability of our method to locate protein domains. For this exercise, we selected 10 pairs of yeast proteins predicted by our method to interact. Using the 14 proteins present in these pairs, we constructed domain-sized amino acid subsequences by sliding a window across each of the protein sequences. Our window was of size 50, and we moved the window in increments of 10 amino acid residues. Using this method, we obtained 1681 subsequences, each 50 amino acids long.

From the model obtained using the full yeast dataset, we predicted which pairs of the subsequences would interact. By examining the positions of these interacting subsequences within the full protein sequences, we could make domain predictions as shown below.



Ordering β-Strands

Using our signature product approach, we have gone on to develop a method for ordering β-strands, which can in turn be used to improve the results of *ab initio* protein folding. The first step in our method is to train a signature product model for predicting β-strand interactions. We trained this model by extracting 9497 proteins from the Protein Data Bank (PDB), such that no two proteins had more than 95% homology. Using the dictionary of protein secondary structure (DPSS) method we extracted 219,012 β-strand pairs. After removing duplicate strands and strands with less than 4 residues, or greater than 100 residues, we obtained 23,143 interacting pairs. We added another 23,143 non-interacting pairs selected at random to arrive at 46,286 β-strand pairs altogether. After training our model, we applied our method to individual proteins, making interaction predictions on every possible pair of β-strands within the β-sheets of the protein, and then considering every possible ordering of these strands. We use the ordering which gave the highest cumulative interaction score, as measured using our β-strand interaction predictor.



Results

Using the product signature method, we trained a β-strand interaction predictor using a randomly selected 75% (or 7122) of the proteins extracted from the PDB. The predictor achieved 81% accuracy on the test set (the remaining 25%, or 2375 proteins). The 2375 proteins in the test set yielded 1386 β-sheets to evaulate our β-strand ordering algorithm. The algorithm ordered perfectly 80.3% of the β-sheets, and ordered 87.3% of the β-sheets correctly according to a normalized β-strand proximity score. The perfectly ordered score of 80.3% was computed according to whether every strand in the sheet was correctly ordered, while the normalized proximity score allowed a strand ordering to be partially correct (e.g. two strands could be switched).

Sandia is a multiprogram laboratory operated by Sandia Corporation, a LockheedMartin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.