BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 17 2004, pages 3273–3276
doi:10.1093/bioinformatics/bth366

POINT: a database for the prediction of protein–protein interactions based on the orthologous interactome

Tao-Wei Huang[1,2,†], An-Chi Tien[1,†], Wen-Shien Huang[2], Yuan-Chii G. Lee[1,3], Chin-Lin Peng[2], Huei-Hun Tseng[1,2], Cheng-Yan Kao[2] and Chi-Ying F. Huang[1,2,4,*]

[1]Division of Molecular and Genomic Medicine, National Health Research Institutes, Taipei 115, Taiwan, Republic of China, [2]Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan, Republic of China, [3]Graduate Institute of Medical Informatics, Taipei Medical University, Taipei 110, Taiwan, Republic of China and [4]Institute of Biotechnology in Medicine, National Yang-Ming University, Taipei 112, Taiwan, Republic of China

Received on November 10, 2003; revised on May 17, 2004; accepted on June 4, 2004
Advance Access publication June 24, 2004

ABSTRACT

Summary: One possible path towards understanding the biological function of a target protein is through the discovery of how it interfaces within protein–protein interaction networks. The goal of this study was to create a virtual protein–protein interaction model using the concepts of orthologous conservation (or interologs) to elucidate the interacting networks of a particular target protein. POINT (the prediction of interactome database) is a functional database for the prediction of the human protein–protein interactome based on available orthologous interactome datasets. POINT integrates several publicly accessible databases, with emphasis placed on the extraction of a large quantity of mouse, fruit fly, worm and yeast protein–protein interactions datasets from the Database of Interacting Proteins (DIP), followed by conversion of them into a predicted human interactome. In addition, protein–protein interactions require both temporal synchronicity and precise spatial proximity. POINT therefore also incorporates correlated mRNA expression clusters obtained from cell cycle microarray databases and subcellular localization from Gene Ontology to further pinpoint the likelihood of biological relevance of each predicted interacting sets of protein partners.

Availability: POINT can be freely accessed at http://insilico.csie.ntu.edu.tw:9999/point/.

Contact: chiying@nhri.org.tw

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

INTRODUCTION

One of the challenges in the post-genomic era is to accelerate the functional analyses of uncharacterized proteins. It is commonly believed that identification of interaction partners for a protein of unknown function should provide novel insights into its biological function. Recently, increasing use of high-throughput two-hybrid analysis from *Saccharomyces cerevisiae* (Walhout and Vidal, 2001), *Drosophila melanogaster* (Giot *et al*., 2003) and *Caenorhabditis elegans* (Li *et al*., 2004; Walhout *et al*., 2000) has generated an enormous amount of data. The knowledge of interactions conserved in other organisms (or interologs) (Walhout *et al*., 2000) ought to represent useful information that will allow the formulation and testing of biological hypotheses. This comparative genomics strategy (Walhout *et al*., 2000; Wojcik and Schachter, 2001) may facilitate functional annotation of uncharacterized proteins.

As the genomic era continues to unfold, it is becoming increasingly apparent that protein interaction networks are extremely complex. There is clearly a need for the exploration of these datasets and the development of a systematic and stepwise approach that can predict and compare the protein–protein interaction networks in different model organisms. Several databases, such as IPPRED (Goffard *et al*., 2003) and STRING (von Mering *et al*., 2003), focus on the prediction of protein–protein interactions. The former is designed to predict such binary interactions that are present in different organisms. The latter predicts association based on phylogenetic profiling, genomic proximity and by neighboring. In contrast, we have previously established a methodology combining various publicly accessible databases to reveal the interaction networks related to Aurora kinases (Tien *et al*.,

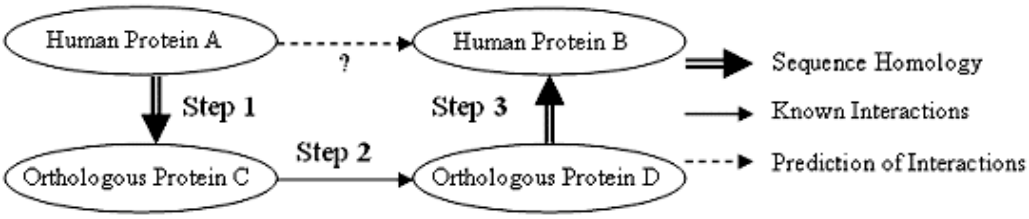

Fig. 1. A schematic illustration of the search module for human protein–protein interactions in POINT.

2004). This proposed interaction model, including binary (direct) interactions and their interaction networks, was based upon the notion that some of the interacting proteins may also be evolutionarily conserved (i.e. from yeast to humans), interacting with each other in the same spatial configurations within cell compartments. In addition, interacting partners may be found within the same gene-expression cluster and, thus, genes with similar expression patterns may respond to the same functional category of proteins, (e.g. Eisen *et al*., 1998; Marcotte *et al*., 1999), and therefore they can be used to aid in interaction-target selection.

Here, we provide the details of a publicly accessible web server, POINT, that visualizes all the potential interacting protein networks among various model organisms for a given query protein to facilitate the selection of possible interacting targets. Additional annotations of cell cycle expression pattern and subcellular localization of a target protein allow researchers to evaluate what the likely biological relevance of each potential interaction may be.

STRUCTURE OF THE DATABASE

POINT has been designed to predict protein–protein interaction networks that are evolutionarily conserved from mouse, fruit fly, worm and yeast and then to human. To construct the predicted human protein–protein interaction networks, a large quantity of protein interactions data was imported into our database from Database of Interacting Proteins (DIP) (Xenarios *et al*., 2002). However, these datasets do not reach saturation level, and this resulted in substantial limitations to the protein interaction maps of potential interologs. Furthermore, protein–protein interactions often occur between protein domains but not full-length proteins. Global sequence alignments by BLAST (Altschul *et al*., 1990) may miss domain-based interactions across organisms. With these two considerations, a given query of a human protein sequence will be searched by BLAST against protein sequences from various different organisms downloaded from PIR-NREF (Wu *et al*., 2003). To satisfy the different needs of users, five proteins with the highest homology in a given organism will be shown in the search results. Next, the user can manually select potential orthologs and then search for proteins that interact with these orthologs; two levels of radial network expansion will be included in the search results. This step reduces the possible absence of predictive interacting proteins when they are mapped back to the human proteins, since the conserved interactions may be either functionally linked or indirect protein–protein interactions (i.e. in a protein complex). Finally, the search tree for the interacting proteins in a given organism can be automatically transformed into human proteins using another homology search by BLAST. The workflow of POINT is illustrated in Figure 1.

Visualization of protein–protein interactions can be represented by two approaches. First, a graph of the protein–protein interactions networks is visualized via a Java applet (Mrowka, 2001) as shown in Figure 2a. Second, the graph provides a tree-view structure with hyperlinks to external databases, such as GO, PIR-NREF, UniGene (Wheeler *et al*., 2004) and UniProt (Apweiler *et al*., 2004), as shown in Figure 2b. In addition, as protein–protein interactions require precise spatial proximity and temporal synchronicity, POINT also provides additional information with respect to such predictive interactions. Each identified protein may be visualized with its annotation from Gene Ontology (GO) (Harris *et al*., 2004), as well as in terms of the cell cycle state using the human cell cycle microarray database (Whitfield *et al*., 2002) or the yeast cell cycle microarray database (Cho *et al*., 1998). This additional information adds valuable parameters, allowing mimicry of interactions within the living cells and can also be utilized to prioritize potential candidates for a given query protein. All of the nodes refer to orthologous proteins. These visualizations represent the overall complexity of the protein networks in different organisms.

STATE OF THE DATABASE

POINT embraces several well-established groups of biological software programs and databases as data sources and these are then used for the evaluation of protein–protein interactions (Table 1). POINT is updated periodically based on the release version of DIP. In addition, the source code for developing this system is open-source and can be freely accessed on request.

FUTURE DIRECTIONS

To provide a better prediction of a protein interaction map in a target organism, it is necessary to rely on a comprehensive analysis of the protein interaction maps of other

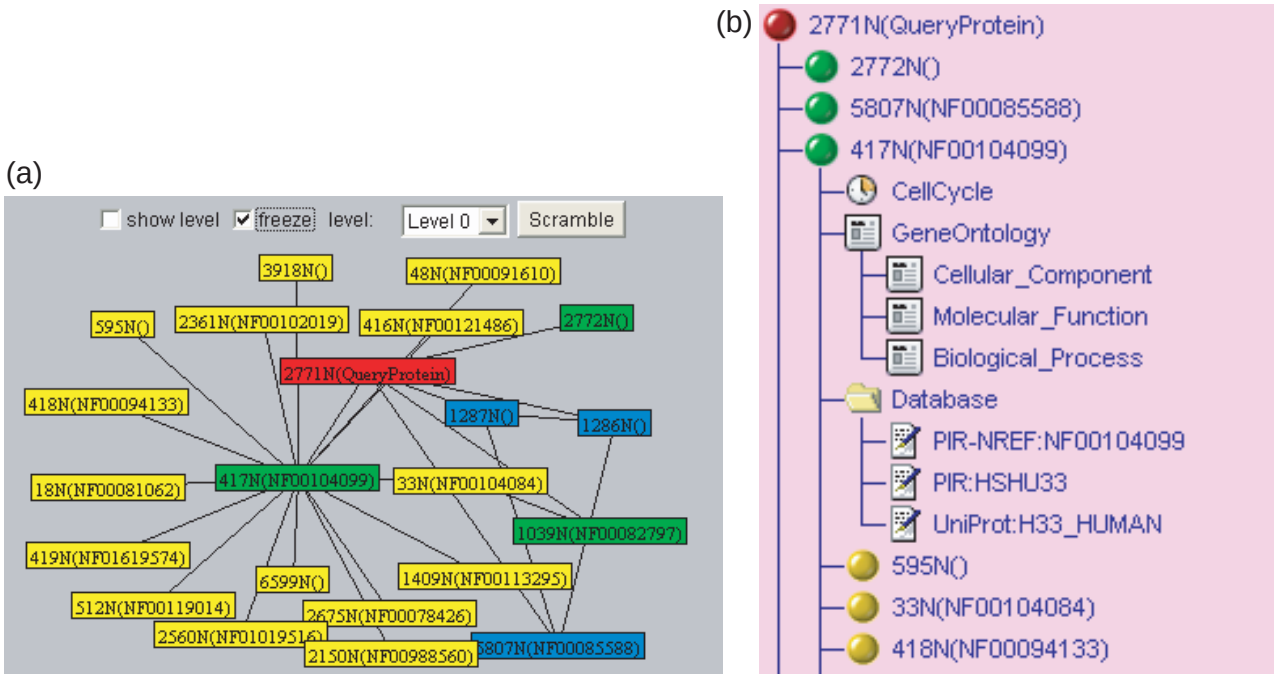

Fig. 2. An example of protein interaction networks in POINT. (**a**) The network representation. All the nodes refer to orthologous proteins in S1(S2) format. The symbol S1 is the mouse, fruit fly, worm or yeast Node ID in DIP, and the symbol S2 is the PIR-NREF human protein sequence accession number. The query protein (2771N) is labeled in red and designated as level 0. The interaction proteins (417N and 2772N; level 1) of the query protein are labeled in green. The proteins (595N and 33N) that interact with level-1 proteins are label in yellow and designated as level 2. Some interaction proteins simultaneously present in both level 1 and 2 are labeled in blue. (**b**) The tree-view representation. Additional external hyperlinks, for example, cell cycle information and gene ontology information, are provided in tree-view representation.

Table 1. List of the biological software and databases used in POINT

Database	Website and version	Purpose
BLAST	http://www.ncbi.nlm.nih.gov/BLAST/ (2.2.8)	The program of searching sequence homology
DIP	http://dip.doe-mbi.ucla.edu/ (2004-04-04)	The data of mouse, fruit fly, worm and yeast protein–protein interactions
GO	http://www.geneontology.org/ (2004-04-26)	The gene ontology of human, mouse, fruit fly, worm and yeast
Human Cell Cycle	http://genome-www.stanford.edu/Human-CellCycle/Hela/	The data of human cell cycle
PIR-NREF	http://pir.georgetown.edu/ (2004-04-26, Release 1.45)	The data of human, mouse, fruit fly, worm and yeast protein sequences
UniGene	http://www.ncbi.nlm.nih.gov/UniGene/(Build#168)	The data of Human EST Clone ID for microarray database
UniProt	http://www.pir.uniprot.org/ (2004-04-26, Release 1.8)	The protein annotations in Swiss-Prot/ TrEMBL
Yeast Cell Cycle	http://genome-www.stanford.edu/cellcycle/data/rawdata/	The data of yeast cell cycle

reference organisms. Several databases, such as BIND (Bader *et al.*, 2001), have extensive collections of protein–protein interaction datasets. These public-accessible datasets will be continuously incorporated into POINT. In addition, we anticipate that additional biological datasets, such as the systematic knockout phenotype data, may provide improved evaluation and help target prioritization. Prediction from more model organisms will be included in the future to offer more information relevant to the different biological requirements of various researchers. In order to determine the most accurate orthologs, the use of conserved domain database such as Pfam and PROSITE will also be integrated. This may

help solve problems with respect to those proteins lacking domain information because the whole protein sequence was identified directly by BLAST.

ACKNOWLEDGEMENTS

We would like to thank DIP, GO, NCBI, PIR-NREF, UniProt for their public-accessible databases and software that provided the foundation for construction of this applied POINT database. Development of the POINT database was supported by grants from the National Science Council (NSC92-3112-B-400-005) and National Health Research Institutes to C.F.H.

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32** (Database issue), D115–D119.

Bader,G.D., Donaldson,I., Wolting,C., Ouellette,B.F., Pawson,T. and Hogue,C.W. (2001) BIND—the biomolecular interaction network database. *Nucleic Acids Res.*, **29**, 242–245.

Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. and Davis,R.W. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Giot,L., Bader,J.S., Brouwer,C., Chaudhuri,A., Kuang,B., Li,Y., Hao,Y.L., Ooi,C.E., Godwin,B., Vitols,E. *et al*. (2003) A protein interaction map of *Drosophila melanogaster. Science*, **302**, 1727–1736.

Goffard,N., Garcia,V., Iragne,F., Groppi,A. and De Daruvar,A. (2003) IPPRED: server for proteins interactions inference. *Bioinformatics*, **19**, 903–904.

Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32** (Database issue), D258–D261.

Li,S., Armstrong,C.M., Bertin,N., Ge,H., Milstein,S., Boxem,M., Vidalain,P.O., Han,J.D., Chesneau,A., Hao,T. *et al*. (2004) A map of the interactome network of the metazoan *C.elegans. Science*, **303**, 540–543.

Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.

Mrowka,R. (2001) A Java applet for visualizing protein–protein interaction. *Bioinformatics*, **17**, 669–671.

Tien,A.C., Lin,M.H., Su,L.J., Hong,Y.R., Cheng,T.S., Lee,Y.C., Lin,W.J., Still,I.H. and Huang,C.Y. (2004) Identification of the substrates and interaction proteins of aurora kinases from a protein-protein interaction model. *Mol. Cell Proteomics*, **3**, 93–104.

von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.

Walhout,A.J., Sordella,R., Lu,X., Hartley,J.L., Temple,G.F., Brasch,M.A., Thierry-Mieg,N. and Vidal,M. (2000) Protein interaction mapping in *C.elegans* using proteins involved in vulval development. *Science*, **287**, 116–122.

Walhout,A.J. and Vidal,M. (2001) Protein interaction maps for model organisms. *Nat. Rev. Mol. Cell Biol.*, **2**, 55–62.

Wheeler,D.L., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E. *et al*. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.*, **32** (Database issue), D35–D40.

Whitfield,M.L., Sherlock,G., Saldanha,A.J., Murray,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. and Botstein,D. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.

Wojcik,J. and Schachter,V. (2001) Protein–protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, **17** (Suppl. 1), S296–S305.

Wu,C.H., Yeh,L.S., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al*. (2003) The protein information resource. *Nucleic Acids Res.*, **31**, 345–347.

Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S.M. and Eisenberg,D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.