Struct2Net: Integrating Structure into Protein-Protein Interaction Prediction

Rohit Singh, Jinbo Xu, and Bonnie Berger

Pacific Symposium on Biocomputing 11:403-414(2006)

STRUCT2NET: INTEGRATING STRUCTURE INTO PROTEIN-PROTEIN INTERACTION PREDICTION

ROHIT SINGH* JINBO XU*‡ BONNIE BERGER†‡

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge MA 02139
E-mail: {rsingh, j3xu, bab}@theory.csail.mit.edu

This paper presents a framework for predicting protein-protein interactions (PPI) that integrates structure-based information with other functional annotations, e.g. GO, co-expression and co-localization, etc. Given two protein sequences, the structure-based interaction prediction technique threads these two sequences to all the protein complexes in the PDB and then chooses the best potential match. Based on this match, structural information is incorporated into logistic regression to evaluate the probability of these two proteins interacting. This paper also describes a random forest classifier which can effectively combine the structure-based prediction results and other functional annotations together to predict protein interactions. Experimental results indicate that the predictive power of the structure-based method is better than many other information sources. Also, combining the structure-based method with other information sources allows us to achieve a better performance than when structure information is not used. We also tested our method on a set of approximately 1000 yeast genes and, interestingly, the predicted interaction network is a scale-free network. Our method predicted some potential interactions involving yeast homologs of human disease-related proteins.
Supplementary Information: `http://theory.csail.mit.edu/struct2net`

1. Introduction

Proteins are the workhorses of the cell, performing a wide variety of functions. Most often, they perform these functions by interacting with other proteins. Indeed, many diseases can be traced to undesirable or malfunctioning protein-protein interactions (e.g.: viral-host interactions[14], prion formation[11]). Clearly, the study of such interactions is very important.

Protein-protein interactions (PPIs) can be studied from two different perspectives. In the traditional view of PPIs, the aim has been to understand the physical mechanism of interaction between two proteins by using experimental and/or computational methods to study each interaction individually.

*These authors contributed equally to the work
†Corresponding author
‡Also in the MIT Dept. of Mathematics

In contrast, the more-recent "high-throughput" view of PPIs treats proteins simply as logical entities and visualizes their interactions as a network, aiming to understand the system of interactions as a whole. This paper describes a computational technique that applies insights gleaned from the older perspective to independently supplement experimental methods designed for the newer, systems-level perspective of PPIs.

We consider the problem of predicting if two proteins interact, given their sequence information and, optionally, other genomic and proteomic information. Such computational prediction of PPIs can supplement experimental methods for elucidating PPIs. When mapping very large interactomes (e.g., human), such PPI predictions– even if only partially accurate– would be valuable in prioritizing the set of interactions to experimentally test. Moreover, experimental techniques are quite error-prone; as prediction methods gain accuracy, they can be used to double-check the results of the experiments.

Contributions: This paper proposes to use structure-based methods, in conjunction with high-throughput information, to predict interactions. We describe a fully-automated structure-based method for computing the likelihood of an interaction, solely from sequence data. A key idea here is that if a potential interaction is sufficiently favorable energetically, it is likely to be true. As part of our method, we introduce a novel algorithm for computing the most-likely structure of the complex formed by two given proteins and describe the use of logistic regression[2] for evaluating if the putative complex corresponds to a true interaction. Furthermore, to the best of our knowledge, this paper is the first to describe a framework for predicting PPIs that integrates structure-based insights with other functional annotation (e.g., co-expression, GO description). Finally, our methods predict new potential interactions involving yeast homologs of human disease-related proteins.

Algorithm Overview: We employ a structure-based method to answer the following question: "*assuming* two given proteins interact, what is the interaction energy of the formed complex[a]?" The method exploits homology between the given protein-pair and complexes with known structure. Then we use logistic regression to identify those pairs for which the interaction energy is low enough and, hence, an interaction is likely. To combine PPI predictions made by our structure-based method with other kinds of functional information we have used a random-forest classifier[7] (see Fig 1).

Related Work: Existing work on predicting PPIs has mostly followed a "guilt-by-association" approach, the idea being that if two proteins share

[a]In this paper, when referring to protein *complexes*, we consider only those with exactly two components.

functional characteristics (co-expression, similar GO annotations etc.) they are likely to interact[15]. These methods employ a variety of functional information, using them to classify an interaction as 'true' or 'false'. Many different machine learning techniques have been used for classification: Bayesian networks[5], random forests[12], probabilistic decision trees[18], and kernel canonical analysis[17]. Qi *et al.*[12], in particular, incorporated a large variety of functional information. More recently, Lin *et al.*[8] have ranked various information sources to identify the strongest predictors of an interaction. However, some of these approaches[5,8] also use high-throughput experimental PPI data itself as a predictor[b]. In contrast, our goal is to predict PPIs *completely independently* of experimental PPI data. In other work, Deng *et al.*[3] have used sequence-based domain signatures, derived from low-throughput data, to identify interacting domains between proteins. None of these methods incorporate structure-based approaches.

Our work is different from existing work in the introduction of structure-based methods as additional predictors of PPIs. The use of such methods provides several advantages. First, these methods can provide *insight* into how, if at all, an interaction happens, unlike guilt-by-association methods which do not. Second, for many protein pairs very little functional annotation is available and structure-based methods might often be the only available predictors. Third, as we show, these methods can be used in addition to existing methods, allowing us to improve upon current performance. We note that Lu, Lu & Skolnick[13] have explored the use of purely structure-based methods to predict PPIs. In comparison, our structure-based method has several advantages (described later) and we describe how it can be integrated with other information sources.

A possible concern might be that current structure prediction methods are not sufficiently accurate and may not work well for every protein-pair. In response, we note that our framework is modular so that better methods can be substituted in, as they become available. Second, our method is homology-based and will improve in performance and coverage as the recent NIH-funded push to elucidate more structures gains momentum.

Another concern might be that just because two protein structures interact *in-silico*, they might not interact *in-vivo*. This risk can be mitigated by combining inferences based on structural-techniques with other kinds of data. Also, note that this concern is equally applicable to existing approaches. Similarly, like many previous approaches, we restrict ourselves

[b]The usual reasoning in such cases is that high-throughput PPI determination methods are noisy enough that they only *indicate* an interaction, not *confirm* it.

to pairwise protein interactions, even though more than two proteins may simultaneously interact *in vivo*.

2. Problem Formulation

We now provide a precise formulation of the two problems we address here:
Problem [StructOnly] Given two proteins p and q, and their associated sequences S_p and S_q, compute the probability that p and q interact.
Problem [Struct&OtherInfo] Given two proteins, p and q, their associated sequences S_p and S_q, and optional annotation information $\{X_p^1, X_p^2, \ldots\}$ and $\{X_q^1, X_q^2, \ldots\}$, compute the probability that p and q interact.

In StructOnly, note that we only require the protein sequences, and not structures. If necessary, the protein sequences can themselves be inferred from the corresponding gene sequences. In Struct&OtherInfo, different kinds of annotation information can be incorporated, as available. Our method for solving this problem can be used with as many information sources as desired, but here we have restricted ourselves to a few:

#	Name	Description
1	Coexpression	Similarity between expression levels of the corresponding genes
2	Colocalization	Co-localization information for the two proteins
3	GO	Similarity between Gene Ontology(GO) terms for the two genes
4	MIPS	Similarity between MIPS terms for the corresponding genes
5	Domain	Seq. motifs indicating the presence of interacting domains
6	Coessentiality	Whether one, both, or none of the corresponding genes are *essential*

Table 1: The various kinds of functional annotation used in Struct&OtherInfo. These benchmark annotations have previously been found to be particularly relevant in PPI predictions (see Supp. Info. for details).

3. Algorithms

3.1. *Problem #1:* StructOnly

Here, we follow a two-staged process (see Fig 1(a)). The advantage of this two-staged process is that as structure-based methods improve in accuracy, better ones can be plugged into the first stage.

3.1.1. Stage 1: Computing Interaction Energies

Here we introduce DBLRAP ("DouBLe RAPTOR"), a novel algorithm that exploits the idea that if the homologs of a pair of proteins interact in a specific way, the latter will also interact in a similar way. The algorithm consists of two major components: (1) construction of the complex template database, and (2) threading the two sequences to each potential complex template.

The complex template database is derived from the latest SCOP[9] database (i.e., SCOP v1.67) as follows: we first check if two protein domains can form a complex as per the following rule. For any pair of SCOP

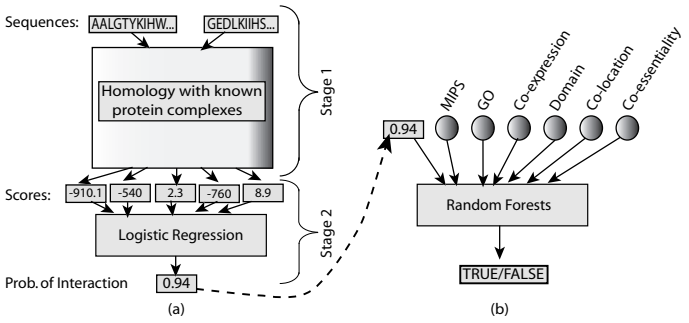


Figure 1: Schematic of our method for (a) StructOnly (b) Struct&OtherInfo

domains with the same PDB ID, we calculate their interfacial contacts, using the same method described in Lu, Lu & Skolnick[13]. If there are more than 10 interfacial contacts between two domains, then we assume that they form a complex. Next, we remove redundant complexes to improve computational efficiency. We use the following clustering method. Suppose we have two protein complexes C_1 and C_2, which are composed of domains A_1 and B_1 and domains A_2 and B_2, respectively. We classify C_1 and C_2 into the same cluster if one of the following two conditions are satisfied: i) A_1 and A_2 are in the same SCOP family and so are B_1 and B_2; ii) A_1 and B_2 are in the same SCOP family and so are A_2 and B_1. We randomly choose one representative from each complex cluster. All the representatives together form a complex template database. In total, the complex template database contains 2443 complexes, which are composed of 4142 unique SCOP domains.

After constructing the complex template database, we then thread each sequence pair to all the complex templates to find the best potential match. We align each sequence pair to the best-matching complex. Using this alignment and the interaction pattern between the complex's constituent subunits, we can also calculate the interfacial energy between our input proteins. The interfacial potential parameters are taken from Lu, Lu, & Skolnick's[13] paper. For computational efficiency, in the actual implementation we did some preprocessing first, the details of which are in the Supp. Info.

In summary, for any given sequence pair (p and q), the threading-based interaction prediction method will generate two alignment scores (E_p, E_q), their associated z-scores (z_p, z_q) , and an interfacial energy (E_{pq}). These are fed into the logistic regression model to predict interaction.

DBLRAP circumvents the docking problem: searching for the optimal orientation of two proteins in a complex. But it has a limitation that the number of complexes with known structures is not yet sufficiently large,

though it is increasing. An alternative approach would be to use homology only to predict structures for individual proteins and then use methods for protein docking to compute and score the optimal relative orientation of the two structures. In theory, this approach should have greater coverage: homology-based structure prediction is possible, and reasonably accurate, for many proteins now. However, our limited exploration indicated that this method does not work very well, possibly because docking programs are not yet sufficiently good.

3.1.2. Stage 2: From Energy Values to Interaction Probabilities

We use binary logistic regression[2] to classify whether a set of scores corresponds to an interaction or not. In binary logistic regression, the goal is to predict a binary output variable Y, given a set of r predictor variables $\mathbf{X} = \{X_1, X_2, \ldots, X_r\}$. For an instance i, suppose y_i and $\mathbf{x_i} = \{x_{i1}, x_{i2}, \ldots, x_{ir}\}$ are the random variables corresponding to Y and $\mathbf{X}$, respectively. Let $\theta_i = P(y_i = 1|\mathbf{x_i})$. In this model, the dependence of θ_i on $\mathbf{x_i}$ is expressed by the logit function:

$$logit(\theta_i) = log(\frac{\theta_i}{1-\theta_i}) = \alpha + \beta^t \mathbf{x_i} = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_r x_{ir} \quad (1)$$

This can be rewritten as:

$$\frac{P(y_i = 1|\mathbf{x_i})}{P(y_i = 0|\mathbf{x_i})} = e^{\alpha + \beta^t \mathbf{x_i}} \quad \text{or} \quad P(y_i = 1|\mathbf{x_i}) = \frac{e^{\alpha + \beta^t \mathbf{x_i}}}{1 + e^{\alpha + \beta^t \mathbf{x_i}}} \quad (2)$$

Logistic regression was performed by the standard Iterative Re-weighted Least Squares algorithm (the R package: `http://www.r-project.org`).

We now describe how we have set up the logistic regression problem for our case. The output variable Y is the probability of interaction of two proteins p and q. The predictor variables come from the first stage. For proteins p and q, DBLRAP provides their interfacial energy E_{pq}, their respective alignment scores E_p and E_q, as well as associated z-scores z_p and z_q. In addition to these, we also put in $\sigma_{pq} = N_p + N_q$ and $\pi_{pq} = \sqrt{N_p N_q}$, where N_p and N_q are the sequence lengths of p and q. Finally we introduce, as separate predictor variables, various functions and combinations of the existing terms: e.g. $\frac{E_p}{N_p}$, $\frac{E_q}{N_q}$, $\frac{E_{pq}}{\sigma_{pq}}$, $\frac{E_{pq}}{\pi_{pq}}$, $\sqrt{E_{pq}}$, etc.

We intentionally built an initial model with an excessively large set of predictor variables: one of our goals was to identify the most informative subset of predictors, using Akaike Information Criterion (AIC) to determine the subset with the optimal trade-off between prediction accuracy and subset-size. The AIC score for a logistic regression model is defined by:

$$AIC = -2\text{log-likelihood} + 2k/N$$

where k = number of predictor variables, N = number of instances in the dataset, and the log-likelihood of data under the model is computed using Eq. 2. The subset of predictor variables with the lowest AIC score was chosen.

The use of logistic regression for prediction confers certain advantages. It allows us to combine multiple scores (interaction energies, z-scores etc.), possibly from different methods. Functions of these scores can also be considered. We can then use logistic regression to identify the most relevant subset of predictors. Compared to Lu, Lu, & Skolnick[13], who only compared the interfacial energy against some threshold, the use of logistic regression allows us to make more sophisticated decisions.

3.2. *Problem# 2:* Struct&OtherInfo

For classification purposes one can associate, with each pair of proteins p and q, a data-vector $D_{pq} = (d_1, \ldots, d_6)$ that contains information from the six non-structure-based information sources described in Table 1. To add structure-based information to this, we simply add one more feature d_7 to D_{pq}. Here, d_7 is the probability of interaction between proteins p and q as computed using logistic regression. Given some training data consisting of known true and likely false interactions, we then train a random forest to classify a possible interaction based on its data-vector (see Fig 1b).

Random Forests: Random forests[7] (RF) generalize the intuition behind decision trees. Given a dataset $\mathbf{D}$ of N data-vectors $D_1, \ldots, D_N$, κ decision trees are constructed. For each tree, only a subset of the feature-space is used to train the tree using the data $\mathbf{D}$. For example, for tree T_{12}, only the features $d_1, d_3,$ and d_7 might be used to create it. Given a test data-vector D_t, the predicted class is determined by running down D_t on each tree and then taking the majority vote over the predicted classes. Random forests can handle missing data. The procedure for handling missing data is somewhat involved; please see the original reference[7] for details.

Our use of random forests is rather straightforward. Our feature space consists of the 7 features described earlier.We then trained a random forest with 500 trees over this space.

Though random forests have only recently been introduced, they have quickly become very popular. They have many desirable characteristics: they rarely overfit the data; they allow classification when features are not independent; they allow for missing values. Lastly, their output is easy to analyze in terms of identifying the strongest predictors and the relationships between the different features. We also note that their usefulness in prediction and analysis of PPIs has previously been demonstrated[12].

Dataset	Interactions			Motivation behind creating the dataset	Post Filtering Interctns.
	Pos.	Notes	Neg.		
Lt	100	From high-quality low-throughput experiments	400	Low-throughput interactions provide "gold-standard" pos.s	69
HtFewAnnot	508	Between 1000 proteins with little functional annotation	2000	Existing guilt-by-assoc. methods do not work well with these	332
HtManyAnnot	489	Between proteins with a lot of functional annotation	300	Test how to combine structure-based methods with other info.	160

Table 2: The construction of three datasets for yeast PPI data. The positive interactions (#'s shown in table) were retrieved from GRID while (putative) negative interactions were generated by randomly pairing two yeast proteins. The difference between the datasets is primarily in how different positive sets were picked. The datasets were filtered to keep only those interactions for which homologous models could be found.

4. Results

Datasets: In this work, we have focused on predicting PPIs in yeast (*S. cerevisiae*). The list of experimentally discovered PPIs for yeast was retrieved from GRID[1]. From this database, three datasets were created: Lt, HtFewAnnot, and HtManyAnnot (see Table 2). The datasets differed in how their positive examples (true interactions) were selected (see Notes in Table 2). Note that because of the significant error-rate[15] in high-throughput experiments, some of the positive examples in HtFewAnnot and HtManyAnnot are likely to be incorrect.

Collecting negative examples (false interactions) is difficult: experimentally confirmed false interactions are rare. As such, we had to design our own— a problem faced by other researchers as well[15,12]. We followed Qi *et al.*'s strategy of considering a random pair of proteins as non-interacting. Since, on average[12], only 1 in 600 possible interactions is true, the chances of a random pair being truly non-interacting are $> 99\%$.

However, not all interactions in the datasets corresponded to protein-pairs for which homologous complexes could be found. Therefore, we had to filter out a subset of the dataset. As discussed before, as more structures become available, the coverage of the homology-based methods will increase and fewer pairs will be filtered out.

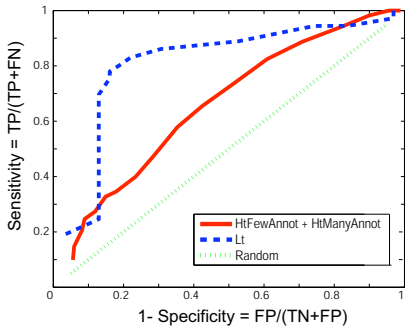
Using Only Structure-based Method (StructOnly):

Using the AIC criterion as described before, we discovered that the subset of predictors of interaction with the optimal balance between model complexity and goodness-of-fit were: $\{\frac{E_{pq}}{\pi_{pq}}, z_p, z_q, \pi_{pq}\}$, where π_{pq} is the square root of the product of sequence lengths of p and q. Of these, $\frac{E_{pq}}{\pi_{pq}}$ ($p < 0.001$) and

z_p, z_q ($p < 0.05$) were the more significant predictors.

In hindsight, it does seem reasonable that E_{pq}/π_{pq} is a stronger predictor of interaction than E_{pq} itself: for large proteins, even relatively weak interactions will have a large (negative) interfacial energy, simply because of there being more interacting entities. Thus, it makes sense that the energy score should be normalized by the sequence length of the two proteins.

We tested our method by 4-fold cross-validation on the LT dataset. In addition, the method was trained on the entire LT dataset and tested on the combined HTFEWANNOT + HTMANYANNOT dataset. By comparing against some threshold value (say $p_{thresh} = 0.5$), the probabilities of interaction predicted by logistic regression can be interpreted as true/false interactions. By varying p_{thresh}, we can plot the sensitivity-vs.-specificity (ROC) curve of the method (see Fig 2a). As can be seen, the structure-based method provides significant signal for prediction purposes. The performance of the method is better on the low-throughput (LT) dataset than on the high-throughput datasets. A possible cause might be that the high-throughput datasets have more errors, i.e., negative examples mis-labeled as positive. Of course, the LT dataset is smaller, and the better performance on it needs more validation. It is also possible that the Skolnick potentials work better for LT dataset. In future work, we plan to explore these issues further.



(a)

Features	Error
All	8.4%
All - Coexpression	28.4%
All - Structure	11.6%
All - Domain	10.9%
All - Coessentiality	8.4%
All - GO	8.4%
All - MIPS	7.7%
All - Colocation	5.8%

(b)

Figure 2: (a) STRUCTONLY: Specificity-vs.-Sensitivity curve when using only the structure-based approach. TP=True Pos., FP=False Pos., TN=True Neg., FN=False Neg. The dotted diagonal line indicates the baseline, a method with zero predictive power. The performance of our method is better for LT than for HTFEWANNOT +HTMANYANNOT. A possible reason might be that the latter datasets themselves might have mislabeled instances. (b)STRUCT&OTHERINFO: Classification error, and its dependence on the various features. "All - X" indicates that all features, except X, have been used for classification. As can be seen, the classification error increases if the structure-based method is not used.

Combining Various Information Sources (STRUCT&OTHERINFO):

We tested our entire framework on the HTMANYANNOT dataset, a dataset specifically chosen for proteins with lots of functional annotation available. We used 5-fold cross-validation to evaluate our method, using the cross-validation error (CVE) as the quality metric.[c]

With average sensitivity = 94.1% and specificity = 92.1%, the overall performance of our method is better than that of existing work, e.g., Zhang *et al.*'s[18] (sensitivity = 81% at specificity = 80%, approximately)[d]. Even when experimental PPI data itself has been used as one of the predictors by others (e.g., Lin *et al.*[8]: sensitivity = 98%, specificity = 92%, approximately), our method— which is *completely* independent of experimental PPI information— performs comparably.

One interesting question is: "do structure-based methods contribute to the predictive power, compared to other features?" To quantify a feature's importance, we removed it from the mix and recomputed the CVE. The difference between this CVE and the baseline CVE (with all the features present) indicates the increase in accuracy offered by including that feature. As the table in Fig 2 shows, coexpression is the most important feature, followed by the information provided by our method. Some of the other features, e.g. colocation, do not seem to be particularly important.

4.1. *Novel Predictions*

Predictions on Less-Characterized Proteins: The proteins in the HTFEWANNOT have very little functional annotation and very few known PPIs (see Supp. Info. for more details). For these, there isn't enough functional annotation for "guilt-by-association" methods to work; in contrast, our structure-based method will still work.

We tested all possible pairs in this set for interaction, using our structure-based method, without any additional functional annotation. The probabilities of interaction, as computed by logistic regression, were used to rank the pairs and the top 2000 pairs were chosen. The network formed by these predicted set of interactions (see Supp. Info. for the predicted set) shows some intriguing properties. It has a scale-free character[15], just like the experimentally-determined yeast PPI network, i.e., the node degree distribution follows the power law. Moreover, the two power-law coefficients

[c]Computing 5-fold cross-validation error (CVE): data was randomly partitioned into five equal parts. Four of the parts constituted the training set while the fifth one made up the test set. The error was computed as the classification error on this test set. By repeating this error computation for each of the classes, five error values were computed and averaged to compute the CVE.

[d]We compared against Zhang *et al.*'s performance in the case when they did not use experimental PPI data as a predictor

are comparable (1.9 for predicted network; 2.3 for the yeast interactome).

In the predicted network, the protein CHS2 is a hub (86 interactions), and the set of its partners is enriched for genes involved in amino acid and amine transporter activity. So, we hypothesized that this protein would have similar functions. This turns out to be true— CHS2 is involved in transferring N-acetylglucosamine to chitin. It is also relevant in disease-treatment; some recent work on developing antiprotozoal drugs has focused on targeting the chitin-synthesis pathway[6]. Similarly, for DSF2—a hitherto uncharacterized gene—the set of its predicted interaction partners is enriched for genes related to DNA transposition and retrotransposons ($p < 0.001$), indicating DSF2's possible function.

Disease-Related Proteins: In the predictions, we also specifically looked for homologs of human disease-related genes. We describe a few findings here; the rest are in Supp. Info.

The human homolog of RAD28 has been implicated in Cockayne Syndrome (related to malfunctions in DNA-repair machinery). Currently, there are only two known PPIs involving RAD28. Our method predicts 19 additional PPIs, and 6 of the predicted partners are involved in DNA repair.

Similarly, the human homolog of PAT1 is Adrenoleukodystrophy– a neurodegenerative disease caused by a malfunctioning fatty-acid transporter protein. There are only three known PPIs involving PAT1; our method predicts 26 more. Moreover, the set of its interaction partners is enriched for proteins involved in lipid and fatty acid transport ($p < 0.01$).

Genome-scale Predictions: We used the structure-based method, without any additional functional annotation, to perform an all-vs-all prediction of the interactions in the yeast genome (see Supp. Info. for predictions). The predicted network has a scale-free character similar to the known yeast interactome and has about 9% overlap with it. This is significantly better than overlap achieved by Lu, Lu & Skolnick's[13] method and is comparable to the overlap between large-scale experimental PPI datasets.

5. Discussion

We have described how structure-based methods can be integrated with other genomic and proteomic information for predicting PPIs. Structure-based methods can be used by themselves when other functional annotation is not available. When used in conjunction with functional annotation, their addition improves prediction accuracy over existing methods.

Our future efforts will focus on (1) applying this method to mammalian genomes, (2) incorporating other kinds of functional annotation (e.g., corre-

lated mutations[10]), and (3) using docking programs as an additional way of computing interfacial energies. As mentioned before, our brief exploration indicated that current docking programs did not perform satisfactorily. However, more work might suggest ways to improve them for our purposes.

Acknowledgments: The authors thank Dr. Ying Xu and Dr. Fengluo Mao at the Univ. of Georgia for allowing us the use of their Linux cluster.

References

1. B.J. Breitkreutz, C. Stark, M. Tyers. The GRID: the General Repository for Interaction Datasets *Genome Biol*, 4(3):R23, 2003
2. T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall
3. M. Deng, S. Mehta, F. Sun, and T. Chen. Inferring domain-domain interactions from protein-protein interactions. *Genome Research*, 12(10):1540–8, 2002.
4. A. Jaimovich, G. Elidan, H. Margalit, and N. Friedman. Towards an Integrated Protein-Protein Interaction Network. *Proceedings of RECOMB*, 2005.
5. R. Jansen *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–53, 2003.
6. E.L. Jarroll and K. Sener. Potential drug targets in cyst-wall biosynthesis by intestinal protozoa. *Drug Resist Update*, 6(5), 2003.
7. L. Breiman. Random Forests. *Machine Learning Journal*, 45(1), 2001.
8. N. Lin, B. Wu, R. Jansen, M. Gerstein, H. Zhao. Information assessment on predicting protein-protein interactions. *BMC Bioinformatics* 18;5:154, 2004.
9. A. Murzin, S. Brenner, T. Hubbard, C. Chothia. SCOP:A structural classification of proteins database *J Mol Biol* 247, 536-540, 1995
10. F. Pazos *et al.* Correlated mutations contain information about protein-protein interaction *J Mol Biol*, 271(4):511-23, 1997
11. D.L. Price, D.R. Borchelt, and S.S. Sisodia. Alzheimer Disease and the Prion Disorders *Proc Natl Acad Sci USA*, 90(14):6381-4, 1993
12. Y. Qi, J. Klein-Seetharaman, and Z. Bar-Joseph. Random Forest Similarity for Protein-Protein Interaction Prediction. *PSB*, 2005.
13. H. Lu, L. Lu, and J. Skolnick. Development of Unified Statistical Potentials Describing Protein- Protein Interactions *Biophysical J*, 84:1895-1901, 2003.
14. P. Uetz, S.V. Rajagopala, Y.A. Dong, and J. Haas. From ORFeomes to protein interaction maps in viruses. *Genome Research*, 14(10B):2029–33, 2004.
15. C. von Mering *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.
16. J. Xu, M. Li, D. Kim, and Y. Xu. RAPTOR: optimal protein threading by linear programming. *J. of Bioinformatics and Comp. Biol.*, 1(1):95–117, 2003.
17. Y. Yamanishi *et al.* Extraction of correlated gene clusters from multiple genomic data. *Bioinformatics*, 19:323–30, 2003.
18. L.V. Zhang, S.L Wong, O.D. King, F.P. Roth. Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics*, 5(38), 2004.