Nucleic Acids Research, 2002, Vol. 30, No. 1 303–305

DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions

Ioannis Xenarios, Łukasz Salwínski, Xiaoqun Joyce Duan, Patrick Higney, Sul-Min Kim and David Eisenberg*

UCLA-DOE Laboratory of Structural Biology and Molecular Medicine, Molecular Biology Institute, PO Box 951570, UCLA, Los Angeles, CA 90095-1570, USA

Received September 18, 2001; Accepted September 20, 2001

ABSTRACT

The Database of Interacting Proteins (DIP: http://dip.doe-mbi.ucla.edu) is a database that documents experimentally determined protein–protein interactions. It provides the scientific community with an integrated set of tools for browsing and extracting information about protein interaction networks. As of September 2001, the DIP catalogs ~11 000 unique interactions among 5900 proteins from >80 organisms; the vast majority from yeast, *Helicobacter pylori* and human. Tools have been developed that allow users to analyze, visualize and integrate their own experimental data with the information about protein–protein interactions available in the DIP database.

INTRODUCTION

During the last 3 years protein interaction databases have grown to the point of becoming both a commonly used reference source for experimental biologists (1–3), as well as a data source enabling studies of the properties and structure of entire protein interaction networks (4). With the recent development of genome-wide experimental methods such as the two-hybrid test, protein chips and mass spectrometric analysis, the number of reported interactions has increased exponentially. On one hand, this leads to a rapid increase of the coverage of the protein interaction map, providing deeper insight into global properties of the interaction networks. On the other hand, the increasing size and complexity of the available dataset challenges the database developers to provide visualization and analysis tools that utilize the information contained in the network structure (5). As the field matures, it is increasingly clear that we must develop data evaluation methods that can estimate uncertainties and identify the most reliable subset of the putative interactions.

STRUCTURE OF THE DATABASE

The structure of the DIP has been designed to capture the essential information about protein–protein interactions available from experimental data. The database is implemented as a relational database composed of four tables (6). *Protein Table* lists proteins participating in an interaction within DIP. It provides, besides the DIP accession number, cross-references to the three major sequence databases (SWISS-PROT, GenBank, PIR) as well as additional information about the proteins such as keyword, localization and cellular function. *Interaction Table* catalogs binary interactions between proteins including, when available, information on the interacting domains and the ranges of amino acids necessary for an interaction. *Method Table* entries capture the experimental technique (such as genome wide two-hybrid screen, immunoprecipitation, affinity binding, antibody blockage) that has been used to determine each interaction and also point to the published sources of experimental data listed in *Reference Table*. *Reference Table* lists all the references to different articles that demonstrate protein interactions and link them to the MEDLINE (National Library of Medicine, MD) database.

STATE OF THE DATABASE

Over the last year, the number of distinct protein–protein interactions in the DIP has nearly tripled. Currently, the database catalogs >10 500 unique protein–protein interactions between >5900 proteins. Table 1 shows the distribution of the data among the organisms most frequently represented in the database. As earlier, we observe that yeast is the predominant organism in DIP accounting for >7900 distinct interactions (70% of the total interactions), it is followed by *Helicobacter pylori* and then human, contributing 1420 and 631 interactions, respectively.

The rapid growth of the database was possible because of two factors. First, the number of articles entered into the database nearly doubled to ~1500 providing a diverse source of protein interaction data. However, most importantly, the database currently contains a set of 6125 interactions identified in a number of genome-wide yeast two-hybrid screens (7–11). The reliability of the large scale data is, in general, lower than that provided by experiments focused on a particular interaction (12). However, we can expect that the rough interaction map generated by genome-wide two-hybrid screens provides the scientific community with leads that can be further verified

*To whom correspondence should be addressed. Tel: +1 310 825 3754; Fax: +1 310 206 3914; Email: david@mbi.ucla.edu

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

Table 1. DIP statistics for September 2001 release

Organisms	Number of proteins	Number of interactions	Number of experiments per interaction			
			1	2	3	>4
Yeast	4162	7975	6892 (5576)	725 (409)	202 (77)	126 (63)
Helicobacter pylori	711	1420	1420			
Human	558	662	486	111	26	23
Other organisms	475	495	396	79	17	7

Statistics are reported for the number of proteins and interactions in the three major organisms. Further sub-classifications of the interactions are shown based on the number of experimental observations. In parentheses are given the number of experiments described by large-scale yeast two-hybrid screen.

by follow-up experiments as well as through computational approaches. As shown in Table 1, >5500 interactions determined uniquely in the genome-wide screens await such an evaluation.

THE JDIP VISUALIZATION TOOL

The increasing size and complexity of the data available in DIP stimulates the development of tools that allow biologists to study and analyze entire networks of protein–protein interactions. Last year we introduced a graphic display of the protein interaction network, centered on any given protein contained in DIP. It provides the means for a fast, visual evaluation of the protein's interaction environment, represented as a static graph. However, it soon became clear that a more interactive approach is necessary; one that allows for a 'navigation' of the protein interaction network. We therefore developed the JDIP tool, both as an applet available within the web interface as well as an independent, cross-platform Java application. Not only does JDIP allow us to curate the data conveniently, but also provides a generic framework for integrating a number of visualization and analysis tools. Currently, besides numerous visualization options, JDIP provides access to a number of genome-wide mRNA expression datasets (13), which can be analyzed after mapping them onto the underlying protein interaction network. The program has been developed to accommodate annotation of network elements with numerical and textual data that can be independent of the DIP. Its XML compliance allows one to annotate the interaction graph with user-specified data and then render the resulting network according to a set of rules specified by the user.

FUTURE DIRECTIONS

At present, the DIP curators have two goals: first, to increase the size of the DIP dataset in the human protein subset and, secondly, to provide additional tools for accessing and analyzing the information contained within the database.

A number of strategies have been employed to identify articles containing information suitable for entering into DIP. These include automated searches of the MEDLINE database both against a strict set of keywords as well as searches based on the data-mining strategy (14). Another approach attempts to transfer a known interaction between a pair of proteins to new pairs, detected as homologs of the interacting partners. In order to confirm such putative interactions, MEDLINE records can be searched for papers with concurrent appearance of each member of the pair. This approach promises to extend the many known interactions between yeast proteins to other organisms, most importantly to human, which currently constitutes only 10% of DIP.

Despite the emergence of automated data-mining strategies, identification of the relevant articles and proteins remains the rate-limiting step in data entry. Therefore, we propose to the scientific community that the authors of articles describing protein–protein interactions insert within their articles a line of condensed text in a format described in Figure 1 that identifies the interactions they discover. Such a simple scheme, similar in idea to that of Bader *et al.* (1) would, at a little overhead to the experimentalist, significantly increase the rate of incorporation of novel data into protein interaction databases.

If the community were to adapt this proposal, the majority of the newly reported interactions could be automatically deposited into any interaction database. At a later stage, those interactions could be reanalyzed and further curated to extract the remaining details.

Another area for DIP improvement encompasses integration of the database with a number of already existing, well established biological databases such as SWISS-PROT, TRANSPATH, KEGG, YPD to allow users to easily access gather most of the information about a single protein. To this we intend to increase the number of cross-references reported for each DIP entry. We also encourage the maintainers of other databases to provide cross-references to DIP entries in a manner similar to the one already present within SWISS-PROT.

DATA SUBMISSION AND CURATION

We seek expert curators to screen entries into the DIP. Scientists are invited to contribute to this database, by submitting interactions directly over the World Wide Web after obtaining a user account. To obtain an account, please contact us at dip@mbi.ucla.edu. Help for editing and submission is available online; questions can also be directed to dip@mbi.ucla.edu or at the fax number and address listed. Please feel free to send email containing published protein–protein interactions, and a curator will enter this information in the DIP.

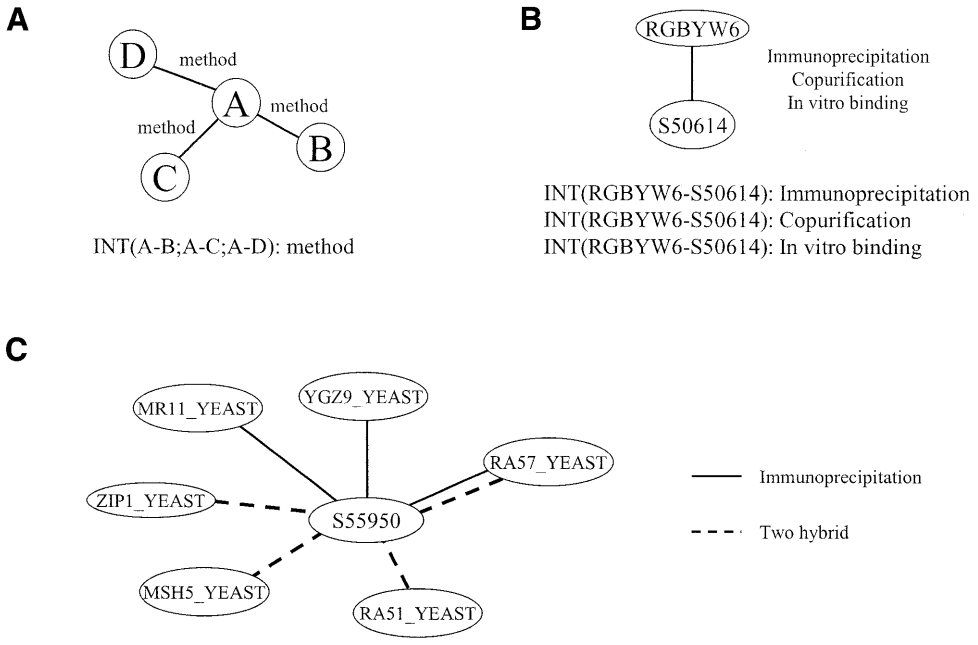

Figure 1. Three illustrations of protein interactions and their descriptions by machine readable text of the sort that would automate database entry. (**A**) A set of four proteins (A, B, C and D) have been studied and shown to interact by a given method. The condensed text is shown below. (**B**) Example of the SWI6–SWI4 interactions observed by Siegmund and Nasmyth (15) using immunoprecipitation, copurification and *in vitro* binding. In this case the protein PIR codes are used RGBYW6 for SWI6p and S50614 for SWI4p. (**C**) Example of the Zip3 interactions demonstrated by two-hybrid screen and immunoprecipitation by Agarwal and Roeder (16). In this case S55950 (Zip3p) can be immunoprecipitated with RA57p (RA57_YEAST), MR11p (MR11_YEAST) and Zip2p (YGZ9_YEAST). Two-hybrid screen was performed and showed interactions of Zip3p with ZIP1p, MSH5p, RAD57p, RAD51p.

ACKNOWLEDGEMENTS

We thank NIH and DOE for support and R. Landgraf for bibliography scanning software.

REFERENCES

1. 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.
2. Costanzo,M.C., Crawford,M.E., Hirschman,J.E., Kranz,J.E., Olsen,P., Robertson,L.S., Skrzypek,M.S., Braun,B.R., Hopkins,K.L., Kondu,P., Lengieza,C., Lew-Smith,J.E., Tillberg,M. and Garrels,J.I. (2001) YPD(TM), PombePD(TM) and WormPD(TM): model organism volumes of the BioKnowledge(TM) library, an integrated resource for protein information. *Nucleic Acids Res.*, **29**, 75–79.
3. Xenarios,I., Fernandez,E., Salwinski,L., Duan,X.J., Thompson,M.J., Marcotte,E.M. and Eisenberg,D. (2001) DIP: the database of interacting proteins: 2001 update. *Nucleic Acids Res.*, **29**, 239–241.
4. Jeong,H., Mason,S.P., Barabasi,A.L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
5. Xenarios,I. and Eisenberg,D. (2001) Protein interaction databases. *Curr. Opin. Biotechnol.*, **12**, 334–339.
6. Xenarios,I., Rice,D.W., Salwinski,L., Baron,M.K., Marcotte,E.M. and Eisenberg,D. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.*, **28**, 289–291.
7. 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.
8. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P., Qureshi-Emili,A., Li,Y., Godwin,B., Conover,D., Kalbfleisch,T., Vijayadamodar,G., Yang,M., Johnston,M., Fields,S. and Rothberg,J.M. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
9. Ito,T., Tashiro,K., Muta,S., Ozawa,R., Chiba,T., Nishizawa,M., Yamamoto,K., Kuhara,S. and Sakaki,Y. (2000) Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA*, **97**, 1143–1147.
10. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
11. Newman,J.R., Wolf,E. and Kim,P.S. (2000) From the cover: a computationally directed screen identifying interacting coiled coils from *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **97**, 13203–13208.
12. Grigoriev,A. (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **29**, 3513–3519.
13. Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.
14. Marcotte,E., Xenarios,I. and Eisenberg,D. (2001) Mining literature for protein–protein interactions. *Bioinformatics*, **17**, 1–7.
15. Siegmund,R.F. and Nasmyth,K.A. (1996) The *Saccharomyces cerevisiae* Start-specific transcription factor Swi4 interacts through the ankyrin repeats with the mitotic Clb2/Cdc28 kinase and through its conserved carboxy terminus with Swi6. *Mol. Cell. Biol.*, **16**, 2647–2655.
16. Agarwal,S. and Roeder,G.S. (2000) Zip3 provides a link between recombination enzymes and synaptonemal complex proteins. *Cell*, **102**, 245–255.