Online Prediction of Interacting Proteins with a User-Specified Protein[¶]

Byungkyu Park and Kyungsook Han*

Department of computer Science and Information Engineering,
Inha University, Incheon 402-751, Korea
khan@inha.ac.kr

Abstract. The protein-protein interaction data available today in databases were either determined by experimental methods or predicted by computational methods. Biologists studying a certain mechanism or function are usually interested in a small-scale network of protein-protein interactions related to the mechanism or function of their interest rather than a genome-wide network of protein-protein interactions. We have previously developed a database server that can predict interactions between the proteins submitted by a user. However, the server cannot find proteins from the databases that potentially interact with a protein given by a user. Finding potentially interacting proteins with a protein of interest is more useful than finding interactions between given proteins, but more difficult partly because it involves intensive computation of searching data in the databases and because different databases have different accession numbers and names for a same protein. This paper describes the new online system for predicting interacting proteins with a user-specified human protein and its preliminary results. We believe that this is the first online system for predicting interacting proteins with a given protein and that it is a useful resource for studying protein-protein interactions.

1 Introduction

A genome-wide network of protein-protein interactions in higher species such as human is very huge and complex, and therefore subnetworks of smaller-scale are usually preferred to a whole network for close examination. However, small-scale subnetworks of protein-protein interactions of interest are not readily available from databases partly because computing all possible subnetworks involves prohibitive computation of searching data in the databases and because different databases have different accession numbers and names for a same protein.

We previously developed a database called the human protein interaction database (HPID; <http://www.hpid.org/>) for predicting potential interactions between proteins submitted by users as well as for providing human protein interaction information pre-computed from existing structural and experimental data [1, 2]. However, the

[¶]This work was supported by the Ministry of Information and Communication of Korea under grant IMT2000-C3-4 and by KOSEF through the Systems Bio-Dynamics Research Center.

* To whom correspondence should be addressed.

previous version of HPID (HPID version 1.0) is not capable of finding potentially interacting partners for a given protein. As an extension of HPID 1.0, we now present a new online prediction system that is capable of predicting interacting proteins with a user-specified human protein. The online prediction system uses the data derived from HPID 1.0 (<http://www.hpid.org/>), BIND (<http://bind.ca/>), DIP (<http://dip.doe-mbi.ucla.edu/>) and HPRD (<http://www.hprd.org/>) for human protein interactions. It also recognizes the protein IDs of EMBL (<http://www.ebi.ac.uk/embl/>), Ensembl (<http://www.ensembl.org/>), HPRD and NCBI (<http://www.ncbi.nlm.nih.gov/>) to search for proteins. This paper discusses the main concepts of the work and its results.

2 Searching Interacting Proteins

The protein interactions were determined on the SCOP [3] Protein Structural Interactome MAP (PSIMAP) [4]. The structural interactions of human proteins were predicted by a homology-based assignment of the domain structures to the whole genome.

Unlike HPID 1.0, HPID 2.0 is not dependent on protein IDs to find interacting partners of proteins. Given a protein sequence, interacting partners with the protein are found by homology search of protein sequences from relevant databases. HPID 2.0 currently considers Ensembl, BIND, DIP, HPRD, and NCBI as relevant databases for homology search. To predict interacting partners of a protein, HPID 2.0 first identifies the domains of the given protein using the information from three databases of superfamily (<http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/>) [5], InterPro (<http://www.ebi.ac.uk/InterProScan/>) and Pfam (<http://www.sanger.ac.uk/>). Fig. 1 shows the overall architecture of the HPID online prediction system.

Fig. 1. The architecture of the HPID online prediction system. Potential interaction proteins are shown in three formats: protein interaction networks visualized by WebInterViewer [6-8], XML or HTML documents

3 Example of Predicting Interacting Proteins

This section shows an example of predicting interacting partners of a protein given by a user. Suppose that a user wants to know interacting proteins with human protein ENSP00000046794. HPID finds 17 interacting partners from the experimental data and 2,101 interacting partners from the prediction data of HPID. 13 out of the 17 interacting partners were also found in the prediction data; that is, the experimental and prediction data have 13 proteins in common. 4 out of the 17 proteins were not found in the prediction data. The reason that the 4 proteins were not found in the prediction data can be explained by the possibility that (1) the proteins have no superfamily assigned to them, or (2) PSIMAP does not have a pair of interacting superfamilies associated with the proteins.

Table 1 shows the superfamilies and functions of the 13 interacting partners that were found both in the experimental and prediction data. The functions of the 13 proteins can be clustered into a few functional groups, which agree with other research results [9, 10]. HPID also provides the information of superfamily, function, sub-cellular localization of potential interaction proteins, which can be further validated by experimental methods.

Table 1. Superfamilies and functions of the 13 proteins interacting with human protein ENSP00000046794

Interaction partners	Superfamily	Gene_Ontology (process, function)
ENSP00000262512	d.93.1	SH3/SH2 adaptor protein activity
ENSP00000315460	d.93.1	SH3/SH2 adaptor protein activity
ENSP00000339007	d.93.1	SH3/SH2 adaptor protein activity, epidermal growth factor receptor binding
ENSP00000339186	d.93.1	SH3/SH2 adaptor protein activity
ENSP00000244007	d.93.1	phosphoinositide phospholipase C activity, receptor signaling protein activity, calcium ion binding, hydrolase activity
ENSP00000343423	d.93.1	phosphoinositide phospholipase C activity, receptor signaling protein activity, calcium ion binding, hydrolase activity
ENSP00000307961	d.93.1	protein tyrosine phosphatase activity, hydrolase activity
ENSP00000264033	g.44.1	transcription factor activity, signal transducer activity, calcium ion binding, ligase activity
ENSP00000302269	d.93.1	transcription factor activity, guanyl-nucleotide exchange factor activity, diacylglycerol binding
ENSP00000288986	d.93.1	receptor binding, cytoskeletal adaptor activity, receptor signaling complex scaffold activity
ENSP00000263405	b.34.2	receptor binding, protein binding
ENSP00000316460	b.34.2	receptor binding, protein binding
ENSP00000221409	d.144.1	

4 Conclusion

Despite a large volume of protein-protein interaction data, small-scale networks of protein-protein interactions of interest are not readily available from databases partly because computing all possible subnetworks involves prohibitive computation of searching data in the databases and because different databases have different accession numbers and names for a same protein. We constructed an online prediction system called HPID 2.0 for searching for interaction partners of a protein given by a user. The online prediction system uses the data derived from HPID 1.0 (<http://www.hpid.org/>), BIND (<http://bind.ca/>), DIP (<http://dip.doe-mbi.ucla.edu/>) and HPRD (<http://www.hprd.org/>) for human protein interactions. It also recognizes the protein IDs of EMBL (<http://www.ebi.ac.uk/embl/>), Ensembl (<http://www.ensembl.org/>), HPRD and NCBI (<http://www.ncbi.nlm.nih.gov/>) to search for relevant proteins. HPID 2.0 is the first online system for predicting interacting proteins with a given protein and will be a useful resource for studying protein-protein interactions.

References

1. Han, K., Park, B., Kim, H., Hong, J., and Park, J.: HPID: The Human Protein Interaction Database. *Bioinformatics* 20 (2004) 2466 – 2470
2. Han, K. and Park, B.: A Database Server for Predicting Protein-Protein Interactions. *Lecture Notes in Computer Science* 3036 (2004) 271-278
3. 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
4. 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
5. Madera, M., Vogel, C., Kummerfeld, S. K., Chothia, C. and Gough, J.: The SUPERFAMILY database in 2004: additions and improvements., *Nucl. Acids. Res.* 32 (2004) Database issue D235-D239
6. Ju, B.-H., Park, B., Park, J. and Han, K.: Visualization and analysis of protein interactions. *Bioinformatics* 19 (2003) 317–318
7. Ju, B.-H. and Han, K.: Complexity management in visualizing protein interaction networks. *Bioinformatics* 19 (2003) i177–i179
8. Han, K. and Ju, B.-H.: A fast layout algorithm for protein interaction networks. *Bioinformatics* 19 (2003) 1882–1887
9. Saito, R., Suzuki, H., and Hayashizaki, Y.: Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucl. Acids. Res.* 30 (2002) 1163-1168
10. Oliver, S.: Guilt-by-association goes global. *Nature* 403 (2000) 601-603