Information of Binding Sites Improves Prediction of Protein-Protein Interaction

Tapan Patel, Manoj Pillay, Rahul Jawa, and Li Liao*

Department of Computer and Information Sciences
University of Delaware, Newark, DE 19716
Email: lliao@cis.udel.edu

Abstract

Protein-Protein interaction is essential to cellular functions. In this work, we describe a simple, novel approach to improve the accuracy of predicting protein-protein interaction by incorporating the binding site information. First, we assess the importance of the seven attributes that are used by Bradford et. al (2005) for predicting protein binding sites. The leave-one-out cross validation experiments and principal component analysis indicate that some attributes such as residue propensity and hydrophobicity are more important than other attributes such as curvedness and shape index in differentiating a binding patch from nonbinding patch. Second, we incorporate those attributes to predict protein-protein interaction by simple concatenation of the attribute vectors of candidate interacting partners. A support vector machine is trained to predict the interacting partners. This is combined with using the attributes directly derived from the primary sequence at the binding sites. The results from the leave-one-out cross validation experiments show significant improvement in prediction accuracy by incorporating the structural information at the binding sites.

1. Introduction

Research in biology and biochemistry has lead to the discovery of various proteins with unknown function that seem to play an important role in biological processes. The accurate annotation of these proteins is often time consuming but can be aided by knowing the precise location of the protein's binding sites and/or its interacting partners. Since almost all proteins carry out their diverse functions by specific protein-protein interactions, the identification of these interacting partners is a wealth of knowledge towards understanding the biochemistry of a particular protein. Currently the high throughput approach to identifying protein-protein interaction (PPI) is the yeast two-hybrid assay [4,5]. However, a typical proteomic project can take over a year to complete and often yields noisy or ambiguous data. This has motivated bioinformatics research in developing computational methods for predicting protein-protein interaction, which can then be quickly tested by *coimmunoprecipitation* or other related experiments.

In [2,3], methods were developed for predicting the binding sites by exploiting characteristics of the surface residues, whereas some methods focus on deriving sequence signatures from PPI and use these signatures for predicting other PPI [10, 11]. In a work by Ben-Hur and Noble [9], kernel methods were developed to predict protein-protein interaction using various sources of data.

In this work, we propose to extend the idea put forth by Bradford & Westhead (2005) [1] to predict interacting partners of a specific protein. Previously it has been shown that binding sites of proteins have specific properties that distinguish them from the rest of the protein. These properties (seven total) were used by Bradford & Westhead (2005) to predict possible binding sites on a protein using a support vector machine.

We now discuss some improvements on their method, namely, assessing the importance of the various attributes used by Bradford & Westhead and using a concatenation approach to predict the binding partners of a given protein.

2. Method

In this section, we first describe the data used in this work, and the schemes for preparing training and testing subsets for cross validation experiments. Then we describe the methods for analyzing the utility of the various structural attributes used in Bradford & Westhead in discriminating binding patches versus nonbinding patches. We end this section by describing a novel method to predict the interacting partners based on the primary sequence and the structural attributes.

* Corresponding author.

2.1 Data

The dataset, adopted from Bradford et al (2005), contains 180 known interacting protein pairs, including 36 enzyme-inhibitor interaction types, 27 hetero-obligate, 87 homo-obligate and 30 non-enzyme-inhibitor transient (NEIT) interactions. A solvent exclude surface was generated for each protein and divided into multiple patches, with a size of about 6~8% of the protein surface. For a patch, each of its vertices was then labeled with seven surface properties. These properties are: shape index and curvedness, conservation, electrostatic potential, hydrophobicity, residue interface propensity and solvent assessable surface area. Each of the seven properties is measured in some way to yield a numerical value, which is then normalized to be in the range [0, 1]. The patch is then represented by a 14 dimension vector, which is formed by the mean and standard deviation for each attribute of the seven properties calculated across the patch. These 14-vectors are then used to train a support vector machine. The data was provided by Bradford which includes a binding patch and a non-binding patch attributes for the 180 proteins (for a total of 360 data points).

Support vector machines [7, 13] are a supervised learning method, which requires training with a set of known examples before being used to predict on data whose classes are unknown. In this study we used SVM^{light} v. 6.01 with a radial kernel function [8]. The performance was evaluated by leave-one-out cross validation. That is, out of the 360 data points (each is a 14-dimension vector), one data point is held for testing while the other 359 data points are used to train the support vector machine. This process is repeated by using different data point for testing. The results are averaged over these leave-one-out experiments.

2.2. Analysis of the binding site attributes

From the original dataset which contains fourteen attributes for binding and non-binding patches for 180 proteins, we generate a training set that discards one particular attribute at a time. We then analyze the effect of removing this particular attribute in classifying a known patch as either binding or non-binding by using a leave-one-out cross validation method. This was repeated one hundred times for each attribute removed. The whole procedure was then performed four more times to obtain an average accuracy which we then plot in Figure 1.

The independent contribution of each attribute in classifying a patch was also measured. To do this, only one of the seven attributes is retained in the vector, used to train the SVM. Similar to the procedure describe above, we follow a leave-one out cross validation method to assess the accuracy of the newly trained SVM. Again, we repeat this process multiple times and record the mean value of the observations. The results for this part are shown in Figure 2.

Aside from the importance of each attribute, we also assessed the significance of using only the mean or only the standard deviation of the seven attributes. Results for this experiment are shown in Figure 3.

In addition to the *ad hoc* assessment as described above, the dataset was more rigorously analyzed using a multivariate statistical method, namely principal component and factor analysis (PCA and FA). First, we plot pairs of attributes to graphically infer possible relationships among the fourteen attributes – the figure is not shown in the paper for the sake of space. The diagonal plots show the distribution of a particular attribute in the dataset. Off-diagonal plots show the co-variation/relationships between pairs of different attributes. From the plots we can readily infer that some attributes are highly dependent on others and may contribute very little to the distinction of binding patch from nonbinding patch. Such attributes are redundant and can be removed from our dataset without loss of accuracy. Next, in order to obtain a quantitative measure of the importance of each attribute, PCA analysis was performed. To expedite the PCA analysis, the data are first standardized, i.e. to have the values normalized into range [0, 1]. It is noted that the attributes 13 and 14 (the mean and standard deviation of conservation) were removed due to a value of -1.0, which was assigned for these attributes because too few homologues were found to compute a score.

The variance and total percentage of variability explained by each component is given in Table 2.

2. 3. Prediction of interacting partners

The main contribution of this work is to incorporate the binding site properties to predict protein-protein interaction. Since it had been shown by Bradford & Westhead (2005) that binding sites of a protein have distinguishing properties from non-binding sites, it is reasonable to believe that these properties should be useful to predict the interacting partners of a particular binding patch.

We approach the protein-protein interaction prediction by two steps: 1) predict the binding patches, 2) identify the binding partners. That is, given two proteins A and B, we first apply Bradford & Westhead's method to identify the binding patches in these two proteins. Then, for each binding patch from protein A, we pair it up with a binding patch from protein B, and calculate a score that measures the likelihood for these two patches to be true binding partners. If there are multiple binding patches in proteins, all possible bipartite pairings will be scored, and the highest score is used for predicting the likelihood of protein A interacting with protein B. One advantage of this two-step approach is to allow us to focus on only these binding patches to see how the properties at these binding sites can impact on the protein-protein interaction.

Of the 180 proteins in the original dataset, 154 proteins have their interacting partner inside the dataset. These 154 proteins make up 77 interacting pairs, which are used as the positive examples. The rest pairings (77x76 – 77 = 5775) among these 154 proteins provide the negative examples. The leave-one-out cross validation experiments are designed as follows. Of the 77 positive examples, 76 are used for training the SVM, and the rest one is reserved for testing. This 76-to-1 ratio is maintained when preparing the negative training and testing data. That is, the 5775 negative examples are split into two subsets with a ratio between their sizes being about 76 to 1. In an alternative scheme for cross validation, we reduce the sample space by considering only those proteins that are in the same family.

For each example, no matter positive or negative, a vector is formed by concatenating the vectors for the vectors for the corresponding proteins in the pair. The concatenated vectors are then provided as input to the SVM for training and testing. The rational for such simple concatenation is that the concatenated vector for the positive example (i.e., interacting partners) will be distinguishable from that for the negative examples.

As a baseline, we tested how only the primary sequences at the binding sites can differentiate interacting pairs from non-interacting pairs. That is, the vectors contain information from the primary sequence. To this end, the primary sequences at the binding sites for these 154 proteins are retrieved from the PDB database [6]. To characterize the binding sites using primary sequences, we profile the sequence into a vector of frequencies for the various amino acid triplets to occur at the site. A window of size 3 is slide across the site and at each residue the occurrence of the possible triplets is updated. Because the small size of the binding patches, many of the 20x20x20 = 8000 possible triplets may never occur, leading to a very sparse vector. Those sparse vectors can cause the SVM training unreliable. To mitigate the problem, amino acids are grouped into 7 classes based on physical-chemical properties: hydrophobic, hydrophilic, positively charged, negatively charged, neutral, able to form hydrogen bond, not able to form hydrogen bond. This reduces the number of triplets to 7x7x7 = 147. In this scheme (A), each example (positive or negative) is represented as a concatenated vector of dimension 2 x 147 = 294. The vector dimension is further reduced by using 5 classes: hydrophobic, positively charged, negatively charged, able to form hydrogen bond, not able to form hydrogen bond. In this scheme (B), each example is represented as a concatenated vector of dimension 2x5x5x5 = 250.

Another baseline is to see how well the prediction goes when only the fourteen structural attributes at the binding sites are used. In this scheme (C), each example is represented as a concatenated vector of dimension 2x14 = 28.

To see how the structural properties at the binding sites can enhance protein-protein interaction prediction, we combine the two baseline cases by concatenating the vectors from the two cases. That is, in this scheme (D), an example is represented by a concatenated vector of dimension 28 + 250 = 278, or (E) 28 + 294 = 322.

In one variation (called scheme F), an example is represented by a vector by adding (instead of concatenating) the two 14-dimension vectors from the corresponding patches. The rational is that the attributes the require compatibility between the interacting partners (such as the curvedness being concave in one and being convex in the other) will stand out in the resulting vector.

Table 1. Description of the fourteen attributes.

Index	Attribute Description
1	Mean Residue propensity
2	Standard Deviation Residue propensity
3	Mean Hydrophobicity
4	Standard Deviation Hydrophobicity
5	Mean Accessible surface area
6	Standard deviation Accessible surface area
7	Mean Shape Index
8	Standard deviation shape index
9	Mean Electrostatic potential
10	Standard deviation electrostatic potential
11	Mean curvedness
12	Standard Deviation curvedness
13	Mean Conservation score
14	Standard deviation conservation

3. Results

From the analysis conducted using the support vector machine approach with different training sets, we come to a conclusion that, each of the 7 properties has disparate weights in contributing towards determining if a patch is a binding patch. For brevity, the fourteen attributes derived from these 7 properties are numbered and listed in Table 1. As shown in Figure 1, it was noted that 86% accuracy of prediction using all the attributes was reduced to 79% when the residue propensity parameter was dropped. In contrast, the prediction is affected minimally if the curvedness is omitted.

These conclusions about the utility of the various attributes are validated by another line of experiments where only one particular attribute was used. Shown in Figure 2 is the prediction accuracy for any one attribute used. It is noted that if only the amino acid propensity is used, the accuracy is still only 76%. Furthermore, the curvedness is the least significant of all. This is in agreement with what is shown in Figure 1 where omitting the curvedness only decreases the accuracy only by 1 percent.

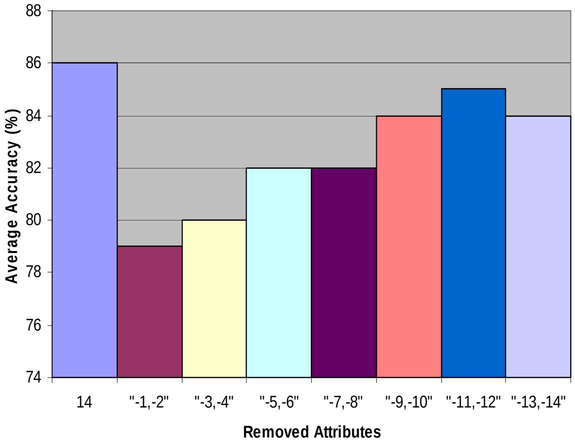

Figure 1: Effect of removing a particular attribute on the accuracy of prediction. The first bar marked with “14” is a control experiment where we used all fourteen attributes.

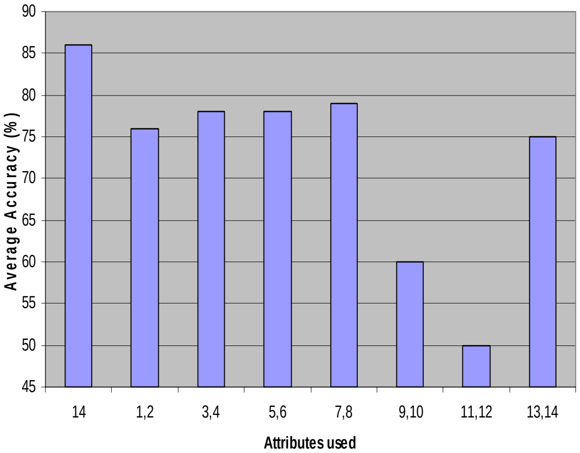

Figure 2: Effect of using only one attribute on the accuracy of prediction. The bar marked with “14” is a control experiment where we used all fourteen attributes.

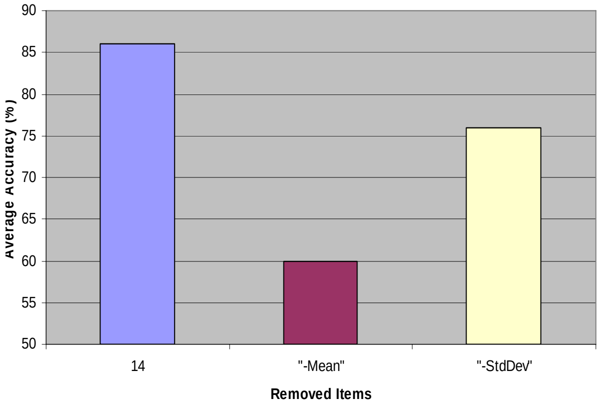

Figure 3: Effect of removing the mean or standard deviation on accuracy. The bar marked with “14” is control experiment where we used all fourteen attributes)

The utility of these 14 attributes is further studied by using principal component analysis method, as described in the Method section. The results are presented in Figure 4 and Tables 2 and 3.

A plot of percent variability explained by each component is given in Figure 4. The line above the Pareto chart shows cumulative percentage. It can be seen from Figure 4 and Table 2 that there a clear break in the amount of variance explained by each component between the second and third components. Components one and two together account for nearly 80% of the total variability in the standardized ratings and may be used to reduce dimensions in order to increase efficiency without much loss of accuracy.

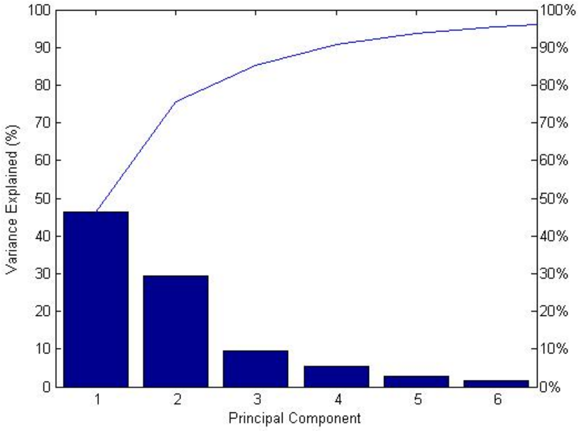

Figure 4: Pareto chart of percent variability explained by each component (components 7-12 were left out for simplicity as they contribute the least to variability).

Table2: Variance explained by each component

Component	Variances	Percent Explained
1	0.0442	46.22
2	0.0281	29.46
3	0.0092	9.64
4	0.0052	5.45
5	0.0028	2.91
6	0.0017	1.74
7	0.0015	1.54
8	0.0012	1.25
9	0.0007	0.74
10	0.0005	0.57
11	0.0004	0.37
12	0.0001	0.11

Table 3: First and second component coefficients

Attributes	First component coefficients	Second component coefficients
1	0.6174	-0.1134
2	0.1329	0.0298
3	0.6647	-0.1125
4	0.0767	0.0101
5	-0.3069	-0.0469
6	-0.1302	-0.0074
7	-0.1552	-0.0031
8	-0.0053	0.0184
9	-0.1211	-0.9806
10	-0.0573	-0.0933
11	-0.0125	-0.0242
12	0.0083	-0.0073

Since the coefficients in PCA are linear combinations of the original data that generate new variables, they can be thought of as weight factors and provide a quantitative measure of importance of each attribute. The smallest coefficient contributes the least amount of information in distinguishing between a binding and nonbinding patch and the largest contributes the most. From Table 3, where the coefficients for the first two components are given, we deduce that the mean amino acid propensity and mean hydrophobicity (attributes 1 and 3) are the most important while the mean curvedness, standard deviation of the curvedness and standard deviation of the shape index (attributes 11, 12 and 8) seem to be the least important. It is interesting to note that the two attributes related to the geometry of a patch (shape index and curvedness) happen to be the two least important attributes for distinguishing a binding patch. Although the mean shape index is of some value, we hypothesize that these four attributes will prove to be more valuable when they are used to predict binding partners using a concatenation approach discussed below. It is worth noting that based on the PCA result, we reduce the dimensionality of our data to 2 (i.e., using only the first and second components), and we achieve 85% accuracy. This is in good standing with the 86% accuracy when all fourteen attributes are used.

Furthermore, our experiments on using only the mean or only the standard deviation (Figure 3) suggests that these two parameters also contribute unequally to the binding property of a patch. We thus hypothesize that using other statistical parameter such as median would further increase the accuracy.

The performance of classification and prediction is evaluated using receiver operating characteristic (ROC) score [12]. A ROC score is the normalized area under a curve that plots the true positives as a function of false positives for varying classification thresholds. That is, all the testing examples are first ranked in a decreasing order based on the output score from the SVM, the higher the output score, the more likely the example is predicted as a positive example. By scanning the ordered rank, at each position if a threshold is imposed, check how many predicted positives are true positive, and how many predicted positives are false positive. At the end of scanning, a curved is generated that plots the true positives as a function of false positives. ROC scores are in the range of [0, 1], with 1 for a perfect classification.

The average ROC scores for the various schemes are reported in Table 4. It can be seen that the simple concatenation of the fourteen attributes from each patch in a pair of patches generate the best performance with an average ROC score of 0.773. In Figure 5, details of the ROC scores over the 154 proteins are given as histograms. The x axis is ROC score, and y axis is the number of proteins whose predicted partner achieves a given ROC score or better. Therefore, the higher up a curve is, the better performance the corresponding scheme is. The results in Figure 5 are consistent with that in Table 4.

Table 4. The average ROC score for various schemes.

Scheme	Average ROC
A	0.643
B	0.600
C	0.773
D	0.750
E	0.745
F	0.704

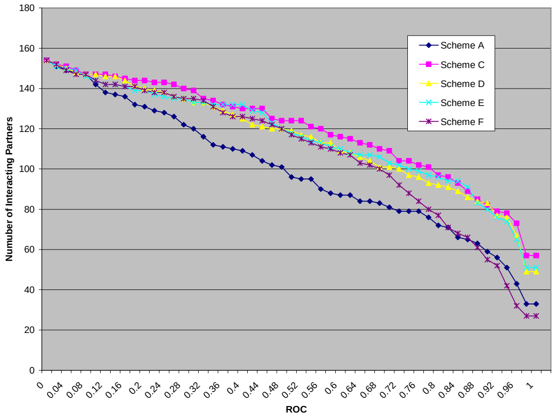

Figure 5. Histogram of ROC scores for predicted interacting partners.

4. Discussion

In this work we first used statistical methods to assess the relative importance of different properties in determining binding site of a protein. We conclude that hydrophobicity, electrostatic potential and residue propensity are among the more importance attributes and surprisingly, attributes describing the geometry of a surface patch are among the least important. We further proposed a simple, novel method to predict interacting partners using these attributes since most proteins bind to their target in a specific “lock and key” configuration. The approach developed here has the added advantage of not only identifying the interacting partners of a protein, but also providing a very specific physical region of protein-protein interaction. As a future work, we plan to incorporate the secondary structure information, which can be obtained by running some standard tools, e.g., PHD. We also plan to investigate using decision tree on the most differentiating attributes, for example, curvedness of the corresponding patches may be the first attribute to look at for the compatibility as binding partners.

Acknowledgements: We are grateful to James R, Bradford and David R. Westhead for making available some of the data in their paper.

References:

[1] Bradford, J. R., and Westhead, D. R. 2005. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, **21**: 1487-1494.

[2] JL Chung, W. Wang, and P. Bourne. 2006. Exploiting Sequence and Structure Homologs to Identify Protein-Protein Binding Sites. *PROTEINS: Structure, Function, and Bioinformatics*, **62**:630-640.

[3] Yan C., Honavar V. and Dobbs, D. 2004. Identifying protein-protein interaction sites from surface residues – A Support Vector Machine Approach. *Neural Computing Applications*. **13:**123-129.

[4]. Ito, T. et al. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA*, **98**: 4569-4574.

[5] Ueta, P. et al 2000. A comprehensive analysis of protein-protein interactions in *Saccharimyces cerevisiae*. *Nature*, **403**: 623-627.

[6] Berman, H.M. et al. 2000. The protein data bank. *Nucleic Acids Res.* **28**: 235-242.

[7] Cristianini N and Shawe-Taylor, J, *Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, UK, 2000.

[8] Joachims T. Marking large-scale SVM Learning Practical. *Advances in Kernel Methods – Support Vector Learning*, B. Scholkopf and C. Burges and A. Smola (e.d), MIT Press, 1999.

[9] Ben-Hur A and Noble W.S. 2005. Kernel methods for predicting protein-protein interactions. *Bioinformatics*, **21**: i38-i46.

[10] Fang, J. et al. 2005. Discover protein sequence signatures from protein-protein interaction data. *BMC Bioinformatics*, **6**:277.

[11] Martin, S. *et al*. 2005. Predicting protein-protein interactions using signature products. *Bioinformatics*, **21**: 218-226.

[12] Gribskov M and Robinson N. Use of receiver operating characteristic analysis to evaluate sequence matching. Computers and Chemistry, 1996, **10**:25-33.

[13] Vapnik V. *The Nature of Statistical Learning Theory*, Springer Verlag, 1995.