A Probabilistic Graph-Theoretic Approach to Integrate Multiple Predictions for the Protein–Protein Subnetwork Prediction Challenge

Hon Nian Chua,[a] Willy Hugo,[b] Guimei Liu,[b] Xiaoli Li,[a] Limsoon Wong,[b] and See-Kiong Ng[a]

[a]*Data Mining Department, Institute for Infocomm Research, Singapore*

[b]*School of Computing, National University of Singapore, Singapore*

The protein–protein subnetwork prediction challenge presented at the 2nd Dialogue for Reverse Engineering Assessments and Methods (DREAM2) conference is an important computational problem essential to proteomic research. Given a set of proteins from the *Saccharomyces cerevisiae* (baker's yeast) genome, the task is to rank all possible interactions between the proteins from the most likely to the least likely. To tackle this task, we adopt a graph-based strategy to combine multiple sources of biological data and computational predictions. Using training and testing sets extracted from existing yeast protein–protein interactions, we evaluate our method and show that it can produce better predictions than any of the individual data sources. This technique is then used to produce our entry for the protein–protein subnetwork prediction challenge.

Key words: **protein–protein interactions; data mining; data integration**

Introduction

Protein–protein interaction data are central to the field of proteomic research. They are widely used in many computational and biological analyses, such as the study of biological pathways, protein complexes, and protein function, that contribute to the elucidation of complex cell machineries. However, despite the advent of such high-throughput detection technologies as the yeast-two-hybrid system and affinity-purification mass spectrometry, there are still no experimental detection methods that can unravel an entire interactome completely. As such, computational prediction approaches have remained a viable alternative to experimental approaches.

The protein–protein subnetwork challenge presented at the 2nd Dialogue for Reverse Engineering Assessments and Methods (DREAM2) conference involves the prediction of the interactions between 47 yeast proteins. Participants are required to rank all possible interactions between the proteins based on decreasing reliability. In order to build on previous work on protein interaction prediction,[1–7] we propose a framework to integrate predictions made by these methods using a probabilistic method. The framework is adapted from the protein function prediction framework described in a recent work[8] shown to outperform large-scale prediction approaches. A variety of existing protein–protein interaction-prediction techniques are integrated, including domain–domain interactions,[1] interaction motifs,[2–4] paralogous interactions,[5,6] and protein function similarity.[7] We also introduce a novel data-centric approach to protein–protein interaction prediction. The method makes use of

Address for correspondence: Hon Nian Chua, 1 Fusionopolis Way, #21-01 Connexis (South Tower), Singapore 138632. Voice: 64082168 hnchua@i2r.a-star.edu.sg

The Challenges of Systems Biology: Ann. N.Y. Acad. Sci. 1158: 224–233 (2009).
doi: 10.1111/j.1749-6632.2008.03760.x

the closed itemset mining technique to identify domain or functional combination pairs associated with interacting proteins to derive complex interaction rules that are not covered by the existing approaches. Using currently known protein–protein interactions from the BioGRID database, we created a set of training and testing data to evaluate our prediction framework. Evaluation of these data show that integrating multiple predictions from the different approaches using our framework significantly outperforms any individual prediction. Our entry for the DREAM 2 protein–protein subnetwork prediction challenge was created by using the technique with all known interactions from BioGRID as training data. Our entry outperformed those of other participants by a clear margin based on evaluation using the area under the precision-recall curve (PRC) and the receiver operating characteristics (ROC) curve.

Datasets

To construct the proposed interaction prediction framework for performance evaluation, we processed the set of protein–protein interactions from the BioGRID database[9] in the following manner:

Remove Non physical Interactions

Genetic interactions are filtered out based on their experimental type so that only physical interactions are used for evaluation. Filtered genetic experimental types includes *Epistatic MiniArray Profile*, *Phenotypic Enhancement*, *Phenotypic Suppression*, *Synthetic Rescue*, *Synthetic Lethality*, *Synthetic Growth Defect*, *Dosage Lethality*, *Dosage Rescue*, and *Dosage Growth Defect*.

Divide into Training and Testing Sets

Proteins involved in the remaining physical interactions are divided randomly into two groups in the ratio 6:4. Physical interactions between proteins in the larger group are used as training data.

Define Gold-Standard Positives for Testing Set

Physical interactions between proteins in the smaller group are used as gold-standard positive (GSP) interactions after further filtering based on these criteria: (1) the interaction must be observed in at least three independent physical experiments (to ensure that the interactions can be replicated and are unlikely to result from experimentation errors or noise); (2) the interaction must involve proteins that are reported to have at least five interaction partners (to ensure that the proteins in the GSP set are reasonably well studied to reduce overestimation of false-positive rates during evaluation). We do not define a set of gold-standard negative interactions, although such a set is defined and used in Ref. 1 to predict domain–domain interactions that are used in our method.

The latest release (version 2.0.33) of yeast interactions from the BioGRID database is obtained from http://www.thebiogrid.org/. This consists of 71,503 unique interactions, of which 38,555 are physical. After processing the data as described above, we obtained 19,215 interactions for training and 446 interactions for testing. The interactions for training and testing are associated with the proteins divided in the ratio 6:4. Interactions for testing are further filtered as described above.

Method

Data Sources of Predicted Interactions

Seven sources of predicted interactions are integrated for predicting protein–protein interactions:

Domain–Domain Interactions

Li and colleagues proposed a probabilistic technique[1] to infer domain–domain interactions using both positive and negative training datasets. Physical interactions from protein

interaction data are used to construct the positive training dataset I. Unlike conventional approaches that use random pairing to generate artificial noninteracting protein pairs as negative training data, the authors generated biologically meaningful noninteracting protein set $\mathcal{N}$ based on the proteins' biological information, namely, proteins are most unlikely to interact if they are from different cellular locations and are involved in different biological processes. The domain information of proteins is obtained from the Pfam[10] database. The probabilistic model assigns interaction probabilities (interacting probability and noninteracting probability) to each domain pair based on its occurrence in the protein–protein interacting set and the negative set.[1]

Given a protein pair $(p_i, p_j) \in I$, we infer that domain $d_{i,r}$ potentially interacts with domain $d_{j,s}$ with a probability of $1/(|p_i| \cdot |p_j|)$, where $|p_i|$ and $|p_j|$ are the number of domains in proteins p_i and p_j respectively; $d_{i,r}$ and $d_{j,s}$ are the rth and sth domains of proteins p_i and p_j, respectively. Let a set of predefined classes be $C = \{I, \mathcal{N}\}$ and all the domain pairs set be DP. For any domain pair $(d_x, d_y) \in DP$, their interacting probability $P((d_x, d_y)|c)$, with Laplacian smoothing and $c \in C$, is defined as:

$$P((d_x, d_y) \mid c) = \frac{1 + freq((d_x, d_y), c)}{|DP| + \sum_{k=1}^{|C|} freq((d_x, d_y), c)} \quad (1)$$

where $freq((d_x, d_y), c)$ is the interacting frequency of (d_x, d_y) in class $c \in C = \{I, \mathcal{N}\}$. For any given domain pair, if its interacting probability $P((d_x, d_y)|I)$ is bigger than noninteracting probability $P((d_x, d_y)|\mathcal{N})$, it will be regarded as interacting domain pair.

Given a new protein pair (p_i, p_j), we predict whether protein p_i interacts with p_j based on the underlying domain–domain interactions between the two proteins. In order to perform classification (i.e., to judge whether the protein pair may interact with each other or not), we compute the posterior probability $P(c|(p_i, p_j))$, $c \in C = \{I, \mathcal{N}\}$. The prior probability $P(c)$ of class c is defined as:

$$P(c) = \frac{\sum P(c, (p_i, p_j)), (p_i, p_j) \in I \cup \mathcal{N}}{|I| + |\mathcal{N}|} \quad (2)$$

The technique then uses the joint probabilities of domain pairs and classes to estimate the probabilities of classes given a protein pair. Our classifier is described as follows:

$$P(c \mid (p_i, p_j)) = \frac{P(c)^* \prod_{1 \le r \le |p_i|, 1 \le s \le |p_j|}^{|p_i||p_j|} P((d_{ir}, d_{js}) \mid c)}{\sum_{k=1}^{|C|} P(c_k)^* \prod_{1 \le r \le |p_i|, 1 \le s \le |p_j|}^{|p_i||p_j|} P((d_{ir}, d_{js}) \mid c_k), c_k \in C = \{I \cup N\}} \quad (3)$$

For a protein pair (p_i, p_j), the class with highest $P(c|(p_i, p_j))$ is assigned as its final class label. In other words, if $I = \text{argmax}_c\ P(c|(p_i, p_j))$, then the protein pair (p_i, p_j) will be classified as an interacting pair. Otherwise, it is classified as noninteracting.

Relative Specificity Similarity

Wu and coworkers[7] proposed a method to predict yeast protein–protein interactions using Gene Ontology's[11] structure and annotations. They proposed a relative specificity similarity (RSS) measure for computing the similarity between the Gene Ontology (GO) annotations of two proteins. The RSS was adapted from another similarity measure proposed in Ref. 12, and takes into account the directed acyclic graph (DAG) structure of the GO terms. Wu and coworkers showed that protein pairs with high RSS scores in both the *cellular component* (CC) and *biological process* (BP) GO namespaces are very likely to coincide with known interaction pairs. Here, we compute the RSS score between all pairs of proteins from the input

dataset as the product of the individual RSS score from the CC and BP namespaces:

$$RSS\,(u, v) = RSS_{CC}\,(u, v) \cdot RSS_{BP}\,(u, v) \quad (4)$$

where RSS_{CC} and RSS_{BP} are the relative specificity similarity scores between protein u and v computed based on the *cellular component* and *biological process* namespaces, respectively.

Gene Ontology Combination Rules

We also propose a novel way of predicting protein–protein interactions based on the GO annotations of each protein. Using a data-centric approach, we identify pairs of GO annotation combinations that are likely to occur in interacting proteins. For closed itemset mining, each protein is transformed into a transaction record, where each GO term annotated to the protein is an item in the transaction. A set of frequently occurring GO term combinations, F, is identified using closed itemset mining with a support threshold of 5. Each pair of GO term combinations C_A and C_B is then scored based on the likelihood that a pair of proteins with annotations C_A and C_B, respectively, are known to interact in training dataset:

$$P(I|C_A, C_B) = \frac{\sum_{(u,v)\in E_{AB}} \delta(u, v)}{|E_{AB}| + 1},\ C_A \in F,\ C_B \in F \quad (5)$$

where E_{AB} refers to the set of unique protein pairs (x,y) for all $A \subseteq GO_x$, $B \subseteq GO_y$, $x \neq y$; GO_u refers to the set of GO terms annotated to protein u; $\delta(u,v) = 1$ if u and v interacts, 0 otherwise.

The likelihood that a given pair of proteins, u and v, interact is then scored using the highest likelihood score of all relevant GO term combination pairs:

$$P(u, v) = \max_{C_A \subseteq GO_u, C_B \subseteq GO_v} P(I \mid C_A, C_B) \quad (6)$$

where GO_u refers to the set of GO terms annotated to protein u; $P(I|\ C_A,C_B)$ is the likelihood score for C_A and C_B described above in equation (5); C_A and C_B are members of set F, that is, $C_A \in F$ and $C_B \in F$.

Closed Pattern Motif Pairs

Liu and colleagues[3] proposed an approach for finding interaction motif pairs based on the observation that proteins usually contain a small number of interaction sites, and the interaction sites of the proteins that have common interacting partners are likely to have similar structures and common sequence motifs.[4] Pairs of interacting motifs are mined from protein sequences and protein interaction networks. The interacting motif pairs are then used to assign a confidence score to protein pairs containing them for protein interaction prediction.

This method comprises four steps. In the first step, spurious interactions from the protein interaction network are removed using a simple measure known as the Czekanowski-Dice distance (CD-Distance), which was shown to be very effective in finding false-positive errors from high-throughput interaction data.[13] The CD-distance between two proteins u and v is defined as:

$$CD\,(u, v) = \frac{2|\mathcal{N}_u \cap \mathcal{N}_v|}{|\mathcal{N}_u \cup \mathcal{N}_v| + |\mathcal{N}_u \cap \mathcal{N}_v|} \quad (7)$$

where $\mathcal{N}_u$ and $\mathcal{N}_v$ are the proteins interacting with u and v, respectively.

In the second step, it identifies groups of proteins that have common interacting partners, called closed pattern (CP) protein groups, from the purified interaction network. A CP protein group contains at least l proteins and has at least k common interacting partners. For each group, it finds sequence motifs from the associated protein sequences using PROTOMAT. To avoid generating too many highly similar motifs, only maximal CP protein groups are considered for motif generation.

In the third step, the interacting confidence scores between every pair of motifs are computed. The confidence of a motif pair (m_1, m_2) is defined as:

$$conf_m(m_1, m_2) = \frac{\mathcal{N}_{\text{int}}\,(m_1, m_2)}{\mathcal{N}_{total}\,(m_1, m_2)} \quad (8)$$

where $\mathcal{N}_{int}\ (m_1,\ m_2)$ is the number of interacting protein pairs containing $(m_1,\ m_2)$ and

$N_{total}(m_1, m_2)$ is the total number of distinct protein pairs containing (m_1, m_2).

In the last step, a confidence score is computed for every protein pair as follows:

$$conf_p(p_1, p_2) = CD(p_1, p_2) \cdot conf_m(p_1, p_2) \quad (9)$$

where $conf_m(p_1, p_2)$ is the maximal confidence of the motif pairs contained in (p_1, p_2).

Correlated Motif Pairs

As another source of interacting motifs for protein interaction prediction, we extended the D-STAR[2] algorithm to mine for correlated motif pairs from the entire *S. cerevisiae* interactome. While the D-STAR algorithm had used the (l,d)-motif model, for simplicity and scalability, we used the simpler regular expression (L,W)-motif model that has been employed by the TEIRESIAS program[14] with considerable success on finding linear motifs in proteins.[15,16] A (l,d)-motif is a nucleotide sequence of length l that matches any short nucleotide pattern that has at most d mutations from it.[21] A (L,W)-motif describes a sequence of at least L consecutive literals and one or more wild-card characters, with literals spanning not more than W positions. This motif model is maximal, and maximizes the coverage of the motif pairs found.

The method starts with mining all (L,W) motifs with a minimum occurrence k in the protein sequences using TEIRESIAS program, where $L = 4$, $W = 7$, and $k = 5$. Each pair of motifs (M_1, M_2) is then scored using the Chi-square statistical measure, as used in Ref. 2:

$$\chi^2_{M_1,M_2} = \frac{(O_{M_1,M_2} - E_{M_1,M_2})^2}{E_{M_1,M_2}} \quad (10)$$

where $O_{M1,M2}$ refers to the set of observed interactions between the protein sets $P(M_1)$ and $P(M_2)$; $P(M_k)$ is the set of proteins whose sequences exhibit occurrence of motif M_k; and $E_{M1,M2}$ refers to the expected number of random interactions between a pair of protein sets with the same size as $P(M_1)$ and $P(M_2)$, respectively.

Since the number of possible (L,W) motifs is enormous, we only consider motif pairs (M_1,M_2) whose protein sets' interaction graph contains a (2,2)-biclique construct. The (2,2)-biclique enforces that there are at least two proteins from $P(M_1)$ and two proteins from $P(M_2)$ that form a full interaction graph. The checking of such construct is done using a simple counting method that removes around 40–50% of the candidate motifs returned by TEIRESIAS. The motifs that are returned by TEIRESIAS are further refined by checking the residues in the wildcard positions of each motif. If these residues share similar properties, we replace the wildcard with a more specific grouping symbol, as described in Ref. 17. This step would guarantee the most specific possible motifs for each protein set. A confidence score is then computed for every protein pair as follows:

$$conf_{cm}(p_1, p_2) = \max_{p_1 \in P(M_1), p_2 \in P(M_2)} \chi^2_{M_1,M_2} \quad (11)$$

Domain Combination Rules

Using Interpro[18] domains obtained from the Saccharomyces Genome Database (SGD),[19] we apply the same technique described above for GO combination rules to domain annotations. The method is exactly the same, with GO annotations replaced by domain annotations in this case.

Paralogous Interactions

Interactions between homologs within the same species have been shown to be conserved,[5,6] and they are used in Ref. 5 as part of a technique to assess the reliability of protein–protein interactions in high-throughput experimental assays. Here we predict the paralogous interactions as follows: (1) the homologs of each protein are inferred by performing the basic local alignment search tool (BLAST)[20] using a E-value threshold of $1e^{-3}$; (2) given two proteins u and v that are not known to interact, we compute the likelihood that u and v interact as:

$$S_{para}(u, v) = \frac{\sum_{(x,y) \in E_{u,v}} \delta(x, y)}{|E_{u,v}| + 1} \quad (12)$$

where $E_{u,v}$ refers to the set of unique protein pairs (x,y) for all $x \in H_u, y \in H_v, x \neq y$; H_u refers to the set of u's homologs; and $\delta(u,v) = 1$ if u and v interact, 0 otherwise.

Integration of Data Sources

In our previous work,[8] we developed a graph-based framework called integrated weighted averaging to integrate data from heterogeneous biological data for protein function prediction. Here, we adapt the framework to integrate the seven data sources to predict protein–protein interactions. The framework involves three key steps:

Graph-Based Model

Each data source is transformed into a list of binary relationships between protein pairs. Each of these sets of protein pairs is then modeled as a graph $G = \{V,E\}$, with each vertex $v \in V$ representing proteins and each edge $(u,v) \in E$ representing a relationship between a pair of proteins u and v, $u \in V$, $v \in V$.

Unified Scoring Scheme

To integrate the information presented by the different data sources in a way that makes sense, the edges in each graph are scored using a common benchmark. Since the primary objective here is to predict protein–protein interactions, the natural benchmark for scoring each edge would be the likelihood of that edge coinciding with a known interaction. However, edges in a graph G may already be weighted in the form of P values, likelihood scores, RSS scores, and others. These weights, which may show some form of correlation to the reliability of the edge, are derived from different context and information, and can be presented in very diverse scales.

To retain the information reflected by these weights in the process of scoring the edges, we subdivide edges from each graph into subtypes based on these edge weights. The range of edge weights is first arranged in ascending order. Starting from the smallest edge weight, the first 100 edges and subsequent edges with the same weight as the 100th edge are placed in the first subgroup. The next 100 edges and subsequent edges with the same weight as the 100th edge are placed in the second subgroup, and so on. Each subgroup k from all graphs is then weighed based on the likelihood of coinciding with a known interaction:

$$p(k) = \frac{\sum_{(u,v)\in E_k} \delta(u,v)}{|E_k| + 1} \qquad (13)$$

where E_k refers to the set of edges in subgroup k; $\delta(u,v) = 1$ if u and v interacts, 0 otherwise.

Graph Integration

The graphs from all subgroups are then combined to form a larger, more complete interaction graph G' that is a superset of the graphs from all the data sources. The weight of each edge (u,v) in G' is computed using a naïve Bayesian approach:

$$r_{u,v} = 1 - \prod_{k \in D_{u,v}} (1 - p(k)) \qquad (14)$$

where $D_{u,v}$ refers to the set of subgroups that contains the edge (u,v), and $p(k)$ is the likelihood weight of subgroup k computed as described in equation (13).

This approach of integration assumes that data sources are independent of each other, which is unlikely in some of the data sources used, such as domain combination pairs and domain interaction pairs. While this assumption may not be very realistic, it is widely known that the naïve Bayes approach works well in many complex real-world applications. Furthermore, due to incompleteness in the data sources used, it may be difficult to accurately model dependence between them. Figure 1 illustrates an example of how the framework works. Graphs G_1, G_2, and G_3 depicts three different data sources, and each graph comes with weighted edges as shown in the first row. In the second row, the edges in each graph have been reweighted based on equation (13). In the last row, the three graphs are integrated into

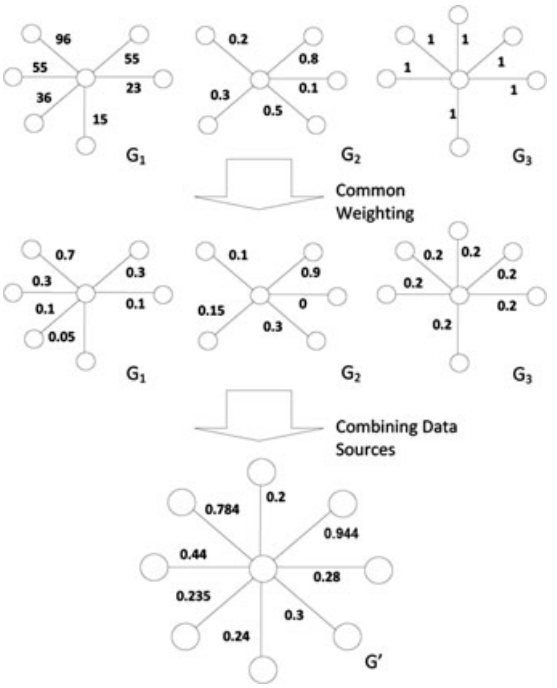

FIGURE 1. Overview of our integration framework for protein–protein interaction prediction.

a more complete graph G', with edge weights computed using equation (14).

Results

Evaluation on Testing Dataset

The computational methods described earlier are applied only on the training data. The GSP test set described above is only used during performance evaluation. Using the test set, we evaluate the performance of individual sources of predicted interactions as well as the combined interactions. Only predicted interactions with both proteins found in the test set are considered. Both precision-recall and ROC are used for the evaluation.

Figures 2 and 3 show the ROC and precision-recall graphs for interactions predicted using (1) domain–domain interaction (DDI); (2) closed patterns motif pairs (CP-MP); (3) relative specificity similarity (RSS) scores; (4) GO combination pairs (GO-CP); (5) domain combination pairs (Dom-CP); (6) correlated motif pairs (Corr-MP); (7) paralogous interactions (Para-Int); and (8) all seven data sources integrated using our prediction framework (Combined). We observe that GO combination pairs yielded the best performance in terms of area under the ROC curve, followed by RSS and CP motif pairs. Similar trends are observed in the precision-recall graphs. Predictions made by integrating the various predictions perform significantly better than those from any individual data source for both evaluation measures. This provides some confidence that our integration framework is able to synergize the predictions from the different approaches to produce more accurate predictions. Table 1 presents the corresponding area under the (precision-recall) curves (AUC) and area under the ROC graph scores for predictions made.

Evaluation on the DREAM Challenge Dataset

The 47 proteins in the DREAM2 challenge dataset have very few known physical interactions between them from the BioGRID database. There are only 19 known unique interactions, of which 15 are self-interactions. This means that there were at most four interactions between the test proteins that are present in our training dataset from BioGRID. Using all known interactions from BioGRID as training data, we used our integration framework to predict all interactions between the 47 proteins. This involves ranking all possible edges between the 47 proteins based on their computed $r_{u,v}$ values in descending order.

Table 2 presents the area under the precision-recall curves and area under the ROC graph scores for predictions made by the various teams who participated in the DREAM2 protein–protein subnetwork challenge, compiled by the organizers of the challenge. We are pleased to find that predictions made by our method outperformed those made by the other teams by a reasonably significant margin. However, the precision of our

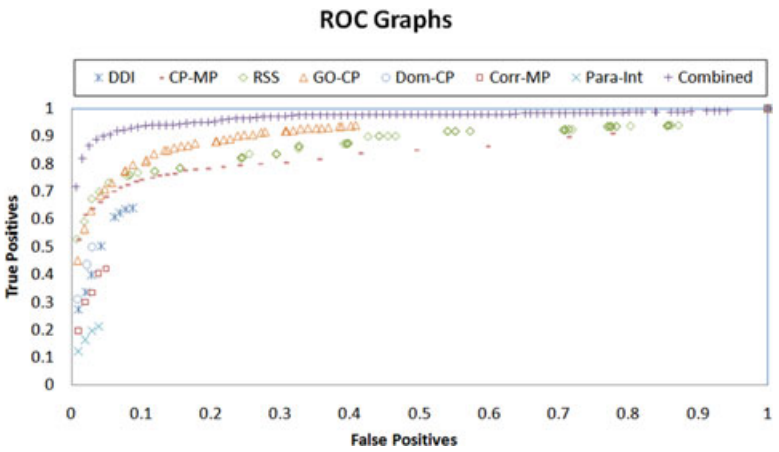

FIGURE 2. Receiver operating characteristic (ROC) graphs for protein–protein interactions predicted using: (1) domain-domain interaction (DDI); (2) closed patterns motif pairs (CP-MP); (3) relative specificity similarity (RSS) scores; (4) GO combination pairs (GO-CP); (5) domain combination pairs (Dom-CP); (6) correlated motif pairs (Corr-MP); (7) paralogous interactions (Para-Int); and (8) all seven data sources integrated using our prediction framework (Combined).

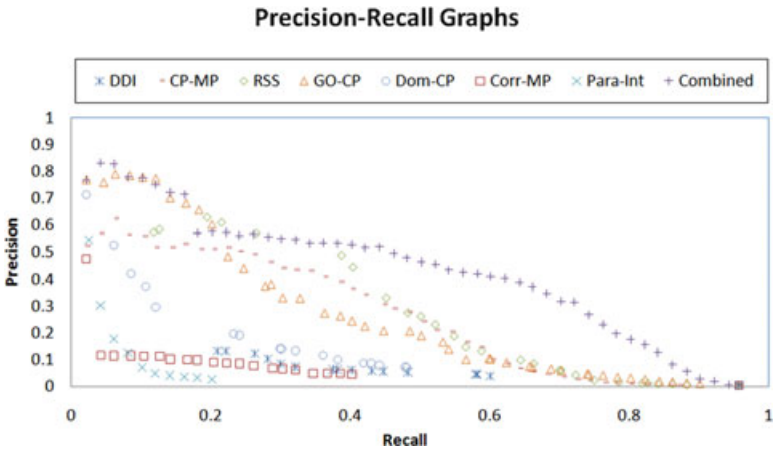

FIGURE 3. Precision-recall graphs for protein–protein interactions predicted using: (1) domain-domain interaction (DDI); (2) closed patterns motif pairs (CP-MP); (3) relative specificity similarity (RSS) scores; (4) GO combination pairs (GO-CP); (5) domain combination pairs (Dom-CP); (6) correlated motif pairs (Corr-MP); (7) paralogous interactions (Para-Int); and (8) all seven data sources integrated using our prediction framework (Combined).

method is still rather low, suggesting that there is much room for improvement in this aspect. The incompleteness of the existing interaction information, as well as the design of the challenge (the testing set for the challenge was produced from yeast two-hybrid interactions, which reflects direct physical interactions, while we included indirect interactions such as affinity capture and copurification) could be possible causes.

Table 3 presents the area under the precision-recall curves and area under the ROC graph scores for predictions made by individual contributing data sources described earlier on the DREAM2 challenge dataset. GO combination pairs performed the best among

TABLE 1. Area under curve (AUC) and receiver operating characteristics (ROC) scores for predictions made using (1) domain-domain interaction (DDI); (2) closed patterns motif pairs (CP-MP); (3) relative specificity similarity (RSS) scores; (4) GO combination pairs (GO-CP); (5) domain combination pairs (Dom-CP); (6) correlated motif pairs (Corr-MP); (7) Paralogous Interactions (Para-Int); and (8) all seven data sources integrated using our prediction framework (Combined) for our cross-validation

Method	AUC	ROC
DDI	0.033761	0.768992
CP-MP	0.126729	0.766510
RSS	0.145970	0.794988
GO-CP	0.119709	0.898615
Dom-CP	0.056730	0.712842
Corr-MP	0.033094	0.671885
Paralogous	0.026646	0.591139
Combination	**0.204618**	**0.956161**

The best figures for each measure are highlighted in bold.

TABLE 2. Area under curve (AUC) and receiver operating characteristics (ROC) scores for predictions made using our method (NetMiner) against competing teams on the DREAM challenge dataset

Team	AUC	ROC
NetMiner	**0.079742**	**0.636008**
Competing teams	0.054518	0.560339
	0.047835	0.492437
	0.050436	0.481776
	0.044577	0.478107

The best figures for each measure are highlighted in bold.

the seven methods using both measures, with slightly better AUC score than the combined prediction. The relatively better performance of GO combination pairs over RSS suggests that there are patterns of interactions between proteins with weak function similarity that GO-CP can identify but RSS cannot. Combining all data sources yielded a significantly higher ROC score than any one source alone, which is consistent with earlier results.

TABLE 3. Area under curve (AUC) and receiver operating characteristics (ROC) scores for predictions made using (1) domain-domain interaction (DDI); (2) closed patterns motif pairs (CP-MP); (3) relative specificity similarity (RSS) scores; (4) GO combination pairs (GO-CP); (5) domain combination pairs (Dom-CP); (6) correlated motif pairs (Corr-MP); (7) paralogous interactions (Para-Int); and (8) all seven data sources integrated using our prediction framework (Combined) on the DREAM challenge dataset

Method	AUC	ROC
DDI	0.043249	0.481266
CP-MP	0.059184	0.546416
RSS	0.059232	0.542788
GO-CP	**0.080694**	0.574670
Dom-CP	0.051473	0.504689
Corr-MP	0.044523	0.494231
Paralogous	0.055691	0.498777
Combination	0.079742	**0.636008**

The best figures for each measure are highlighted in bold.

Conclusions

In this work, we have devised a framework to integrate predictions from various prediction techniques (existing and new ones) using a novel graph-based probabilistic approach. The method was used to produce our entry in the DREAM2 challenge, which outperformed those of other participants in the protein–protein subnetwork challenge. We have also shown that our approach works well on training and testing data built from existing protein–protein interaction data. Our results show that it is advantageous to integrate multiple prediction approaches to produce more complete and accurate predictions, and our framework provides a systematic way for integrating the diverse data for better results. From our study, we find that prediction results can vary greatly with the data sources used. Since the focus of the conference lies in reverse engineering, it may be more appropriate to include in the challenge formulation the data sources from which to make predictions, and if possible positive and negative gold-standard

sets of interactions. By limiting participants to similar sets of input, it is possible to benchmark methodology objectively. A method that can sieve relevant information from multiple sources of noisy and incomplete data and effectively combine them to make predictions will be very valuable.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Li, X.L., S.H. Tan & S.K. Ng. 2006. Improving domain-based protein interaction prediction using biologically-significant negative dataset. *Int. J. Data Mining Bioinform.* **1:** 138–149.
2. Tan, S.H., W. Hugo, W.K. Sung & S.K. Ng. 2006. A correlated motif approach for finding short linear motifs from protein interaction networks. *BMC Bioinform.* **7:** 502.
3. Liu, G., J. Li, S. Lukman & L. Wong. 2007. Predicting protein interactions using interacting motif pair (poster). *The 18th International Conference on Genome Informatics*.
4. Li, H., J. Li & L. Wong. Discovery motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics* **22:** 989–996.
5. Deane, C.M., . Salwiński, I. Xenarios & D. Eisenberg. 2002. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics* **1:** 349–356.
6. Mika, S. & B. Rost. 2006. Protein-protein interactions more conserved within species than across species. *PLoS Comput. Biol.* **2:** 379.
7. Wu, X., L. Zhu, J. Guo, *et al*. 2006. Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. **34:** 2137–2150.
8. Chua, H.N., W.K. Sung & L.S. Wong. 2007. An efficient strategy for extensive integration of diverse biological data for protein function prediction. *Bioinformatics* **3:** 3364–3373.
9. Breitkreutz, B.J., C. Stark & M. Tyers. 2003. The GRID: the General Repository for Interaction Datasets. *Genome Biol.* **4:** R23.
10. Bateman, A. *et al*. 2004. The Pfam protein families database. *Nucleic Acids Res.* **32:** D138–D141.
11. Ashburner, M. *et al*. 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.
12. Wu, H., Z. Su, F. Mao, *et al*. 2005. Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Res.* **33:** 2822–2837.
13. Chen, J., H.N. Chua, W. Hsu, *et al*. 2006. Increasing confidence of protein-protein interactomes. *Proceedings of the 17th International Conference on Genome Informatics*, 284–297.
14. Rigoutsos, I. & A. Floratos. 1998. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics* **14:** 55–67.
15. Neduva, V.R. *et al*. 2005. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol.* **3:** e405.
16. Davey, N.E., D.C. Shields & R.J. Edwards. 2006. SLiMDisc:short, linear motif discovery, correcting for common evolutionary descent. *Nucleic Acids Res.* **34:** 3546–3554.
17. Aasland, R. *et al*. 2002. Normalization of nomenclature for peptide motifs as ligands of modular protein domains. *FEBS Lett.* **513:** 141–144.
18. Mulder, N.J. *et al*. 2007. New developments in the InterPro database. *Nucleic Acids Res.* **35:** D224–228.
19. SGD project. Saccharomyces Genome Database. ftp://ftp.yeastgenome.org/yeast/ (11/9/2007).
20. Altschul, S.F., W. Gish, W. Miller, *et al*. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.
21. Pevzner, P.A & S.H. Sze. 2000. Combinatorial approaches to finding subtle signals in DNA sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8:** 269–278.