HSPPIP: An Online Tool for Prediction of Protein–Protein Interactions in Humans

Yu Xue[1], Changjiang Jin[1], and Xuebiao Yao[1,2]

[1] School of Life Science, University of Science and Technology of China, Hefei, Anhui 230027, P.R. China
[2] Department of Physiology, Morehouse School of Medicine, Atlanta, GA 30310, USA
{yxue, jcjin, yaoxb}@ustc.edu.cn

Abstract. Recently, protein-protein interaction prediction (PPIP) has been emerging as an appealing question. Although several in silico approaches have been developed to delineate the potential protein-protein interaction (PPI), there are few online tools of human PPIP for further experimental design. Here we present an online service, hsPPIP (Protein-Protein Interaction Predicting of Homo Sapiens), to predict or evaluate the potential PPIs in human. The annotations of functional domain (Interpro) and GO (Gene Ontology) for proteins are employed as two informative features, and are integrated by the naïve Bayesian approach. The prediction accuracy is comparable to the existing tools. Based on the hypothesis that the features correlated with PPIs are conserved in different organisms, the web server hsPPIP is established and could predict the PPIs of human dynamically. hsPPIP is implemented in PHP+MySQL and can be freely accessed at: http://973-proteinweb.ustc.edu.cn/hsppip/.

1 Introduction

Dissecting the Protein-Protein Interaction (PPI) network in a cell is the foundation for elucidating all kinds of cellular mechanisms. In vivo or in vitro PPI identifications are usually time-consuming and labour intensive. Although high-throughput approaches have generated many data sets of PPIs, they are fairly noisy and need further confirmation by in vivo or in vitro experiments. So for its statistical accuracy and convenience, in silico PPI prediction (PPIP) might be a great help and insightful for further experimental consideration.

In this work, we have developed an easy-to-use online tool, hsPPIP (Protein-Protein Interaction Predicting of Homo Sapiens), to predict or evaluate the potential PPIs of human. We choose the annotations of functional domain (Interpro) [1] and GO (Gene Ontology) [2] for proteins as two features associated with PPIs. The naïve Bayesian approach is adopted to integrate the features together to improve the prediction accuracy. Since there is increasing evidence that interacting protein pairs are usually co-conserved through evolution and the PPI network structures are also conserved in different organisms [3, 4], we hypothesize that the features correlated with PPIs are also conserved in different organisms. So we might predict human PPIs

D.-S. Huang, K. Li, and G.W. Irwin (Eds.): ICIC 2006, LNBI 4115, pp. 603 – 611, 2006.
© Springer-Verlag Berlin Heidelberg 2006

based on the associated features from budding yeast. hsPPIP is especially fast and convenient for the construction of small scale human PPI networks. Although there have been several other web servers or databases for human PPIP, we propose that hsPPIP could be a useful tool for the experimentalists.

2 Materials and Methods

2.1 Training Data Set

The PPI data set for training was collected from six public PPI database, the MIPS CYGD (Comprehensive Yeast Genome Database) PPI (Mar. 2004) [11], BIND (Biomolecular Interaction Network Database) (Apr. 2004) [12], GRID (General Repository for Interaction Database) (Mar. 2004) [13], SGD (Saccharomyces Genome Database) (Apr. 2004) [14], MINT (Molecular INTeraction database) (Apr. 2004) [15], and Database of Interacting Proteins (DIP) (Mar. 2004) [16]. We only used budding yeast PPI data in the current analysis. The PPI information of each PPI database is listed in table 1. We combined the six PPI data sets into a non-redundant PPI data set including 21, 295 unique protein pairs. We also manually validated the data set.

2.2 Gold Standards and Testing Data

To test our method, we adopt both positive and negative control data sets reported previously [5]. The positive control contains protein pairs from the same MIPS complex (pos_MIPS), and the negative control is protein pairs with different subcellular localization and couldn't be found in verified PPI list (L_neg). As in the literature [5], the positive control is regarded as the set of positive PPIs (*P*), and the negative control is regarded as the set of negative PPIs (*N*).

For comparison, we also randomly generate 200, 000 protein pairs (Random200K), in which the real PPIs are expected to be only a small fraction. The gold standard and testing data sets are listed in table 1.

2.3 Collection of Human and Yeast Proteins

The protein sequences were downloaded from ExPASy (ftp://cn.expasy.org). The hsPPIP used the protein accession numbers as standard entries for prediction. All accession numbers of human and yeast proteins are retrieved and stored in the local MySQL database for further processing.

2.4 Protein Annotation by Functional Domain (Interpro) and Gene Ontology (GO)

Interpro is an integrated resource of annotation for protein families, functional domains and motifs [1]. Gene Ontology Consortium [2] is to produce a controlled vocabulary as the knowledge of gene and protein roles in cells. We downloaded the files containing all Interpro and GO annotations for the proteins from the EBI repository (ftp://ftp.ebi.ac.uk/pub/databases/), and mapped the Interpro and GO

annotations to human and yeast proteins, respectively. The information is also stored in our local MySQL database and could be searched and browsed. Users could search or browse our local database for each protein entry in human or yeast.

We excluded three GO annotations from the analysis: GO:0000004 (biological_process unknown), GO:0008372 (cellular_component unknown), and GO:0005554 (molecular_function unknown), since such unknown and ambiguous assignments might introduce noises into the analysis.

Table 1. The total data sets we used for this work. Training data set is retrieved and combined from six public PPI databases into a non-redundant PPI data set, including 21, 295 protein pairs. The gold standard is taken from the previous work [5] with both positive and negative control. We also randomly generate 200, 000 protein pairs as an additional testing data set.

Data type	Data set		# Protein pairs	Used for …
Training data	CYGD (MIPS) (2004-03-21)		11,862	
	BIND (2004-04-01)		5,875	
	GRID (2004-03-18)		19,038	
	SGD (2004-04-01)		4,624	Training
	MINT (2004-04-01)		14,014	
	DIP (2004-03-07)		15,393	
	TOTAL (non-redundant PPI data set)		21,295	
Gold standards	Positive	Protein pair in the same MIPS complex	8,617	
	Negative	Protein pair separated by localization	2,705,844	Testing
Testing data	Random	Random PPI (200K)	200,000	

2.5 Algorithm

PPIP (protein-protein interaction prediction) with functional domain-based strategies is widely used [17, 18, 19, 20]. In this work we adopt the Naïve Bayesian approach to integrate two features, functional domains and GO annotations, for PPIP. We considered if two proteins P_i, P_j have *a, b* Interpro annotations and *c*, *d* GO annotations respectively, then the probability of P_i and P_j to be interacting pair is shown below:

$$P(PPI_{ij}=1) = 1 - \prod_{(I_k,I_l)\in(P_i\times P_j)}(1-P(I_{ab}=1)) \times \prod_{(G_m,G_n)\in(P_i\times P_j)}(1-P(G_{cd}=1)) \cdot \quad (1)$$

PPI_{ij}=1: Protein P_i interacts with Protein P_j.
I_{ab}= 1: Interpro annotation I_a and I_b are interacting functional domain.
G_{cd}= 1: GO annotation G_c and G_d are functional associated.
a, b: Numbers of Interpro annotations in P_i and P_j, respectively.
c, d: Numbers of GO annotations in P_i and P_j, respectively.

$(I_{a'}, I_{b'}) \in (P_i \times P_j)$: Interpro pair (I_a, I_b) is included in protein pair $Pi \times Pj$.

$(G_{c'}, G_{d'}) \in (P_i \times P_j)$: GO pair (G_c, G_d) is included in protein pair $Pi \times Pj$.

The $P(I_{ab}=1)$ and $P(G_{cd}=1)$ could be obtained from training our non-redundant PPI data set. And the equation for calculating the two *probabilities* could be proposed as:

$$P(I_{ab}=1)=\frac{Int_{ab}}{N_{ab}}. \tag{2}$$

$$P(G_{cd}=1)=\frac{Int_{cd}}{N_{cd}}. \tag{3}$$

Int_{ab} : Number of PPIs that include (I_a, I_b);
N_{ab}: Number of all potential PPIs that include (I_a, I_b).
Int_{cd} : Number of PPIs that include (G_c, G_d);
N_{cd}: Number of all potential PPIs that include (G_c, G_d).

3 Results and Discussion

3.1 Testing the Accuracy of hsPPIP

Table 1 gives the basic information about the PPI data sets used in this work. We test our approach with four data sets: our training data set (nr_ppi), random PPI (Random200K), positive control (pos_MIPS) and negative control (L_neg). For random PPI, we randomly pick out two proteins of budding yeast to form 200,000 pairs. Obviously, this data set contains some real PPI, but its overall characteristics should be similar to the negative control. The *cut-off value* of hsPPIP is changed from 0-0.9 with per 0.1 per step. The result is diagramed in figure 1.

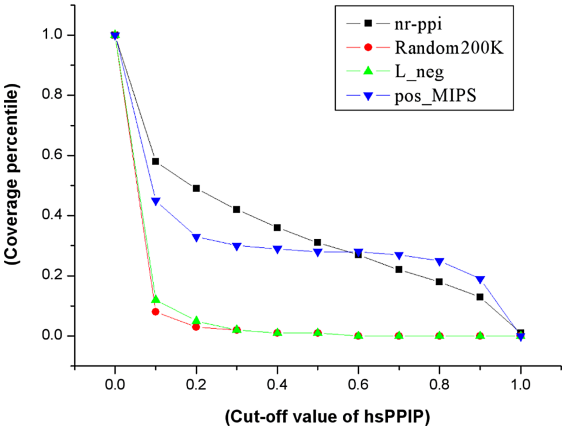

Fig. 1. Prediction accuracy of hsPPIP. We use four data sets to test the prediction accuracy: the training data set (nr-ppi), positive control (pos_MIPS), random PPIs (Random200K) and negative control (L_neg). It's clear that the positive control is much similar with the training data. And both the negative and random PPIs get little hits. The default cut-off value of hsPPIP is taken as 0.3.

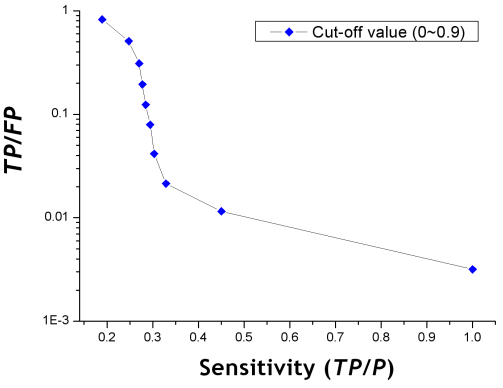

Fig. 2. The curve of TP/FP vs. Sensitivity (TP/P) under certain cut-off value of hsPPIP. We change the cut-off value of hsPPIP from 0-0.9 with per 0.1 per step. With the cut-off value of hsPPIP as 0.3, the TP/FP is 0.041 and Sensitivity (TP/P) is about 30.32%. If the cut-off value is taken as 0.7, the TP/FP is 0.309 and Sensitivity (TP/P) is about 27.04%, which could be comparable with the previous work.

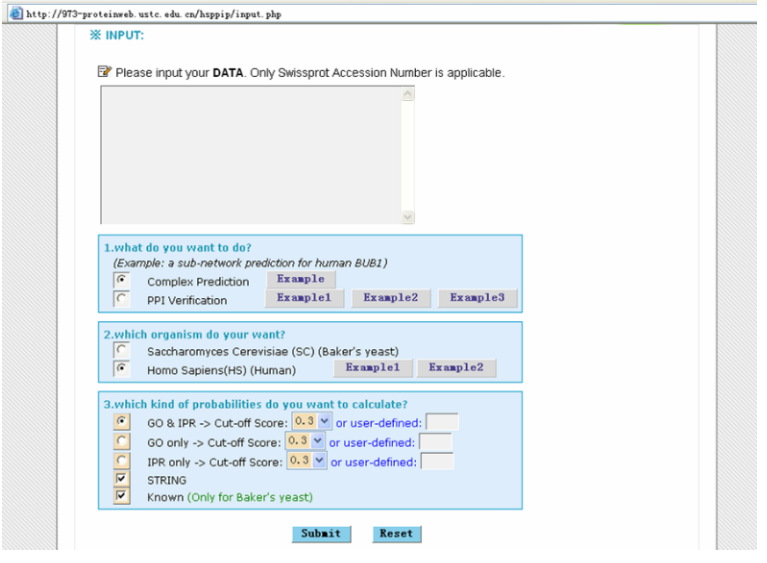

Fig. 3. The prediction page of hsPPIP WWW server. For convenience, hsPPIP supports PPIP in both budding yeast and human. The default cut-off value is taken as 0.3. Two accessing methods are provided. The first is Complex Prediction, which gives out all the potential PPIs among the given list of proteins with the predicted probabilities higher than the cut-off. The second is PPI Verification, which verifies the given list of PPIs with the predicted probabilities higher than the cut-off. Users could either choose Interpro & GO or one of the two features for PPIP.

It's obvious that the training data set gets the best accuracy. And the curve of positive control is much similar with the training data set. The curves of both the negative control and random PPI are similar and very different to our positive control and training data set. So we propose that hsPPIP could distinguish true positive PPIs from random PPIs accurately. The curve of the positive inclines to smooth when *cut-off value* is greater than 0.3. And even when *cut-off value* is great than 0.9, our approach could still predict ~20% PPIs in positive control properly. For convenience, we set the default *cut-off value* of hsPPIP as 0.3. Users could choose their favorite *cut-off value*.

We also perform an additional test with *TP/FP* vs. Sensitivity (*TP/P*) and compare the prediction results with previous work [5]. The *P* is total positive PPIs and *TP* is true positive PPIs which hsPPIP could predict properly. The *FP* is false positive predictions from hsPPIP. We change the *cut-off value* of hsPPIP from 0-0.9 with per 0.1 per step. The result is diagramed in figure 2.

With the *cut-off value* of hsPPIP as 0.3, the *TP/FP* is ~0.041 and Sensitivity (*TP/P*) about 30.32%. If the *cut-off value* is taken as 0.7, the *TP/FP* is ~0.309 and Sensitivity (*TP/P*) about 27.04%, which could be comparable with the previous work (Ref. 5, Figure 2C, *TP/FP*: *~0.3-0.4,* Sensitivity (*TP/P*): ~27%).

3.2 The Implementation of Web Server

The web server of hsPPIP has been implemented in PHP+MySQL. The web server is developed mainly for human PPIP. But for convenience, we also include the PPIP of budding yeast. In our system, only protein's Swissprot accession numbers are available. In addition, the information for proteins included in hsPPIP could be visited by either searching or browsing. The correlated probabilities of two features associated with PPIs, Interpro and GO annotations, are pre-computed and stored in a MySQL database. There are two approaches to access hsPPIP. Firstly, for complex prediction, users could submit a list of their required proteins with the Swissprot accession numbers. hsPPIP will find all the potential PPIs among the given list of proteins with predicted probabilities higher than the cut-off. Secondly, for PPI verification, users could submit a list of protein pairs, and hsPPIP will find all the potential PPIs whose probabilities are higher than the cut-off. The prediction page of hsPPIP is shown in Figure 3.

The hsPPIP could export the prediction results into multiple file formats. The plain text format could be downloaded and viewed locally. The .dot format file could be imported into the professional graphic software Graphviz (http://www.research.att.com/sw/tools/graphviz/) to get a publication-quality figure. Users could modify the color setting for nodes (proteins) and edges (interactions) of the .dot file. Moreover, the .sif format file could be generated for Cytoscape [6]. We also adopt a Java applet program [7] (a modified version for hsPPIP) to visualize the prediction results directly in the web browser.

3.3 An Example of Usage

Here we use human Bub1 protein (Swissprot accession number: O43683) and its putative interacting partners as an instance to diagram the usage of hsPPIP. Human Bub1 protein is one of the spindle checkpoint components localized to kinetochore during mitosis, blocking the onset of anaphase until all chromosomes are attached to

microtubules properly [8]. Aberrant organization of spindle checkpoint complex will contribute the genomic instability and aneuploidy, which could induce a variety of cancers with high rate [9]. Although many experimental studies were performed to depict the PPIs implicated in spindle checkpoints, it's still unclear how many proteins will interact with Bub1 and Bub1-related sub-network.

Firstly, we used STRING web server [10] to get the putative interacting partners of human Bub1 with default parameters (see in figure 4A). The sub-network contains eleven human proteins that interact with each other. Then we retrieve the Swissprot accession number of the eleven proteins and input them into hsPPIP for Complex Prediction. The default *cut-off value* for hsPPIP is 0.3. And the prediction results are visualized by Java applet in figure 4B. Only protein with at least one interaction with other proteins is shown. Moreover, the .dot file of prediction results from hsPPIP is modified directly and imported into Graphviz with the output format .png (see in figure 4C). In addition, .sif file format could be visualized by Cytoscape (see in figure 4D). By comparing the prediction results of hsPPIP to STRING, we find the prediction results of hsPPIP are similar with of STRING. Some putative PPIs are existed in STRING while missed by hsPPIP, such as Mxd3 protein (Q9BW11) with its binding partners. But there are several putative PPIs not existed in STRING but

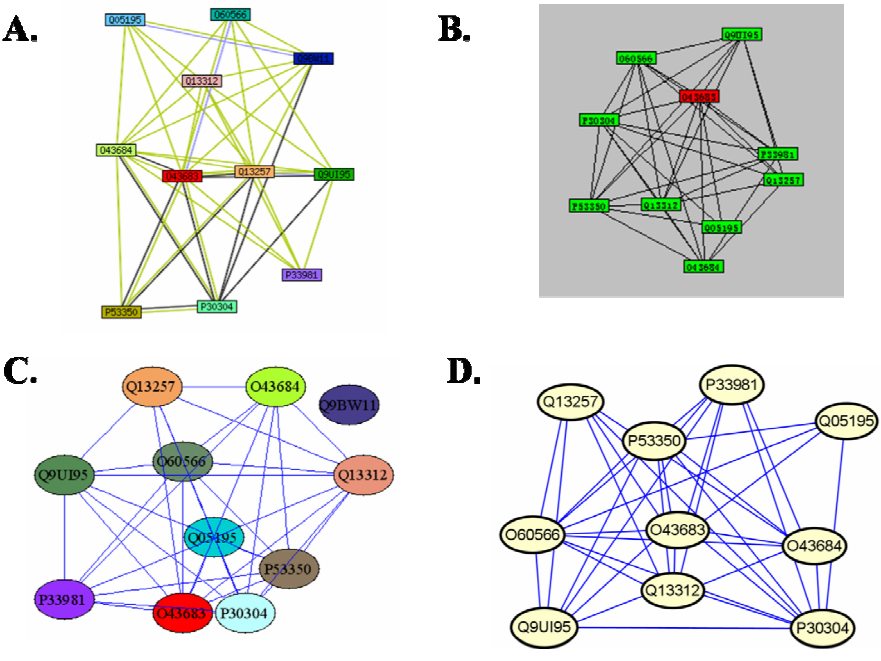

Fig. 4. Predicted sub-network for human Bub1 protein (O43683). (A) The prediction results of STRING, one of the most popular web servers for PPIP. (B) The prediction results of hsPPIP is visualized by java applet online. (C) The prediction results of hsPPIP is modified directly and then imported to Graphviz for visualization. (D) The prediction results of hsPPIP is viewed in Cytoscape.

predicted by hsPPIP. For instance, another spindle checkpoint protein human BubR1 (O60566) is predicted to interact with MAD (Q05195) by hsPPIP, but STRING doesn't find this potential interaction. And the predicted PPIs for human Bub1 may be insightful and needs the further experimental verification.

4 Conclusions

In this work, we provide a web server hsPPIP for protein-protein interaction prediction in human. For convenience, the hsPPIP could also predict the PPIs in budding yeast. The prediction results could be downloaded, modified locally, or visualized online directly. We also use human Bub1 protein to predict its potential PPIs and construct its sub-network by hsPPIP. Compared to another popular web server STRING, the prediction results are similar. Users could submit their favorite proteins in a list to predict the PPIs among them. The prediction results from hsPPIP may be helpful and insightful for further experimental design.

Acknowledgements

The work is supported by grants from Chinese Natural Science Foundation (39925018 and 30121001), Chinese Academy of Science (KSCX2-2-01), Chinese 973 project (2002CB713700), Beijing Office for Science (H020220020220) and American Cancer Society (RPG-99-173-01) to X. Yao. X. Yao is a GCC Distinguished Cancer Research Scholar.

References

1. Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A.: InterPro, Progress and Status in 2005. Nucleic Acids Res. 33 (2005) 201-205
2. Harris, M. A.: The Gene Ontology (GO) Database and Informatics Resource. Nucleic Acids Res., 32 (2004) 258-261
3. Hahn, M. W., Kern, A. D.: Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks. Mol Biol Evol. 22 (2004) 803-806
4. Yu H., Luscombe, N. M., Lu ,H. X., Zhu, X., Xia, Y., Han, J. D., Bertin, N., Chung, S., Vidal, M., Gerstein, M.: Annotation Transfer Between Genomes: Protein-protein Interologs and Protein-DNA regulogs. Genome Res. 14 (2004) 1107-1118
5. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., Gerstein, M.: A Bayesian Networks Aapproach for Predicting Protein-Protein Interactions from Genomic Data. Science, 302 (2003) 449-453
6. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 13 (2003) 2498-2504
7. Mrowka, R.: A Java Applet for Visualizing Protein-protein Interaction. Bioinformatics, 17 (2001) 669-671
8. Shah, J. V., Botvinick, E., Bonday, Z., Furnari,,F., Berns, M., Cleveland, D. W.: Dynamics of Centromere and Kinetochore Proteins, Implications for Checkpoint Signaling and Silencing. Curr. Biol. 14 (2004) 942-952

9. Cahill, D. P., Lengauer, C., Yu, J., Riggins, G. J., Willson, J. K., Markowitz, S. D., Kinzler, K. W., Vogelstein, B.: Mutations of Mitotic Checkpoint Genes in Human Cancers. Nature, 392 (1998) 300-303
10. Von, M. C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., Bork, P.: String: Known and Predicted Protein-protein Associations, Integrated and Transferred Across Organisms. Nucleic Acids Res. 33 (2005) 433-437
11. Mewes, H. W., Amid, C., Arnold, R.: MIPS: Analysis and Annotation of Proteins from Whole Genomes. Nucleic Acids Res. 32 (2004) 41-44
12. Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M.: The Biomolecular Interaction Network Database and Related Tools 2005 Update. Nucleic Acids Res.33 (2005) 418-424
13. Breitkreutz, B. J., Stark, C., Tyers, M.: The GRID: The General Repository for Interaction Datasets. Genome Biol. 4 (3) (2003) 233-245
14. Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A.: SGD: Saccharomyces Genome Database. Nucleic Acids Res. 26 (1998) 73-79
15. Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., Cesareni, G.: MINT: a Molecular INTeraction database. FEBS Lett. 513 (2002) 135-140
16. Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., Eisenberg, D.: The Database of Interacting Proteins: 2004 Update. Nucleic Acids Res. 32 (2004) 449-451
17. Sprinzak, E., Margalit, H.: Correlated Sequence-signatures as Markers of Protein-protein Interaction. J. Mol Biol. 311 (2001) 681-692
18. Kim, W. K., Park, J., Suh, J. K.: Large Scale Statistical Prediction of Protein-protein Interaction by Potentially Interacting Domain (PID) Pair. Genome Inform Ser Workshop Genome Inform. 13 (2002) 42-50
19. Obenauer, J. C., Yaffe, M. B.: Computational Prediction of Protein-protein Interactions. Methods Mol Biol., 261 (2004) 445-468
20. Han, D. S., Kim, H. S., Jang, W. H., Lee, S. D., Suh, J. K.: PreSPI: A Domain Combination Based Prediction System for Protein-protein Interaction. Nucleic Acids Res. 32 (2004) 6312-6320