



RESEARCH ARTICLE

BIO INFORMATICS

EXPERIMENTAL AND COMPUTATIONAL TECHNIQUES IN PROTEIN INTERACTIONS

V. SIVASAKTHI AND SUDHA RAMAIAH*

Bioinformatics Division, School of Biosciences & Technology, VIT University, Vellore, India.



SUDHA RAMAIAH

Bioinformatics Division, School of Biosciences & Technology, VIT University, Vellore, India.

ABSTRACT

Proteins are biochemical compounds consisting of one or more polypeptides typically folded into a globular or fibrous form in a biologically functional way. Protein-protein interactions are intrinsic to virtually every cellular process. Interactions between proteins have been studied through a number of high-throughput experiments and have also been predicted through an array of computational methods that leverage the vast amount of sequence of data generated in the last decade. Here we review the important experimental and computational methods for the prediction of interactions and functional linkages between proteins. An overview of some of the databases and tools that are useful for a study of protein-protein interactions have also been discussed.



KEYWORDS

Protein interaction, Experimental methods, Yeast two hybrid and Computational methods.

INTRODUCTION

Proteins are the main catalysts, structural elements, signaling messengers and molecular machines of biological tissues¹. Proteins interact with each other within a cell, and those interactions give rise to the biological function and dynamical behavior of cellular systems. Generally, the protein interactions are temporal, spatial, or condition dependent in a specific cell, where only a small part of interactions usually take place under certain conditions. It is now becoming clear that protein interactions determine the outcome of most cellular processes²⁻⁵. Therefore, identifying and characterizing protein-protein interactions and their networks is essential for understanding the mechanisms of biological processes at molecular level. Various methods have been used to identify protein-protein interactions. These interactions may range from direct physical interactions inferred from experimental methods to functional linkages predicted on the basis of computational analyses. Experimental methods based on microarrays and yeast two-hybrid, as well as computational methods based on protein sequences and structures have been developed and widely used. To overcome the difficulties in experimentally identifying PPIs, a wide range of computational methods have been used to identify protein-protein functional linkages and interactions. These methods range from identifying a single pair of interacting proteins to the identification and analysis of a large network of thousands of proteins as that of an entire proteome of a given cell.

EXPERIMENTAL METHODS FOR IDENTIFYING PROTEIN INTERACTION

Traditionally, protein interactions have been studied individually by genetic, biochemical and biophysical techniques. However, the speed with which the new proteins are being discovered or predicted has created a need for high-throughput interaction-detection methods. Consequently, in the last two years, methods have been introduced that can globally tackle the problem, resulting in a vast amount of interaction data⁶. Protein-protein interactions fall into a few broad categories. Fragment complementation assays such as the yeast two-hybrid (Y2H) system are based on split proteins that are functionally reconstituted by fusions of interacting proteins. Biophysical methods include structure determination and mass spectrometric (MS) identification of proteins in complexes.

Apart from these techniques, protein array, Tandom affinity Purification (TAP) and synthetic lethality methods are also used for screening large number of proteins in a cell. Only the Y2H and protein complex purification combined with MS have been used on a larger scale.

(i) Yeast two-hybrid method:

Y2H is based on the fact that many eukaryotic transcription activators have at least two distinct domains, one that directs binding to a promoter DNA sequence (BD) called bait and another that activates transcription (AD) called prey. It was demonstrated that splitting BD and AD inactivates the transcription, but the transcription can be restored if a DNA-binding domain is physically associated with an activating domain⁷. According to the Y2H method, a protein of interest is fused to bait.

This protein is cloned in an expression plasmid, which is then transfected into a yeast cell. A similar procedure creates a chimeric sequence of another protein fused to prey. If two proteins physically interact, the reporter gene is activated. Fields and Song generate a novel genetic system to study these interactions by taking advantage of the properties of the GAL4 protein of the yeast *Saccharomyces cerevisiae*. The most broadly used Y2H systems are GAL4/LexA-based, where the GAL4 protein controls in yeast the expression of the LacZ gene encoding Beta-galactosidase. This GAL4 protein is a transcriptional activator required for the expression of genes encoding enzymes of galactose utilization¹³ and systems for screening the interactions between membrane proteins⁸⁻¹². Two-hybrid can be applied in high-throughput mode across the entire proteome of an organism to produce a comprehensive protein-protein interaction map^{14, 15}. In this Y2H method, two main approaches used to screening genome are matrix-based, a matrix of prey clones is created where each clone expresses a particular prey protein in one well of a plate and library-based, each bait is screened against an undefined prey library containing random cDNA fragments or open reading frames (ORFs).

(ii) MS Method:

MS is a powerful Mass Spectrometry method of studying macromolecular interactions in vitro. Mass spectrometric measurements are carried out in the gas phase on ionized analytes. A mass analyser in a mass spectrometer consists of an ion source, that measures the mass-to-charge ratio (m/z) of the ionized analytes, and a detector in MS registers the number of ions at each m/z value. Electrospray ionization (ESI)¹⁶ and matrix-assisted laser desorption/ionization (MALDI)¹⁷ are the two

techniques most commonly used to volatilize and ionize the proteins or peptides for mass spectrometric analysis.

MALDI and ESI have greatly advanced our ability to characterize large, thermally labile molecules by providing an efficient means of generating intact, gas-phase ions. These two techniques have been used to gain molecular weight information on biological samples with unprecedented speed, accuracy, and sensitivity.

Developments in instrumentation¹⁸ coupled with newer sampling methods have enabled higher levels of sensitivity, increased mass range, and better mass accuracy and promoted an increasing number of MS-based applications in the study of covalent and noncovalent protein structure. Both approaches offer unique and complementary capabilities. Electrospray ionization mass spectrometry (ESI-MS) has demonstrated utility for the detection and study of weakly bound, non covalent complexes, including protein interactions with inhibitors, cofactors, metal ions, carbohydrates, other peptides and proteins, enzyme-substrate pairings, and nucleic acid complexes¹⁹.

MALDI-MS is normally used to analyse relatively simple peptide mixtures. ESI and MALDI-MS commonly use quadrupole and time-of-flight (TOF) mass analyzers, respectively. ESI with quadrupole mass analyzers typically has accuracy on the order of 0.01%, and ESI with the quadrupole ion trap mass analysis offers the additional advantage of allowing collision-induced dissociation experiments to be performed without having multiple analyzers. As integrated liquid-chromatography ESI-MS systems (LC-MS) are preferred for the analysis of complex sample²⁰.

(iii) TAP method of complex purification:

Rigaut et al., 1999²¹ and Puig et al.²² invented Tandem Affinity Purification (TAP).

Several proteins are characterized²¹⁻²⁴ initially. This technique involves creating a fusion protein with a designed piece, the TAP tag on the end. The protein of interest with the TAP tag first binds to beads coated with IgG, the TAP tag is then broken apart by an enzyme, and finally a different part of the TAP tag binds reversibly to beads of a different type. After the protein of interest has been washed through two affinity columns, it can be examined for binding partners. Starting from a relatively small number of cells, active macromolecular complexes can be isolated and used for multiple applications.

Variations of the method to specifically purify complexes containing two given components or to subtract undesired complexes can easily be implemented. The TAP method is initially developed in yeast but can be successfully adapted to various organisms²⁵. Its simplicity, high yield, and wide applicability make the TAP method a very useful procedure for purification and proteome exploration^{25, 26, 15}. Several large-scale studies of protein complexes have been performed using TAP-MS and Y2H methods^{3, 27-29}.

COMPUTATIONAL METHODS FOR PROTEIN INTERACTION PREDICTION

An experimental approach studying protein interactomes have certain limitations that can be overcome by the computational methods for predicting protein interactions. Especially, the high-throughput methods are believed to contain many false positives, i.e. interactions that are identified by experiments but never take place in the cell³⁰⁻³³. These computational methods can be very useful for choosing potential targets for experimental screening or for validating experimental data and can provide information about interaction details which

may not be apparent from the experimental techniques.

(i) Domain fusion:

The Rosetta Stone or domain fusion method was proposed by Eisenberg and co-workers³⁴. The method is based on the hypothesis that if domains A and B exist as fused in a single polypeptide AB in another organism, then A and B are functionally linked. The Empirical information from structural data also used to refine and assess domain fusion based protein interaction predictions. These interactions can then be integrated with downstream biochemical and genetic assays to generate more reliable protein interaction data sets³⁵. Domain fusion analysis has shown itself to be well suited for predicting protein physical interactions in human, a complex eukaryotic proteome, for which it is inherently difficult to efficiently implement other *ab initio* methods to predict PPIs³⁶.

(ii) Phylogenetic profile methods:

The phylogenetic profile (PP) method is an *in silico* method based on the hypothesis that functionally linked and potentially interacting nonhomologous proteins co-evolve and have orthologs in the same subset of fully sequenced organisms³⁷⁻⁴². The phylogenetic profiles method is based on the assumption that there is a strong selective pressure on proteins that functionally interact with each other so that they are inherited together during speciation events³⁸. These phylogenetic methods predict many functional interactions between proteins and help to identify specific functions for numerous proteins. The refined phylogenetic profiles method⁴³ based on the selection of the reference organisms set and the criteria for homology identification shows greater performance and potentially provides

more reliable functional linkages compared to previous methods.

Phylogenetic profiles can also be identified for protein domains instead of entire proteins⁴⁴. A profile is constructed for each domain and the presence or absence of the domain in different genomes is recorded which in turn can give information about domain interactions. Some drawbacks of PP include its high computational cost, its dependence on high information profiles, and homology detection between distant organisms.

(iii) Methods based on co-evolution:

Co-evolution can be defined as the joint evolution of ecologically interacting species⁴⁵ and it implies the evolution of a species in response to selection imposed by another. Co-evolution requires the existence of mutual selective pressure on two or more species⁴⁶.

The protein feature most intuitively related to co-evolution is the similarity of the phylogenetic trees of interacting protein families. Recent studies that have quantified the relationship between tree similarities and protein interactions in large data sets⁴⁶ have demonstrated that such similarity is not anecdotal⁴⁷.

The significance of the similarity of the trees of two protein families is evaluated in the

context of the similarities to the trees of the rest of the proteome. Taking the complete co-evolutionary context into account, it improves the detection of interacting proteins. This procedure not only corrects the interdependence between the pairwise co-evolutions but it also corrects the other factors that influence tree similarity. It is clear that if a given protein interacts with many different partners, the changes in its amino-acid sequence will be a complex combination of the effects produced by the interactions with all these partners. In this sense, the full network of molecular interactions in a cell can be seen⁴⁶.

DATABASES AND TOOLS FOR ANALYSIS PPI:

Large variety of databases exist to give information of binary protein interactions and the higher order interactions in protein complexes. This vast amount of data is considered as a rich source of information, from which new biological insight can be gained.

In this section, we review some of the important databases that are used in PPIs, as well as some useful tools for analysis of PPIs. Protein interaction databases have also been reviewed⁴⁸. Some important databases and their resources are given in table 1.

Table 1

Databases	URL	Resources
HPRD	Human protein functions, PPIs, post-translational modifications, enzyme–substrate relationships and disease associations	http://www.hprd.org



DIP	Experimentally determined interactions between proteins	http://dip.doe-mbi.ucla.edu
MINT	Experimentally verified PPI mined from the scientific literature by expert curators	http://mint.bio.uniroma2.it/mint
BioGRID	Protein and genetic interactions from major model organism species	http://www.thebiogrid.org
BIND	Peer-reviewed bio-molecular interaction database containing published interactions and complexes	http://bind.ca
Predictome	Experimentally derived and computationally predicted functional linkages	http://visant.bu.edu
STRING	Protein functional linkages from experimental data and computational predictions	http://string.embl.de
ProLinks	Protein functional linkages	http://mysql5.mbi.ucla.edu/cgi-bin/functionator/pronav
IntAct	Interaction data abstracted from literature or from direct data depositions by expert curators	http://www.ebi.ac.uk/intact
COGs	Orthology data and	http://www.ncbi.nlm.nih.gov/COG

	phylogenetic profiles	
	physical interactions	
IBIS	observed in experimentally-determined structures for protein	http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi

(i) DIP:

The Database of Interacting Proteins (DIP) <http://dip.doe-mbi.ucla.edu> database⁴⁹ contains experimentally determined protein interactions and also includes a core subset of interactions that have passed a quality assessment⁵⁰. Interaction data are obtained from the literature, PDB and high-throughput experimental methods such as Y2H, DNA and protein microarrays and TAP-MS analysis of protein complexes. DIP is a member of the International Molecular Exchange Consortium (IMEx), a group of the major public providers of interaction data. The curation of data stored within DIP can be done manually or using computational approaches that utilize the knowledge about the protein-protein interaction networks extracted from the most reliable, core subset of the DIP data.

(ii) STRING:

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins <http://string.embl.de/>)^{51,52} contains information from numerous sources, including computational methods, experimental data, and public text collections. It is freely accessible and it is regularly updated. The latest version 8.3 contains information about 2.5 millions proteins from 630 species. The graphical user interface is appealing and user-friendly, backed by an excellent visualization engine.

Medusa (<http://coot.embl.de/medusa/>) a general graph visualisation tool is a front end to the STRING protein interaction database⁵².

(iii) HPRD:

Human Protein Reference Database (HPRD; <http://www.hprd.org/>)⁵³ contains information relevant to the function of human proteins in health and disease. All the information in HPRD has been manually extracted from the literature by expert biologists who read, interpret and analyze the published data. HPRD has been created using an object oriented database in Zope, an open source web application server that provides versatility in query functions and allows data to be displayed dynamically. HPRD data is available for download in tab delimited and XML file formats.

(iv) MINT:

A Molecular INTERaction database designed to store data on functional interactions between proteins. MINT consists of entries extracted from the scientific literature by expert curators assisted by 'MINT Assistant', software that targets abstracts containing interaction information and presents them to the curator in a user-friendly format. The interaction data can be easily extracted and viewed graphically through MINT Viewer⁵⁴. MINT also includes, as an integrated addition, HomoMINT, a database of



interactions between human proteins inferred from experiments with ortholog proteins in model organisms (<http://mint.bio.uniroma2.it/mint/>)⁵⁵.

CONCLUSION

Cellular function can only be understood by considering the individual properties of cellular components (proteins, genes, etc.) in the context of their complex relationships. Therefore, the study of these interactions and complexes is established in the 'post-genomic' era⁵⁶. Data on protein-protein interaction pose computational challenges to assess the data quality and organize data into a consistent, easily accessible database that is useful for further studies. These data give the

physiological properties of living cell determined from biological networks.

In this paper, we have reviewed widely used experimental and computational techniques for identifying and characterizing protein interactions. Each technique can provide a piece in the puzzle of mechanisms of protein recognition⁵⁷. The vast amount of sequence data have been generated for the predictions of interactions and functional associations. By integrating experimental methods for determining PPIs and computational methods for prediction, a lot of useful data on PPIs have been generated, including a number of high-quality databases. The global overview presented by interaction maps is no doubt useful, but the finer details of the interactions may be significantly important to make testable predictions about biological systems⁵⁸.

REFERENCES

1. Eisenberg D, et al, Protein function in the post-genomic era. *Nature*, 405(6788):823-826, (2000).
2. Giot L, et al, A protein interaction map of *Drosophila melanogaster*. *Science* 302(5651): 1727-1736, (2003).
3. Li S, et al, A map of the interactome network of the metazoan *C. elegans*. *Science* 303: 540-543, (2004).
4. Krogan NJ, et al., Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440: 637-643, (2006).
5. Gavin AC, et al, Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440: 631- 636, (2006).
6. Anna Panchenko and Teresa Przytycka, Protein-protein Interactions and Networks: Identification, Computer Analysis, and Prediction (Computational Biology) Publisher: Sr | 2008 | ISBN: 184800124X |.
7. Karthik Raman, Construction and analysis of protein-protein interaction networks. *Automated Experimentation*, 2:2, (2010).
8. Toby GG and Golemis EA, Using the yeast interaction trap and other two-hybrid-based approaches to study protein-protein interactions. *Methods* 24: 201-217, (2001).
9. Lee JW and Lee SK, Mammalian two-hybrid assay for detecting protein-protein interactions in vivo. *Methods Mol Biology*, 261: 327-336, (2004).
10. Walhout AJ and Vidal M, High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods* 24: 297-306, (2001).
11. Aronheim A, et al, Isolation of an AP-1 repressor by a novel method for detecting protein-protein interactions. *Mol Cell Biol* 17: 3094-3102, (1997).
12. Mohler WA and Blau HM, Gene expression and cell fusion analyzed by lacZ Complementation in mammalian cells.



- Proc Natl Acad Sci U S A 93: 12423–2427, (1996).
13. Fields et al, A novel genetic system to detect protein-protein interactions. *Nature* 340(6230):245-246, (1989).
 14. T. Ito et al., A comprehensive two-hybrid analysis to explore the yeast protein interactomes, *Proc Natl Acad Sci U S A* 98(8):4277-4278, (2001).
 15. Uetz, P, et al, A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403(6770): 623-627, (2000).
 16. Fenn JB, et al, Electro spray ionization—principles and practice. *Mass Spectrom Review* 9:37-70, (1990).
 17. Hillenkamp F, et al, Matrix–assisted laser desorption / ionization mass spectrometry of biopolymers. *Anal Chem* 63: 1193-1201, (1991).
 18. Siuzdak G, *Mass Spectrometry for Biotechnology*. San Diego: Academic Press, (1996).
 19. Joseph A. Loo, *Electrospray ionization mass spectrometry: A technology for studying noncovalent macromolecular complexes*. *International journal of mass spectrometry* 200: 175-186, (2000).
 20. Kriwacki RW, et al, Combined Use of Proteases and Mass Spectrometry in Structural Biology. *Journal of biomolecular technology* 5:5-15, (1998).
 21. B Rigaut G., et al. A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology* 17 (10): 1030–1032, (1999).
 22. Puig et al, The Tandem Affinity Purification (TAP) Method. A General Procedure of Protein Complex Purification, *Methods* 24 (3): 218–229, (2001).
 23. Caspary F, et al, Partial purification of the yeast U2 snRNP reveals a novel yeast pre-mRNA splicing factor required for pre-spliceosome assembly. *The EMBO Journal* 18 (12): 3463–3474, (1999).
 24. Bouveret E, et al, A Sm-like protein complex that participates in mRNA degradation, *The EMBO Journal* 19 (7):1661–1671, (2000).
 25. Fromont-Racine M., et al, Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.* 16(3):277-282, (1997).
 26. Ito T, et al, 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(3):1143-1147, (2000).
 27. Gavin AC, et al, Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440: 631–636, (2006).
 28. Gavin AC, et al, Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415: 141–147, (2002).
 29. Ho Y, et al, Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415: 180–183, (2002).
 30. Legrain P, et al, From the analysis of protein complexes to proteome-wide linkage maps. *Current Opinion in Biotechnology* 11: 402–407, (2000).
 31. Mrowka R., et al, Is there a bias in proteome research? *Genome Res.* 11:1971–1973, (2001).
 32. Von Mering C, et al, Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417:399–403, (2002).
 33. Deane CM, et al, Protein interactions: two methods for assessment of the reliability of high throughput observations. *Molecular and cellular Proteomics*, 1(5): 349–356, (2002).



34. Marcotte EM, et al, Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science*, 285(5428):751-753, (1999).
35. Jer-Ming Chia and Prasanna R Kolatkar, Implications for domain fusion protein-protein interactions based on structural information. *BMC Bioinformatics*, 5:161,(2004).
36. Ian Morilla, et al, Assessment of protein domain fusions in human protein interaction network prediction: Application to human kinetochore model. *New biotechnology*, 27(6):755-765, (2010).
37. Galperin MY and Koonin EV, Who's your neighbor? New computational approaches for functional genomics. *Nature Biotechnology* 18: 609–613, (2000).
38. Pellegrini M, et al, Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc Natl Acad Sci U S A* 96: 4285–4288, (1999).
39. Eisenberg D, et al, Protein function in the post-genomic era. *Nature* 405: 823–826, (2000).
40. Date SV, and Marcotte EM, Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnology*, 21: 1055–1062, (2003).
41. Koonin EV and Galperin MY, Sequence—Evolution—Function. *Computational approaches in comparative genomics*. Boston: Kluwer Academic Publishers. 488 p, (2002).
42. Snitkin ES, et al, Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC Bioinformatics* 7: 420, (2006).
43. Jingchun Sun , et al., Refined phylogenetic profiles method for predicting protein–protein interactions. *Oxford Journals* 21(16): 3409-3415,(2005).
44. Pagel P, et al, A domain interaction map based on phylogenetic profiling. *Molecular Biology* 344: 1331–1346, (2004).
45. Thompson J: *The Co evolutionary Process*, University of Chicago Press:387p,(1994).
46. Florencio Pazos and AlfonsoValencia, Protein co-evolution, co- adaptation and interactions. *The European Molecular Biology Organization* , 27(20): 2648–2655 ,(2008).
47. Chern-Sing Goh, et al, Co-evolution of Proteins with their Interaction Partners, *Molecular biology* 299:283-293,(2008).
48. Shoemaker BA and Panchenko AR Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Computational Biology*, 3(3):e42., (2007).
49. Xenarios I, et al, DIP: The Database of Interacting Proteins. *Nucleic Acids Research* 29:239-241, (2001).
50. Salwinski L, et al, The Database of Interacting Proteins. *Nucleic Acids Res* 32: D449–D451, (2004).
51. Von Mering C, et al, STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, 33(1):433-437, (2005).
52. Hooper SD and Bork P, Medusa: A simple tool for interaction graph analysis. *Bioinformatics*, 21(24):4432-4433, (2005).
53. Peri S et al, Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13(10):2363-2371, (2003).
54. Z. Andreas, et al, MINT: a Molecular INTetaction database, *FEBS Letters* 513 , pp. 135–140, (2002).
55. Chatr-aryamontri A and Ceol A, et al, MINT: the Molecular INTERaction database. *Nucleic acids research* 35(database issue):D572-4, (2007).



56. Walhout AJ and Vidal M, Protein interaction maps for model organisms. *Nature Reviews, Molecular cell biology* 2(1):55-62, (2001).
57. Panchenko AR and Shoemaker BA, Protein-protein interactions, Structure and systems approaches to analyze diverse genomic data. ISMB tutorial 2006 Available:
http://www.ncbi.nlm.nih.gov/CBBresearch/Panchenko/ismb_tutorial2006.ppt.
58. Bray D, Molecular networks: the top-down view. *Science* 301(5641):1864-1865, (2003).