Improving domain-based protein interaction prediction using biologically-significant negative dataset

Xiao-Li Li*, Soon-Heng Tan
and See-Kiong Ng

Knowledge Discovery Department,
Institute For Infocomm Research,
21 Heng Mui Keng Terrace, Singapore 119613
E-mail: xlli@i2r.a-star.edu.sg
E-mail: soonheng@i2r.a-star.edu.sg
E-mail: skng@i2r.a-star.edu.sg
*Corresponding author

Abstract: We propose a domain-based classification method to predict protein-protein interactions using probabilities of putative interacting domain pairs derived from both experimentally-determined interacting protein pairs and carefully-chosen non-interacting protein pairs. Multi-species comparative results for protein interaction prediction show that such careful generation of biologically-meaningful negative training data can improve classification performance.

Keywords: protein-protein interaction prediction; domain-domain interaction; biologically-significant negative set; randomly selected negative set; classification; F-measure.

Reference to this paper should be made as follows: Li, X-L., Tan, S-H. and Ng, S-K. (xxxx) 'Improving domain-based protein interaction prediction using biologically-significant negative dataset', *Int. J. Data Mining and Bioinformatics*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Xiao-Li Li received his PhD Degree in Computer Science in 2001 from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He is currently working as a research scientist at the Knowledge Discovery Department at the Institute for Infocomm Research. He is interested in the research areas such as data mining, machine learning and bioinformatics.

Soon-Heng Tan is a research officer at the Institute for Infocomm Research, Singapore. He received his BSc in Molecular Biology from National University of Singapore (NUS) and he is currently completing a MSc Degree in Computer Science at NUS under the supervision of A/Prof. Ng. His research interest is in bioinformatics, focusing on data mining and knowledge discovery from biological interaction data.

See-Kiong Ng is currently the Department Manager of the Knowledge Discovery Department at the Institute for Infocomm Research, Singapore. He is also Adjunct Associate Professor at the School of Computer Engineering, Nanyang Technological University, Singapore. He obtained both his BS and PhD Degrees in Computer Science from Carnegie Mellon University and his

Masters Degree from University of Pennsylvania. In terms of research, He is currently interested in unravelling the underlying functional mechanisms of protein interaction networks, using computational techniques from diverse research areas of machine learning, information extraction, and natural language processing.

1 Introduction

Cellular processes are biochemical events that are typically achieved by the interactions of proteins with one another. The elucidation of protein interactions is therefore the an essential step toward understanding the biology of cellular processes. Many experimental methods have been developed to detect protein-protein interactions, however, none of the current experimental methods is adequate to interrogate the entire interactome (Ng and Tan, 2004; von Mering et al., 2002). It is therefore useful to develop complementary computational methods for predicting new protein-protein interactions.

Several computational techniques have been proposed to predict protein-protein interactions. For example, potential protein interactions can be derived from gene context analysis such as gene neighbourhood (Dandekar et al., 1998; Overbeek et al., 1999), gene fusion (Enright et al., 1999; Marcotte, 1999), and gene co-occurrences and phylogenetic profiles (Huynen and Bock, 1998; Pellegrini et al., 1999). Alternatively, the physiochemical properties or tertiary structure of proteins can also be used for predicting interactions (Bock and Gough, 2001; Martin et al., 2004).

Recently, however, there is an increased focus on using *protein domains* to predict protein-protein interactions (Deng et al., 2002; Han et al., 2003, 2004; Ng et al., 2003; Wan and Jong, 2002). Protein domains are evolutionarily-conserved functional subunits in proteins found across different proteins. They are often found to participate in intermolecular interactions with one another. The existence of certain domains in proteins can therefore suggest the possibility of interaction between two proteins. As such, the analysis of many protein-protein interactions can be reduced to understanding the underlying domain-domain interactions between two proteins.

Domain-based protein interaction prediction methods generally consist of two main steps:

1 inferring domain-domain interactions from known protein interactions

2 predicting protein interactions based on the inferred domain-domain interaction information.

A few domain-based interaction detection techniques have recently been proposed. Deng et al. (2002) described a Maximum Likelihood estimation technique to infer domain-domain interactions that was then used to predict protein interactions. Wan and Jong (2002) presented an alternative statistical scoring system as a measure of the interaction probability between domains. Ng et al. (2003) devised an integrative approach to infer the protein domain interactions from other data sources in addition to experimentally determined protein interactions. Han et al. designed a probabilistic framework that takes domain combinations instead of single domains as basic units of protein interactions (Han et al., 2003, 2004).

These proposed techniques can be grouped into two main paradigms in terms of the way they infer domain-domain interactions. The domain interactions that are used for predicting protein-protein interactions are learned either (1) from an interacting protein set or positive set only (Deng et al., 2002; Ng et al., 2003; Wan and Jong, 2002), or (2) from both an interacting protein set and an artificially generated non-interacting protein set as negative set; the latter being generated by randomly pairing the proteins (Han et al., 2003, 2004). In the case of (1) where learning is conducted only from an interacting protein set, many false positive domain pairs may be derived because these domain pairs may occur in the (unavailable) negative set with high frequency. In the case of (2), the use of a putative negative data set helps alleviate this problem. However, using artificially generated non-interacting protein set as negative set is inadequate for inferring domain-domain interactions because the randomly generated negative dataset may contain unknown interacting protein pairs. In addition, if the artificially generated negative dataset is subsequently used in evaluating the performance of classifier, it will lead to inaccurate computation of the actual sensitivity and specificity of the technique.

In this paper, we propose a novel probabilistic technique to infer domain-domain interactions using both positive and negative training datasets. Our probabilistic model was able to outperform other domain-based techniques in predicting potential protein interactions. Unlike conventional approaches that use random pairing to generate artificial non-interacting protein pairs as negative training data, we generate biologically meaningful non-interacting protein pairs 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. We showed that the performance of classifier is improved with the more confident negative dataset. We also showed that improvements were consistently obtained across interaction data from multiple species, namely, yeast, fly, and worm (Li et al., 2005)[1].

2 Methods

Our proposed approach classifies a protein pair to be either interacting or non-interacting based on inferred underlying domain-domain interactions. The approach consists of four steps as follows:

- we pre-process the biological annotations for each training protein
- we generate a biologically significant negative set *N* (non-interacting protein pairs) based on the biological annotations
- we infer domain-domain interactions based on the interacting proteins pair set *I* and the negative set *N*
- finally, we build a classifier based on the interacting probabilities of domain pairs.

Below, we present the methods for these four steps in turn.

2.1 Pre-process biological annotations

Recent progress in genomic sequencing, computational biology, and ontology development has presented an opportunity to investigate biological systems from a more information-driven perspective. In our approach, we propose to make use of the increasing availability of functional annotations of proteins to generate a more biologically significant negative training data set – unlike conventional approaches that conveniently use random pairing to generate artificial non-interacting protein pairs as negative training data – as a means to improve classification performance. In our earlier work (Li et al., 2005) we have used the functional annotations from the MIPS database (http://mips.gsf.de/genre/proj/yeast/index.jsp) for yeast proteins. In this work, in order to cover the proteins from species other than yeast, we use the annotations from the more comprehensive Gene Ontology (GO) (http://www.geneontology.org/) as the source of biological information to help us create a more biologically meaningful set of negative training data.

GO consists of biological annotations under three main categories: molecular functions, biological processes and cellular components. A molecular function is a biological activity, such as catalytic or binding activities, at the molecular level. A biological process is a series of events accomplished by one or more ordered assemblies of molecular functions. A cellular component is a component of a cell but with the proviso that it is part of some larger object.

Physically, proteins that are in different cellular locations are less likely to interact because of their situations. Biologically, genes (and their product proteins) from the different biological process are also less likely to interact than those within the same biological process. As such, we use both biological process and cellular locations as a dual constraint to generate our negative set. Proteins are most unlikely to interact if they are from different cellular locations and biological processes.

GO terms are organised in structures called Directed Acyclic Graphs (DAGs), where a generic biological process is progressively broken down into more specific terms or types. As such, two related proteins may be annotated with different GO terms that are located in different levels of the ontology. To facilitate easy comparison of the biological annotations of the training proteins, we have chosen a fixed cut-off level of level three; in other words, any GO term beyond the cut-off level is considered as the same location or biological process as the corresponding parent GO term at level three.

As such, we must first perform a preprocessing step to normalise the different GO annotations of the training proteins to the cut-off level so that all proteins are annotated with GO terms at level three. Given two proteins, we will regard that they are not involved in the same biological process or cellular location only if their level-three GO cellular locations and biological processes are different. Such strict selection strategy (higher level cut off) creates a much purer negative set for training our classifier.

2.2 Generate the negative set

After annotating the training proteins with GO cellular locations and biological processes at the cut-off level, we are now ready to generate a biologically meaningful negative set by pairing those proteins involved in different biological process and from different cellular locations. Algorithm 1 shows how to generate non-interacting protein pairs (negative set).

Algorithm 1: Generate non-interacting protein pairs

```
 1: **Input**: interacting set $I$, protein set $P$;
 2: **Output**: negative set N;
 3: BEGIN
 4: Set $N = \emptyset$;
 5: **for** all the protein $p_i \in P$ **do**
 6:   Search $p_i$'s locations (l) and biological processes(b);
 7: **end for**
 8: Combine all protein pairs into a set $PS$: $PS = \{(p_i, p_j)|p_i \in P, p_j \in P, i \neq j\}$;
 9: **repeat**
10:   **for** each protein pair $(p_i, p_j) \in PS$ **do**
11:     **if** $(p_i, p_j) \notin I$ **then**
12:       **if** $((p_i.l \neq p_j.l) \wedge (p_i.b \neq p_j.b))$ **then**
13:         $N = N \cup \{(p_i, p_j)\}$;
14:       **end if**
15:     **end if**
16:     $PS = PS - \{(p_i, p_j)\}$;
17:   **end for**
18: **until** $(PS = \emptyset)$
19: END
```

In Algorithm 1, for each protein in *P*, the set of proteins of interest, we retrieve the biological information about its locations and biological processes (Steps 5–6) from the GO. Then, from Steps 9–18, we check each protein pair (p_i, p_j) in protein pair set *PS*: if it is already in the interacting protein set *I*, we eliminate it from *PS*; otherwise, if p_i and p_j are located at different cellular locations and from different biological processes, we add them into negative set *N*.

Note that in Step 12, a protein (p_i or p_j) may be located at multiple locations and involved in multiple biological process. We consider (p_i, p_j) to be non-interacting only if none of p_i's locations and biological processes match the p_j's locations and biological processes.

2.3 *Infer domain-domain interactions*

The objective of this next step is to assign interaction probabilities to each domain pair based on its occurrence in the protein-protein interacting set *I* and the negative set *N*. For 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 $l/(|p_i| \times |p_j|)$, 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 *r*th and *s*th domains of proteins p_i and p_j respectively.

Given that a domain pair (d_x, d_y) may occur in many interacting protein pairs of *I*, the interacting frequency of (d_x, d_y) in *I* is defined as:

$$N((d_x, d_y), I) = \sum_{i=1}^{|I|} \lambda_i(d_x, d_y) \times \frac{1}{|\, p_x^i \,| \times |\, p_y^i \,|} \tag{1}$$

where (p_x^i, p_y^i) is the *i*th protein pair in *I* and $\lambda_i(d_x, d_y)$ is the total number of occurrences of the domain pair (d_x, d_y) in (p_x^i, p_y^i). We compute $N((d_x, d_y), N)$, the interacting frequency of (d_x, d_y) in *N*, in a similar way:

$$N((d_x, d_y), N) = \sum_{i=1}^{|N|} \lambda_i (d_x, d_y) \times \frac{1}{|p_x^i| \times |p_y^i|}. \tag{2}$$

Let a set of pre-defined classes be $C = \{I, N\}$ and all the domain pairs set be DP. For any domain pair $(d_x, d_y) \in DP$, their interacting probability $P((d_x, d_y)|c_e)$, with Laplacian smoothing and $c_e \in C$, is defined as:

$$P((d_x, d_y) | c_e) = \frac{1 + N((d_x, d_y), c_e)}{|DP| + \sum_{k=1}^{|C|} N((d_x, d_y), c_e)}. \tag{3}$$

For a domain pair (d_x, d_y), the greater the interacting probability $P((d_x, d_y)|I)$, the more frequent it occurs in the interacting set I. However, since such a domain pair may also be chanced occurrences in class I, it is necessary to check its interacting probability in N: $P((d_x, d_y)|N)$. Obviously, if $P((d_x, d_y)|I)$ is significantly larger than $P((d_x, d_y)|N)$, then the domain pair (d_x, d_y) is likely to be a genuine domain-domain interaction. Otherwise, if $P((d_x, d_y)|N)$ is similar or even bigger than $P((d_x, d_y)|I)$, then the domain pair is unlikely to be interacting. In other words, to check if a domain pair (d_x, d_y) interacts, we compute its interacting probabilities in both interacting set I and negative set N.

Note that the purity of N can affect the accuracy of inferred domain-domain interactions. If N were generated from randomly paired proteins, the false negative protein pairs in N will result in the inference of many domain pairs that should have occurred only in interacting protein set (i.e., positive class). This will result in assigning inaccurate interacting probabilities to domain pairs and subsequently affect the accuracy of the eventual classifier to infer protein interactions.

2.4 Build a protein interaction classifier

Given a protein pair (p_i, p_j), in order to perform classification (i.e., to judge whether the proteins may interact with each other or not), we compute the posterior probability $P(c_e|(p_i, p_j))$, $c_e \in C$. The prior probability $P(c_e)$ of class c_e is defined as:

$$P(c_e) = \frac{\sum p(c_e, (p_i, p_j)), (p_i, p_j) \in I \cup U}{|I| + |N|}. \tag{4}$$

Based on equations (4) and (3), our proposed technique 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:

$$P(c_e | (p_i, p_j)) = \frac{p(c_e) \times \prod_{m=1}^{|p_i| \times |p_j|} p((d_{i,r}, d_{j,s}) | c_e)}{\sum_{k=1}^{|C|} p(c_e) \times \prod_{m=1}^{|p_i| \times |p_j|} p((d_{i,r}, d_{j,s}) | c_e)}. \tag{5}$$

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

3 Evaluation

In this section, we evaluate the proposed technique for predicting protein interactions. In order to evaluate the overall performance of our proposed technique, we have performed comprehensive experiments with the interaction data of the species *yeast* (*S. cerevisiae*), *worm* (*C. elegans*) and *fly* (*D. melanogaster*).

3.1 Data preparation

Interacting proteins are retrieved from DIP (http://dip.doe-mbi.ucla.edu/dip/) – a comprehensive catalogue of about 55,708 experimentally determined protein-protein interactions in over 110 organisms. In DIP (10/03/2004 version), *yeast* has 15,461 protein interactions among 4,773 proteins, *worm* has 4,030 protein interactions among 2,638 proteins, and *fly* has 20,988 interactions among 7,068 proteins.

Positive and negative datasets are employed to train a classifier and to evaluate the performance of our method. For each species (*yeast*, *worm* and *fly*), we select all interactions in DIP to construct our positive dataset *I*. The negative set of non-interacting protein pairs used in this work is constructed using Algorithm 1 described in the previous section. Proteins are paired up only if they are not from the same cellular location and biological process. This results in a negative set of 1,025,063 protein pairs for *yeast*, 439,057 protein pairs for *fly* and 5,128 protein pairs for *worm*. To avoid size bias between the positive and negative datasets, for each species, we randomly assembled a negative set *N* with the same number of protein pairs as *I*.

The domain information of proteins are obtained from the *Pfam* database (Corpet et al., 2002), which contains a large collection of multiple sequence alignments and profile hidden Markov models of protein domains. Both *Pfam-A* and *Pfam-B* are used to ensure sufficient coverage.

3.2 Experimental results

Yeast is a well-studied model organism that is generally used to evaluate the prediction performance of protein interaction systems. As such, we first report the results of our technique on species *yeast* in details and compare it with other techniques to illustrate the effectiveness of our proposed technique.

For *yeast*, we infer the domain-domain interactions from both positive set *I* and negative set *N*. Each domain pair gets an interacting probability for *I* and *N* using equation (3).

For illustration, Table 1 shows the top ten interacting and top ten non-interacting domain pairs respectively. The top interacting domain pairs have maximal values of $P((d_x, d_y)|I)/P((d_x, d_y)|N)$. In other words, these are domain pairs with the biggest $P((d_x, d_y)|I)$ and smallest $P((d_x, d_y)|N)$ values. The top non-interacting domain pairs show those with significant occurrence in non-interacting protein pairs. As we know, not all domain pairs derived from protein-protein interactions are truly interacting as some could occur in interacting proteins by chance. This could lead to false positive domain-domain predictions if we learn from the positive class *I* only.

Table 1 Top ten interacting and non-interacting domain pairs

Interacting domain pairs	*Non-interacting domain pairs*
(PF00400, PF00400)	(PF00400, PF00624)
(PF00118, PF00400)	(PF00069, PF00399)
(PF00069, PF07714)	(PF00153, PF00624)
(PF00076, PF00514)	(PF00624, PF01239)
(PF00271, PF00400)	(PF00076, PF00624)
(PF00076, PF00400)	(PF00153, PF00399)
(PF00806, PF07714)	(PF00399, PF00400)
(PF00270, PF00400)	(PF00399, PF07714)
(PF00400, PF01423)	(PF00271, PF00624)
(PF00400, PF00514)	(PF00005, PF00399)

For example, the domain pairs (PF07714, PF00400) and (PF00069, PF00400) occurred 170 and 66 times in *I* respectively. If we just learn from *I*, it is natural to conclude that they are interacting domain pairs as they have high occurrence in interacting set. However, with the help of our biologically refined negative class *N*, we were able to eliminate them since both domain pairs also occurred 901 and 860 times in *N* respectively. Furthermore, since our negative class *N* is more biologically significant than randomly paired proteins, we can estimate the interacting probabilities of each domain pair more precisely, resulting in a more accurate classifier.

For evaluation, we use the inferred domain-domain interactions to classify protein pairs. A 5-fold cross validation is performed to test the accuracy of the classifier described in equation (5). We compare our results with the reported results using the ‘Hybrid Classification’ technique from Han et al. (2003) and the ‘Possibility Ranking’ technique from Han et al. (2004), both of which used positive and negative training datasets for improved protein interaction prediction in *yeast*. Table 2 shows the comparison results of four domain-based protein interaction prediction techniques in terms of sensitivity and specificity. The first two techniques were from Han et al. (2003, 2004). The other two are our probabilistic technique on two different negative sets, namely, randomly paired negative set (random pairs) and the biologically significant negative set (biological refinement).

Table 2 Comparative results of different domain-based protein interaction prediction methods on *yeast*

Prediction method	*Specificity*	*Sensitivity*
Hybrid Classification (Han et al., 2003)	56.00	86.00
Possibility Ranking (Han et al., 2004)	75.00	84.36
Our method with random pairs	84.19	86.80
Our method with biological refinement	91.56	91.16

Compared with the techniques in Han et al. (2003, 2004), our probabilistic technique was able to achieve much higher specificity at similar sensitivity regardless of whether it has been trained with random protein pairs as the negative set or the refined negative set assembled using domain knowledge. Our classifier that was trained with the biologically refined negative dataset gave the best performance, obtaining an increase of 7.4% and 4.4% in specificity and sensitivity respectively as compared to the same probabilistic classifier trained with negative dataset of randomly paired proteins. This shows that the use of biological domain knowledge for negative dataset construction can benefit the prediction performance of the eventual classifier built on the training data.

The techniques from Han et al. (2003, 2004) were not tested with cross-validation. They randomly selected 20% DIP data as test set and the remaining 80% as training set. Then they repeated their experiments three times and got the average results. In fact, as reported in Han et al. (2003), their specificities was rather fluctuating according to the selected test sets. Our method is more robust as our results fluctuated only within 3% in each division of cross validation.

We also investigated how the size of negative set may affect the performance of our classifier. We systematically increase the size of *N* by 2–10 times. The results with respect to varying sizes of *N* are shown in Table 3. Generally, with the increase in the size of the negative set, the specificity of our classifier increases while the sensitivity decreases. One reason that sensitivity has decreased is that the imbalance of positive and negative training set makes our classifier biased towards the negative class *N*. However, we believe that it is possible to get better performance through intelligently selecting the negatives, and we will leave this problem as our future study.

Table 3 Performance of our classifier with different sizes of *N*

Size ratio	*2*	*4*	*6*	*8*	*10*
Specificity	92.18	94.32	95.12	96.35	96.43
Sensitivity	89.63	83.82	79.12	74.27	73.30

In addition to applying our method on the standard model organism *yeast*, we also investigate the applicability of our approach in two other species, *fly* and *worm*, that are relatively less well-studied. Surprisingly, we were also able to achieve a high 80.77% and 87.07% for *fly* and *worm* respectively in terms of F-measure. Given that these two species are not well studied (with incomplete interactomes and biological annotations), the classification accuracy for this two species obtained by our method is therefore quite impressive.

Finally, in order to systematically investigate the effectiveness of different methods to construct negative training set for protein interaction prediction, we perform the comparison test among four different methods:

- using randomly selected negative
- using biological process only
- using location only
- using both location and biological process.

Figure 1 shows the comparison results for the four different methods to construct the negative set in terms of F-measure. Overall, the results depicted in the figure shows that our technique that uses both location and biological process to construct negative set achieves the best results. In *yeast*, this technique was able to achieve 5.88%, 5.18% and

1.58% higher F-measure than other three methods of constructing negative sets (techniques 1–3) respectively. Similarly, compared with other three methods, our technique in *fly* can achieve 8.39%, 3.93% and 2.32% higher improvement and in *worm* we can achieve 13.17%, 9.43% and 2.79% higher improvement. Our results from the three species (of varying sizes of interaction data sets as well as functional annotations) are consistently better than other three methods, indicating we can build a more accurate classifier by combining the location and biological process to construct much purer negative set, and that the performance improvement is not a species-specific artifact.

Figure 1 Comparison of four different methods to construct negative training sets for domain-based protein interaction prediction

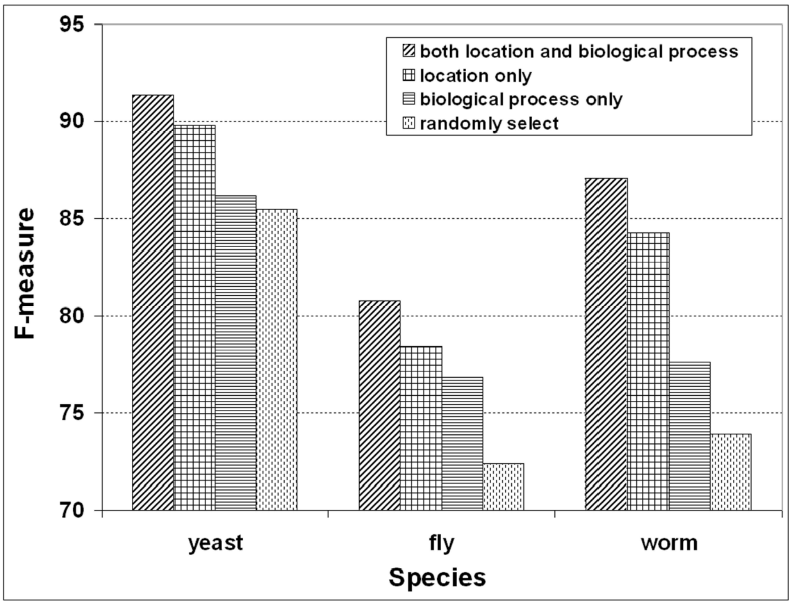

Figure 1 also shows that even using only location or biological process GO annotations to construct negative set can still get significantly better results than using randomly selected negative set in all the three species, signifying the advantages of using a biologically meaningful negative training set over a randomly generated one. Interestingly, compared with biological process annotations, the cellular location information can achieve better results. This is reasonable because the physical constraint is a stricter restriction, while proteins in different biological processes may interact with each other as some proteins do play multiple roles.

4 Conclusions

In this paper, we predict protein-protein interactions based on domain information. Our learning algorithm first constructs a biologically meaningful negative set based on biological annotations from GO. It then infers the underlying domain interactions based on their probabilities in both interacting class and non-interacting class. A probabilistic classifier for predicting protein interactions is then built upon the inferred probabilistic domain interactions. Our experimental results in multiple species – *yeast*, *fly*, and *worm* – show that our probabilistic approach is effective and outperforms other similar domain-based techniques for protein interaction prediction. Our results also suggest that it is advantageous to generate biologically meaningful non-interacting protein pairs based on the proteins' biological annotations instead of the conventional approaches that use random pairing to generate artificial non-interacting protein pairs as negative training data, as such careful generation of negative training data set was found to improve classification performance.

One inherent limitation of using domain-based methods to predict protein interactions is that not all proteins have domain information. As such, the domain-based methods cannot be used to predict interactions for those proteins that are currently without domain information. Therefore, one possible future work that we will be investigating is to integrate other biological features such as 'motif' and 'amino acid composition' with the domain information for more comprehensive and accurate predictions of protein interactions.

References

Bock, J.R. and Gough, D.A. (2001) 'Prediction of protein-protein interaction from primary structure', *Bioinformatics*, Vol. 17, pp.455–460.

Corpet, F., Servant, F., Gouzy, J. and Kahn, D. (2000) 'Prodom and prodom-cg: tools for protein domain analysis and whole genome comparisons', *Nucleic Acids Res.*, Vol. 28, pp.267–269.

Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998) 'Conservation of gene order: a finger print of proteins that physically interact', *Trends Biochem. Sci.*, Vol. 23, pp.324–328.

Deng, M., Sun, F., Metha, S. and Chen, T. (2002) 'Inferring domain-domain interactions from protein-protein interactions', *Genome Research*, Vol. 12, pp.1540–1548.

Enright, A.J., Illiopoulos, I., Kyrpides, N.C. and Ouzounis, C.A. (1999) 'Protein interaction maps for complete genomes based on gene fusion events', *Nature*, Vol. 402, pp.86–90.

Han, D., Kim, H., Jang, W. and Lee, S. (2004) 'Domain combination based protein-protein interaction possibility ranking method', *IEEE Fourth Symposium on Bioinformatics and Bioengineering (BIBE2004)*, pp.434–441.

Han, D., Kim, H., Seo, J. and Jang, W. (2003) 'Domain combination based probabilistic framework for protein-protein interaction predication', *Genome Informatics*, Vol. 14, pp.250–259.

Huynen, M.A. and Bock, P. (1998) 'Measuring genome evolution', *Natl Acad. Sci.*, Vol. 95, pp.5849–5856.

Li, X.L., Tan, S.H. and Ng, S.K. (2005) 'Protein interaction prediction using inferred domain interactions and biologically-significant negative dataset', *ICCSA (3)*, LNCS (3482), Vol. 13, pp.318–326.

Marcotte, E.M. (1999) 'Detecting protein function and protein-protein interactions from genome sequences', *Science*, Vol. 285, pp.751–753.

Martin, S., Roe, D. and Faulon, J.L. (2004) 'Predicting protein-protein interactions using signature products', *Bioinformatics*, Vol. 20, pp.1–9.

Ng, S.K. and Tan, S.H. (2004) 'Discovering protein-protein interactions', *Journal of Bioinformatics and Computational Biology*, Vol. 1, No. 4, pp.711–741.

Ng, S.K., Zhang, Z. and Tan, S.H. (2003) 'Integrative approach for computationally inferring protein domain interactions', *Bioinformatics*, Vol. 19, pp.923–929.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. and Maltsev, N. (1999) 'The use of gene clusters to infer functional coupling', *Natl Acad. Sci.*, Vol. 96, pp.2896–2901.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeastes, T.O. (1999) 'As signing protein functions by comparative genome analysis: protein phylogenetic profiles', *Natl. Acad. Sci.*, Vol. 96, pp.4285–4288.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. (2002) 'Comparative assessment of large-scale data sets of protein-protein interactions', *Nature*, Vol. 417, No. 6887, pp.399–403.

Wan, K.K. and Jong, P. (2002) 'Large scale statistical prediction of protein-protein interaction by potentially interacting domain (pid) pair', *Genome Informatics*, Vol. 13, pp.45–50.

Websites

http://mips.gsf.de/genre/proj/yeast/index.jsp.

http://www.geneontology.org/.

http://dip.doe- mbi.ucla.edu/dip/.

Note

[1] This manuscript is based on our recent conference paper (Li et al., 2005), with additional materials on the use of GO annotations for defining biologically-significant negative datasets, and a more comprehensive evaluation of our proposed approach with results based on interaction data from multiple species.