Chapter 30
A Review on Protein-Protein Interaction Network Databases

Chandra Sekhar Pedamallu and Linet Ozdamar

30.1 Introduction

Cells are the structural and functional units of all known living organisms. These carry out numerous functions, from DNA replication, cell replication, protein synthesis, and energy production to molecule transport, to various inter- and intracellular signaling. Many of these fundamental processes require cascades of biochemical reactions that are catalyzed by possibly interacting protein enzymes. Other proteins provide structural support for the cells, form scaffolds for intracellular localization, or serve as chaperones or as transporters. The large-scale study of all cellular proteins is known as Proteomics [2,5]. One of the main goals of proteomics is to map the interactions of proteins. Interactomes, study of interaction networks, for dozens of model organisms have been established experimentally. Functions of proteins can be defined by their complex interactions and by their positions in interaction networks. Protein-protein interaction information plays a vital role in basic biological research; it also helps in the discovery of novel drug targets for the treatment of various chronic diseases. Experimental probing of protein-protein interactions requires labor-intensive techniques, such as co-immunoprecipitation, or affinity chromatography [26]. High-throughput experimental techniques, such as yeast two-hybrid [34] and mass spectrometry [11] are also available for large-scale detection of protein-protein interactions, and for the exploration of protein sequences, structures, and relationships in complete genomes [26].

C.S. Pedamallu (✉)
Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA
The Broad Institute of MIT and Harvard, Boston, MA, USA
e-mail: pcs.murali@gmail.com

L. Ozdamar
Department of Systems Engineering, Yeditepe University, Istanbul, Turkey
e-mail: linetozdamar@lycos.com

A.A. Pinto and D. Zilberman (eds.), *Modeling, Dynamics, Optimization and Bioeconomics I*, Springer Proceedings in Mathematics & Statistics 73,
DOI 10.1007/978-3-319-04849-9_30,

Following these advances, numerous computational methods have been developed to predict protein-protein interaction networks based on sequence or (and) structural features of the proteins. These computational approaches use phylogenetic profiling [21, 28], homologous interacting partner analysis [1], structural pattern comparisons [4, 15, 19], Bayesian network modeling [13], data mining techniques [20, 36] and so on.

There are several surveys on computational methods that predict protein-protein interaction networks (e.g., [23, 26, 31]). Complementing efforts have been made to centralize protein-protein interaction data through the construction of databases, such as STRING [29], MINT [7], BioGRID [27], FPPI [35], HAPPI [8], PIP [18], DIP [25], POINeT [16] and IntAct [3]. These databases can be classified as general databases and specialized databases. In this paper, we attempt to provide a summary of available databases.

30.2 Protein-Protein Interaction Databases

PPI databases can be grouped into two categories, (1) General databases that contain interactions networks from a wide variety of organisms; (2) Specialized databases that contain interaction networks from specific organisms. These databases are used during efforts of protein-protein interaction predictions.

30.2.1 General Databases

30.2.1.1 Search Tool for the Retrieval of Interacting Genes [29]

Search Tool for the Retrieval of Interacting Genes (STRING) is a comprehensive database that provides both experimental as well as predicted interaction information. Each of interactions in STRING are provided with a confidence score, and accessory information such as protein domains and 3D structures are made available, all within a stable and consistent identifier space. Other features that are included in STRING are interactive network viewer that can cluster networks on demand, updated on-screen previews of structural information including homology models, extensive data updates and strongly improved connectivity and integration with third-party resources. The current version of STRING covers more than 5,214,234 proteins from 1,133 organisms range from Bacteria, Archaea to Homo sapiens.

The resource can be reached at http://string-db.org.

30.2.1.2 Molecular INTeration Database [7]

Molecular INTeration database (MINT) is a public repository for molecular interactions reported in peer-reviewed journals. It mainly focuses on experimentally verified protein-protein interactions mined from the scientific literature by expert curators. The interactions curated and validated in MINT are automatically imported, according to its properties, by one or more sister databases that include human interactions database—HomoMINT (http://mint.bio.uniroma2.it/domino), domain-peptide interactions DOMINO (http://mint.bio.uniroma2.it/domino), all virus-virus and virus-host interactions databases, VirusMINT (http://mint.bio.uniroma2.it/virusmint).

The scoring function used for the MINT is based on size of experiment, type of experiment, evidence of direct interaction (i.e. two-hybrid) with respect to experimental support, number of interaction partners detected in single purification, sequence similarity of ortholog proteins (in case of human proteome in HomoMINT), and the number of publications supporting the interaction. The resulting score ranges between 0 to 1 (well supported evidence). This database contains interaction networks from Homo sapiens, C. elegans, Bacteria, and various Viruses.

The resources can be reached at http://mint.bio.uniroma2.it/mint/Welcome.do.

30.2.1.3 Biological General Repository for Interaction Datasets [27]

Biological General Repository for Interaction Datasets (BioGRID) is a public database that archives and disseminates genetic and protein interaction data from model organisms and humans (http://www.thebiogrid.org). It currently holds 347,966 interactions (170,162 genetic, 177,804 protein) curated from both high-throughput data sets and individual focused studies, as derived from over 23,000 publications in the primary literature. All interaction data are freely provided through the search index and available via download in a wide variety of standardized formats. This database contains interaction networks from Homo sapiens, C. elegans, Plant, Mouse, and different bacterial species.

The resources can be reached at http://thebiogrid.org/.

30.2.1.4 IntAct [3]

IntAct is an open data molecular interaction database abstracted from the literature or from direct data depositions by expert curators following a deep annotation model providing a high level of detail. It contains over 268,920 binary interactions, 57,741 proteins and 13,802 experiments. The search interface allows the user to iteratively develop complex queries, exploiting the detailed annotation with hierarchically controlled vocabularies. This database contains interaction information from a wide variety of organisms that includes but not limited to Homo sapiens, Mus musculus,

Drosophila melanogaster, Caenorhabditis elegans, Escherichia coli and Arabidosis thaliana.

The resources can be reached at http://www.ebi.ac.uk/intact/.

30.2.1.5 POINeT [16]

POINeT is an integrated web service that processes protein-protein interaction searching, analysis and visualization. It merges protein-protein interaction and tissue-specific expression data from multiple resources including DIP (http://dip.doe-mbi.ucla.edu/), MINT [7], BIND (http://www.bind.ca), HPRD (http://www.hprd.org), MIPS (http://mips.gsf.de/proj/ppi/), CYGD, BioGRID [27] and NCBI interaction (ftp://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/interactions.gz). The tissue-specific PPIs and the number of research papers supporting the PPIs can be filtered with user-adjustable threshold values and are dynamically updated in the viewer. The network constructed in POINeT can be readily analyzed with, for example, the built-in centrality calculation module and an integrated network viewer. Nodes in global networks can also be ranked and filtered using various network analysis formulas, i.e., centralities. To prioritize the sub-network, a ranking filtered method (S3) is developed to uncover potential novel mediators in the midbody network. This database contains interaction information from wide variety of organisms.

The resources can be reached at http://poinet.bioinformatics.tw/.

30.2.1.6 Reactome [9, 12]

Reactome is a database of pathways and reactions (pathway steps) in human biology that have been curated by expert biologist researchers which is extensively cross-referenced to other resources e.g. NCBI, Ensembl, UniProt, UCSC Genome Browser, HapMap, KEGG (Gene and Compound), ChEBI, PubMed and GO. It includes many events in biology that involve changes in state, such as binding, activation, translocation and degradation, in addition to classical biochemical reactions. Reactome contains inferred orthologous reactions for over 20 non-human species including mouse, rat, chicken, puffer fish, worm, fly, yeast, rice, Arabidopsis and E.coli.

The resources can be reached at http://www.reactome.org/ReactomeGWT/entrypoint.html.

30.2.1.7 iRefWeb [30]

iRefWeb provides a web interface to protein interaction data consolidated from ten public databases that includes BIND, BioGRID, CORUM, DIP, IntAct, HPRD, MINT, MPact, MPPI and OPHID. It provides an overview of the consolidated protein-protein interaction landscape and shows how it can be automatically cropped

to help generate meaningful organism-specific interactomes. iRefWeb presents aggregated interactions for a protein of interest, and various statistical summaries of the data across databases, such as the number of organism-specific interactions, proteins and cited publications.

The resources can be reached at http://wodaklab.org/iRefWeb and http://wodaklab.org/iRefWeb/.

30.2.1.8 Database of Interacting Proteins [25]

Database of interacting proteins (DIP) is a database that catalogs experimentally determined protein-protein. It combines information from a variety of sources to create a single, consistent set of protein-protein interactions. The data stored within the DIP database were curated, both, manually by expert curators and also automatically using computational approaches that utilize the knowledge about the protein-protein interaction networks extracted from the most reliable, core subset of the DIP data. DIP contains 23,201 proteins from 372 organisms including but not limited to Homo sapiens, Mouse, E. Coli, Rat, Bakers Yeast, and 71,276 interactions.

The resources can be reached at http://dip.doe-mbi.ucla.edu/dip/Main.cgi.

30.2.1.9 CORUM [24]

CORUM is a collection of experimentally verified mammalian protein complexes. All information presented in the database is obtained from individual experiments published in scientific data; however, data from high-throughput experiments are excluded. The majority of protein complexes in CORUM originate from human. The resources can be reached at http://mips.helmholtz-muenchen.de/genre/proj/corum/index.html.

30.2.2 Specialized Databases

30.2.2.1 Arabidopsis Thaliana Protein Interactome Database [22]

The AtPID (Arabidopsis thaliana Protein Interactome Database) represents protein-protein interaction networks, domain architecture, ortholog information and GO annotation in the Arabidopsis thaliana proteome. The protein-protein interaction pairs are predicted by integrating several methods with the Naive Bayesian Classifier. AtPID contains 28,062 putative PPIs.

The resources can be reached at http://www.megabionet.org/atpid/webfile/

30.2.2.2 Human Protein Reference Database [14]

Human Protein Reference Database (HPRD) is a database of curated proteomic information pertaining to human proteins. It integrates information pertaining to domain architecture, post-translational modifications, interaction networks and disease association for each protein in the human proteome. All the information in HPRD has been manually extracted from the literature by expert biologists. HPRD contains 39,194 protein-protein interactions from 30,047 protein entries.

The resources can be reached at http://www.hprd.org/.

30.2.2.3 Fusarium Graminearum Protein-Protein Interaction Database [35]

Fusarium graminearum protein-protein interaction (FPPI) database provides comprehensive information of protein-protein interactions (PPIs) of Fusarium graminearum. The PPIs are predicted based on both interologs from several PPI databases of seven species and domain-domain interactions experimentally determined based on protein structures. FPPI database contains 223,166 interactions among 7,406 proteins and 27,102 interactions among 3,745 proteins in the core PPI set.

The resources can be reached at http://csb.shu.edu.cn/fppi/.

30.2.2.4 Human Annotated and Predicted Protein Interaction Database [8]

The Human Annotated and Predicted Protein Interaction (HAPPI) database is developed to integrate publicly available human protein interaction data from BIND, OPHID, MINT, IntAct, HPRD, and STRING databases into a data warehouse. In the data warehouse, various types of sequence, structure, pathway, and literature annotation data from established bioinformatics resources such as NCBI, PubMed, UniProt, HUGO, EBI, PDB were also integrated. HAPPI contains 601,757 protein-protein interactions and associations from 70,829 curated human proteins.

The resources can be reached at http://discern.uits.iu.edu:8340/HAPPI/index.html.

30.2.2.5 Human Protein-Protein Interaction Prediction Database [18]

The Human protein-protein interaction prediction database (PIPs) is a resource for studying protein-protein interactions in human. It contains predictions of more than 37,000 high probability interactions of which more than 34,000 are not reported in the interaction databases HPRD, BIND, DIP or OPHID. The interactions in PIPs were calculated by a Bayesian method that combines information from expression, orthology, domain co-occurrence, post-translational modifications and sub-cellular

location. The predictions also take into account the topology of the predicted interaction network.

The resources can be reached at http://www.compbio.dundee.ac.uk/www-pips.

30.3 Brief Note on Application of Protein-Protein Interaction Networks

Study of protein-protein interaction networks is fundamental to understanding the key biological systems across different organisms. Protein-protein interaction networks can be used to protein function prediction (e.g. [17, 33]), drug discovery in cancer, autoimmune diseases, etc. [10, 32], understanding of successful plant defense mechanisms against pathogens [6] and etc.

30.4 Conclusion

Protein-protein interaction networks helps in understanding biological processes in living cells. There are several experimental and computational approaches developed to identify and predict these interaction networks experimentally and *in silico* respectively. Moreover, there are several complementing efforts made to centralize protein-protein interaction data through the construction of databases from experimental and computational protein-protein interactions networks. In this paper, we attempt to provide a summary of most widely used protein-protein interactions databases. These databases will serve as a platform for researchers to mine the data in a systematic fashion and employ them to predict the protein function, identifying important proteins in diseases and so on.

References

1. Aloy, P., Russell, R.B.: InterPreTS: protein interaction prediction through tertiary structure. Bioinformatics **19**(1), 161–162 (2003)
2. Anderson, N.L., Anderson, N.G.: Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis **19**(11), 1853–1861 (1998)
3. Aranda, B., Achuthan P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A.T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S.N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K., Hermjakob, H.: The IntAct molecular interaction database in 2010. Nucleic Acids Res. **38**(Database issue), D525–D531 (2010)
4. Aytuna, A.S., Keskin, O., Gursoy, A.: Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. Bioinformatics **21**(12), 2850–2855 (2005)

5. Blackstock, W.P., Weir, M.P.: Proteomics: quantitative and physical mapping of cellular proteins. Trends Biotechnol. **17**(3), 121–127 (1999)
6. Bogdanove, A.J.: Protein-protein interactions in pathogen recognition by plants. Plant Mol. Biol. **50**(6), 981–989 (2002)
7. Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., Cesareni, G.: MINT, the molecular interaction database: 2009 update. Nucleic Acids Res. **38**(Database issue), D532–D539 (2010)
8. Chen, J.Y., Mamidipalli, S., Huan, T.: HAPPI: an online database of comprehensive human annotated and predicted protein interactions. BMC Genomics **10**(Suppl 1), S16 (2009)
9. Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P., Stein, L.: Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. **39**(Database issue), D691–D697 (2011)
10. Drews, J.: Drug discovery: a historical perspective. Science **287**(5460), 1960–1964 (2000)
11. Figeys, D., McBroom, L.D., Moran, M.F.: Mass spectrometry for the study of protein-protein interactions. Methods **24**(3), 230–239 (2001)
12. Hermjakob, H., D'Eustachio, P., Stein, L.: Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. **39**(Database issue), D691–D697 (2011)
13. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M.: A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science **302**(5644), 449–453 (2003)
14. Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D.S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C.J., Kanth, S., Ahmed, M., Kashyap, M.K., Mohmood, R., Ramachandra, Y.L., Krishna, V., Rahiman, B.A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., Pandey, A.: Human protein reference database–2009 update. Nucleic Acids Res. **37**(Database issue), D767–D772 (2009)
15. Keskin, O., Ma, B., Nussinov, R.: Hot regions int protein-protein interactions: The organization and contribution of structurally conserved hot spot residues. J. Mol. Biol. **345** 1281–1294 (2004)
16. Lee, S.A., Chan, C.H., Chen, T.C., Yang, C.Y., Huangm K.C., Tsai, C.H., Lai, J.M., Wang, F.S., Kao, C.Y., Huang, C.Y.: POINeT: protein interactome with sub-network analysis and hub prioritization. BMC Bioinformatics **21**(10), 114 (2009)
17. Marcotte, E.M., Pellegrini, M., Ng H.,-L., Rice, D.W., Yeates, T.O., Eisenberg, D.: Detecting protein function and protein-protein interactions from genome sequences. Science **285**(5428), 751–753 (1999)
18. McDowall, M.D., Scott, M. S., Barton, G.J.: PIPs: human protein-protein interaction prediction database. Nucleic Acids Res. **37**(Database issue), D651–D656 (2009)
19. Ogmen, U., Keskin, O., Aytuna, A.S., Nussinov, R., Gursoy, A.: PRISM: protein interactions by structural matching. Nucleic Acids Res. **33** (Web Server issue), W331–W336 (2005)
20. Paradesi, M.S.R., Caragea, D., Hsu, W.H.: Incorporating graph features for predicting protein-protein interactions. In: Li, X.-L., Ng, S.-K. (eds.) Biological Data Mining in Protein Interaction Networks. IGI Publishers, USA (2008)
21. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O.: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA **96**, 4285–4288 (1999)
22. Peng, L., Weidong, Z., Yuhua, L., Feng, X., Jigang ,W., Tieliu, S.: AtPID: the overall hierarchical functional protein interaction network interface and analytic platform for Arabidopsis. Nucleic Acids Res. **39**(suppl 1), D1130–D1133 (2011)
23. Pitre, S., Alamgir, M., Green, J.R., Dumontier, M., Dehne, F., Golshani, A.: Computational methods for predicting protein-protein interactions. Adv. Biochem. Eng. Biotechnol. **110**, 247–267 (2008)

24. Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., Mewes, H.W.: CORUM: the comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res. **38**(Database issue), D497–D501 (2010)
25. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D.: The database of interacting proteins: 2004 update. Nucleic Acids Res. **32**(Database issue), D449–D451 (2004)
26. Skrabanek, L., Saini, H.K., Bader, G.D., Enright, A.J.: Computational prediction of protein-protein interactions. Mol. Biotechnol. **38**(1), 1–17 (2008)
27. Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J.M., Winter, A., Dolinski, K., Tyers, M.: The BioGRID interaction database: 2011 update. Nucleic Acids Res. **39**(Database issue), D698–D704 (2011)
28. Sun, J., Xu J., Liu, Z., Liu, Q., Zhao, A., Shi, T., Li, Y.: Refined phylogenetic profiles method for predicting protein-protein interactions. Bioinformatics **21**(16),3409–3415 (2005)
29. Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L.J., von Mering, C.: The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. **39**(Database issue), D561–D568 (2011)
30. Turner, B., Razick, S., Turinsky, A.L., Vlasblom, J., Crowdy, E.K., Cho, E., Morrison, K., Donaldson, I.M.,Wodak, S.J.: iRefWeb: Interactive analysis of consolidated protein interaction data and their supporting evidence. Database. 2010: baq023 (2010)
31. Valencia, A., Pazos, F.: Computational methods for the prediction of protein interactions. Curr. Opin. Struct. Biol. **12**(3), 368–373 (2002)
32. Valkov E., Sharpe T., Marsh M., Greive S., Hyvönen M.: Targeting protein-protein interactions and fragment-based drug discovery. Top Curr. Chem. **317**, 145–179 (2012)
33. Vazquez , A., Flammini, A., Maritan, A., Vespignani, A.: Global protein function prediction from protein-protein interaction networks. Nat. Biotechnol. **21**, 697–700 (2003)
34. Young, K.H.: Yeast two-hybrid: so many interactions, (in) so little time . . .. Biol. Reprod. **58**, 302–311 (1998)
35. Zhao, X.M., Zhang, X.W., Tang, W.H., Chen, L.: FPPI: fusarium graminearum protein-protein interaction database. J. Proteome Res. **8**(10), 4714–4721 (2009)
36. Zhou, D., He, Y.: Extracting interactions between proteins from the literature. J. Biomed. Inform. **14**(2), 393–407 (2008)