A Protein Interaction Verification System Based on a Neural Network Algorithm

Min Su Lee Seung Soo Park Min Kyung Kim
Department of Computer Science and Engineering, Ewha Womans University, Korea
ssue@ewhain.net sspark@ewha.ac.kr minkykim@ewha.ac.kr

Abstract

Large amounts of protein-protein interaction data have been identified using various genome-scale screening techniques. Although interaction data is a valuable resource, high-throughput datasets are prone to higher false positive rates.

We developed a new reliability assessment system for protein-protein interaction dataset of yeast that can identify real interacting protein pairs from noisy dataset. The system is based on a neural network algorithm, and utilizes three characteristics of interacting proteins: 1) interacting proteins share similar functional category, 2) interacting proteins must locate in close proximity, at least transiently, and 3) an interacting protein pair is tightly linked with other proteins in the protein interaction network.

The statistical evaluation of the neural network classifier by 10-fold cross-validation shows that it performs well with 96% of accuracy on the average. We experimented our classifier with pure 5,564 interactions. The classifier distinguished the yeast two-hybrid dataset into 2,831 true positives and 2,733 false positives.

1. Introduction

One of the key issues in proteomics is the analysis of protein-protein interactions (PPI). PPI knowledge is the fundamental basis of studying cellular process and mechanism of disease, and is especially useful in predicting unknown functions of protein [1-3]. PPIs have been studied individually to elucidate the mechanism of focused process. Recently, large quantities of protein interaction data are collected due to the evolution of high-throughput experiments. They include genome-scale Yeast Two-Hybrid assays (Y2H) [4,5] and mass spectrometry methods [6,7].

Vast amount of data produced by high-throughput experiments allow for efficient identifications of unknown PPIs information. However, they are prone to higher false positive rates than small-scale studies [8-10]. von Mering et al. estimate that approximately half the interactions obtained from high-throughput data may be false positives [8]. Containing false positive data requires an additional task to validate the reliability of each candidate PPI pair.

The intersection of multiple high-throughput PPI datasets can be effective in obtaining more credible interacting protein pairs. However, the coverage of intersection is very small in the huge amount of PPI dataset [8]. Some studies have been made on the assumption that interacting proteins whose transcripts being co-expressed are more likely to be credible [11,12]. However, recent research shows that interactions in genome-wide datasets have only a weak relationship with gene expression owing to different degradation rates [13,14]. These methods need whole genome-scale PPI dataset to assess the reliability of each PPI pair. Moreover, it is very ambiguous for biologists to define the cutoff value to classify between true positives and false positives. Hence, a new model to assess the reliability of individual protein interaction pair is needed.

In this paper, we developed a new reliability assessment system for PPI dataset that can distinguish real interacting protein pairs from noisy dataset. The system uses a neural network algorithm based on the three characteristics of interacting proteins. First, interacting proteins share similar functional category. Second, interacting proteins must locate in close proximity, at least transiently. Third, an interacting protein pair is tightly linked with other proteins in the protein interaction network. We use these three characteristics in the classification scheme to assess the reliability of PPIs.

The rest of this paper is organized as follows. Section 2 presents the system architecture and methods for assessing the reliability of PPIs. Section 3 shows the analytical results of the proposed system, and finally we conclude the paper in Section 4.

2. System Architecture

To separate true positives and false positives from putative PPI dataset, we have developed a classification system based on the neural network algorithm (Figure 1). Our classification system consists of a PPI database with attributes annotated, a computation module for each PPI pair, a neural network algorithm for filtering false positives, and a PPI classifier generated by the algorithm.

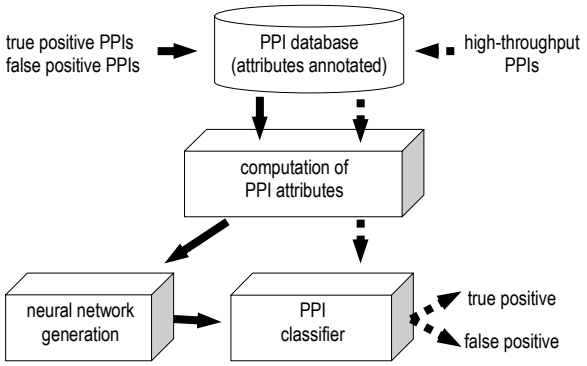

Figure 1. System architecture for classification of high-throughput PPIs into true positives and false positives. The solid line means the workflow of the classifier construction and the dotted line represents the workflow of the classifying PPI data.

2.1. Dataset

Our system first trains from a collection of protein pairs and their attributes. The training database consists of the true positive PPI dataset from Munich Information Center for Protein Sequence (MIPS) [15] and the false positive PPI dataset from Bader *et al.*'s work [16]. MIPS PPI dataset is usually regarded as a trusted PPI reference. We select more reliable true positives from MIPS dataset whose confidence scores by Bader *et al.* are more than 0.5 (3,231/15,628 pairs) and false positives from Bader's dataset with the confidence score of less than 0.2 (4,061/47,783 pairs) without redundant pair. The Bader *et al.*'s confidence scores were examined by statistical and topological correlation between the paired proteins in protein interaction networks constructed from published Y2H and Co-IP data using logistic regression model [16].

2.2. Computation of PPI attributes

The "computation of PPI attributes" module calculates three attribute values for the training PPI database, namely, similarity of functional category, possibility of co-localization, and topological properties within the interaction network structure.

2.2.1. Similarity of functional categories. Since most proteins function within complexes, interacting proteins share similar functional category. The similarity of functional category between interacting proteins is calculated based on the functional category (FunCat) of MIPS database [15,17]. The FunCat is described with a hierarchical tree structure (Figure 2(a)). It consists of 28 main functional categories with up to six levels of increasing specificity. A unique two-digit number is assigned to each category hierarchy. The levels of categories are separated by dots (eg. 01.01.03.02.01). The similarity of the functional category between two interacting proteins is determined by the level of the Lowest Common Ancestor (LCA) of the two proteins. Higher level of LCA implies more similar functional category. Since a majority of proteins are included in more than one functional category, we compute all LCAs from combinations of functional categories in each PPI pair. Then, the LCA with the maximum level is selected as the similarity value for the functional category between the pair. Finally, the similarity weight $w_F(p_1, p_2)$ for the functional category of the two proteins p_1 and p_2 is calculated as follows.

$$w_F(p_1, p_2) = \underset{\forall F_i \in P_1, \forall F_j \in P_2}{MAX} \{2^{Level_{LCA}(F_i, F_j)}\}$$

The similarity weight of the functional categories ranges from 1 to 64.

2.2.2. Frequency of co-localization. Interacting protein pairs must locate in close proximity, at least transiently. Hence the co-localization may be an effective means for evaluating hypothetical interactions. Huh *et al.* determined the subcellular localizations of each interacting protein pair and the fraction of total number of interactions occurring for each localization pairs [18]. Interactions are strongly enriched between proteins that co-localize, but the degree of enrichment varies widely by compartment. Hence the greater co-localization weight of two putative interacting proteins means the better evidence of physical interaction. The co-localization weight matrix is generated by the fold enrichment observed for each localization pair as compared with the randomized data set. (Figure 2(b)) The matrix divides subcellular localization into 22 categories, such as bud, nucleus, and golgi.

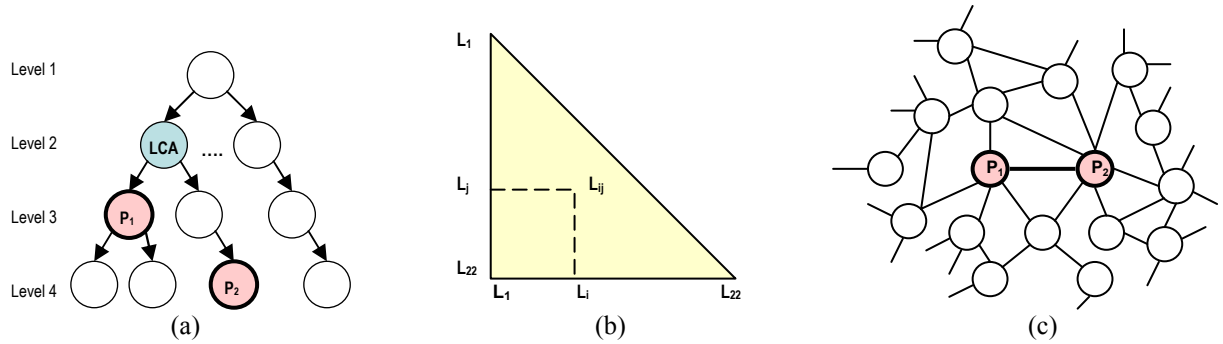

Figure 2. Calculating three attribute values for interacting proteins. (a) Tree structure for functional categories. (b) Co-localization matrix of interacting protein pairs. (c) Topology of interacting protein pair within the interaction networks.

Like the functional category case, most proteins move through several subcellular locations. Hence we find all combinations of localization weight in PPI pair and select the maximum weight as a delegate weight $w_L(p_1, p_2)$ for proteins p_1 and p_2 as follows.

$$w_L(p_1, p_2) = \underset{\forall L_i \in P_1, \forall L_j \in P_2}{MAX} \{Matrix_{Localization}(L_i, L_j)\}$$

The frequency of co-localization is distributed from 0 to 290.

2.2.3. Topological properties within the interaction network. False positive interaction pairs may be resulted in sticky proteins which tend to interact with unrelated proteins in vitro. The scale-free nature of biological networks suggests that highly connected proteins are a real feature of protein interaction networks [19-20]. Hence the reliable interacting protein pair must be tightly connected within interaction network, and their many other interacting partners have further interactions. Saito *et al*. proposed interaction generality measure (IG2) using five groups of topology of the protein interaction network around the target interacting pair [21]. For each interacting pair, IG2 weight is calculated with number of common proteins, alternative pathways, and several types of interaction that interact with a target interacting pair by applying principal component analysis (Figure 2(c)).

The IG2 value is distributed from -6.35 to 53. Lower IG2 value implies more tightly connected pairs in the interaction networks.

2.4. Neural Network Algorithm

Neural network learning methods provides a robust approach to approximating real-valued, discrete-valued, and vector-valued target functions. Neural network algorithm is quite robust to noise in the training data. Hence it is well-suited to assessment of noisy high-throughput experimental dataset. Although neural network algorithm learning times are relatively long, evaluating the learned network is typically very fast.

Neural network algorithm creates reliability assessment classifier for PPI dataset by constructing a multi-layer perceptron network of neurons based on the three input attributes – similarity of functional categories, frequency of co-localization, and topological properties within the interaction network. Given each state of the target class – true positive or false positive, the algorithm calculates probabilities for each possible state of the input attribute. The algorithm iteratively compares the predicted class of the PPI pair with known actual class of the pair. The errors from initial classification of the target class of the first iteration of the whole PPI pairs is fed back into the network, and used to modify the network's performance for the next iteration, and so on. These probabilities are used to predict an outcome of the target class, based on the input attributes.

3. Experimental Result

The system first calculates three input attribute from PPI dataset which consists of true positive and false positive PPI pairs. The neural network classifier is trained based on these attributes and target classes. Then, the system performs the assessment of input PPI pairs based on the aforementioned three characteristics on the neural network.

The statistical evaluation of the system by 10 fold cross-validation shows that it performs well with 96.18% of accuracy, 94.06% of sensitivity, 97.63% of specificity, 96.42% of true positive rate, and 96.03% of false positive rate on the average (Table 1). 10 fold cross-validation of the training set (using 90% of the training set to predict target classes for the remaining 10%) indicated that the model was not biased.

We experimented our classifier with pure 5,564 non-redundant interaction pairs from Ito and Uetz et al.

Table 1. 10 fold validation of the neural network classifier

	0	1	2	3	4	5	6	7	8	9	average
accuracy	95.78	95.78	95.36	95.99	97.26	97.47	96.62	95.15	95.57	96.84	96.18
error rate	4.22	4.22	4.64	4.04	2.74	2.53	3.38	4.85	4.43	3.16	3.82
sensitivity	92.78	93.16	93.65	94.71	95.79	96.81	94.30	92.31	92.78	94.36	94.06
specificity	97.86	97.54	96.49	96.84	98.24	97.90	98.22	97.13	97.50	98.57	97.63
TP rate	96.77	96.20	94.65	95.21	97.33	96.81	97.33	95.74	96.26	97.87	96.42
FP rate	95.14	95.52	95.82	96.50	97.21	97.90	96.17	94.76	95.12	96.15	96.03

The classifier distinguished the whole yeast two-hybrid dataset into 2,831 true positive interaction pairs and 2,733 false positive pairs.

4. Conclusion

In this paper, we presented an assessment scheme for the reliability of candidate interacting proteins based on the neural network algorithm. We used three biological attributes related to PPI, namely, similarity of functional category, frequency of co-localization, and topological properties within the interaction network. The proposed scheme shows good performance in distinguishing true interacting protein pairs from noisy PPI dataset. Our neural network classifier can be used to predict candidate interaction protein pairs.

Proteomics studies which are based on interaction data should be started with reliable interaction data. The proposed reliability verification system for PPI pairs may be very useful for this purpose.

5. References

[1] Vazquez, A., Flammini, A., Maritan A., and Vespignani A., 'Global protein function prediction from protein-protein interaction networks', *Nat Biotechnol. 2003, 21*, 697-700.

[2] Spirin, V., and Mirny, L. A., 'Protein complexes and functional modules in molecular networks', *PNAS 2003, 100*, 12123-12128.

[3] Deng M., Tu, Z., Sun, F., and Chen, T., 'Mapping Gene Ontology to proteins based on protein-protein interaction data', *Bioinformatics 2004, 20*, 895-902.

[4] Uetz P., Giot L., Cagney G., Mansfield T.A., *et al*. 'A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae', *Nature 2000, 403*, 623–627

[5] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., *et al*. 'A comprehensive two-hybrid analysis to explore the yeast protein interactome', *PNAS 2001, 98*, 4569–4574.

[6] Gavin, A. C., Bosche, M., Krause, R., *et al*. 'Functional organization of the yeast proteome by systematic analysis of protein complexes', *Nature 2002, 415*, 141–147.

[7] Ho, Y., Gruhler, A., Heilbut, A., *et al*. 'Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry', *Nature 2002, 415*, 180–183.

[8] von Mering, C., Krause, R., Snel, B., Cornell, M., *et al*. 'Comparative assessment of large-scale data sets of protein-protein interactions', *Nature 2002, 417*, 399-403.

[9] Deane C. M., Salwinski L., Xenarios I., and Eisenberg D., 'Protein interactions: two methods for assessment of the reliability of high throughput observations', *Mol Cell Proteomics 2002, 1*, 349-356.

[10] Sprinzak, E., Sattath, S., and Margalit, H. J., 'How reliable are experimental protein-protein interaction data?', *Mol Biol. 2003, 327*, 919-923.

[11] Ge, H., Liu, Z., Church, G. M., and Vidal, M., 'Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae', *Nat Genet. 2001, 29*, 482-486.

[12] Kemmeren, P., van Berkum, N. L., Vilo, J., Bijma, T., *et al*. 'Protein interaction verification and functional annotation by integrated analysis of genome-scale data', *Mol Cell 2002, 9*, 1133-1143.

[13] Gygi, S., Rochon, Y., Franza, B. R., and Aebersold, R., 'Correlation between protein and mRNA abundance in yeast', *MCB 1999, 19*, 1720-1730.

[14] Jasen R., Greenbaum, D., and Gerstein, M., 'Relating whole-genome expression data with protein-protein interaction', *Genome Res. 2002, 12*, 37-46.

[15] Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., *et al*. 'MIPS: a database for genomes and protein sequences', *Nucleic Acids Res. 2002, 30*, 31-34.

[16] Bader, J. S., Chaudhuri, A., Rothberg, J. M., and Chant, J. 'Gaining confidence in high-throughput protein interaction network', *Nat Biotech. 2004, 22*, 78-85.

[17] Ruepp, A., Zollner, A., Maier, D., Albermann, K., et al. 'The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes', *Nucleic Acids Res. 2004, 32*, 5539-5545.

[18] Huh, W. K, Falvo, J. V., Gerke, L. C., et al. 'Global analysis of protein localization in budding yeast' *Nature 2003, 425*, 686-691.

[19] Jeong, H., Mason, S. P., Barabasi, A. L., Oltvai, Z. N., 'Lethality and centrality in protein networks', *Nature* 2001, 411, 41-42.

[20] Wolf, Y. I., Karev, G., Koonin, E. V., 'Scale-free networks in biology: new insights into the fundamentals of evolution?', *Bioessays* 2002, 24, 105-109

[21] Saito, R., Suzuki, H., and Hayashizaki Y., 'Construction of reliable protein-protein interaction networks with a new interaction generality measure' *Bioinformatics 2003, 19*, 756-763.