Prediction of Protein-Protein Interactions Using Protein Signature Profiling

Mahmood A. Mahdavi and Yen-Han Lin*

Department of Chemical Engineering, University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada.

Protein domains are conserved and functionally independent structures that play an important role in interactions among related proteins. Domain-domain interactions have been recently used to predict protein-protein interactions (PPI). In general, the interaction probability of a pair of domains is scored using a trained scoring function. Satisfying a threshold, the protein pairs carrying those domains are regarded as "interacting". In this study, the signature contents of proteins were utilized to predict PPI pairs in *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Homo sapiens.* Similarity between protein signature patterns was scored and PPI predictions were drawn based on the binary similarity scoring function. Results show that the true positive rate of prediction by the proposed approach is approximately 32% higher than that using the maximum likelihood estimation method when compared with a test set, resulting in 22% increase in the area under the receiver operating characteristic (ROC) curve. When proteins containing one or two signatures were removed, the sensitivity of the predicted PPI pairs increased significantly. The predicted PPI pairs are on average 11 times more likely to interact than the random selection at a confidence level of 0.95, and on average 4 times better than those predicted by either phylogenetic profiling or gene expression profiling.

Key words: protein-protein interaction, protein signature, ROC curve

Introduction

Protein-protein interaction (PPI) is the key element of any biological process in a living cell. Proteins interact through their functional subunits (*1*). Protein domains, active sites, and motifs (collectively called signatures) are sub-sequence functional and conserved patterns that are essential to the functioning of individual cells and are the interfaces used in interactions at the protein level (*2*). With the completion of genome sequences of many organisms, genome-wide characterization of protein signatures is now practical. Although proteins are specified by unique amino acid sequences, the signature content of a protein sequence is crucial to determining interactions in which the particular protein is involved.

Protein signature (domain) information has been used to predict PPI. Naively, when two proteins are known to interact, their homologs in other organisms are assumed to interact based on comparative analysis (*3*). Domain contents of the interacting partners are utilized as input to predict more accurate interactions in another organism (*4*). Intermolecular or intramolecular interactions among protein families that share one or more domains are implemented to infer interactions among proteins (*5*). Domain-domain relationships are used to predict interactions at the protein level. In the association method (*6*), interacting domains are learned from a dataset of experimentally determined interacting proteins, where one protein contains one domain and its interacting partner contains the other domain. The probabilistic model of maximum likelihood estimation (MLE) outperforms the association method through taking the experimental errors into account. Following a recursive calculation procedure, in MLE method probabilities for domain-domain interactions are predicted based on the observation of interactions between their corresponding proteins. Then the prediction is extended to the protein level, assuming that two proteins interact if and only if at least one pair of domains from the two proteins interact (*7*). Potentially interacting domain (PID) pairs are extracted from an ex-

***Corresponding author.**
E-mail: yenhan.lin@usask.ca

perimentally confirmed protein pair dataset using the PID matrix score as a measure of domain interaction probability. The information for interacting proteins could be enriched about 30 folds with the PID matrix (*8*). In another study, the strengths of protein pairs are incorporated into the association method to enrich the predicting probability (*9*). As many domain structures are shared by different organisms, the integration of data from multiple sources may strengthen the reliability of domain associations and protein interactions (*10*). Moreover, the combination of protein interaction data from multiple species, molecular sequences, and gene ontology is used to construct a set of high-confidence domain-domain interactions (*11*).

In all above-mentioned methods, if a probability score meets a certain threshold, then the domains and subsequently related proteins are considered "interacting". However, these methods do not distinguish between single-unit proteins and multi-unit proteins. To overcome this limitation, a method based on domain combination was proposed, which predicts protein interactions according to the interactions of multi-domain pairs or the interactions of domain groups (*12*). As an alternative approach, machine learning techniques have been used to train support vector machines, called descriptors (*13*). Signatures have been used to train descriptors and each descriptor reflects the amino acid sequence of a protein that, in turn, consists of several signatures. However, in this machine learning approach, signature is defined as one single amino acid and its two neighbors (three consecutive letters), which is totally different from the definition utilized in this article as functional conserved patterns. Recently, interactomes (*14–17*) and databases, such as the Database of Interacting Proteins (*18*), have been used as reliable sources for mining interacting domains, which may contribute to inferring uncharacterized interacting proteins (*19*). Therefore, signature contents of proteins play a crucial role in predicting protein interactions. Signature-based PPI prediction techniques rely on statistically significant related signatures. When the interaction probability score between two signatures (in two different proteins) is greater than a threshold value, such a relationship is extended to the corresponding proteins and the potential interaction is inferred.

Close assessment of the protein pairs whose signatures possess high interaction probability scores shows that many of these protein pairs share at least one common signature. Sprinzak and Margalit (*6*) reported 40 overrepresented signature pairs in the protein interaction dataset of yeast. More than half of those signature pairs (22 out of 40 pairs) contained similar signatures and the rest of them were functionally close signatures. Non-identical pairs could not pass the threshold, even though the threshold was considered very loose. Okada *et al* (*20*) studied the role of common domains in the extraction of accurate functional associations in interacting partners. It has been shown that, when two proteins share a similar domain structure, their interaction probability score is higher than that of two proteins with non-similar domains (*21*). In a study of interacting signatures in the SCOP database, interacting signature pairs were predicted based on the finding that they use significantly higher surfaces to form these interactions, and interestingly, like-like interacting signatures were observed in high frequency (*2*). As reported by Ramini and Marcotte (*22*), proteins sharing common signatures possess a high possibility of being co-evolved in a correlated manner. These common signatures contribute to the similarity of protein families detected through optimal alignment between protein family similarity matrices. Therefore, common signatures between interacting proteins enhance the interaction probability of two proteins.

In this study, we propose a new genome-wide approach to predict PPI based on the observation that proteins with common signatures are more likely to interact. The signature content of a protein is represented by a binary profile, called the protein signature profile, and then the similarity between two profiles is scored using a binary similarity function. Imposing a threshold based on a pre-determined significance level, two proteins are considered "interacting" if they satisfy the significant threshold value. Furthermore, by removing proteins with one or two known signatures from the dataset, the false positive rate of the predicted PPI dataset reduces significantly. Different from the domain-based methods that score the relationship between two protein domains and extrapolate such a relationship to infer PPI, our approach directly scores protein relationships based on the signature content of each individual protein and the extent of commonality in signature patterns. The more signatures in common, the higher the similarity score will be between two different profiles. We applied this approach to three organisms: *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans* (worm), and *Homo sapiens* (human). Although at the time being, a relatively small portion of genes in each genome have been identified with their signatures, the proposed

approach is capable of covering the entire genome as more genes with known signature contents are discovered.

Results

The protein signature profiling (PSP) approach is illustrated in Figure 1 and it was implemented to predict PPI pairs for *S. cerevisiae*, *C. elegans*, and *H. sapiens*. Three predicted PPI datasets for each organism were generated by removing none, one, or two known protein signatures in their sequences, respectively. The predicted PPI pairs and their corresponding binary similarity scores are presented in Additional File 1. To evaluate the applicability of the PSP approach, sensitivity and specificity analysis was conducted and the predicted results for *S. cerevisiae* were compared with those confirmed by the MLE method over the same dataset (*7*). Furthermore, fold value analysis was performed to compare the predicted results with those confirmed by two non-signature-based methods, phylogenetic profiling (PP) (*23*) and gene expression profiling (GEP) (*24*). In either case, the PSP approach has higher true positive rates.

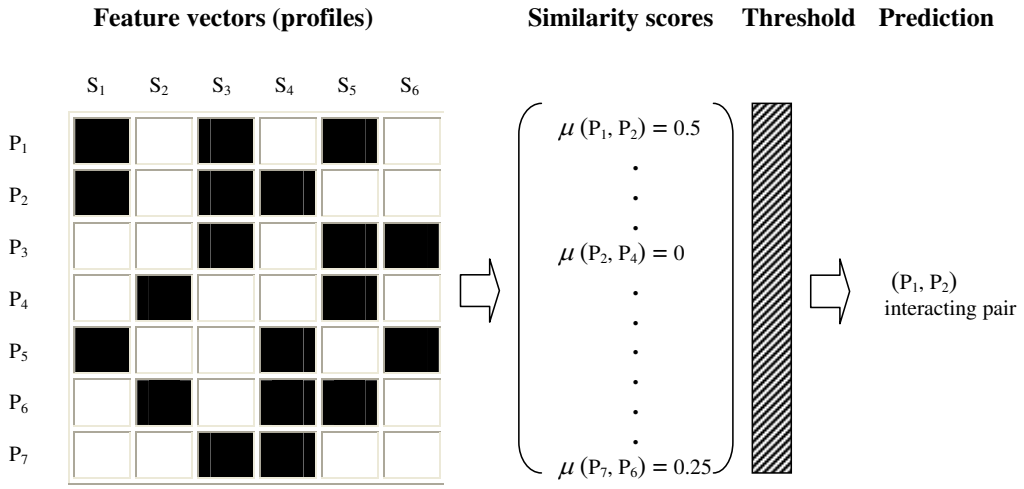

Fig. 1 Schematic of the protein signature profiling approach to predict PPI pairs. As illustrated, proteins P_1 and P_2 contain signatures (S_1, S_3, S_5) and (S_1, S_3, S_4), respectively. Using Equation 2, a binary similarity score of 0.5 is calculated (see Materials and Methods). If the score is greater than a significant threshold value estimated at a pre-specified confidence level (see Figure 4), then P_1 and P_2 are considered as an interacting pair. The same procedure was repeated for all proteins in *S. cerevisiae*, *C. elegans*, and *H. sapiens*. A complete list of all PPI pairs for the examined species can be found in Additional File 1.

Sensitivity and specificity analysis

The receiver operating characteristic (ROC) curve was implemented to evaluate the efficacy of the prediction of PPI pairs between our approach and the MLE method over the same dataset. Both methods use signature content information to score a relationship between two proteins. To implement the MLE method, the experimental dataset was randomly split into two parts: a training set and a test set. Domain interactions were predicted based on the observed protein interactions in the training set, and protein partners were identified based on the assumption that two proteins interact if and only if one pair of domains from two proteins interact. Then the PPI dataset predicted by the MLE method was compared with both the training set and the test set. More details on the MLE implementation and the numerical results of comparing the PSP-predicted PPI dataset with experimental PPI datasets, as well as the MLE-predicted PPI dataset with both the training and the test sets are presented in Additional File 2.

The ROC curve portrays the trade-off between the true positive rate (sensitivity) and the false positive rate (1−specificity) for different threshold values. The true positive rate is defined as the fraction of experimentally confirmed PPI pairs (all positives) that are correctly predicted. Likewise, the false positive rate is defined as the fraction of all potential interactions that are predicted and do not match with experimental pairs. Therefore, the two rates can be formulated as follows:

$$\text{True positive rate (Sensitivity)} = \frac{TP}{TP+FN}$$

$$\text{False positive rate (1−Specificity)} = \frac{FP}{FP+TN}$$

where TP is the number of experimentally confirmed PPI pairs that are predicted (matched), FN is the number of experimentally confirmed PPI pairs that are not predicted, FP is the number of predicted PPI pairs that do not match experimentally confirmed pairs, and TN is the number of potential PPI pairs that are neither experimentally confirmed nor computationally predicted.

The area under the ROC curve (AUC) is a quantitative indicator for ranking the performance of PPI prediction among various PPI predicting methods. At an AUC of 1, a perfect PPI prediction is obtained. The closer the area is to 0.5, the poorer the prediction. As shown in Figure 2, the AUC of the PSP-predicted dataset in the case of no protein removal is 0.549. The AUC of the MLE-predicted dataset is 0.511 when compared with the test set and is 0.686 when compared with the training set. Approximately 68% of PSP-predicted PPI pairs have the highest similarity score of 1, indicating a complete matching signature profile between two query proteins. Among this portion of predicted PPI pairs, many of them contain only one or two known protein signatures. As a result, a high false positive rate was observed through our method as compared with that calculated by the MLE method. This is attributed to the low number of known signatures in these proteins. To reduce false positive rates of predicted PPI pairs, and thus increase the accuracy of the PPI prediction, proteins with one or two signatures were removed consecutively, and the proposed approach was then applied to the remaining proteins in the dataset. As illustrated in Figure 2, the increase in AUC was observed for both cases. The AUC increased to 0.583 when proteins with one known signature were removed and eventually increased to 0.651 when proteins with two known signatures were also deleted from the dataset. Thus, the AUC values indicate that the performance of this approach can be ranked higher than that of the MLE method with the test set, although lower than that of the MLE method with the training set.

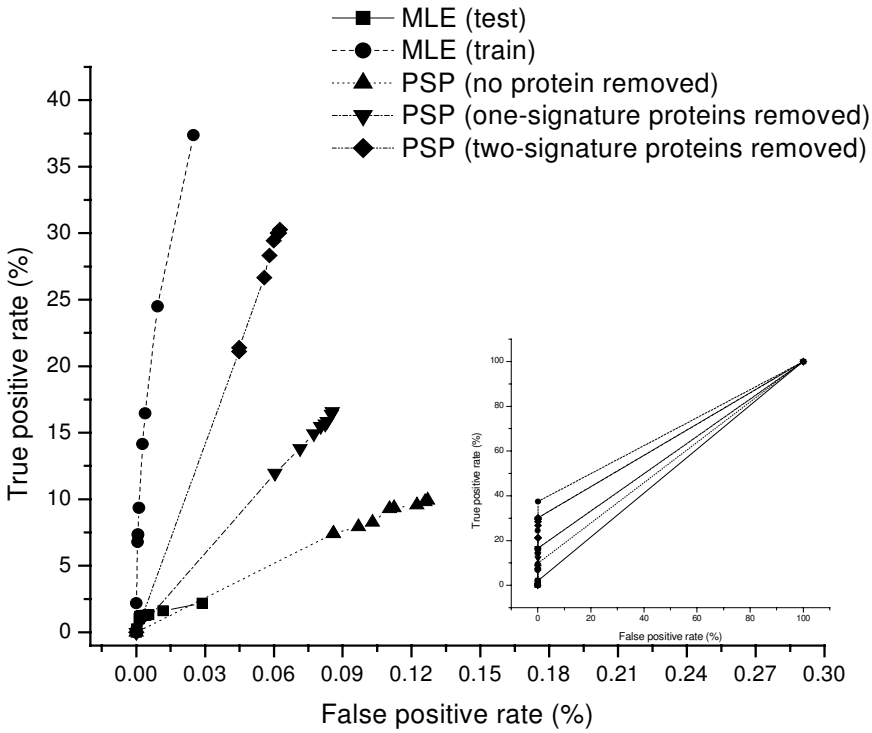

Fig. 2 Changes of the ROC curves subjected to the removal of proteins containing one or two known signatures and their comparison with results obtained by MLE method. Overall, PSP prediction results can be ranked between the MLE results compared with the training set [MLE (train)] and the MLE results compared with the test set [MLE (test)]. However, when proteins with one or two known signatures were removed from the dataset, the AUC increased and the curve approached the MLE (train) results. In this figure each data point represents a threshold value > 0 and ≤ 1. The inset figure illustrates the complete ROC curve for each case, and in this figure each data point represents a threshold value ≥ 0 and ≤ 1. See Additional File 2 for numerical data.

Our signature protein dataset does not allow us to remove proteins with more than two known signatures from the dataset due to the low number of proteins with a high number of known signatures; however, it is expected that with the availability of more information on signature content of proteins, the true positive rate of the proposed approach will drastically increase along with a low false positive rate. Nevertheless, the examination of the ROC curve indicates that the PSP approach presents a competitive or even better result compared with other currently available domain-based methods such as the MLE method with a test set.

Fold value analysis

The PPI pairs predicted by the PSP approach were also compared with those predicted by two non-signature-based methods, PP and GEP. Based on genomic information, the PP method has been reported as one of the most promising computational methods to predict PPI pairs (*25*); whereas the GEP method utilizes conserved co-expression patterns of genes to predict interacting protein pairs (*26*). For the PP method, to construct phylogenetic profiles among proteins, query proteins were blasted against a reference genome database consisting of 90 species. Proteins with matching patterns of presence or absence in reference genomes were paired. The detail information on implementation of this method is in Additional File 2. For the GEP method, the co-expression patterns were constructed based on normalized DNA microarray data confirmed from the Stanford Microarray Database (*27*). Using the EXPANDER program (*28*), genes were clustered according to their expression profiles. Genes clustered in the same group were considered as interacting pairs (see Additional File 2).

To quantify the statistical significance of the predicted PPI pairs among the three profiling methods, a statistical parameter, called fold, was used to facilitate the comparison. Fold is the ratio of the fraction of the experimentally confirmed dataset predicted by a method to the fraction of total potential PPI pairs predicted by the same method:

$$Fold = \frac{k_0/K}{n/M}$$

where k_0 is the number of predicted PPI pairs matched with the experimentally confirmed dataset (matched), K is the size of the experimentally confirmed dataset (observed), n is the predicted PPI pairs satisfying a threshold value (predicted), and M is the total number of potential PPI pairs; $M = m(m-1)/2$ where m is the number of proteins. The m value for *S. cerevisiae*, *C. elegans*, and *H. sapiens* is 2,242, 1,402, and 8,667, respectively. Fold is the probability of true interaction in predicted PPI pairs compared with the random prediction. The greater the fold value, the higher the probability of interaction will be, as compared with the probability of interaction based on random pairing.

Figure 3 illustrates the changes in fold values among PSP, PP, and GEP methods applied to *S. cerevisiae*, *C. elegans*, and *H. sapiens*. Generally speaking, PSP can predict PPI pairs with higher probability of interaction than other two methods. As one/two-signature proteins were removed, the fold values of PPI pairs predicted by PSP increased significantly than those predicted by PP and GEP methods. This suggests that as proteins possessing lower number of known signatures were deleted from the predicted PPI pairs, the probability of predicting false relationships would be noticeably reduced. As a result, more PPI pairs with a high confidence level can be predicted.

Discussion

In this study, we propose that common protein signatures could be used to predict interactions between two proteins. Different from other domain-based approaches such as the MLE method that utilizes a dataset of observed PPI pairs as a learning set to train a scoring function in order to calculate the domain interaction probability and subsequently predict protein interactions, the proposed approach does not require any learning set. In fact, the entire data can be used as a query dataset. The PSP approach predicts interactions upon the extent of similarity between the signature contents of the two proteins, while domain-based methods predict interactions between protein domains and assume that two proteins are interacting if one pair of domains from the two proteins interacts.

It is worthwhile to note that as signatures are small fragments of sequences, interactions between their corresponding proteins may not be predictable by simple homology. Homology-based methods rely on whole sequence alignment of primary structures, and protein interactions are predicted when the similarity between sequences is less than a threshold E-value calculated by BLAST. In signature-based methods, however, interactions among signatures are

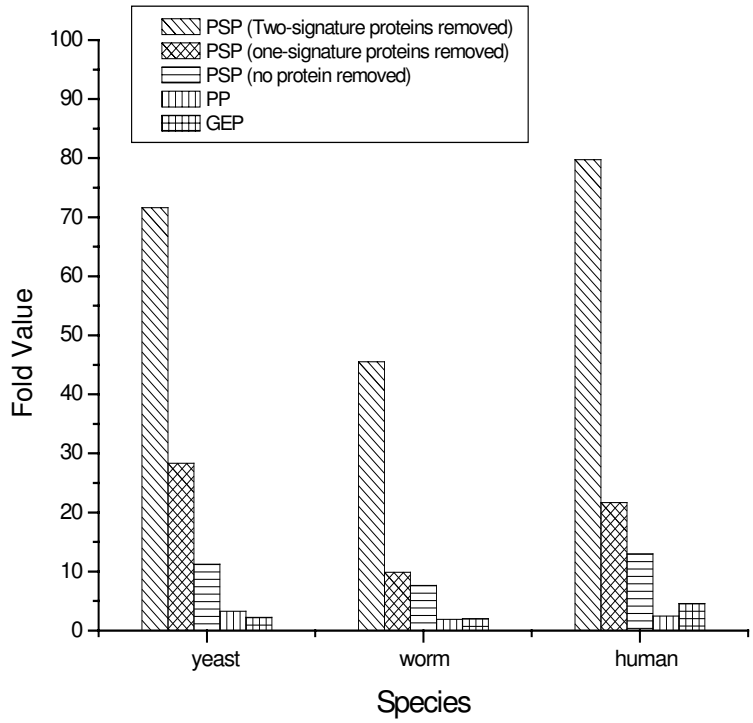

Fig. 3 Comparison of fold value changes among three different PPI prediction methods applied to *S. cerevisiae* (yeast), *C. elegans* (worm), and *H. sapiens* (human). See Additional File 2 for support information.

predicted based solely on observed protein interactions by means of statistical evaluations. Moreover, when two proteins share one or more signatures in common, it does not necessarily mean that the remainders of their sequences are similar. They may contain other non-similar signatures as well, making detecting their relationship impossible through simple homology. To assess the extent of overlap between PSP prediction and homology-based prediction, KOG database was used (*29*). The KOG database includes orthologous and paralogous proteins of eukaryotic species. Each group is associated with a conserved and specific function. Our investigation illustrated that 89.5% of the PPI pairs predicted by PSP could not be detected by homology-based methods when using BLAST.

Proteins involving in PPI that are predicted by homology-based techniques or the PP method often refer to "functionally interacting proteins", implying that these proteins cooperate to perform a given task without necessarily involving in physical contact. Experimental PPI detection techniques, such as yeast two-hybrid and large-scale affinity purification with mass spectrometry, attempt to discover direct physical interactions between proteins. However, there is a limited overlap between sets of interacting proteins identified by functional and physical relationships (*30*). Given the incomplete coverage of experimental results, there is clearly the need to develop large-scale computational sets of interacting proteins to be validated by future experiments. The proposed PSP method is a computational approach that predicts functionally interacting proteins based on the signature content of proteins. It is a new effort to predict robust protein interaction datasets that have better matches with physical interactions in compiled experimental datasets compared with those PPI pairs predicted by PP and GEP methods regarding the fold value.

Fold is an appropriate parameter that examines the ability of a computational method to predict experimentally confirmed PPIs. Furthermore, when the MLE-predicted dataset was compared with the test set of physical interactions, it was observed that the PSP approach outperformed the MLE method in terms of overlap with experimental data regarding the AUC. Both methods use signature content information of proteins to predict PPI, and the ROC curve well ranks them over the prediction of experimentally confirmed PPIs.

The significant threshold values are associated with the confidence level and the size of predicted PPI pairs. The significant threshold value in each confidence level is calculated by $(-0.1)\log(P)$. P,

an absolute probability, is defined as the ratio of confidence level (= 1−significance level) over the size of predicted PPI pairs, and "0.1" is the scaling factor that scales the threshold value to its corresponding binary similarity score between 0 and 1. Figure 4 portrays a significant threshold value with respect to each confidence level for three investigated organisms. For instance, at a confidence level of 0.95 (that is, a 1 in 20 chances of being false positive), the significant threshold value of choosing a binary similarity score for *S. cerevisiae*, *C. elegans*, and *H. sapiens* is 0.56, 0.53, and 0.72, respectively. At these threshold values, the predicted PPI pairs will possess a significance level of 0.05. In other words, there is a 95% probability that the predicted PPI pairs are not resulted from random events.

At a confidence level of 0.975, the corresponding significant threshold value is 0.6 for *S. cerevisiae*. From Figure 2, the true positive rate for the case of two-signature proteins removed, one-signature proteins removed, and no proteins removed under the PSP approach (see legend shown in the figure) is 28.33%, 14.92%, and 8.25%, respectively; whereas the true positive rate for the MLE results at the same confidence level is 7.37% and 1.18% when they are compared with the training dataset and the test dataset, respectively. This indicates that the PSP approach is more sensitive than the MLE method, and the sensitivity of the approach can be manipulated by means of deleting proteins containing less signature content. As a result, more experimentally confirmed PPI pairs are predicted.

Other than depicting the absolute relationship of fold value variations among different PPI profiling methods, Figure 5 presents the effect of removing proteins with different signature contents on the relative changes of fold values. As seen in the figure, by removing proteins with two signature contents from the predicted PPI pairs, the relative fold change of PSP vs. PP is 22.03, 23.60, and 32.41 for *S. cerevisiae*, *C. elegans*, and *H. sapiens*, respectively; whereas the relative fold change of PSP vs. GEP is 32.11, 22.66, and 17.45 for *S. cerevisiae*, *C. elegans*, and *H. sapiens*, respectively. The removal of proteins with a low number of known signatures improves the performance of the approach significantly. Even at the case of no protein removal, the PSP approach still outperforms the two non-signature-based profiling methods.

New putative PPI can be deduced from our results. In the case of yeast, the experimental dataset contains 1,438 proteins, while our analysis is focused on 2,242 proteins whose signature contents are available. Interactions involved with the additional 804 (= 2242−1438) proteins may point out a direction for further experimental validation. For example, proteins YBR208C and YGL062W are found interacting using our approach but they are not reported in the experimental dataset. Note that YBR208C contains seven domains, six of which are shared by YGL062W. Both proteins function as carboxylases. One may postulate a potential interaction between YBR208C and YGL062W. Such a clue may be used to guide a follow-up experiment.

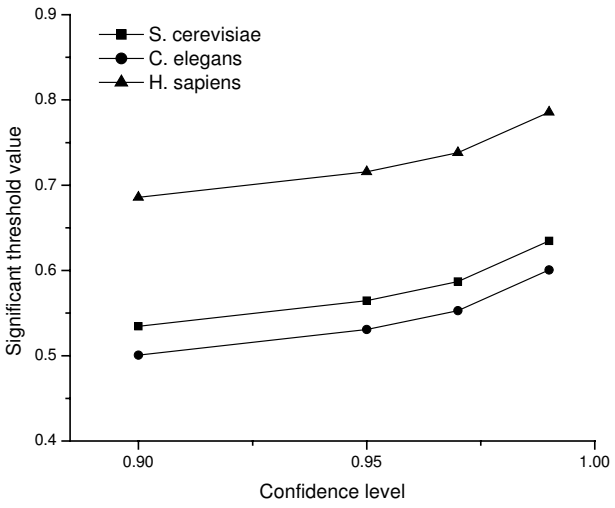

Fig. 4 The relationship between confidence level and significant threshold value with respect to a binary similarity score.

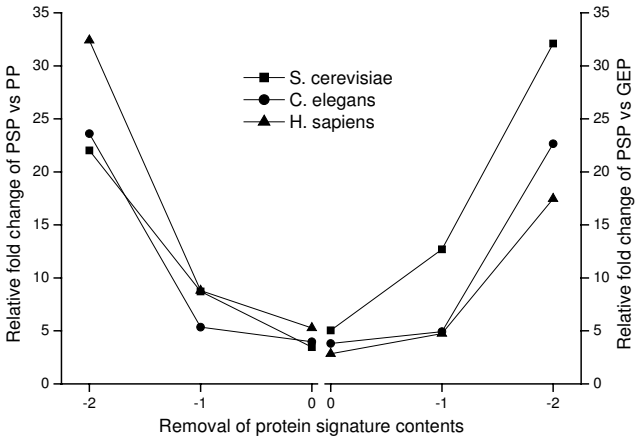

Fig. 5 The effect of removing proteins with one or two known signatures on the relative fold change. "–2", proteins containing two known signatures were removed; "–1", proteins containing one known signature were removed; "0", no removal.

Protein signature-based methods including our approach embed more intuitive biological reflection than others such as the PP method. Upon the notion that proteins interact through their conserved interfaces, not the whole sequence, the PP method may not be able to identify true interacting partners. It relies on identifying orthologs of a query sequence in a set of genomes based on whole sequence alignment. Instead, PSP identifies interacting partners based solely on the pattern of functional interfaces involved in protein interactions. The GEP method provides information on co-expression of genes in different biological events. Although this information is a strong indication that genes with similar expression profiles may have functional relationships, it provides a relatively lower degree of contribution to the prediction of physical interactions.

Conclusion

Proteins interact with each other through their functionally independent, structurally conserved, and biologically related signatures. These properties have established new insights into PPI prediction. Many existing domain-based prediction methods calculate the interaction probability score between two signatures. The scoring function is trained based on a learning dataset and subsequently applied to predict protein interactions. In contrast, the proposed PSP approach does not require training information, and proteins are directly paired based on their signature contents, providing that they have at least one signature in common. When proteins with a low number of known signature contents (one or two signatures) were removed from the dataset, it resulted in more predicted PPI pairs at a high confidence level. Thus, with the availability of more and more proteins with known signature contents across organisms, the coverage and accuracy of protein interacting pairs predicted by this approach is expected to increase. The predicted PPI pairs can, for instance, be incorporated into metabolic pathway reconstruction, or be used to reveal existing knowledge gaps in the association of proteins and pathways.

Meterials and Methods

Signature content information

The signature content of each protein sequence was confirmed from PROSITE database (*31*), which is a database of protein families and domains consisting of biologically significant sites, patterns, motifs, and domains. The entire PROSITE database (Release 19.27, May 2006) was downloaded and three files were created for the three organisms of interest. Each file contains the signatures found in one genome, including 884 signatures in *S. cerevisiae*, 738 signatures in *C. elegans*, and 1,354 signatures in *H. sapiens*.

Experimental PPI datasets

To evaluate and compare the predicted PPI by means of the proposed approach, datasets containing experimentally confirmed PPI pairs were compiled to serve as a common reference. The dataset for yeast contains 16,507 pairs that were confirmed from three sources: von Mering *et al* (*32*) (1,920 pairs), BIND database (*33*) (10,618 pairs), and CYGD database (*34*) (10,472 pairs). Combination of these three sources after removing duplicated pairs resulted in 16,507 pairs, comprising 4,391 proteins. Those proteins that were not found to have any signature in PROSITE were eliminated. As a result, 3,745 pairs remained in the final dataset, consisting of 1,438 proteins.

The worm dataset was constructed from BIND and Li *et al* (*14*), which reported 4,960 and 6,629 protein pairs, respectively. These pairs were confirmed by the yeast two-hybrid technique and were manually curated. After removing duplicated pairs, the dataset consisted of 7,081 pairs, comprising 3,390 proteins. Those proteins not having known signatures in PROSITE were dropped off, resulting in 344 pairs remained in the worm dataset, consisting of 220 proteins.

The human dataset is a combination of BIND and HPRD (*35*), containing 2,332 and 23,187 interactions, respectively. These pairs were confirmed using either mass spectrometry or yeast two-hybrid technique, and were manually curated. Merging the two sources resulted a dataset of 25,000 interactions, consisting of 5,726 proteins. Only 13,319 pairs involved in 3,975 proteins that contain known signatures in PROSITE were eventually used as the final human dataset.

Signature content representation

A protein is characterized by the signatures existing in its sequence. Hence, each protein can be represented by a vector of n features, called signature profile, where each feature corresponds to a signature and n is the number of signatures identified in the proteome

of an organism (for example $n = 884$ in yeast). Let $P_i = [S_{i_1}, S_{i_2}, \ldots, S_{i_n}]$ represent the feature vector of protein P_i with n signatures. $S_{i_1} = 1$ if signature S_1 exists in protein P_i and $S_{i_1} = 0$ otherwise. Therefore, each genome is represented by an m-dimensional vector where m is the number of proteins. In this study, $m = 2{,}242$ in yeast, 1,402 in worm, and 8,667 in human. A similarity measure was implemented to calculate the similarity between signature profiles (feature vectors). Binary similarity function (*36*) is used in this study to measure the similarity between a pair of signature profiles:

$$\mu(P_i, P_j) = \frac{\sum_{l=1}^{n} (P_i \wedge P_j)_l}{\sum_{l=1}^{n} (P_i \vee P_j)_l} \quad (1)$$

where μ is the similarity score between profiles P_i and P_j. This score is calculated over n signatures contained in proteins of a genome of interest. If protein P_i contains x signatures, protein P_j contains y signatures, and both proteins contain z signatures in common, the score can then be calculated as follows:

$$\mu(P_i, P_j) = \frac{z}{x + y - z} \quad (2)$$

Note that $0 \leq \mu \leq 1$. The value of μ increases when there are more common signatures between the two proteins. If the similarity score is higher than a threshold, the two proteins are considered as an "interacting pair".

Acknowledgements

MAM received financial supports from Iran Ministry of Science, Research, and Technology, and the Natural Sciences and Engineering Research Council of Canada.

Authors' contributions

YHL conceived the idea of this study. MAM contributed to the design and planning of the research, data collection and analysis. Both authors contributed to data interpretation, edited and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

Additional Files

Additional File 1: A list of predicted PPI pairs using protein signature profiling. The list includes predicted PPI pairs for cases of no protein removed, one-signature proteins removed, and two-signature proteins removed for *S. cerevisiae*.

Additional File 2: Detailed description of implementation of predicted PPI pairs using MLE, PP, and GEP methods. This file also includes numerical data of the comparison of predicted datasets with experimental datasets.

(The Additional Files are avilable from the corresponding author upon request.)

References

1. Ponting, C.P. and Russell, R.R. 2002. The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.* 31: 45-71.
2. Littler, S.J. and Hubbard, S.J. 2005. Conservation of orientation and sequence in protein domain-domain interactions. *J. Mol. Biol.* 345: 1265-1279.
3. Bansal, A.K. 1999. An automated comparative analysis of 17 complete microbial genomes. *Bioinformatics* 15: 900-908.
4. Wojcik, J. and Schächter, V. 2001. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics* 17: S296-305.
5. Park, J., *et al.* 2001. Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.* 307: 929-938.
6. Sprinzak, E. and Margalit, H. 2001. Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol.* 311: 681-692.
7. Deng, M., *et al.* 2002. Inferring domain-domain interactions from protein-protein interactions. *Genome Res.* 12: 1540-1548.
8. Kim, W.K., *et al.* 2002. Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair. *Genome Inform.* 13: 42-50.
9. Hayashida, M., *et al.* 2004. A simple method for inferring strengths of protein-protein interactions. *Genome Inform.* 15: 56-68.
10. Liu, Y., *et al.* 2005. Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics* 15: 3279-3285.
11. Lee, H., *et al.* 2006. An integrated approach to the prediction of domain-domain interactions. *BMC Bioinformatics* 7: 269.

12. Han, D.S., *et al.* 2004. PreSPI: design and implementation of protein-protein interaction prediction service system. *Genome Inform.* 15: 171-180.
13. Martin, S., *et al.* 2005. Predicting protein-protein interactions using signature products. *Bioinformatics* 21: 218-226.
14. Li, S., *et al.* 2004. A map of the interactome network of the metazoan *C. elegans*. *Science* 303: 540-543.
15. Rain, J.C., *et al.* 2001. The protein-protein interaction map of *Helicobacter pylori*. *Nature* 409: 211-215.
16. Uetz, P. and Pankratz, M.J. 2004. Protein interaction maps on the fly. *Nat. Biotechnol.* 22: 43-44.
17. Rhodes, D.R., *et al.* 2005. Probabilistic model of the human protein-protein interaction network. *Nat. Biotechnol.* 23: 951-959.
18. Salwinski, L., *et al.* 2004. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* 32: D449-451.
19. Riley, R., *et al.* 2005. Inferring protein domain interactions from database of interacting proteins. *Genome Biol.* 6: R89.
20. Okada, K., *et al.* 2005. Accurate extraction of functional associations between proteins based on common interaction partners and common domains. *Bioinformatics* 21: 2043-2048.
21. Ng, S.K., *et al.* 2003. InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res.* 31: 251-254.
22. Ramani, A.K. and Marcotte, E.M. 2003. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.* 327: 273-284.
23. Pellegrini, M., *et al.* 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* 96: 4285-4288.
24. van Noort, V., *et al.* 2003. Predicting gene function by conserved co-expression. *Trends Genet.* 19: 238-242.
25. Marcotte, E.M., *et al.* 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* 402: 83-86.
26. Fraser, H.B., *et al.* 2004. Coevolution of gene expression among interacting proteins. *Proc. Natl. Acad. Sci. USA* 101: 9033-9038.
27. Ball, C.A., *et al.* 2005. The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res.* 33: D580-582.
28. Shamir, R., *et al.* 2005. EXPANDER—an integrative program suite for microarray data analysis. *BMC Bioinformatics* 6: 232.
29. Tatusov, R.L., *et al.* 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41.
30. Bork, P., *et al.* 2004. Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* 14: 292-299.
31. Hulo, N., *et al.* 2006. The PROSITE database. *Nucleic Acids Res.* 34: D227-230.
32. von Mering, C., *et al.* 2002. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417: 399-403.
33. Alfarano, C., *et al.* 2005. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.* 33: D418-424.
34. Güldener, U., *et al.* 2005. CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res.* 33: D364-368.
35. Peri, S., *et al.* 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* 13: 2363-2371.
36. Rawat, S., *et al.* 2006. Intrusion detection using text processing techniques with a binary-weighted cosine metric. *J. Inf. Assur. Secur.* 1: 43-50.