BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 6 2005, pages 832–834
doi:10.1093/bioinformatics/bti115

Databases and ontologies

The MIPS mammalian protein–protein interaction database

Philipp Pagel[1], Stefan Kovac[1], Matthias Oesterheld[1], Barbara Brauner[1], Irmtraud Dunger-Kaltenbach[1], Goar Frishman[1], Corinna Montrone[1], Pekka Mark[2], Volker Stümpflen[1], Hans-Werner Mewes[1,2], Andreas Ruepp[1] and Dmitrij Frishman[1,2,*]

[1]Institute for Bioinformatics/MIPS, GSF—National Research Center for Environment and Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany and [2]Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, 85350 Freising, Germany

Received on September 3, 2004; revised and accepted on October 21, 2004
Advance Access publication November 5, 2004

ABSTRACT

Summary: The MIPS mammalian protein–protein interaction database (MPPI) is a new resource of high-quality experimental protein interaction data in mammals. The content is based on published experimental evidence that has been processed by human expert curators. We provide the full dataset for download and a flexible and powerful web interface for users with various requirements.

Availability: The MPPI database is located at http://mips.gsf.de/proj/ppi/

Contact: d.frishman@wzw.tum.de

1 INTRODUCTION

Protein–protein interactions (PPI) determine biological processes at many levels of cellular complexity—from basic metabolism to cell differentiation. Their importance is reflected by the number of protein interaction experiments described in the life science literature and the increasing interest in high-throughput techniques such as yeast two-hybrid (Ito *et al*., 2001; Uetz *et al*., 2000) and large scale mass spectroscopy of protein complexes (Ho *et al*., 2002; Gavin *et al*., 2002). Computational analyses of experimental data as well as *in silico* predictions are important tools in the effort to increase our understanding of cellular architecture. In addition to the necessity of a complete and in-depth knowledge of PPI networks for the understanding of cellular biology, they are highly interesting for target selection aimed at pharmaceutical applications.

Until recently, most of the databases and large scale experiments on PPI were derived from microorganisms, most prominently *Saccharomyces cerevisiae*. While yeast is the best established model organism, many open questions concerning higher eukaryotes involve features not present in this organism. Especially, work with potential medical implications often requires mammalian models. Despite its practical relevance, comparatively little PPI data from mammals has been available in public databases like BIND (Bader *et al*., 2003), DIP (Salwinski *et al*., 2004) and MINT (Zanzoni *et al*., 2002). Recent efforts by database maintainers and experimental researchers have started to greatly improve this situation. Scientific literature is rich with experiments demonstrating such interactions utilizing a large number of technical approaches. Our goal was to harvest this abundance of available literature and generate a systematic, manually curated database of mammalian PPI (MPPI) to serve both the bioinformatics community as well as the wet lab scientist who wants to quickly find relevant links between the protein of interest and known binding partners.

2 ANNOTATION STRATEGY

The first and foremost principle of our MPPI database is to favor quality over completeness. Therefore, we decided to include only published experimental evidence derived from individual experiments as opposed to large-scale surveys. High-throughput data may be integrated later, but will be marked to distinguish it from evidence derived from individual experiments.

Our next design decision was to choose an appropriate organism as the primary model organism for the database. Although both mouse and human immediately come to mind as the ideal choices, a human or mouse PPI database would unnecessarily limit the project and ignore common lab practice. Due to the great structural and sequence similarity among mammalian orthologous proteins, it is quite common to perform interaction experiments using, e.g., endogenous protein *X* in a human cell line together with recombinant protein *Y* derived from sheep sequence thus crossing species boundaries. Such cross-species experiments represent a large fraction of the available evidence in literature. Taking this into account, we decided not to restrict our database to a single species but rather allow any mammalian protein in our dataset. Nevertheless, for systematic analysis it can be desirable to map the data to one reference genome. We chose to use *Mus musculus*—the most widely used mammalian model—as our reference species and provide links to our PEDANT (Frishman *et al*., 2003) mouse genome database whenever possible.

Given the large number of genes in mammalian genomes and the high percentage of yet uncharacterized and putative proteins, the classical gene-by-gene strategy which has commonly been used in

*To whom correspondence should be addressed.

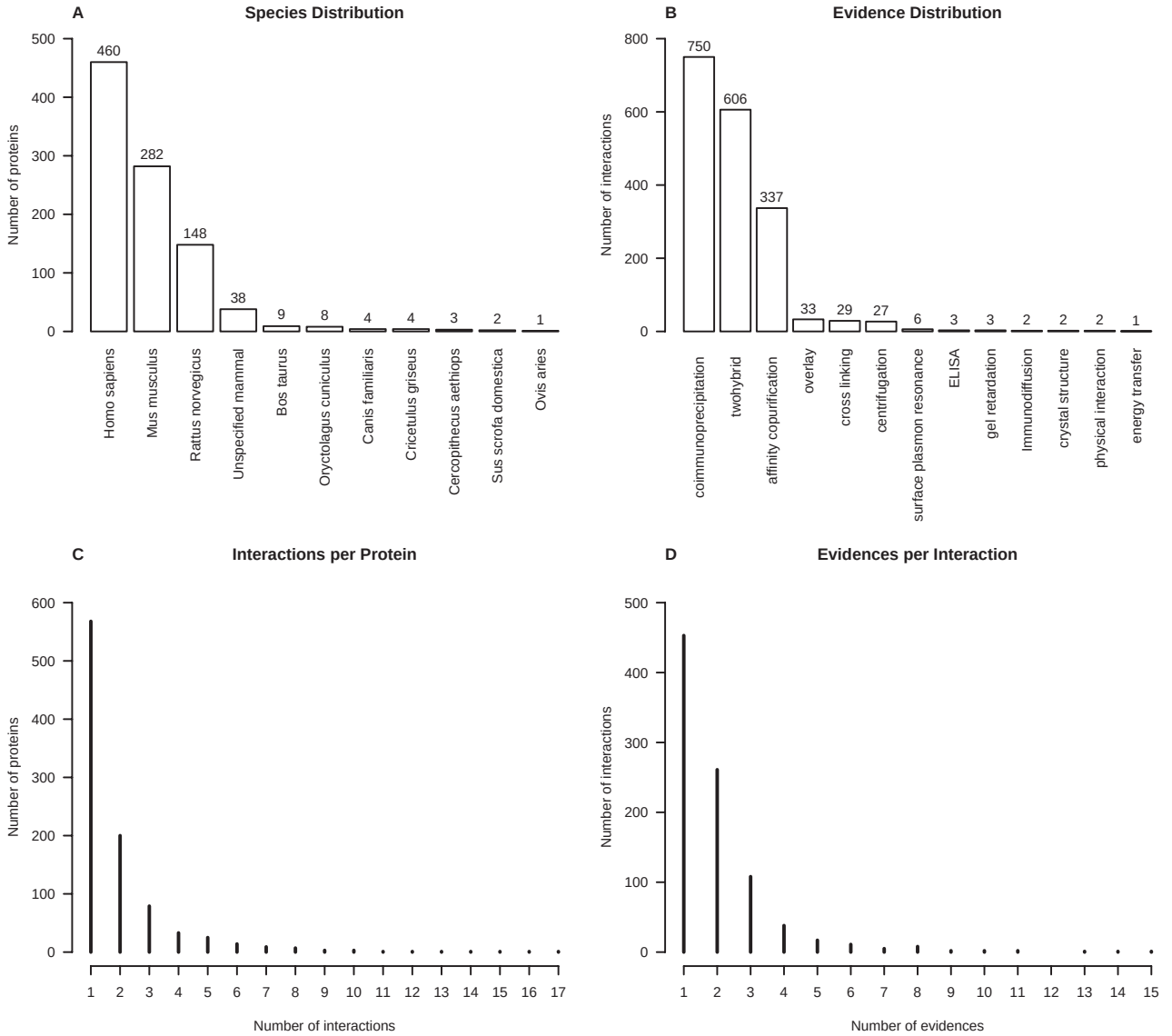

Fig. 1. Statistics: **(A)** Three species account for >90% of the proteins in our data. **(B)** Co-IP, two-hybrid methods and co-purification clearly dominate the evidence entries. **(C)** While most proteins in our database have only one annotated interaction, up to 17 binding partners can be found for some. **(D)** For many interactions there is more than one experimental evidence in our dataset.

the annotation of small genomes is not a good solution. Instead of finding literature about each gene product we decided to reverse the approach and locate the gene for each literature reference at hand. Relevant publications were identified in PubMed searches using keywords such as 'mammalian', 'mouse', 'two-hybrid', 'coimmunoprecipitation', 'binds to', ... in various combinations.

3 IMPLEMENTATION AND DATA

All data are stored in a MySQL database and are accessible through a web interface implemented as Perl CGI scripts. The interface was designed to be as intuitive as possible for the occasional user while allowing complex Boolean queries for advanced requirements. We provide three different search forms and two result formats. Another feature is the graphical representation of a protein with all its neighbors.

For detailed analysis, the entire dataset is available for download in the recently defined PSI-MI standard format (Hermjakob *et al*., 2004).

In addition to the proteins involved in an interaction we provide information on specific details such as the PubMed reference, experimental technique used, probable binding sites and functional role of the interaction. Links to external databases such as Swiss-Prot are provided for most proteins.

Currently, our dataset contains >1800 evidence entries for PPI among >900 proteins from 10 mammalian species. The data was extracted from >370 articles. On average, each protein in the database is involved in 1.92 interactions and each interaction is supported

by 1.98 evidence entries. Figure 1 gives a graphical overview of the composition of our data.

As the importance of protein-interaction data in higher eukaryotes—and especially mammals—has been recognized by many researchers, several efforts to improve the amount of available data have been undertaken. The human protein reference database (Peri *et al*., 2003) aims at a comprehensive annotation of the human proteome and includes information about a large number of protein interactions. While their dataset is significantly larger than ours we believe that our data is complementary to the HPRD set because the overlap is comparatively small (less than 30% of our PMIDs appear in HPRD at the time of writing) and especially because we provide much more detailed information on the interactions and do not limit our data to one species. Other efforts are underway in many of the well-known PPI databases. Large-scale interaction experiments have been performed for *Caenorhabditis elegans* (Li *et al*., 2004) and *Drosophila* (Giot *et al*., 2003) but little such data exist for mammals at this time.

ACKNOWLEDGEMENTS

We would like to thank Ulrich Güldener and Martin Münsterkötter from the MIPS yeast database group and Philip Wong for helpful comments. This work was funded by a grant from the German Federal Ministry of Education and Research (BMBF) within the BFAM framework (031U112C).

REFERENCES

Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the biomolecular interaction network database. *Nucleic Acids Res.*, **31**, 248–250.

Frishman,D., Mokrejs,M., Kosykh,D., Kastenmuller,G., Kolesov,G., Zubrzycki,I., Gruber,C., Geier,B., Kaps,A., Albermann,K. *et al*. (2003) The PEDANT genome database. *Nucleic Acids Res.*, **31**, 207–211.

Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

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.

Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al*. (2004) The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nature Biotechnol.*, **22**, 177–183.

Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al*. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.

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.

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.

Peri,S., Navarro,J.D., Amanchy,R., Kristiansen,T.Z., Jonnalagadda,C.K., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T.K., Gronborg,M. *et al*. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, **13**, 2363–2371.

Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–451.

Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al*. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett.*, **513**, 135–140.