A Database Server for Predicting Protein-Protein Interactions*

Kyungsook Han** and Byungkyu Park

School of Computer Science and Engineering, Inha University, Inchon 402-751, Korea

Abstract. Large-scale protein interactions are known for several species due to the recent improvements in experimental methods for detecting protein interactions. However, direct determination of all the interactions between the human proteins is difficult even with current high-throughput methods. This paper describes a database server called HPID (http://www.hpid.org) that (1) provides structural interactions between human proteins precomputed from existing structural and experimental data and (2) predicts structural interactions between proteins submitted by users. The structural interactions were obtained by finding known structural interactions of PDB in SCOP domains and then by finding homologs of the domains in target proteins. Based on the structural interactions, we constructed two protein interaction maps, one for human and another for yeast. We believe this is the first attempt to map a whole human interactome at the superfamily level and to compare a human protein interaction map with other species' interaction map.

1 Introduction

One of today's challenges in bioinformatics is to identify all the interactions of human proteins. Large-scale protein interactions are known for several organisms due to the development of high-throughput methods for detecting protein interactions, such as two-hybrid method and mass spectrometry. However, determination of genome-wide protein interactions by experimental methods is limited to low-order organisms such as yeast and Helicobacter pylori [1, 2]. The genes of the human genome are known, but direct determination of all the interactions between the human proteins is still difficult even with high-throughput methods. An intrinsic problem with high-throughput methods is that protein interactions detected by the methods include many false positives. In fact, more than half of current high-throughput data are estimated to be spurious [3]. Considering these constraints, it is important to develop computational methods that can predict protein interactions, and compare different sets of predicted or experimental data of protein interactions.

* This work was supported by the Ministry of Information and Communication of Korea under grant IMT2000-C3-4. We would like to thank Hyongguen Kim and Jinsun Hong for providing the sub-cellular localization information of human proteins.

** To whom correspondence should be addressed. Email: khan@inha.ac.kr

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3036, pp. 271–278, 2004.

It has been widely conjectured that core structural protein interactions are conserved among different organisms. The number of distinct protein domains known so far is around 1,000 (http://scop.mrc-lmb.cam.ac.uk/, SCOP version used here is 1.57 unless otherwise stated) [4]. Therefore, it is inevitable that the same kind of protein domains are involved in diverse types of protein-protein interactions [5]. We have previously predicted protein interactions in human by homologous interactions in yeast, and compared them [6]. However, the predicted interactions between human proteins are estimated to contain many false positives partly because they are derived from the experimental data of yeast protein interactions and the experimental data themselves contain many false positives.

As an improvement of our previous study [6], we attempted to predict structural protein interactions of human and yeast and to compare them. The structural protein interactions are expected to be more accurate than the interactions obtained from the previous study for the following reason. Protein interactions are predicted by finding known structural interactions of PDB [7] in SCOP domains [4] and then by finding homologs of the domains in human proteins. X-ray crystallography and NMR (Nuclear Magnetic Resonance) were main methods for determining the structure data of PDB, and they are more precise experimental techniques than the high-throughput methods for detecting protein interactions.

The aim of this comparative study is to estimate the extent of protein superfamilies in human structural interactome and examine how much overlap exists between the two very diverse eukaryotes. The overall procedure from assigning protein folds to the whole predicted proteome of the complete genomes to visualizing the large network of interactions forms a systematic methodology that can be applied to other genomes. This pilot study can provide the major problems associated in such a bioinformatics analysis and a rough insight on the comparative structural interactomics.

2 Prediction Method

Structural interactions were determined based on the Protein Structural Interaction Map (PSIMAP) [8], which classifies interactions between all known structural protein superfamilies. Structures were assigned to the whole genome (predicted coding regions in the genome) by homology search. The level of homology interaction applied is at the SCOP superfamily. This means that the estimated structural interactome of human does not describe the protein-protein or domain-domain interactions at the molecular level, but at the protein family level. This section starts with definitions of the main concepts in the work.

Definition 1. Given a set *P* of proteins and a set *S* of protein structures in an organism, a set *B* of pairs of the form (protein, structure) shows the structure assignment to proteins in the organism.

$$B = \{(p,s) \mid p \in P, s \in S\} \qquad (1)$$

Fig. 1 shows the data schema of the species_proteins_node table, species_proteins table, and protein_structures table in HPID. The species_proteins_node table represents the set *B*, and the protein_structures table is constructed with the data from the SCOP [4] and Pfam [13] databases.

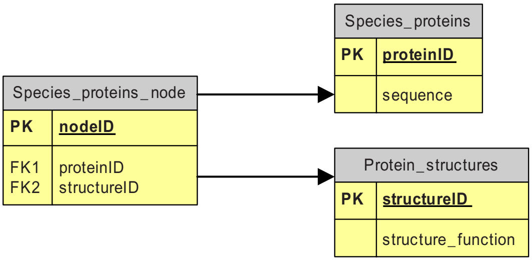


Fig. 1. Data schema for the species_proteins_node table, species_proteins table, and protein_structures table.

Definition 2. If structures s_1 and s_2 interact to each other, and there exist structure assignments $(p_1, s_1) \in B$, and $(p_2, s_2) \in B$ such that $p_1, p_2 \in P$ and $s_1, s_2 \in S$, then proteins p_1 and p_2 interact to each other.

The edge table in Fig. 2 represents the data schema for protein-protein interactions, obtained by definition 2. The global framework of HPID is shown in Fig. 3.

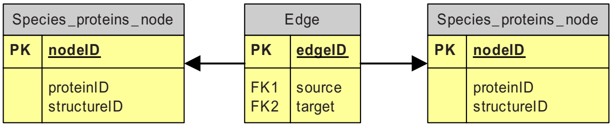


Fig. 2. Data schema for protein-protein interactions.

For the homology search, we constructed a composite database with 27,049 human proteins (http://www.ensembl.org/, v7.29.1), 3,877 yeast proteins, SCOP proteins [4] and NRDB90 [9]. PSI-BLAST [10] was run on the composite DB with SCOP domain sequences as query sequences, the e-value threshold of 0.0005, and the maximum number of profile search rounds of 10 (iterations above 5 does not bring a very high number of distant homologs). The output of PSI-BLAST was parsed by our MS C# program to extract human proteins and yeast proteins matched to SCOP domain sequences and the start and end positions of the matched parts.

Following algorithms describe the procedure for determining the reliability of the predicted protein interactions. 42% of the predicted interactions between human proteins were 'reliable' and the rest 58% were 'unknown'. Fig. 4 shows the data objects used by the algorithms.

```
Check-Reliable-Assignment (protein_ID)
superfamily1 ← Online_prediction_result.Rows[protein_ID][superfamily]
for (int i=0; i < EnsMart.Rows.Length; i++)
    if (EnsMart.Rows[i][protein_ID] == protein_ID)
        superfamily2←Get-Superfamily (Get-PDB_ID-in-Pfam (EnsMart.Rows[i][Pfam_an]))
        If (superfamily2 != null)
                if (superfamily1 == superfamily2) return "reliable"
                else return "unknown"
```

```
Get-Superfamily (PDB_ID)
for (int i=0; i < Pfam2SCOP.Rows.Length; i++)
    if (PDB_ID == Pfam2SCOP.Rows[i][PDB_ID])
                return Pfam2SCOP.Rows[i][superfamily]
```

```
Get-PDB_ID-in-Pfam (Pfam_an)
for (int i=0; i < EnsMart.Rows.Length; i++)
    if (EnsMart.Rows[i][Pfam_an] == Pfam_an)
                return EnsMart.Rows[i][PDB_ID]
```

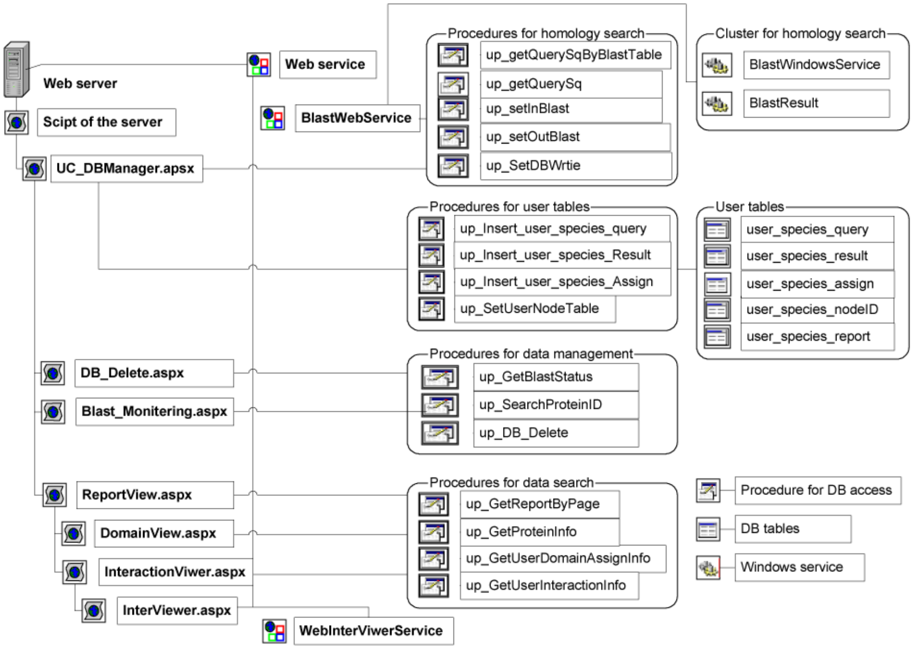
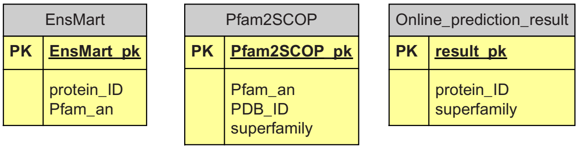


Fig. 3. The framework of the database server.



EnsMart	
PK	EnsMart_pk
	protein_ID Pfam_an

Pfam2SCOP	
PK	Pfam2SCOP_pk
	Pfam_an PDB_ID superfamily

Online_prediction_result	
PK	result_pk
	protein_ID superfamily

Fig. 4. The data objects of EnsMart, Pfam2SCOP, and Online_prediction_result.

The database allows the user to infer potential interactions between proteins submitted by the user. Registration is required to use the online prediction service since prediction results are maintained for individual users. The only required information for registration is the email address, and a user ID and password for the user to use when logging onto the database to view prediction results. When a registered user logs onto to the database server, the status of the user's previous job is displayed regarding whether there is an error in the submitted protein sequences, and whether homology search is complete, in progress, or has not been started yet.

3 Results and Discussion

A protein superfamily was assigned to a human protein (or yeast protein), in a conservative manner, when the matched part is 70% or longer of the original protein superfamily. When multiple superfamilies were matched to a same location of a protein, a superfamily with the highest matching score was assigned to that location and overlap of superfamilies was not allowed. 46% (12,550 proteins) of the total 27,049 human proteins were assigned one or more superfamilies. One human protein was assigned 152 superfamilies and others were assigned 52 or fewer superfamilies (Fig. 5A). 39% (1,509 proteins) of the total 3,877 yeast proteins were assigned at least one superfamily but no more than 6 superfamilies (Fig. 5B).

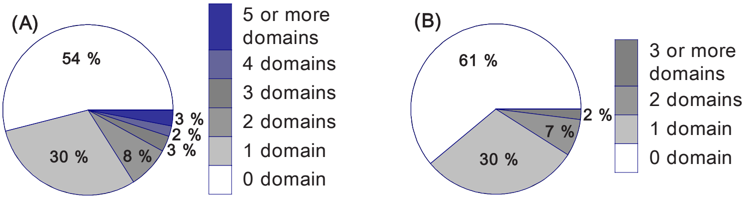
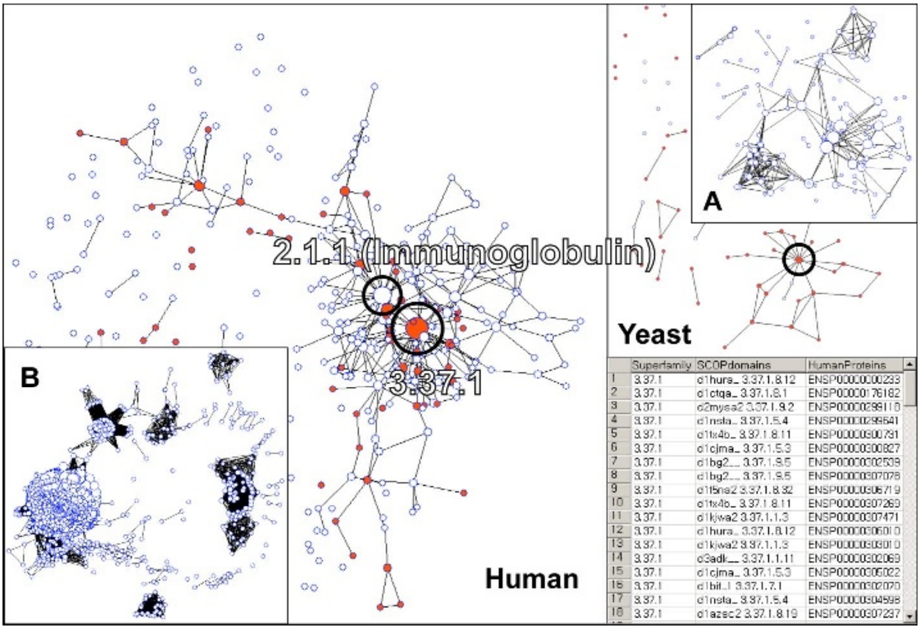


Fig. 5. (A) Superfamily assignment to 27,049 human proteins. (B) Superfamily assignment to 3,877 yeast proteins.

In order to assess the reliability of predicted interactions, we scored interactions based on the identity value of matched parts. In human protein interactions the average identity (X) was 34% with standard deviation (σ) of 23% whereas the average identity was 32% with standard deviation of 25% in yeast protein interactions. An interaction (p1, p2) between proteins p1 and p2 was declared to have a high identity score when both proteins p1 and p2 were assigned a superfamily with an identity $\geq$ X+σ. As shown in Table 1, 220,066 interactions between 2,424 human proteins had a high identity (≥57%=34%+23%), whereas 1,127 interactions between 184 yeast proteins had a high identity (≥57%=32%+25%). The 220,066 human protein interactions with a high identity correspond to 617 interactions at the superfamily level (i.e., at the PSIMAP data) and the 1,127 yeast protein interactions with a high identity correspond to 157 interactions at the superfamily level. Yeast and human are evolutionarily distant species but 74.5% (117 interactions) of the 157 yeast interactions at the superfamily-level were also found in human protein interactions.

Table 1. Protein interactions (including self-loops) in human and yeast. 74.5% (117 interactions) of the 157 yeast interactions at the superfamily-level were also found in human protein interactions.

	human		yeast	
	# proteins (# superfamilies)	# interactions	# proteins (# superfamilies)	# interactions
total interactions	12,550	7,000,943	1,509	88,855
interactions w/high identity at protein level	2,424	220,066	184	1,127
interactions w/high identity at superfamily level	358	617	102	157



	Superfamily	SCOPdomains	HumanProteins
1	3.37.1	d1hura_ 3.37.1.8.12	ENSP00000000233
2	3.37.1	d1ctqa_ 3.37.1.8.1	ENSP00000176182
3	3.37.1	d2mysa2 3.37.1.9.2	ENSP00000299110
4	3.37.1	d1nsla_ 3.37.1.5.4	ENSP00000299641
5	3.37.1	d1tx4b_ 3.37.1.8.11	ENSP00000300791
6	3.37.1	d1cjma_ 3.37.1.5.3	ENSP00000300827
7	3.37.1	d1bg2__ 3.37.1.9.5	ENSP00000302539
8	3.37.1	d1bg2__ 3.37.1.9.5	ENSP00000307078
9	3.37.1	d1f5na2 3.37.1.8.32	ENSP00000306719
10	3.37.1	d1tx4b_ 3.37.1.8.11	ENSP00000307269
11	3.37.1	d1kjwa2 3.37.1.1.3	ENSP00000307471
12	3.37.1	d1hura_ 3.37.1.8.12	ENSP00000306010
13	3.37.1	d1kjwa2 3.37.1.1.3	ENSP00000303010
14	3.37.1	d3adk__ 3.37.1.1.11	ENSP00000302069
15	3.37.1	d1cjma_ 3.37.1.5.3	ENSP00000305022
16	3.37.1	d1bil_l 3.37.1.7.1	ENSP00000302070
17	3.37.1	d1nsla_ 3.37.1.5.4	ENSP00000304598
18	3.37.1	d1azec2 3.37.1.8.19	ENSP00000307237

Fig. 6. The large maps visualize protein interactions at the superfamily level with a high identity in human and yeast (the last row of Table 1). Interaction maps at the protein level are shown in boxes A and B for human and yeast, respectively (the second row of Table 1). Red nodes represent superfamilies shared by human and yeast. A box at the lower right corner lists proteins with superfamily '3.37.1' assigned.

Fig. 6 shows protein interactions with a high identity both at the superfamily level (the last row of Table 1) and at the protein level (the second row of Table 1), visualized by WebInterViewer [11]. Node '3.37.1' has the largest number of interacting superfamilies in both human and yeast. 3.37.1 is a p-loop containing hydrolase. It is one of the most important protein structures occurring in all 4 superkingdoms of life. Its functions are tightly related to energy metabolism such as ATP synthase and signal transduction such as G-protein containing pathways. Therefore, it is not surprising that it has diverse superfamily level structural interactions. Proteins with superfamily

'3.37.1' assigned were also listed in the lower right corner of Fig. 6. Node '2.1.1' of Fig. 6 (IG superfamily. b.1.1 in original SCOP versions) represents the immunoglobulin superfamily, which was found in human, but was missing in yeast assignment. This corroborates a well-known fact that immunoglobulin exists more predominantly in the high-order species only.

Fig. 7 shows many homologs of human proteins that have known homologs in PDB as structural interaction pairs. One of the points of it is that the seemingly complicated interaction patterns in human can be reduced dramatically to a simple basic backbone of family-family interactions. In smaller genomes, the same interaction patterns are found but with much fewer homologs associated with the graph. It follows a scale-free network [12] in parameters. However, the result is a theoretical estimation of human structural interactome. It is analogous to genome draft showing the magnitude of the problem and a rough map for higher resolution mapping of protein-protein interaction.

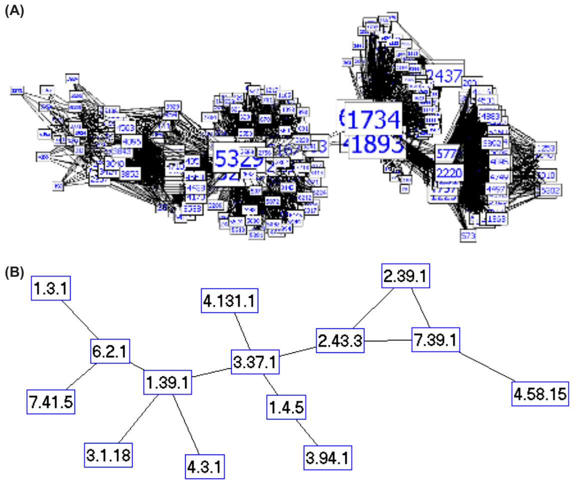


Fig. 7. (A) The second largest connected component in the human protein interaction network. (B) When grouping proteins directly interacting, the apparently complex network in (A) is reduced to a simple network in (B). The simplified network corresponds to a subnetwork of PSIMAP, consisting of superfamilies assigned to the human proteins in (A)

4 Conclusion

We constructed a database server for (1) providing structural interactions between human proteins precomputed from existing structural and experimental data and for (2) predicting structural interactions between proteins submitted by users. The structural interactions between human proteins were compared with those between yeast proteins. They revealed a significant overlap in structural interactome (75%), showing the high rate of conservation of structural interaction at the protein superfamily level. Even though the portion of genes assigned to the whole genomes at present, this indicates that the functional diversification of humans, on the large, is not correlated from different or new interactions derived from new superfamilies. It is more likely that the core structural interactions in life are tightly conserved and the functional diversification and species differentiation are more associated with complex regulatory differentiation. It is possibly related to subtle differentiation in interactions with a basic set of protein structures and their interaction types. As illustrated in Fig. 7, a complicated network of interacting proteins with many homologs can be dramatically reduced to a single backbone network using the family-family interaction concept. The methodology applied here covers many different computational steps and forms a pipeline of structural interactome analysis. Albeit partial, we believe this is the first bioinformatics attempt to map a whole human interactome and to compare a human protein interaction map with other species' interaction map.

References

1. Gavin, A.-C., Bosche, M., Krause, R., et al.: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415 (2002) 141-147
2. Rain, J.-C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A., Legrain, P.: The protein-protein interaction map of Helicobacter pylori. Nature 409 (2001) 211-215
3. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Field, S., Bork, P.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417 (2002) 399-403
4. Lo Conte, L. Brenner, S.E. Hubbard, T.J.P., Chothia, C., Murzin, A.G.: SCOP database in 2002: refinements accommodate structural genomics. Nucl. Acids. Res. 30 (2002) 264-267
5. Park, J., Bolser, D.: Conservation of protein interaction network in evolution. Genome Informatics 12 (2001) 135-140
6. Kim, H., Park, J. Han, K.: Predicting protein interactions in human by homologous interactions in yeast. LNCS 2637 (2003) 159-169
7. Westbrook, J., Feng, Z., Chen, L., Yang, H., Berman, H.M.: The Protein Data Bank and structural genomic. Nucl. Acids. Res. 31 (2003) 489-491
8. Park, J., Lappe, M., Teichmann, S.: Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the pdb and yeast. J. Mol. Biol. 307 (2001) 929–938
9. Lappe, M., Park, J., Niggemann, O., Holm, L.: Generating protein interaction maps from incomplete data: application to fold assignment. Bioinformatics 17 (2001) 149-156
10. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids. Res. 25 (1997) 3389-3402
11. Han, K., Ju, B.-H.: fast layout algorithm for protein interaction networks. Bioinformatics 19 (2003) 1882-1887
12. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabasi, A.L.: The large-scale organization of metabolic networks. Nature 407 (2000) 651-654
13. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller. L., Eddy, S.R., Griffiths-Jones. S., Howe. K.L., Marshall, M., Sonnhammer. E.L.: The Pfam Protein Families Database. Nucleic Acids Research 30 (2002) 276-280