CHAPTER 6.2

Protein–protein interactions: analysis and prediction

D. Frishman[1,2], M. Albrecht[3], H. Blankenburg[3], P. Bork[4,5], E. D. Harrington[4], H. Hermjakob[6], L. Juhl Jensen[4,7], D. A. Juan[8], T. Lengauer[3], P. Pagel[1], V. Schachter[9] and A. Valencia[8f]

[1]Department of Genome Oriented Bioinformatics, Technische Universität Munchen, Wissenschaftzentrum Weihenstephan, Freising, Germany
[2]Institute for Bioinformatics and Systems Biology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany
[3]Department of Computational Biology and Applied Algorithmics, Max-Planck-Institute for Informatics, Saarbrücken, Germany
[4]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany
[5]Max-Delbrück-Centre for Molecular Medicine, Berlin-Buch, Berlin, Germany
[6]European Molecular Biology Laboratory Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[7]Novo Nordisk Foundation Center for Protein Research, Panum Institute, Copenhagen, Denmark
[8]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Melchor Fernandez Almagro, Madrid
[9]Computational Systems Biology Group – Genoscope – CEA, Evry, France

1 Introduction

Proteins represent the tools and appliances of the cell – they assemble into larger structural elements, catalyze the biochemical reactions of metabolism, transmit signals, move cargo across membrane boundaries and carry out many other tasks. For most of these functions proteins cannot act in isolation but require close cooperation with other proteins to accomplish their task. Often, this collaborative action implies physical interaction of the proteins involved. Accordingly, experimental detection, *in silico* prediction and computational analysis of protein–protein interactions (PPI) have attracted great attention in the quest for discovering functional links among proteins and deciphering the complex networks of the cell.

Proteins do not simply clump together – binding between proteins is a highly specific event involving well defined binding sites. Several criteria can be used to further classify interactions (Nooren and Thornton 2003). Protein interactions are not mediated by covalent bonds and, from a chemical perspective, they are always

Corresponding author: Dmitrij Frishman, Department of Genome Oriented Bioinformatics, Technische Universität Munchen, Wissenschaftzentrum Weihenstephan, 85350 Freising, Germany (e-mail: d.frishman@wzw.tum.de)

reversible. Nevertheless, some PPI are so persistent to be considered irreversible (obligatory) for all practical purposes. Other interactions are subject to tight regulation and only occur under characteristic conditions. Depending on their functional role, some protein interactions remain stable for a long time (e.g. between proteins of the cytoskeleton) while others last only fractions of a second (e.g. binding of kinases to their targets). Protein complexes formed by physical binding are not restricted to so called binary interactions which involve exactly two proteins (dimer) but are often found to contain three (trimer), four (tetramer), or more peptide chains. Another distinction can be made based on the number of distinct proteins in a complex: homo-oligomers contain multiple copies of the same protein while hetero-oligomers consist of different protein species. Sophisticated "molecular machines" like the bacterial flagellum consist of a large number of different proteins linked by protein interactions.

2 Experimental methods

The focus of this chapter is on the computational methods for analyzing and predicting protein–protein interactions. Nevertheless, some basic knowledge about experimental techniques for detecting these interactions is highly useful for interpreting results, estimating potential biases, and judging the quality of the data we use in our work.

Many different types of methods have been developed but the vast majority of interactions in the literature and public databases come from only two classes of approaches: co-purification and two-hybrid methods. Co-purification methods (Rigaut et al. 1999) are carried out *in vitro* and involve three basic steps. First, the protein of interest is "captured" from a cell lysate – e.g. by attaching it to an immobile matrix. This may be done with specific antibodies, affinity tags, epitope tags along with a matching antibody, or by other means. Second, all other proteins in the solution are removed in a washing step in order to purify the captured protein. Under suitable conditions, protein–protein interactions are preserved. In the third step, any proteins still attached to the purified protein are detected by suitable methods (e.g. Western-blot or mass spectrometry). Hence, the interaction partners are co-purified, as the name of the method implies.

The two-hybrid technique (Fields and Song 1989) uses a very different approach – it exploits the fact that transcription factors such as Gal4 consist of two distinct functional domains. The DNA-binding domain (BD) recognizes the transcription factor (TF) binding site in the DNA and attaches the protein to it while the activation domain (AD) triggers transcription of the gene under the control of the factor. When expressed as separate protein chains, both domains remain fully functional: the BD still binds the DNA but lacks a way of triggering transcription. The AD could trigger transcription but has no means of binding to the DNA. For a two-hybrid test, two proteins X and Y are *fused* to these domains resulting in two hybrids: X-BD and Y-AD. If X binds to Y, the

resulting protein complex turns out to be a fully functional transcription factor. Accordingly, an interaction is revealed by detecting transcription of the *reporter gene* under the control of the TF. In contrast to co-purifications, the interaction is tested *in vivo* in the two-hybrid system (usually in yeast, but other systems exist).

The above description refers to small-scale experiments testing one pair of proteins at a time, but both approaches have successfully been extended to large-scale experiments testing thousands of pairs in very short time. While such high-throughput data is very valuable, especially for computational biology which often requires comprehensive input data, a word of caution is necessary. Even with the greatest care and a maximum of thoughtful controls, high-throughput data usually suffer from a certain degree of false-positive results as well as false-negatives compared to carefully performed and highly optimized individual experiments.

The ultimate source of information about protein interactions is provided by high-resolution three-dimensional structures of interaction complexes, such as the one shown in Fig. 1. Spatial architectures obtained by X-ray crystallography or NMR spectroscopy provide atomic-level detail of interaction interfaces and allow for mechanistic understanding of interaction processes and their functional implications. Additional kinetic, dynamic and structural aspects of protein interactions can be elucidated by electron and atomic force microscopy as well as by fluorescence resonance energy transfer.

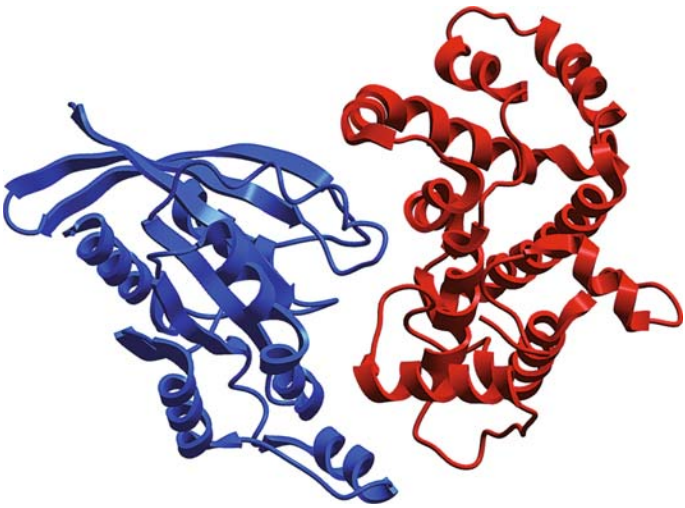


Fig. 1 Structural complex between RhoA, a small GTP protein belonging to the Ras superfamily, and the catalytic GTPase activating domain of RhoGAP (Graham et al. 2002)

3 Protein interaction databases

A huge number of protein–protein interactions has been experimentally determined and described in numerous scientific publications. Public protein interaction databases that provide interaction data in form of structured, machine-readable datasets organized according to well documented standards have become invaluable resources for bioinformatics, systems biology and researchers in experimental laboratories. The data in these databases generally originate from two major sources: large-scale datasets and manually curated information extracted from the scientific literature. As pointed out above, the latter is considered substantially more reliable and large bodies of manually curated PPI data are often used as the gold standard against which predictions and large-scale experiments are benchmarked. Of course, these reference data are far from complete and strongly biased. Many factors, including experimental bias, preferences of the scientific community, and perceived biomedical relevance influence the chance of an interaction to be studied, discovered and published. In the manual annotation process it is not enough to simply record the interaction as such. Additional information such as the type of experimental evidence, citations of the source, experimental conditions, and more need to be stored in order to convey a faithful picture of the data. Annotation is a highly labor intensive task carried out by specially trained database curators.

PPI databases can be roughly divided in two classes: specialized databases focusing on a single organism or a small set of species and general repositories which aim for a comprehensive representation of current knowledge. While the former are often well integrated with other information resources for the same organism, the latter strive for collecting all available interaction data including datasets from specialized resources. The size of these databases is growing constantly as more and more protein interactions are identified. As of writing (November 2007), global repositories are approaching 200,000 pieces of evidence for protein interactions in various species.

All of these databases offer convenient web interfaces that allow for interactively searching the database. In addition, the full datasets are usually provided for download in order to enable researchers to use the data in their own computational analyses. Table 1 gives an overview of some important PPI databases.

4 Data standards for molecular interactions

Until relatively recently, molecular interaction databases like the ones listed in Table 1 acted largely independently from each other. While they provided an extremely valuable service to the community in collecting and curating available molecular interaction data from the literature, they did so largely in an uncoordinated manner. Each database had its own curation policy, feature set, and data formats. In 2002, the Proteomics Standards Initiative (PSI), a work group of the Human Proteome Organization (HUPO), set out to

Table 1 A selection of protein–protein interaction databases

Name	Focus	URL	Reference
BioGrid	global	www.thebiogrid.org	(Stark et al. 2006)
BIND/BOND	global	bond.unleashedinformatics.com	(Bader et al. 2003)
DIP	global	dip.doe-mbi.ucla.edu	(Salwinski et al. 2004)
IntAct	global	www.ebi.ac.uk/intact/	(Kerrien et al. 2007a)
MINT	global	mint.bio.uniroma2.it	(Chatr-aryamontri et al. 2007)
HPRD	Human	www.hprd.org	(Mishra et al. 2006)
IM	*D. melanogaster, C. jejunii*	proteome.wayne.edu/PIMdb.html	(Pacifico et al. 2006)
MPact/MIPS	*S. cerevisiae*	mips.gsf.de/genre/proj/mpact/	(Guldener et al. 2006)
MPPI	Mammals	mips.gsf.de/proj/ppi/	(Pagel et al. 2005)

improve this situation, with contributions from a broad range of academic and commercial organizations, among them BIND, Cellzome, DIP, GlaxoSmithKline, Hybrigenics SA, IntAct, MINT, MIPS, Serono, and the Universities of Bielefeld, Bordeaux, and Cambridge. In a first step, a community standard for the representation of protein–protein interactions was developed, the PSI MI format 1.0 (Hermjakob et al. 2004). Recently, version 2.5 of the PSI MI format has been published (Kerrien et al. 2007b), extending the scope of the format from protein–protein interactions to molecular interactions in general, allowing to model for example protein-RNA complexes.

The PSI MI format is a flexible XML format representing the interaction data to a high level of detail. N-ary interactions (complexes) can be represented as well as experimental conditions and technologies, quantitative parameters and interacting domains. The XML format is accompanied by detailed controlled vocabularies in OBO format (Harris et al. 2004). These vocabularies are essential for standardizing not only the syntax, but also the semantics of the molecular interaction representation. As an example, the "yeast two-hybrid technology" described above is referred to in the literature using many different synonyms, for example Y2H, 2H, "yeast-two-hybrid", etc. While all of these terms refer to the same technology, filtering interaction data from multiple different databases based on this set of terms is not trivial. Thus, the PSI MI standard provides a set of now more than 1000 well-defined terms relevant to molecular interactions. Figure 2 shows the IntAct advanced search tool with a branch of the hierarchical PSI MI controlled vocabulary. Figure 3 provides a partial graphical representation of the annotated XML schema, combined with an example dataset in PSI MI XML format, reprinted from Kerrien et al. (2007b).

For user-friendly distribution of simplified PSI data to end users, the PSI MI 2.5 standard also defines a simple tabular representation (MITAB), derived from the BioGrid format (Breitkreutz et al. 2003). While this format necessarily excludes details

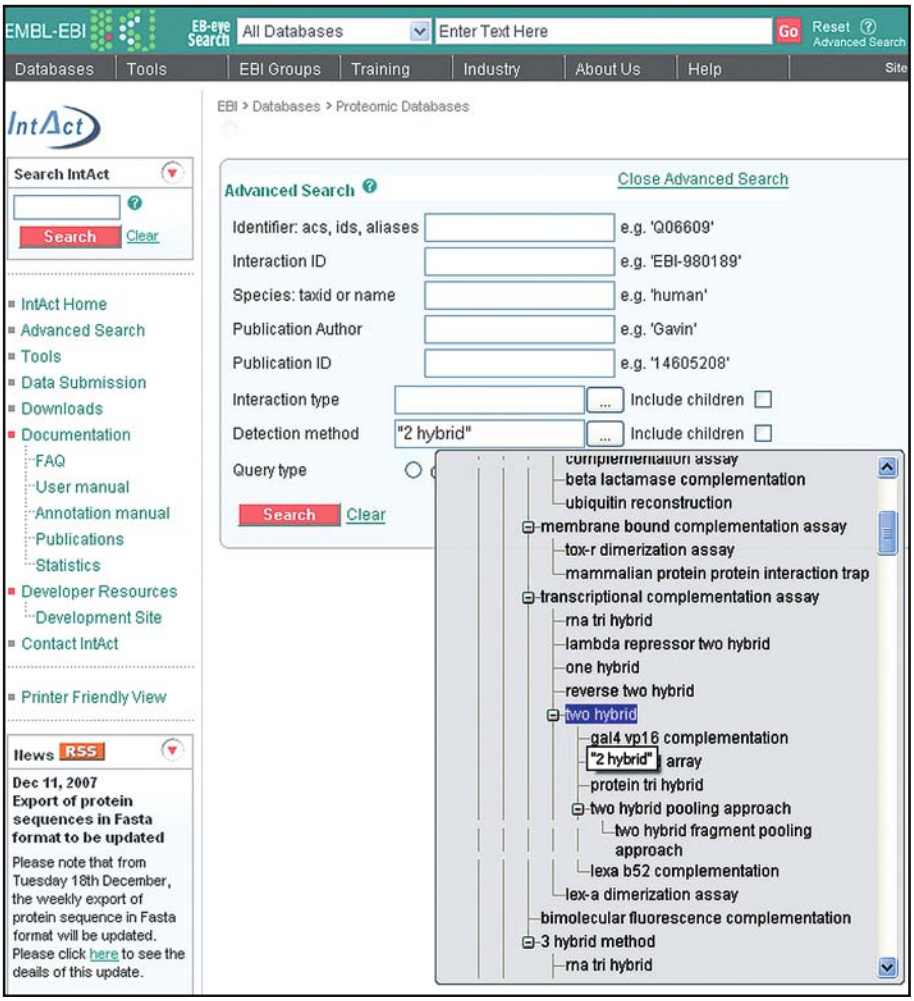


Fig. 2 IntAct advanced search

of interaction data like interacting domains, it provides a means to efficiently access large numbers of basic binary interaction records.

The PSI MI format is now widely implemented, with data available from BioGrid, DIP, HPRD, IntAct, MINT, and MIPS, among others. Visualization tools like Cytoscape (Shannon et al. 2003) can directly read and visualize PSI MI formatted data. Comparative and integrative analysis of interaction data from multiple sources has become easier, as has the development of analysis tools which do not need to provide a plethora of input parsers any more. The annotated PSI MI XML schema, a list of tools and

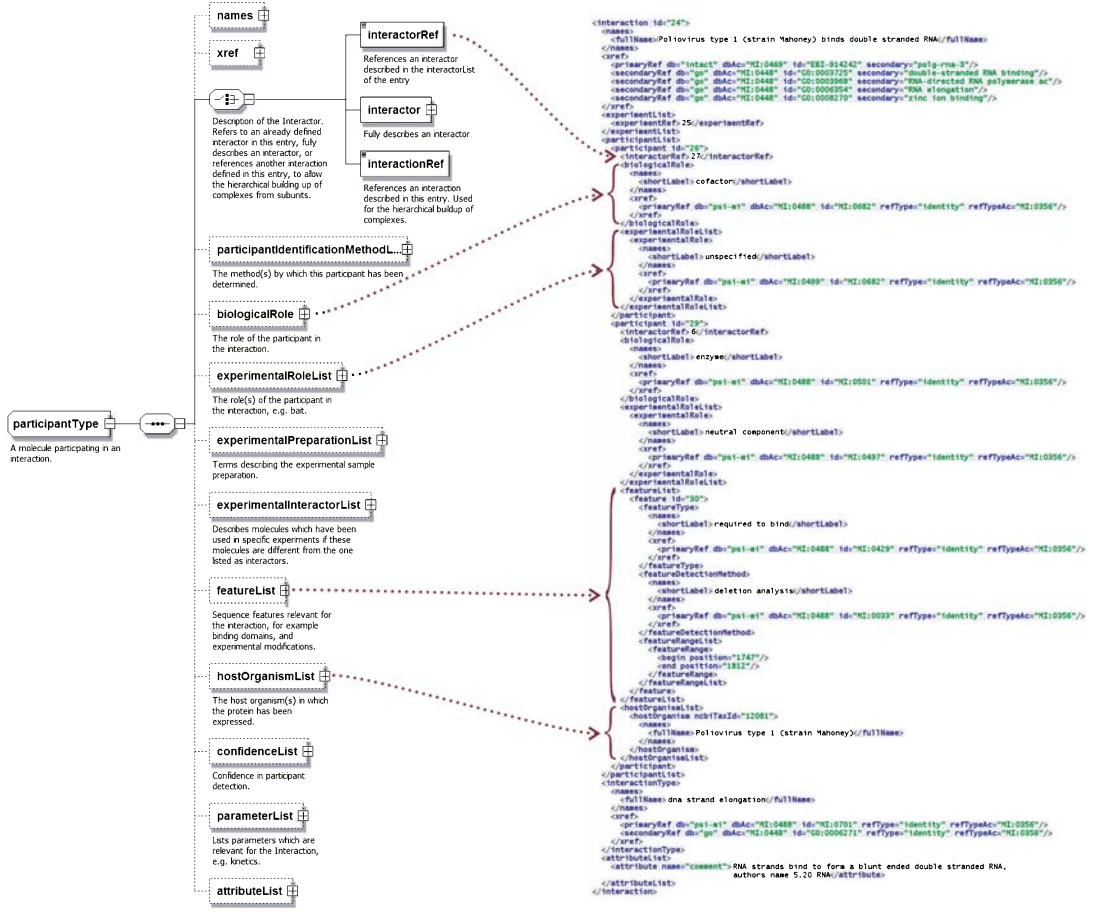


Fig. 3 Partial graphical representation of the annotated PSI MI XML schema, combined with an example dataset in PSI MI XML format (reprinted from Kerrien et al. (2007b))

databases implementing it, as well as further information, are available from http://www.psidev.info/.

However, the development and implementation of a common data format is only one step towards the provision of consistent molecular interaction data to the scientific community. Another key step is the coordination of the data curation process itself between different molecular interaction databases. Without such synchronization, independent databases will often work on the same publications and insert the data into their systems, according to different curation rules, thus doing redundant work on some publications, while neglecting others. Recognizing this issue, the DIP, IntAct, and MINT molecular interaction databases are currently synchronizing their curation efforts in the context of the IMEx consortium (http://imex.sf.net). These databases are now applying the same curation rules to provide a consistent high level of curation quality, and are synchronizing their fields of activity, each focusing on literature curation from a non-overlapping set of scientific journals. For these journals, the databases aim to insert all published interactions into the database shortly after publication. Regular data exchange of all newly curated data between IMEs databases is currently in the implementation phase.

To support the systematic representation and capture of relevant molecular interaction data supporting scientific publications, the HUPO Proteomics Standards Initiative has recently published "The minimum information required for reporting a molecular interaction experiment (MIMIx)" (Orchard et al. 2007b), detailing data items considered essential for the authors to provide, as well as a practical guide to efficient deposition of molecular interaction data in IMEx databases (Orchard et al. 2007a).

The IMEx databases are also collaborating with scientific journals and funding agencies, to increasingly recommend data producers to deposit their data in an IMEx partner database prior to publication. Database deposition prior to publication not only ensures public availability of the data at the time of publication, but also provides important quality control, as database curators often assess the data in much more detail than reviewers. The PSI journal collaboration efforts are starting to show first results. Nature Biotechnology, Nature Genetics, and Proteomics are now recommending that authors deposit molecular interaction data in a relevant public domain database prior to publication, a key step to a better capture of published molecular interaction data in public databases, and to overcome the current fragmentation of molecular interaction data.

5 The IntAct molecular interaction database

As an example of a molecular interaction database implementing the PSI MI 2.5 standard, we will provide a more detailed description of the IntAct molecular interaction database (Kerrien et al. 2007a), accessible at http://www.ebi.ac.uk/intact. IntAct

is a curated molecular interaction database active since 2002. IntAct follows a full text curation policy, publications are read in full by the curation team, and all molecular interactions contained in the publication are inserted into the database, containing basic facts like the database accession numbers of the proteins participating in an interaction, but also details like experimental protein modifications, which can have an impact on assessments of confidence in the presence or absence of interactions. Each database record is cross-checked by a senior curator for quality control. On release of the record, the corresponding author of the publication is automatically notified (where an email address is available), and requested to check the data provided. Any corrections are usually inserted into the next weekly release. While such a detailed, high quality approach is slow and limits coverage, the provision of high quality reference datasets is an essential service both for biological analysis, and for the training and validation of automatic methods for computational prediction of molecular interactions.

As it is impossible for any single database, or even the collaborating IMEx databases, to fully cover all published interactions, curation priorities have to be set. Any direct data depositions supporting manuscripts approaching peer review have highest priority. Next, for some journals (currently Cell, Cancer Cell, and Proteomics) IntAct curates all molecular interactions published in the journal. Finally, several special curation topics are determined in collaboration with external communities or collaborators, where IntAct provides specialized literature curation and collaborates in the analysis of experimental datasets, for example around a specific protein of interest (Camargo et al. 2006).

As of November 2007, IntAct contains 158.000 binary interactions supported by ca. 3,000 publications. The IntAct interface implements a standard "simple search" box, ideal for search by UniProt protein accession numbers, gene names, species, or PubMed identifiers. The advanced search tool (Fig. 2) provides field-specific searches as well a specialized search taking into account the hierarchical structure of controlled vocabularies. A default search for the interaction detection method "2 hybrid" returns 30,251 interactions, while a search for "2 hybrid" with the tickbox "include children" activated returns more than twice that number, 64,589 interactions. The hierarchical search automatically includes similarly named methods like "two hybrid pooling approach", but also "gal4 vp16 complement". Search results are initially shown in a tabular form based on the MITAB format, which can also be directly downloaded. Each pairwise interaction is only listed once, with all experimental evidence listed in the appropriate columns. The final column provides access to a detailed description of each interaction as well as a graphical representation of the interaction in is interaction neighborhood graph. For interactive, detailed analysis, interaction data can be loaded into tools like Cytoscape (see below) via the PSI 2.5 XML format.

All IntAct data is freely available via the web interface, for download in PSI MI tabular or XML format, and computationally accessible via web services. IntAct software is open source, implemented in Java, with Hibernate (www.hibernate.org/)

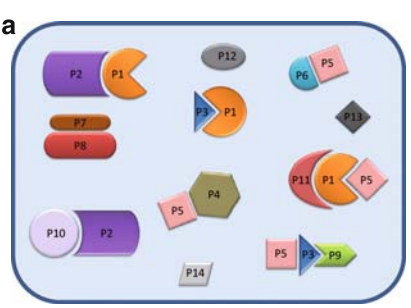
for the object-relational mapping to Oracle™ or Postgres, and freely available under the Apache License, version 2 from http://www.ebi.ac.uk/intact.

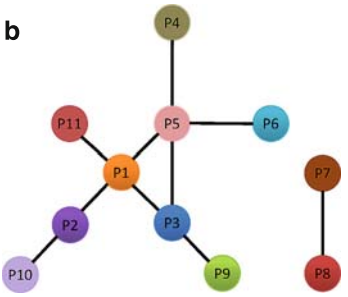
6 Interaction networks

On a global scale, protein–protein interactions participate in the formation of complex biological networks which, to a large extent, represent the paths of communication and metabolism of an organism. These networks can be modeled as graphs making them amenable to a large number of well established techniques of graph theory and social network analysis. Even though interaction networks do not directly encode cellular processes nor provide information on dynamics, they do represent a first step towards a description of cellular processes, which is ultimately dynamic in nature. For instance, protein-interaction networks may provide useful information on the dynamics of complex assembly or signaling. In general, investigating the topology of protein interaction, metabolic, signaling, and transcriptional networks allows researchers to reveal the fundamental principles of molecular organization of the cell and to interpret genome data in the context of large-scale experiments. Such analyses have become an integral part of the genome annotation process: annotating genomes today increasingly means annotating networks.

A protein–protein interaction network summarizes the existence of both stable and transient associations between proteins as an (undirected) graph: each protein is represented as a node (or vertex), an edge between two proteins denotes the existence of an interaction. Interactions known to occur in the actual cell (Fig. 4a) can thus be represented as an abstract graph of interaction capabilities (Fig. 4b). As such a graph is limited by definition to binary interactions, its construction from a database of molecular interactions may involve arbitrary choices. For instance, an n-ary interaction measured by co-purification can be represented using either the clique (all binary interactions between the n proteins are retained) or the spoke model (only edges connecting the "captured" protein to co-purified proteins are retained).

Once a network has been reconstructed from protein interaction data, a variety of statistics on network topology can be computed, such as the distribution of vertex degrees, the distribution of the clustering coefficient and other notions of density, the distribution of shortest path length between vertex pairs, or the distribution of network motifs occurrences (see (Barabasi and Oltvai 2004) for a review). These measures can be used to describe networks in a concise manner, to compare, group or contrast different networks, and to identify properties characteristic of a network or a class of network under study. Some topological properties may be interpreted as 'traces' of underlying biological mechanisms, shedding light on their dynamics, their evolution, or both and helping connect structure to function (see the "Network Modules" section below). For instance, most interaction networks seem to exhibit scale-free topology (Jeong et al.





c

Interactor 1	Interactor 2	Evidence	Reliability of the evidence	Validity
P1	P2	LIT	high	Correct
P1	P2	3D	high	Correct
P1	P2	LOC	low	Correct
P1	P11	Y2H	low	Correct
P1	P3	Y2H	low	Correct
P1	P5	3D	high	Correct
P1	P5	LOC	low	Correct
P3	P5	APMS	low	Correct
P3	P9	PROF	low	Correct
P3	P9	Y2H	low	Correct
P5	P6	Y2H	low	Correct
P6	P7	Y2H	low	Wrong
P7	P8	APMS	low	Correct
P7	P8	LIT	high	Correct
P8	P9	Y2H	low	Wrong

3D: Three-dimensional structure of the complex available
LIT: Known from literature
PROF: predicted by phylogenetic profiling
LOC: predicted by conserved chromosomal localization
Y2H: experimentally determined by large-scale yeast two-hybrid essay
APMS: experimentally determined by affinity purification/mass spectrometry

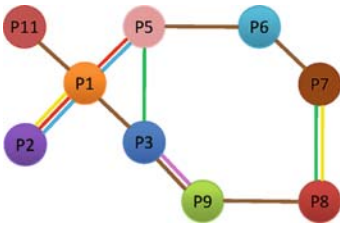


Fig. 4 Graph representation of interaction networks. (**a**) Hypothetical protein interactions in the living cell. Interacting proteins are denoted as P1, P2, etc. (**b**) A graph representation of the protein interactions shown in **a**. Each node represents a protein, and each edge connects proteins that interact. (**c**) Information on protein interactions obtained by different methods. (**d**) Protein interaction network derived from experimental evidence shown in **c**. As in **a**, each node is a protein, and edges connect interactors. Edges a colored according to the source of evidence: red – 3D, green – APMS, brown – Y2H, magenta – PROF, yellow – LIT, blue – LOC

2001; Yook et al. 2004), i.e. their degree distribution (the probability that a node has exactly k links) approximates a power law $P(k) \sim k^{-\gamma}$, meaning that most proteins have few interaction partners but some, the so-called "hubs", have many.

As an example of derived evolutionary insight, it is easy to show that networks evolving by growth (addition of new nodes) and preferential attachment (new nodes are more likely to be connected to nodes with more connections) will exhibit scale-free topology (degree distribution approximates a power-law) and hubs (highly connected nodes). A simple model of interaction network evolution by gene duplication, where a duplicate initially keeps the same interaction partners as the original, generates preferential attachment, thus providing a candidate explanation for the scale-free nature and the existence of hubs in these networks (Barabasi and Oltvai 2004).

A corresponding functional interpretation of hubs and scale-free topology has been proposed in terms of robustness. Scale-free networks are robust to component failure, as random failures are likely to affect low degree nodes and only failures affecting hub nodes will significantly change the number of connected components and the length of shortest paths between node pairs. Deletion analyses have, perhaps unsurprisingly, confirmed that highly connected proteins are more likely to be essential (Winzeler et al. 1999; Giaever et al. 2002; Gerdes et al. 2003).

Most biological interpretations that have been proposed for purely topological properties of interaction networks have been the subject of heated controversies, some of which remain unsolved to this day (e.g. (He and Zhang 2006; Yu et al. 2007) on hubs). One often cited objection to any strong interpretation is the fact that networks reconstructed from high-throughput interaction data constitute very rough approximations of the "real" network of interactions taking place within the cell. As illustrated in Fig. 4c, interaction data used in a reconstruction typically result from several experimental methods, often complemented with prediction schemes. Each specific method can miss real interactions (false negatives) and incorrectly identify other interactions (false positives), resulting in biases that are clearly technology-dependent (Gavin et al. 2006; Legrain and Selig 2000). Assessing false-negative and false-positive rates is difficult since there is no 'gold standard' for positive interactions (protein pairs that are known to interact) or, more importantly, for negative interactions (protein pairs that are known not to interact). Using less-than-ideal benchmark interaction sets, estimates of 30-60% false positives and 40-80% false negatives have been proposed for yeast-two-hybrid and co-purification based techniques (Aloy and Russell 2004). In particular, a comparison of several high-throughput interaction datasets on yeast, showing low overlap, has confirmed that each study covers only a small percentage of the underlying interaction network (von Mering et al. 2002) (see also "Estimates of the number of protein interactions" below).

Integration of interaction data from heterogeneous sources towards interaction network reconstruction can help compensate for these limitations. The basic principle is fairly simple and rests implicitly on a multigraph representation: several interaction networks to be integrated, each resulting from a specific experimental or predictive method, are defined over the same set of proteins. Integration is achieved by merging them into a single network with several types of links – or edge colors– each drawn from one of the component networks. Some edges in the multigraph may be incorrect, while some existing interactions may be missing from the multigraph, but interactions confirmed independently by several methods can be considered reliable. Figure 4d shows the multigraph that corresponds to the evidence from Fig. 4c and can be used to reconstruct the actual graph in Fig. 4b.

In practice, integration is not always straightforward: networks are usually defined over subsets of the entire gene or protein complement of a species, and meaningful integration requires that the overlap of these subsets be sufficiently large.

In addition, if differences of reliability between network types are to be taken into account, an integrated reliability scoring scheme needs to be designed (Jansen et al. 2003; von Mering et al. 2007) with the corresponding pitfalls and level of arbitrariness involved in comparing apples and oranges. Existing methods can significantly reduce false positive rates on a subset of the network, yielding a subnetwork of high-reliability interactions.

7 Visualization software for molecular networks

The tremendous amounts of available molecular interaction data raise the important issue of how to visualize them in a biologically meaningful way. A variety of tools have been developed to address this problem; two prominent examples are VisANT (Hu et al. 2005) and Cytoscape (Shannon et al. 2003). A recent review of further network visualization tools is provided by Suderman and Hallett (2007). In this section, we focus on Cytoscape (http://www.cytoscape.org) and demonstrate its use for the investigation of protein–protein interaction networks. For a more extensive protocol on the usage of Cytoscape, see (Cline et al. 2007).

Cytoscape is a stand-alone Java application that is available for all major computer platforms. This software provides functionalities for (i) generating biological networks, either manually or by importing interaction data from various sources, (ii) filtering interactions, (iii) displaying networks using graph layout algorithms, (iv) integrating and displaying additional information like gene expression data, and (v) performing analyses on networks, for instance, by calculating topological network properties or by identifying functional modules.

One advantage of Cytoscape over alternative visualization software applications is that Cytoscape is released under the open-source Lesser General Public License (LGPL). This license basically permits all forms of software usage and thus helps to build a large user and developer community. Third-party Java developers can easily enhance the functionality of Cytoscape by implementing own plug-ins, which are additional software modules that can be readily integrated into the Cytoscape platform. Currently, there are more than forty plug-ins publicly available, with functionalities ranging from interaction retrieval and integration across topological network analysis, detection of network motifs, protein complexes, and domain interactions, to visualization of subcellular protein localization and bipartite networks. A selection of popular Cytoscape plug-ins is listed in Table 2. In the following, we will describe the functionalities of Cytoscape in greater detail.

The initial step of generating a network can be accomplished in different ways. First, the user can import interaction data that are stored in various flat file or XML formats such as BioPax, SBML, or PSI-MI, as described above. Second, the user can directly retrieve interactions from several public repositories from within Cytoscape. A number

Table 2 Brief descriptions of popular Cytoscape plug-ins with web links to their project sites

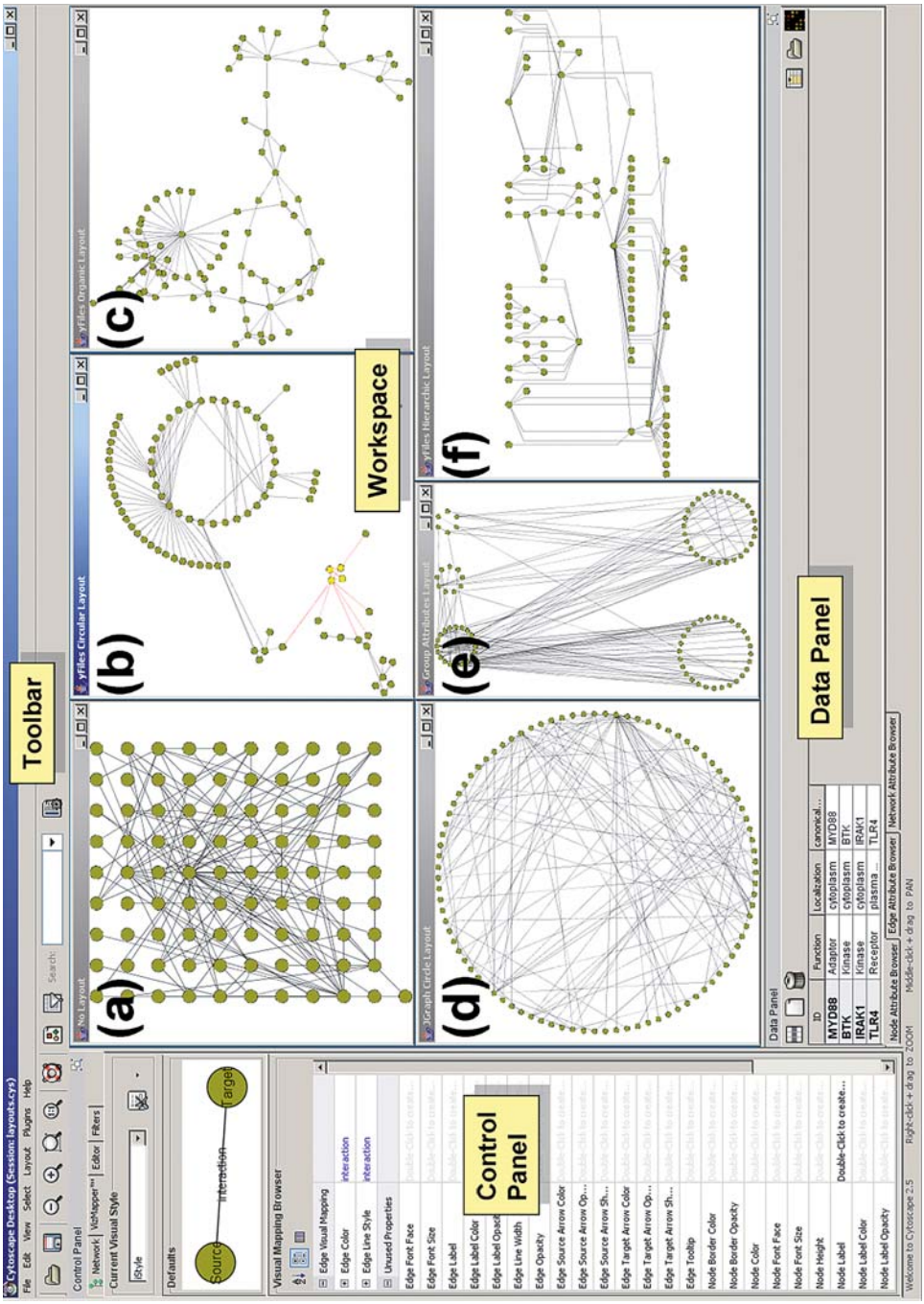
Plug-in	Description	Project web site
Agilent Literature Search	Network generation based on text-mining of scientific publications	http://cytoscape.org/plugins/
APID2NET	Network generation and analysis based on the Agile Protein Interaction DataAnalyzer (APID)	http://bioinfow.dep.usal.es/apid/apid2net.html
BiLayout	Generation of bipartite network layouts	http://bilayout.bioinf.mpi-inf.mpg.de/
BiNGO	Determination of overrepresented Gene Ontology (GO) terms	http://www.psb.ugent.be/cbd/papers/bingo/
BiNoM	Manipulation of networks represented in standardized formats like SBML and BioPAX	http://bioinfo-out.curie.fr/projects/binom/
BubbleRouter	Incremental layout generation based on various attributes	http://www.genmapp.org/BubbleRouter/manual.htm
CABIN	Exploratory analysis and integration of multiple interaction networks	http://www.sysbio.org/capabilities/compbio/cabin.stm
Cerebral	Layout generation based on subcellular protein localizations	http://www.pathogenomics.ca/cerebral/
DomainGraph	Decomposition of protein networks into domain-domain interaction networks	http://domaingraph.bioinf.mpi-inf.mpg.de
Enhanced Search	Sophisticated search functionality within a network	http://conklinwolf.ucsf.edu/genmappwiki/Google_Summer_of_Code_2007/Maital
GenePro	Analysis of functional modules and clusters	http://genepro.ccb.sickkids.ca/
GOlorize	Network visualization based on Gene Ontology (GO) categories (only in combination with BiNGO plug-in)	http://www.pasteur.fr/recherche/unites/Biolsys/GOlorize/
GroupTool	Combination of nodes and edges into groups	http://www.rbvi.ucsf.edu/Research/cytoscape/
jActiveModules	Determination of expression activated subnetworks and modules	http://cytoscape.org/plugins/
MCODE	Determination of highly connected clusters and putative complexes	http://baderlab.org/Software/mcode
MetaNode-Plugin2	Abstraction of nodes into meta nodes that can be expanded or collapsed	http://www.rbvi.ucsf.edu/Research/cytoscape/
MiMIplugin	Network generation based on the Michigan Molecular Interaction Database (MiMI)	http://mimi.ncibi.org/cytoscape/
MiSink	Network generation based on the Database of Interacting Proteins (DIP)	http://dip.doe-mbi.ucla.edu/dip/Software.cgi
NamedSelection	Temporary storage of node and edge selections	http://www.rbvi.ucsf.edu/Research/cytoscape/
NetworkAnalyzer	Computation of topological network parameters	http://med.bioinf.mpi-inf.mpg.de/networkanalyzer/
StructureViz	Linkage to macromolecular structures and sequences provided by UCSF Chimera	http://www.cgi.ucsf.edu/Research/cytoscape/structureViz/

of plug-ins exists that facilitate querying certain databases for interactions related to specific genes/proteins or species (APID2NET, MiMIplugin, MiSink; Table 2). Third, the user can utilize a text-mining plug-in that builds networks based on associations found in publication abstracts (Agilent Literature Search; Table 2). While these associations are not as reliable as experimentally derived interactions, they can be helpful when the user is investigating species that are not well covered yet in the current data repositories. Fourth, the user can directly create or manipulate a network by manually adding or removing nodes (genes, proteins, domains, etc.) and edges (interactions or relationships). In this way, expert knowledge that is not captured in the available data sets can be incorporated into the loaded network.

Generated networks can be further refined by applying selections and filters in Cytoscape. The user can select nodes or edges by simply clicking on them or framing a selection area. In addition, starting with at least one selected node, the user can incrementally enlarge the selection to include all direct neighbor nodes. Cytoscape also provides even sophisticated search and filter functionality for selecting particular nodes and edges in a network based on different properties; in particular, the Enhanced Search plug-in (Table 2) improves the built-in search functionality of Cytoscape. Filters select all network parts that match certain criteria, for instance, all human proteins or all interactions that have been detected using the yeast two-hybrid system. Once a selection has been made, all selected parts can be removed from the network or added to another network.

The main purpose of visualization tools like Cytoscape is the presentation of biological networks in an appropriate manner. This can usually be accomplished by applying graph layout algorithms. Sophisticated layouts can assist the user in revealing specific network characteristics such as hub proteins or functionally related protein clusters. Cytoscape offers various layout algorithms, which can be categorized as circular, hierarchical, spring-embedded (or force-directed), and attribute-based layouts (Fig. 5). Further layouts can be included using the Cytoscape plug-in architecture, for example, to arrange protein nodes according to their subcellular localization or to their pathways assignments (BubbleRouter, Cerebral; Table 2).

Some layouts may be more effective than others for representing molecular networks of a certain type. The spring-embedded layout, for instance, has the effect of exposing the inherent network structure, thus identifying hub proteins and clusters of tightly connected nodes. It is noteworthy that current network visualization techniques have limitations, for example, when displaying extremely large or dense networks. In such cases, a simple graphical network representation with one node for each interaction partner, as it is initially created by Cytoscape, can obfuscate the actual network organization due to the sheer number of nodes and edges. One potential solution to this problem is the introduction of meta-nodes (MetaNode plug-in; Table 2). A meta-node combines and replaces a group of other nodes. Meta-nodes can be collapsed to increase clarity of the visualization and expanded to increase the level of detail (Fig. 6).


Toolbar
Workspace
Control Panel
Data Panel
(a)
(b)
(c)
(d)
(e)
(f)

An overview of established and novel visualization techniques for biological networks on different scales is presented in (Hu et al. 2007).

All layouts generated by Cytoscape are zoomable, enabling the user to increase or decrease the magnification, and they can be further customized by aligning, scaling, or rotating selected network parts. Additionally, the user can define the graphical network representation through visual styles. These styles define the colors, sizes, and shapes of all network parts.

A powerful feature of Cytoscape is its ability of visually mapping additional attribute values onto network representations. Both nodes and edges can have arbitrary attributes, for example, protein function names, the number of interactions (node degree), expression values, the strength and type of an interaction, or confidence values for interaction reliability. These attributes can be used to adapt the network illustration by dynamically changing the visual styles of individual network parts (Fig. 7). For example, this feature enables highlighting trustworthy interactions by assigning

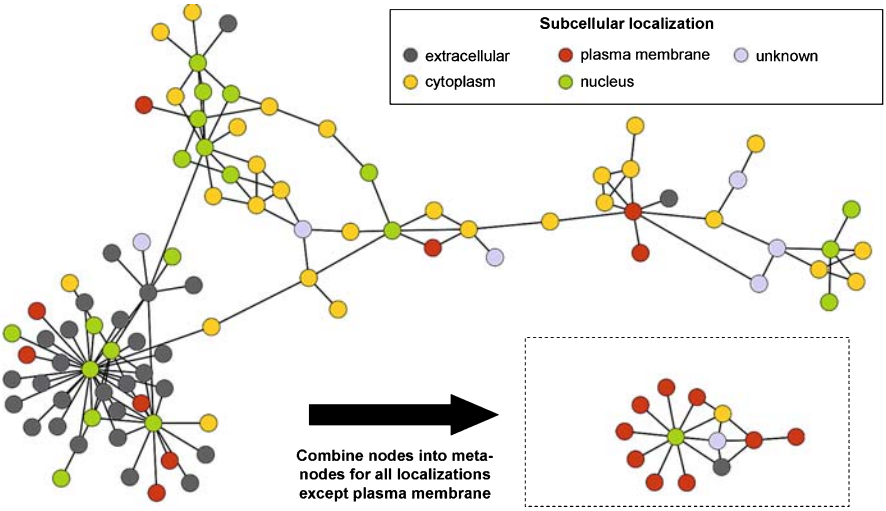


Fig. 6 Combination of nodes into meta-nodes using the Cytoscape plug-in *MetaNode* (Table 2). All protein nodes with subcellular localizations different from plasma membrane are combined into meta-nodes. These meta-nodes can be collapsed or expanded to increase clarity or detailedness, respectively

Fig. 5 The Cytoscape desktop. The workspace (middle) shows six identical networks with different layouts. The toolbar (top) contains basic control buttons for zooming and filtering/searching. The Control Panel (left) displays the VizMapper that defines the graphical network representation. The Data Panel (bottom) lists node attributes of the four selected nodes (yellow) in network (**b**). The different network layouts are: (**a**) grid, (**b**) circular with several circles, (**c**) spring-embedded or force-directed, (**d**) circular with one circle, (**e**) attribute-based, (**f**) hierarchical

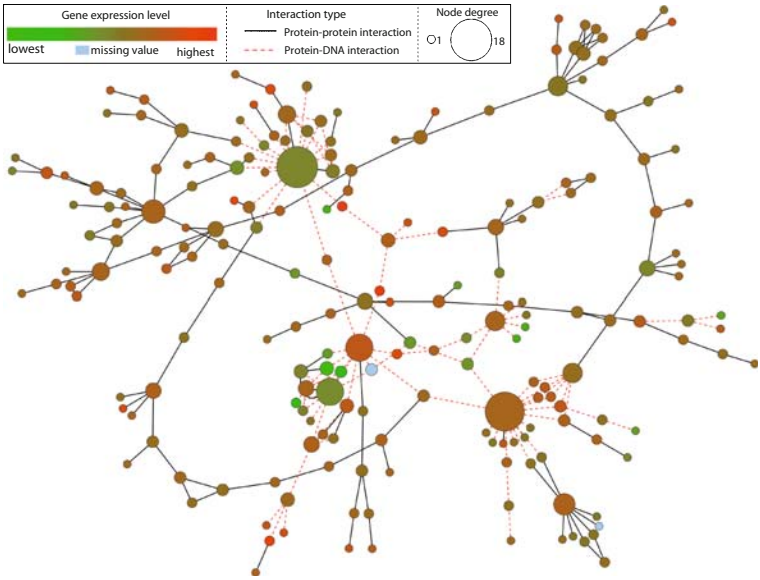


Fig. 7 Visual representation of a subset of the GAL4 network in yeast. The protein nodes are colored with a red-to-green gradient according to their expression value; green represents the lowest, red the highest value, and blue a missing value. The node size indicates the number of interactions (node degree); the larger a node, the higher is its degree. The colors and styles of the edges represent different interaction types; solid black lines represent protein-protein, dashed red lines protein-DNA interactions

different line styles or sizes to different experiment types (discrete mapping of an edge attribute), to spot network hubs by changing the size of a node according to its degree (discrete or continuous mapping of a node attribute), or to identify functional network patterns by coloring protein nodes with a color gradient according to their expression level (continuous mapping of a node attribute). Hence, it is possible to simultaneously visualize different data types by overlaying them with a network model.

In order to generate new biological hypotheses and to gain insights into molecular mechanisms, it is important to identify relevant network characteristics and patterns. For this purpose, the straightforward approach is the visual exploration of the network. Table 2 lists a selection of Cytoscape plug-ins that assist the user in this analysis task, for instance, by identifying putative complexes (MCODE), by grouping proteins that show a similar expression profile (jActiveModules), or by identifying overrepresented GO terms (BiNGO, GOlorize). However, the inclusion of complex data such as time-series results or diverse Gene Ontology (GO) terms into the network visualization might not be feasible without further software support. Particularly in case of huge, highly connected, or dynamic networks, more advanced visualization techniques will be required in the future.

In addition to the visual presentation of interaction networks, Cytoscape can also be used to perform statistical analyses. For instance, the NetworkAnalyzer plug-in (Assenov et al. 2008) computes a large variety of topology parameters for all types of networks. The computed simple and complex topology parameters are represented as single values and distributions, respectively. Examples of simple parameters are the number of nodes and edges, the average number of neighbors, the network diameter and radius, the clustering coefficient, and the characteristic path length. Complex parameters are distributions of node degrees, neighborhood connectivities, average clustering coefficients, and shortest path lengths. These computed statistical results can be exported in textual or graphical form and are additionally stored as node attributes. The user can then apply the calculated attributes to select certain network parts or to map them onto the visual representation of the analyzed network as described above (Fig. 7). It is also possible to fit a power law to the node degree distribution, which can frequently indicate a so-called scale-free network with few highly connected nodes (hubs) and many other nodes with a small number of interactions. Scale-free networks are especially robust against failures of randomly selected nodes, but quite vulnerable to defects of hubs (Albert 2005).

8 Estimates of the number of protein interactions

How many PPIs exist in a living cell? The yeast genome encodes approximately 6300 gene products which means that the maximal possible number of interacting protein pairs in this organism is close to 40 million, but what part of these potential interactions are actually realized in nature? For a given experimental method, such as the two-hybrid essay, the estimate of the total number of interactions in the cell is given by

$$N_{int} = N_{measured} \times R_{fp} \times R_{fn}{}^{-1}$$

where $N_{measured}$ is the number of interactions identified in the experiment, and R_{fp} and R_{fn} are false positive and false negative rates of the method. R_{fn} can be roughly estimated based on the number of interactions known with confidence (e.g., those confirmed by three-dimensional structures) that are being recovered by the method. Assessing R_{fp} is much more difficult because no experimental information on proteins that *do not* interact is currently available. Since it is known that proteins belonging to the same functional class often interact, one very indirect way of calculating R_{fn} is as the fraction of functionally related proteins not found to be interacting.

An even more monumental problem is the estimation of the total number of unique structurally equivalent interaction *types* existing in nature. An interaction type is defined as a particular mutual orientation of two specific interacting domains. In some cases homologous proteins interact in a significantly different fashion while in other cases proteins lacking sequence similarity engage in interactions of the same type.

In general, however, interacting protein pairs sharing a high degree of sequence similarity (30–40% or higher) between their respective components almost always form structurally similar complexes (Aloy et al. 2003). This observation allows utilization of available atomic resolution structures of complexes for building useful models of closely related binary complexes.

The total number of interaction *types* can then be estimated as follows:

$$N_{types} = N_{measured} \times R_{fp} \times R_{fn}^{-1} \times C \times E_{All\text{-}species}$$

where the interaction similarity multiplier C reflects the clustering of all interactions of the same type, and $E_{All\text{-}species}$ extrapolates from one biological species to all organisms. Aloy and Russel (2004) derived an estimate for C by grouping interactions between proteins that share high sequence similarity, as discussed above. C depends on the number of paralogous sequences encoded in a given genome. For small prokaryotic organisms it is close to 1 while for larger and more redundant genomes it adopts smaller values, typically in the range of 0.75–0.85. The multiplier for all species $E_{All\text{-}species}$ can be derived by assessing what fraction of known protein families is encoded in a given genome. Based on the currently available data this factor is close to 10 for bacteria, which means that a medium size prokaryotic organism contains around one tenth of all protein families. For eukaryotic organisms $E_{All\text{-}species}$ lies between 2 and 4. For the comprehensive two-hybrid screen of yeast by (Uetz 2000) in which 936 interactions between 987 proteins were identified, Aloy and Russell (2004) estimated C, $R_{fp,}$ and R_{fn}^{-1}, and $E_{All\text{-}species}$ to be 0.85, 3.92, 0.55, and 3.35 respectively, leading to an estimated 1715 different interaction types in yeast alone, and 5741 over all species. Based on the two-hybrid interaction map of the fly (Giot 2003) the number of all interaction types in nature is estimated to be 9962. It is thus reasonable to expect the total number of interaction types to be around 10,000, and only 2000 are currently known.

9 Multi-protein complexes

Beyond binary interactions, proteins often form large molecular complexes involving multiple subunits (Fig. 8). These complexes are much more than a random snapshot of a group of interacting proteins – they represent large functional entities which remain stable for long periods of time. Many such protein complexes have been elucidated step by step over time and recent advances in high-throughput technology have led to large-scale studies revealing numerous new protein complexes. The preferred technology for this kind of experiment is initial co-purification of the complexes followed by the identification of the member proteins by mass spectrometry.

As the bakers yeast *S. cerevisiae* is one of the most versatile model organisms used in molecular biology, it is not surprising that the first large-scale complex datasets were obtained in this species (Gavin et al. 2002; Ho et al. 2002; Gavin et al. 2006; Krogan et al.

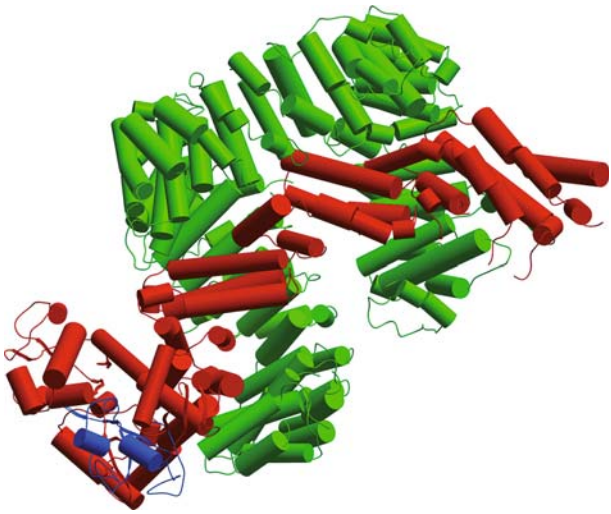


Fig. 8 Ternary complex between the Cand1 protein (green) and the catalytic core of the ubiquitin ligase consisting of cullin (red) and Roc1 (blue) (Goldenberg et al. 2004)

2006). The yeast protein interaction database MPact (Guldener et al. 2006) provides access to 268 protein complexes based on careful literature annotation composed of 1237 different proteins plus over 1000 complexes from large-scale experiments which contain more than 2000 distinct proteins. These numbers contain some redundancy with respect to complexes, due to slightly different complex composition found by different groups or experiments. Nevertheless, the dataset covers about 40% of the *S.cerevisiae* proteome. While many complexes comprise only a small number of different proteins, the largest of them features an impressive 88 different protein species.

A novel manually annotated database, CORUM (Ruepp et al. 2008) contains literature-derived information about 1750 mammalian multi-protein complexes. Over 75% of all complexes contain between three and six subunits, while the largest molecular structure, the spliceosome, consists of 145 components (Fig. 9).

10 Network modules

Modularity has emerged as one of the major organizational principles of cellular processes. Functional modules are defined as molecular ensembles with an autonomous function (Hartwell et al. 1999). Proteins or genes can be partitioned into modules based on shared patterns of regulation or expression, involvement in a common metabolic or regulatory pathway, or membership in the same protein complex or subcellular structure. Modular representation and analysis of cellular processes allows for inter-

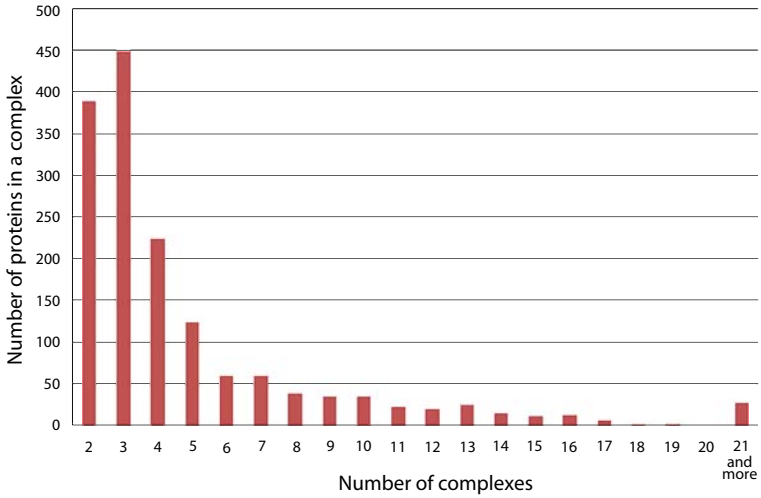


Fig. 9 Number of proteins in the CORUM complexes

pretation of genome data beyond single gene behavior. In particular, analysis of modules provides a convenient framework for studying the evolution of living systems (Snel and Huynen 2004). Multiprotein complexes represent one particular type of functional modules in which individual components engage in physical interactions to execute a specific cellular function.

Algorithmically, modular architectures can be defined as densely interconnected groups of nodes on biological networks (for an excellent review of available methods see (Sharan et al. 2007). Statistically significant functional subnetworks are characterized by a high degree of local clustering. The density of a cluster can be represented as a function $Q(m,n) = 2m/(n(n-1))$, where m is the number of interactions between the n nodes of the cluster (Spirin and Mirny 2003). Q thus takes values between 0 for a set of unconnected nodes and 1 for a fully connected cluster (clique). The statistical significance of Q strongly depends on the size of the graph. It is obvious that random clusters with $Q = 1$ involving just three proteins are very likely while large clusters with $Q = 1$ or even with values below 0.5 are extremely unlikely. In order to compute the statistical significance of a cluster with n nodes and m connections Spirin and Mirny calculate the expected number of such clusters in a comparable random graph and then estimate the likelihood of having m or more interactions within a given set of n proteins given the number of interactions that each of these proteins has. Significant dense clusters identified by this procedure on a graph of protein interactions were found to correspond to functional modules most of which are involved in transcription regulation, cell-cycle/cell-fate control, RNA processing, and protein transport. However, not all of them constitute physical protein complexes and, in general, it is not possible to predict

whether a given module corresponds to a multiprotein complex or just to a group of functionally coupled proteins involved in the same cellular process.

The search for significant subgraphs can be further enhanced by considering evolutionary conservation of protein interactions. With this approach protein complexes are predicted from binary interaction data by network alignment which involves comparing interaction graphs between several species (Sharan et al. 2005). First, proteins are grouped by sequence similarity such that each group contains one protein from each species, and each protein is similar to at least one other protein in the group. Then a composite interaction network is created by joining with edges those pairs of groups that are linked by at least one conserved interaction. Again, dense clusters on such network alignment graph are often indicative of multiprotein complexes.

An alternative computational method for deriving complexes from noisy large-scale interaction data relies on a "socio-affinity" index which essentially reflects the frequency with which proteins form partnerships detected by co-purification (Gavin et al. 2006). This index was shown to correlate well with available three-dimensional structure data, dissociation constants of protein–protein interactions, and binary interactions identified by the two-hybrid techniques. By applying a clustering procedure to a matrix containing the values of the socio-affinity index for all yeast protein pairs found to associate by affinity purification, 491 complexes were predicted, with over a half of them being novel and previously unknown. However, dependent on the analysis parameters distinct complex variants (isoforms) are found that differ from in terms of their subunit composition. Those proteins present in most of the isoforms of a given complex constitute its core while variable components present only in a small number of isoforms can be considered "attachments" (Fig. 10). Furthermore, some stable, typically smaller protein groups can be found in multiple attachments in which case they are

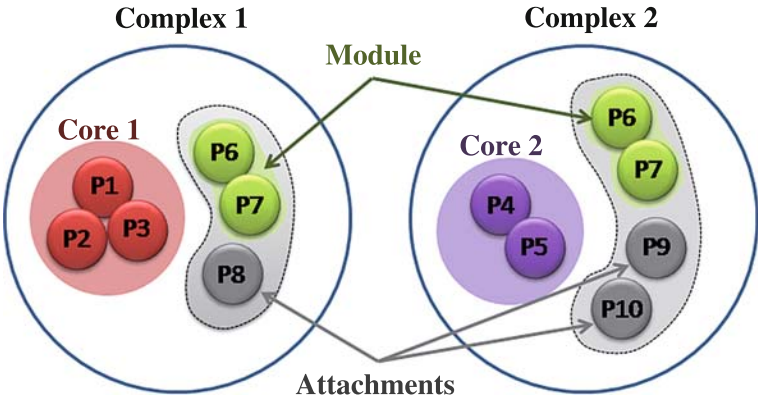


Fig. 10 Definitions of complex cores, attachments, and modules. Redrawn and modified with permission from (Gavin et al. 2006)

called "modules". Stable functional modules can thus be flexibly used in the cell in a variety of functional contexts. Proteins frequently associated with each other in complex cores and modules are likely to be co-expressed and co-localized.

11 Diseases and protein interaction networks

In this section, we offer a computational perspective on utilizing protein network data for molecular medical research. The identification of novel therapeutic targets for diseases and the development of drugs has always been a difficult, time-consuming and expensive venture (Ruffner et al. 2007). Recent work has charted the current pharmacological space using different networks of drugs and their protein targets (Paolini et al. 2006; Keiser et al. 2007; Kuhn et al. 2008; Yildirim et al. 2007) based on biochemical relationships like ligand binding energy and molecular similarity or on shared disease association. Above all, since many diseases are due to the malfunctioning of proteins, the systematic determination and exploration of the human interactome and homologous protein networks of model organisms can provide considerable new insight into pathophysiological processes (Giallourakis et al. 2005).

Knowledge of protein interactions can frequently improve the understanding of relevant molecular pathways and the interplay of various proteins in complex diseases (Fishman and Porter 2005). This approach may result in the discovery of a considerable number of novel drug targets for the biopharmaceutical industry, possibly affording the development of multi-target combination therapeutics. Observed perturbations of protein networks may also offer a refined molecular description of the etiology and progression of disease in contrast to phenotypic categorization of patients (Loscalzo et al. 2007). Molecular network data may help to improve the ability of cataloging disease unequivocally and to further individualize diagnosis, prognosis, prevention, and therapy. This will require a network-based approach that does not only include protein interactions to differentiate pathophenotypes, but also other types of molecular interactions as found in signaling cascades and metabolic pathways. Furthermore, environmental factors like pathogens interacting with the human host or the effects of nutrition need to be taken into account.

After large-scale screens identified enormous amounts of protein interactions in organisms like yeast, fly, and worm (Goll and Uetz 2007), which also serve as model systems for studying many human disease mechanisms (Giallourakis et al. 2005), experimental techniques and computational prediction methods have recently been applied to generate sizable networks of human proteins (Cusick et al. 2005; Stelzl and Wanker 2006; Assenov et al. 2008; Ramírez et al. 2007). In addition, comprehensive maps of protein interactions inside pathogens and between pathogens and the human host have been compiled for bacteria like *E. coli*, *H. pylori*, *C. jejuni*, and other species (Noirot and Noirot-Gros 2004), for many viruses such as herpes viruses, the Epstein-

Table 3 Selection of pathogenic organisms for which comprehensive protein interaction maps are available

Organism	References
Bacteria	
Escherichia coli	(Butland et al. 2005)
Helicobacter pylori	(Colland et al. 2001)
Campylobacter jejuni	(Parrish et al. 2007)
Viruses	
Herpesvirus family	(Uetz et al. 2006)
Epstein-Barr virus	(Calderwood et al. 2007)
SARS coronavirus	(von Brunn et al. 2007)
HIV-1	(Wheeler et al. 2007)
Hepatitis C virus	(Flajolet et al. 2000)
Parasite	
Plasmodium falciparum	(LaCount et al. 2005)

Barr virus, the SARS coronavirus, HIV-1, the hepatitis C virus, and others (Uetz et al. 2004), and for the malaria parasite *P. falciparum* (Table 3). Those extensive network maps can now be explored to identify potential drug targets and to block or manipulate important protein–protein interactions.

Furthermore, different experimental methods are also used to expand the known interaction networks around pathway-centric proteins like epidermal growth factor receptors (EGFRs) (Tewari et al. 2004; Oda et al. 2005; Jones et al. 2006), Smad and transforming growth factor–β (TGFβ) (Colland and Daviet 2004; Tewari et al. 2004; Barrios-Rodiles et al. 2005), and tumor necrosis factor-α (TNFα) and the transcription factor NF-κB (Bouwmeester et al. 2004). All of these proteins are involved in sophisticated signal transduction cascades implicated in various important disease indications ranging from cancer to inflammation. The immune system and Toll-like receptor (TLR) pathways were the subject of other detailed studies (Oda and Kitano 2006). Apart from that, protein networks for longevity were assembled to research ageing-related effects (Xue et al. 2007).

High-throughput screens are also conducted for specific disease proteins causative of closely related clinical and pathological phenotypes to unveil molecular interconnections between the diseases. For example, similar neurodegenerative disease phenotypes are caused by polyglutamine proteins like huntingtin and over twenty ataxins. Although they that are not evolutionarily related and their expression is not restricted to the brain, they are responsible for inherited neurotoxicity and age-dependent dementia only in specific neuron populations (Ralser et al. 2005). Yeast two-hybrid screens revealed an unexpectedly dense interaction network of those disease proteins forming interconnected subnetworks (Fig. 11), which suggests common pathways affected in disease (Goehler et al. 2004; Lim et al. 2006). Some of the protein–protein interactions may be

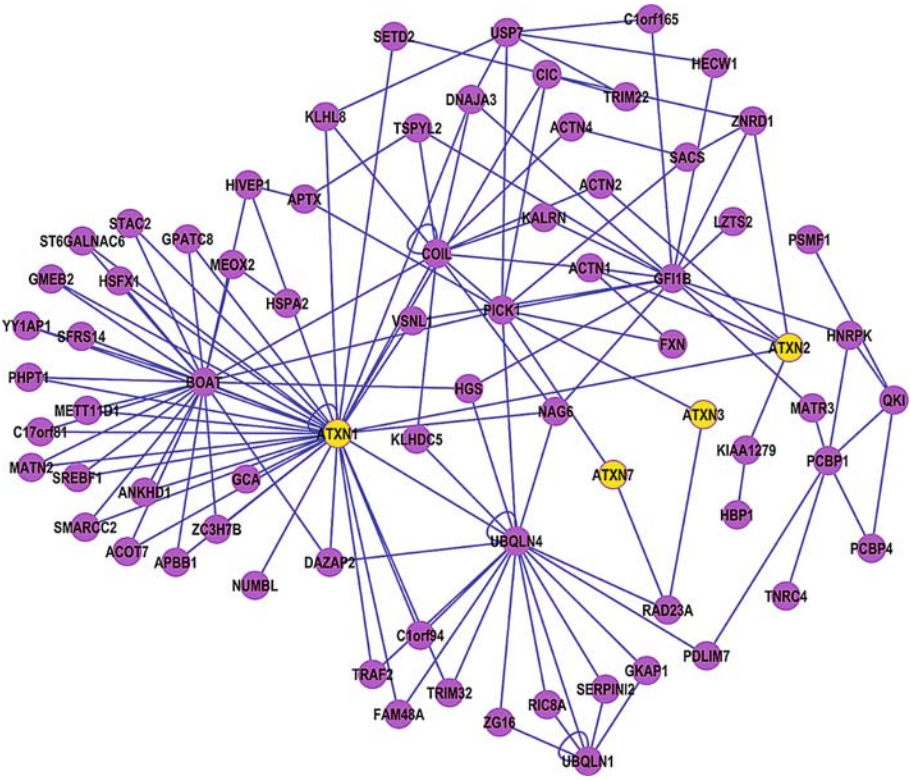


Fig. 11 Part of the protein interaction network around the four yellow-colored ataxins causative of neurodegenerative diseases

involved in mediating neurodegeneration and thus may be tractable for drug inhibition, and several interaction partners of ataxins could additionally be shown to be potential disease modifiers in a fly model (Kaltenbach et al. 2007).

A number of methodological approaches concentrate on deriving correlations between common topological properties and biological function from subnetworks around proteins that are associated with a particular disease phenotype like cancer. Recent studies report that human disease-associated proteins with similar clinical and pathological features tend to be more highly connected among each other than with other proteins and to have more similar transcription profiles (Gandhi et al. 2006; Xu and Li 2006; Goh et al. 2007). This observation points to the existence of disease-associated functional modules. Interestingly, in contrast to disease genes, essential genes whose defect may be lethal early on in life are frequently found to be hubs central to the network.

Further work focused on specific disease-relevant networks. For instance, to analyze experimental asthma, differentially expressed genes were mapped onto a protein

interaction network (Lu et al. 2007). Here, highly connected nodes tended to have smaller expression changes than peripheral nodes. This agrees with the general notion that disease-causing genes are typically not central in the network. Similarly, a comprehensive protein network analysis of systemic inflammation in human subjects investigated blood leukocyte gene expression patterns when receiving an inflammatory stimulus, a bacterial endotoxin, to identify functional modules perturbed in response to this stimulus (Calvano et al. 2005). Topological criteria and gene expression data were also used to search protein networks for functional modules that are relevant to type 2 diabetes mellitus (Liu et al. 2007) or to different types of cancer (Jonsson and Bates 2006; Cui et al. 2007; Lin et al. 2007; Pujana et al. 2007). Moreover, it was recently demonstrated that the integration of gene expression profiles with subnetworks of interacting proteins can lead to improved prognostic markers for breast cancer outcome that are more reproducible between patient cohorts than sets of individual genes selected without network information (Chuang et al. 2007).

In drug discovery, protein networks can help to design selective inhibitors of protein–protein interactions which target specific interactions of a protein, but do not affect others (Wells and McClendon 2007). For example, a highly connected protein (hub) may be a suitable target for an antibiotic whereas a more peripheral protein with few interaction partners may be more appropriate for a highly specific drug that needs to avoid side effects. Thus, topological network criteria are not only useful for characterizing disease proteins, but also for finding drug targets. The diversity of interactions of a targeted protein could also help in predicting potential side effects of a drug. Apart from that, it is remarkable that some potential drugs have been found to be less effective than expected due to the intrinsic robustness of living systems against perturbations of molecular interactions (Kitano 2007). Furthermore, mutations in proteins cause genetic diseases, but it is not always easy to distinguish protein interactions impaired by mutated binding sites from other disease causes like structural instability induced by amino acid mutations.

Nowadays many genome-wide association and linkage studies for human diseases suggest genomic loci and linkage intervals that contain candidate genes encoding SNPs and mutations of potential disease proteins (Kann 2007). Since the resultant list of candidates frequently contain dozens or even hundreds of genes, computational approaches have been developed to prioritize them for further analyses and experiments. In the following, we will demonstrate the variety of available prioritization approaches by explicating three recent methods that utilize protein interaction data in addition to the inclusion of other sequence and function information. All methods capitalize on the above described observation that closely interacting gene products often underlie polygenic diseases and similar pathophenotypes (Oti and Brunner 2007).

Using protein–protein interaction data annotated with reliability values, Lage et al. (2007) first predict human protein complexes for each candidate protein. They then score the pairwise phenotypic similarity of the candidate disease with all proteins within each complex that are associated with any disease. The scoring function basically

measures the overlap of the respective disease phenotypes as recorded in text entries of OMIM (Online Mendelian Inheritance in Man) (Hamosh et al. 2005) based on the vocabulary of UMLS (Unified Medical Language System) (Bodenreider 2004). Lastly, all candidates are prioritized by the probability returned by a Bayesian predictor trained on the interaction data and phenotypic similarity. Therefore, this method depends on the premise that the phenotypic effects caused by any disease-affected member in a predicted protein complex are very similar to each other.

Another prioritization approach by Franke et al. (2006) does not make use of overlapping disease phenotypes and primarily aims at connecting physically disjoint genomic loci associated with the same disease using molecular networks. At the beginning, their method Prioritizer performs a Bayesian integration of three different network types of gene/protein relationships. The latter are derived from functional similarity using Gene Ontology annotation, microarray coexpression, and protein–protein interaction. This results in a probabilistic human network of general functional links between genes. Prioritizer then assesses which candidate genes contained in different disease loci are closely connected in this gene-gene network. To this end, the score of each candidate is initially set to zero, but it is increased iteratively during network exploration by a scoring function that depends on the network distance of the respective candidate gene to candidates inside another genomic loci. This procedure finally yields separate prioritization lists of ranked candidate genes for each genomic loci.

In contrast to the integrated gene-gene network used by Prioritizer, the Endeavour system (Aerts et al. 2006) directly compares candidate genes with known disease genes and creates different ranking lists of all candidates using various sources of evidence for annotated relationships between genes or proteins. The evidence can be derived from literature mining, functional associations based on Gene Ontology annotations, co-occurrence of transcriptional motifs, correlation of expression data, sequence similarity, common protein domains, shared metabolic pathway membership, and protein–protein interactions. At the end, Endeavour merges the resultant ranking lists using order statistics and computes an overall prioritization list of all candidate genes.

Finally, it is important to keep in mind that current datasets of human protein interactions may still contain a significant number of false interactions and thus biological and medical conclusions derived from them should always be taken with a note of caution, in particular, if no good confidence measures are available.

12 Sequence-based prediction of protein interactions

A comprehensive atlas of protein interactions is fundamental for a better understanding of the overall dynamic functioning of the living organisms. These insights arise from the integration of functional information, dynamic data and protein interaction networks. In order to fulfill the goal of enlarging our view of the protein interaction network,

several approaches must be combined and a crosstalk must be established among experimental and computational methods. This has become clear from comparative evaluations which show similar performances for both types of methodologies. In fact, over the recent years this field has grown into one of the most appealing fields in bioinformatics. Evolutionary signals result from restrictions imposed by the need to optimize the features that affect a given interaction and the nature of these features can differ from interaction to interaction. Consequently, a number of different methods have been developed based a range of different evolutionary signals. This section is devoted to a brief review of some of these methods.

12.1 Phylogenetic profiling

These techniques are based on the similarity of absence/presence profiles of interacting proteins. In its original formulation (Gaasterland and Ragan 1998; Huynen and Bork 1998; Pellegrini et al. 1999; Marcotte et al. 1999a) the phylogenetic profiles were codified as 0/1 vectors for each reference protein according to the absence/presence of proteins of the studied family in a set of fully sequenced organisms (see Fig. 12a). The vectors for different reference sequences are compared by using the Hamming distance (Pellegrini et al. 1999) between vectors. This measure counts the number of differences between two binary vectors. The rationale for this method is that both interacting proteins must be present in an organism and that reductive evolution will remove unpaired proteins in the rest of the organisms. Proposed improvements include the inclusion of quantitative measures of sequence divergence (Marcotte et al. 1999b; Date and Marcotte 2003) and the ability to deal with biases in the taxonomic distribution of the organisms used (Date and Marcotte 2003; Barker and Pagel 2005). These biases are due to the intuitive fact that evolutionarily similar organisms will share a higher number of protein and genomic features (in this case presence/absence of an orthologue).

To reduce this problem, Date et al. used Mutual Information from sequence divergent profiles for measuring the amount of information shared by both vectors. Mutual Information is calculated as:

$$MI(P1, P2) = H(P1) + H(P2) - H(P1, P2),$$

where $H(P1) = \sum p(P1) \ln p(P1)$ is the marginal entropy of the probability distribution of protein P1 sequence distances and $H(P1, P2) = -\sum\sum p(P1, P2) \ln p(P1, P2)$ is the joint entropy of the probability distributions of both protein P1 and P2 sequence distances. The corresponding probabilities are calculated from the whole distribution of orthologue distances for the organisms. In this way, the most likely evolutionary distances between orthologues from a pair of organisms will produce smaller entropies and consequently smaller values of Mutual Information. This formulation should implicitly reduce the effect of taxonomic biases. In an interesting work, published recently by Barker et al. (2007), the authors showed that detection of correlated gene-

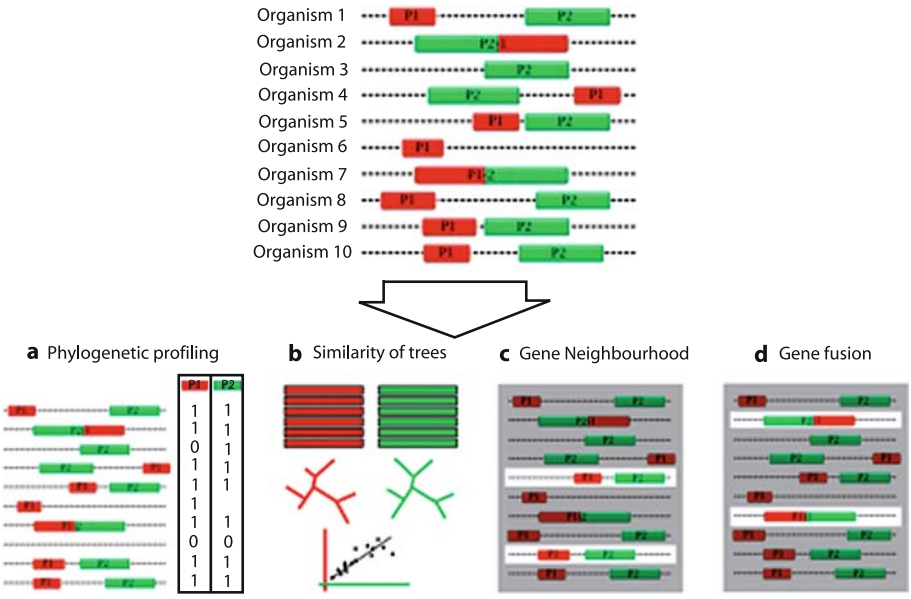


Fig. 12 Prediction of protein interactions based on genomic and sequence features. Information coming from the set of close homologs of the proteins P1 and P2 from the Organism 1 in other organisms can be used to predict an interaction between these proteins. (**a**) Phylogenetic profiling. Presence/absence of a homolog of both proteins in different organisms is coded as the corresponding two '1/0' profiles (most simple approach) and an interaction is predicted for very similar profiles. (**b**) Similarity of phylogenetic trees. Multiple sequence alignments are built for both sets of proteins and phylogenetic trees are derived from the proteins with a possible partner present in its organism. Proteins with highly similar trees are predicted to interact. (**c**) Gene neighbourhood conservation. Genome closeness is checked for those genes coding for both sets of homologous proteins. Interaction is predicted if gene pairs are recurrently close to each other in a number of organisms. (**d**) Gene fusion. Finding the proteins containing different sequence regions homologous to each of the two proteins is used to predict an interaction between them

gain/gene-loss events improves the predictions by reducing the number of false positives due to taxonomic biases.

The phylogenetic profiling approach has been shown to be quite powerful, because its simple formulation has allowed the exploration of a number of alternative interdependencies between proteins. This is the case for enzyme "displacement" in metabolic pathways detected as anti-correlated profiles (Morett et al. 2003), and for complex dependence relations among triplets of proteins (Bowers et al. 2004). Phylogenetic profiles have also been correlated with bacterial traits to predict the genes related to particular phenotypes (Korbel et al. 2005). The main drawbacks of these methods are the difficulty of dealing with essential proteins (where there is no absence information) and the requirement for the genomes under study to be complete (to establish the absence of a family member).

12.2 Similarity of phylogenetic trees

Similarity in the topology of phylogenetic trees of interacting proteins has been qualitatively observed in a number of cases (Fryxell 1996; Pages et al. 1997; Goh et al. 2000). The extension of this observation to a quantitative method for the prediction of protein interactions requires measuring the correlation between the similarity matrices of the explored pairs of protein families (Goh et al. 2000). This formulation allows systematic evaluation of the validity of using the original observation as a signal of protein interaction (Pazos and Valencia 2001).

The general protocol for these methods is illustrated in Fig. 12b. It includes the building of the multiple sequence alignment for the set of orthologues (one per organism) related to every query sequence, the calculation of all protein pair evolutionary distances (derived from the corresponding phylogenetic trees) and finally the comparison of evolutionary distance matrices of pairs of query proteins using Pearson's correlation coefficient. Protein pairs with highly correlated distance matrices are predicted to be more likely to interact.

Although this signal has been shown to be significant, the underlying process responsible for this similarity is still controversial (Chen and Dokholyan 2006). There are two main hypotheses for explaining this phenomenon. The first hypothesis suggests that this evolutionary similarity comes from the mutual adaptation (co-evolution) of interacting proteins and the need to retain interaction features while sequences diverge. The second hypothesis implicates external factors. In this scenario, the restrictions imposed by evolution on the functional process implicating both proteins would be responsible for the parallelism of their phylogenetic trees.

Although the relative importance of both factors is still not clear, the predictive power of similarities in phylogenetic trees is not affected. Indeed, a number of developments have improved the original formulation (Pazos et al. 2005; Sato et al. 2005). The first advance involved managing the intrinsic similarity of the trees because of the common underlying taxonomic distribution (due to the speciation processes). This effect is analogous to the taxonomic biases discussed above. In these cases, the approach followed was to correct both trees by removing this common trend. For example, Pazos et al. subtracted the distances of the 16S rRNA phylogenetic tree to the corresponding distances for each protein tree. The correlations for the resulting distance matrices were used to predict protein interactions.

Additionally some analyses have focused on the selection of the sequence regions used for the tree building (Jothi et al. 2006; Kann et al. 2007). For example, it has been shown that interacting regions, both defined as interacting residues (using structural data) and as the sequence domain involved in the interaction, show more clear tree similarities than the whole proteins (Mintseris and Weng 2005; Jothi et al. 2006). Other interesting work showed that prediction performance can be improved by removing poorly conserved sequence regions (Kann et al. 2007).

Finally, in a very recent work (Juan et al. 2008) the authors have suggested a new method for removing noise in the detection of tree similarity signals and detecting different levels of evolutionary parallelism specificity. This method introduces the new strategy of using the global network of protein evolutionary similarity for a better calibration of the evolutionary parallelism between two proteins. For this purpose, they define a protein 'co-evolutionary' profile as the vector containing the evolutionary correlations between a given protein tree and all the rest of the protein trees derived from sequences in the same organism. This co-evolutionary profile is a more robust and comparable representation of the evolution of a given protein (it involves hundreds of distances) and can be used to deploy a new level of evolutionary comparison. The authors compare these co-evolutionary profiles by calculating Pearson's correlation coefficient for each pair. In this way, the method detects pairs of proteins for which high evolutionary similarities are supported by their similarities with the rest of proteins of the organism. This approach significantly improves the predictive performance of the tree similarity-based methods so that different degrees of co-evolutionary specificity are obtained according to the number of proteins that might be influencing the co-evolution of the studied pair. This is done by extending the approach of Sato et al. (2006), that uses partial correlations and a reduced set of proteins for determining specific evolutionary similarities. Juan et al. calculated the partial correlation for each significant evolutionary similarity with respect to the remaining proteins in the organism and defined levels of co-evolutionary specificity according to the number of proteins that are considered to be co-evolving with each studied protein pair. With this strategy, it's possible to detect a range of evolutionary parallelisms from the protein pairs (for very specific similarities) up to subsets of proteins (for more relaxed specificities) that are highly evolution dependent. Interestingly, if specificity requirements are relaxed, protein relationships among components of macro-molecular complexes and proteins involved in the same metabolic process can be recovered. This can be considered as a first step in the application of higher orders of evolutionary parallelisms to decode the evolutionary impositions over the protein interaction network.

12.3 Gene neighbourhood conservation

This method exploits the well-known tendency of bacterial organisms to organize proteins involved in the same biochemical process by clustering them in the genome. This observation is obviously related to the operon concept and the mechanisms for the coordination of transcription regulation of the genes present in these modules. These mechanisms are widespread among bacterial genomes. Therefore the significance of a given gene proximity can be established by its conservation in evolutionary distant species (Dandekar et al. 1998; Overbeek et al. 1999).

The availability of fully sequenced organisms makes computing the intergenic distances between each pair of genes easy. Genes with the same direction of transcrip-

tion and closer than 300 bases are typically considered to be in the same genomic context (see Fig. 12c). The conservation of this closeness must be found in more than two highly divergent organisms to be considered significant because of the taxonomic biases.

While this signal is strong in bacterial genomes, its relevance is unclear in eukaryotic genomes. This is the main drawback of these methodologies. In fact, this signal only can be exploited for eukaryotic organisms by extrapolating genomic closeness of bacterial genes to their homologues in eukaryotes. Obviously, this extrapolation leads to a considerable reduction in the confidence and number of obtained predictions for this evolutionary lineage. However, conserved gene pairs that are transcribed from a shared bidirectional promoter can be detected by similar methods and can found in eukaryotes as well as prokaryotes (Korbel et al. 2004)

12.4 Gene fusion

A further use of evolutionary signals in protein function and physical interaction prediction has been the tendency of interacting proteins to be involved in gene fusion events. Sequences that appear as independently expressed ORFs in one organism become 'fused' as part of the same polypeptide sequence in another organism. These fusions are strong indicators of functional and structural interaction that have been suggested to increase the effective concentration of interacting functional domains (Enright et al. 1999; Marcotte et al. 1999b). This hypothesis proposes that gene fusion could remove the effect of diffusion and relative correct orientation of the proteins forming the original complex.

These fusion events are typically detected when sequence searches for two non-homologous proteins obtain a significant hit in the same sequence. Cases matching to the same region of the hit sequence are removed (these cases are schematically represented in Fig. 12d).

In spite of the strength of this signal, gene fusion seems to not be a habitual event in bacterial organisms. The difficulty of distinguishing protein interactions belonging to large evolutionary families is the main drawback of the automatic application of these methodologies.

13 Integration of experimentally determined and predicted interactions

As described above, there are many both experimental techniques and computational methods for determining and predicting interactions. To obtain the most comprehensive interaction networks possible, as many as possible of these sources of interactions should be integrated. The integration of these resources is complicated by the fact that

the different sources are not all equally reliable, and it is thus important to quantify the accuracy of the different evidence supporting an interaction.

In addition to the quality issues, comparison of different interaction sets is further complicated by the different nature of the datasets: yeast two-hybrid experiments are inherently binary, whereas pull-down experiments tend to report larger complexes. To allow for comparisons, complexes are typically represented by binary interaction networks; however, it is important to realize that there is not a single, clear definition of a "binary interaction". For complex pull-down experiments, two different representations have been proposed: the matrix representation, in which each complex is represented by the set of binary interactions corresponding to all pairs of proteins from the complex, and the spoke representation, in which only bait-prey interactions are included (von Mering et al. 2002). The binary interactions obtained using either of these representations are somewhat artificial as some interacting proteins might in reality never touch each other and others might have too low an affinity to interact except in the context of the entire complex bringing them together. Even in the case of yeast two-hybrid assays, which inherently report binary interactions, not all interactions correspond to direct physical interactions.

The database STRING ("Search Tool for the Retrieval of Interacting Genes/Proteins") (von Mering et al. 2007) represents an effort to provide many of the different types of evidence for functional interactions under one common framework with an integrated scoring scheme. Such an integrated approach offers several unique advantages: 1) various types of evidence are mapped onto a single, stable set of proteins, thereby facilitating comparative analysis; 2) known and predicted interactions often partially complement each other, leading to increased coverage; and 3) an integrated scoring scheme can provide higher confidence when independent evidence types agree.

In addition to the many associations imported from the protein interaction databases mentioned above (Bader et al. 2003; Salwinski et al. 2004; Guldener et al. 2006; Mishra et al. 2006; Stark et al. 2006; Chatr-aryamontri et al. 2007), STRING also includes interactions from curated pathway databases (Vastrik et al. 2007; Kanehisa et al. 2008) and a large body of predicted associations that are produced de novo using many of the methods described in this chapter (Dandekar et al. 1998; Gaasterland and Ragan 1998; Pellegrini et al. 1999; Marcotte et al. 1999c). These different types of evidence are obviously not directly comparable, and even for the individual types of evidence the reliability may vary. To address these two issues, STRING uses a two-stage approach. First, a separate scoring scheme is used for each evidence type to rank the interactions according to their reliability; these raw quality scores cannot be compared between different evidence types. Second, the ranked interaction lists are benchmarked against a common reference to obtain probabilistic scores, which can subsequently be combined across evidence types.

To exemplify how raw quality scores work, we will here explain the scoring scheme used for physical protein interactions from high-throughput screens. The two funda-

mentally different types of experimental interaction data sets, complex pull-downs and binary interactions are evaluated using separate scoring schemes. For the binary interaction experiments, e.g. yeast two-hybrid, the reliability of an interaction correlates well with the number of non-shared interaction partners for each interactor. STRING summarizes this in the following raw quality score:

$$S_1 = \log((N_1+1) \cdot (N_2+1)),$$

where N_1 and N2 are the numbers of non-shared interaction partners. This score is similar to the IG1 measure suggested by Saito et al. (2002). In the case of complex pull-down experiments, the reliability of the inferred binary interactions correlates better with the number of times the interactors were co-purified compared to what would be expected at random:

$$S_2 = \log((N_{12} \cdot N)/((N_1+1) \cdot (N_2+1))),$$

where N_{12} is the number of purifications containing both proteins, N_1 and N_2 are the numbers of purifications containing either protein 1 or 2, and N is the total number of purifications. For this purpose, the bait protein was counted twice to account for bait–prey interactions being more reliable than prey–prey interactions. These raw quality scores are calculated for each individual high-throughput screen. Scores vary within one dataset, because they include additional, intrinsic information from the data itself, such as the frequency with which an interaction is detected. For medium sized data sets that are not large enough to apply the topology based scoring schemes, the same raw score is assigned to all interactions within a dataset. Finally, very small data sets are pooled and considered jointly as a single interaction set.

We similarly have different scoring schemes for predicted interactions based on co-expression in microarray expression studies, conserved gene neighborhood, gene fusion events and phylogenetic profiles. Based on these raw quality scores, a confidence score is assigned to each predicted association by benchmarking the performance of the predictions against a common reference set of trusted, true associations. STRING uses as reference the functional grouping of proteins maintained at KEGG (Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al. 2008). Any predicted association for which both proteins are assigned to the same "KEGG pathway" is counted as a true positive. KEGG pathways are particularly suitable as a reference because they are based on manual curation, are available for a number of organisms, and cover several functional areas. Other benchmark sets could also be used, for example "Biological Process" terms from Gene Ontology (Ashburner et al. 2000) or Reactome pathways (Vastrik et al. 2007). The benchmarked confidence scores in STRING generally correspond to the probability of finding the linked proteins within the same pathway or biological process.

The assignment of probabilistic scores for all evidence types solves many of the issues of data integration. First, incomparable evidence types are made comparable by

assigning a score that represents how well the evidence type can predict a certain type of interactions (the type being specified by the reference set used). Second, the separate benchmarking of interactions from, for example, different high-throughput protein interaction screens accounts for any differences in reliability between different studies. Third, use of raw quality scores allows us to separate more reliable interactions from less reliable interactions even within a single dataset. The probabilistic nature of the scores also makes it easy to calculate the combined reliability of an interaction given multiple lines of evidence. It is computed under the assumption of independence for the various sources, in a naïve Bayesian fashion.

In addition to having a good scoring scheme, it is crucial to make the evidence for an interaction transparent to the end users. To achieve this, the STRING interaction network is made available via a user-friendly web interface (http://string.embl.de). When performing a query, the user will first be presented with a network view, which provides a first, simplified overview (Fig. 13). From here the user has full control over parameters such as the number of proteins shown in the network (nodes) and the minimal reliability required for an interaction (edge) to be displayed. From the network, the user also has the ability to drill down on the evidence that underlies any given interaction using the dedicated viewer for each evidence type. For example, it is possible to inspect the publications that support a given interaction, the set of protein that were

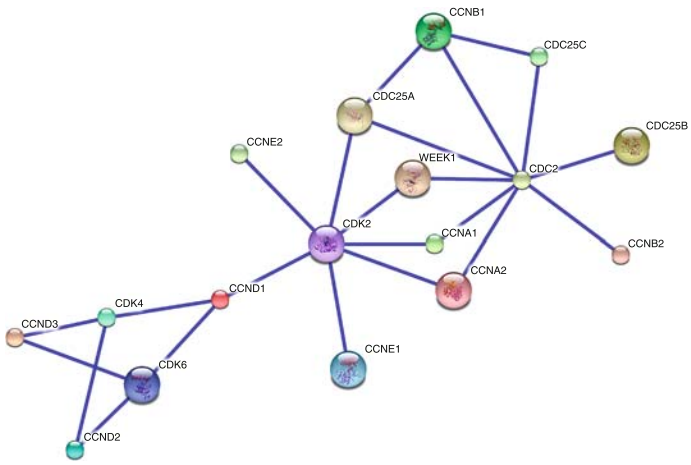


Fig. 13 Protein interaction network of the core cell-cycle regulation in human. The network was constructed by querying the STRING database (von Mering et al. 2007) for very high confidence interactions (conf. score > 0.99) between four cyclin-dependent kinases, their associated cyclins, the WEE1 kinase and the CDC25 phosphatases. The network correctly recapitulates CDC2 interacts with cyclin-A/B, CDK2 with cyclin-A/E, and CDK4/6 with cyclin-D. It also shows that the WEE1 and CDC25 phosphatases regulate CDC2 and CDK2 but not CDK4 and CDK6. Moreover, the network suggests that CDC25A phosphatase regulates CDC2 and CDK2, whereas CDC25B and CDC25C specifically regulate CDC2

co-purified in a particular experiment and the phylogenetic profiles or genomic context based on which an interaction was predicted.

14 Domain–domain interactions

Protein binding is commonly characterized by specific interactions of evolutionarily conserved domains (Pawson and Nash 2003). Domains are fundamental units of protein structure and function (Aloy and Russell 2006), which are incorporated into different proteins by genetic duplications and rearrangements (Vogel et al. 2004). Globular domains are defined as structural units of fifty and more amino acids that usually fold independently of the remaining polypeptide chain to form stable, compact structures (Orengo and Thornton 2005). They often carry important functional sites and determine the specificity of protein interactions (Fig. 14). Essential information on

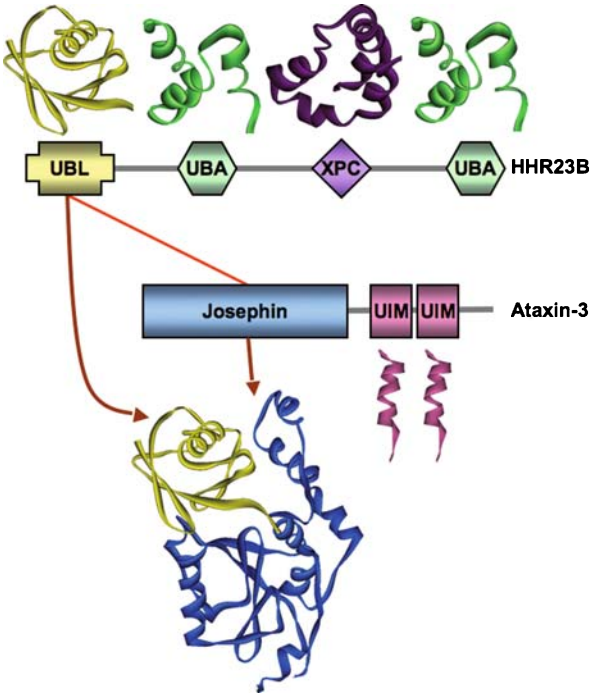


Fig. 14 Exemplary interaction between the two human proteins HHR23B and ataxin-3. Each protein domain commonly adopts a particular 3D structure and may fulfill a specific molecular function. Generally, the domains responsible for an observed protein-protein interaction need to be determined before further functional characterizations are possible. In the depicted protein-protein interaction, it is known from experiments that the ubiquitin-like domain UBL of HHR23B (yellow) forms a complex with de-ubiquitinating Josephin domain of ataxin-3 (blue) (Nicastro et al. 2005)

the cellular function of specific protein interactions and complexes can often be gained from the known functions of the interacting protein domains. Domains may contain binding sites for proteins and ligands such as metabolites, DNA/RNA, and drug-like molecules (Xia et al. 2004). Widely spread domains that mediate molecular interactions can be found alone or combined in conjunction with other domains and intrinsically disordered, mainly unstructured, protein regions connecting globular domains (Dunker et al. 2005). According to Apic et al. (2001) multi-domain proteins constitute two thirds of unicellular and 80% of metazoan proteomes. One and the same domain can occur in different proteins, and many domains of different types are frequently found in the same amino acid chain.

Much effort is being invested in discovering, annotating, and classifying protein domains both from the functional (Pfam (Finn et al. 2006), SMART (Letunic et al. 2006), CDD (Marchler-Bauer et al. 2007), InterPro (Mulder et al. 2007) and structural (SCOP (Andreeva et al. 2004), CATH (Greene et al. 2007)) perspective. Notably, it may be confusing that the term 'domain' is commonly used in two slightly different meanings. In the context of domain databases such as Pfam and SMART, a domain is basically defined by a set of homologous sequence regions, which constitute a domain family. In contrast, a specific protein may contain one or more domains, which are concrete sequence regions within its amino acid sequence corresponding to autonomously folding units. Domain families are commonly represented by Hidden Markov Models (HMMs), and highly sensitive search tools like HMMER (Eddy 1998) are used to identify domains in protein sequences.

Different sources of information about interacting domains with experimental evidence are available. Experimentally determined interactions of single-domain proteins indicate domain–domain interactions. Similarly, experiments using protein fragments help identifying interaction domains, but this knowledge is frequently hidden in the text of publications and not contained in any database. However, domain databases like Pfam, SMART, and InterPro may contain some annotation obtained by manual literature curation. In the near future, high-throughput screening techniques will result in even larger amounts of protein fragment interaction data to delineate domain borders and interacting protein regions (Colland and Daviet 2004).

Above all, three-dimensional structures of protein domain complexes are experimentally solved by X-ray crystallography or NMR and are deposited in the PDB database (Berman et al. 2007). Structural contacts between two interacting proteins can be derived by mapping sequence positions of domains onto PDB structures. Extensive investigations of domain combinations in proteins of known structures (Apic et al. 2001) as well as of structurally resolved homo- or heterotypic domain interactions (Park et al. 2001) revealed that the overlap between intra- and intermolecular domain interactions is rather limited. Two databases, iPfam (Finn et al. 2005) and 3did (Stein et al. 2005), provide pre-computed structural information about protein interactions at the level of Pfam domains.

Analysis of structural complexes suggests that interactions between a given pair of proteins may be mediated by different domain pairs in different situations and in different organisms. Nevertheless, many domain interactions, especially those involved in basic cellular processes such as DNA metabolism and nucleotide binding, tend to be evolutionarily conserved within a wide range of species from prokaryotes to eukaryotes (Itzhaki et al. 2006). In yeast, Pfam domain pairs are associated with over 60% of experimentally known protein interactions, but only 4.5% of them are covered by iPfam (Schuster-Bockler and Bateman 2007).

Domain interactions can be inferred from experimental data on protein interactions by identifying those domain pairs that are significantly overrepresented in interacting proteins compared to random protein pairs (Deng et al. 2002; Ng et al. 2003a; Riley et al. 2005; Sprinzak and Margalit 2001) (Fig. 15). However, the predictive power of such an approach is strongly dependent on the quality of the data used as the source of information for protein interactions, and the coverage of protein sequences in terms of domain assignments. Basically, the likelihood of two domains, D_i and D_j, to interact

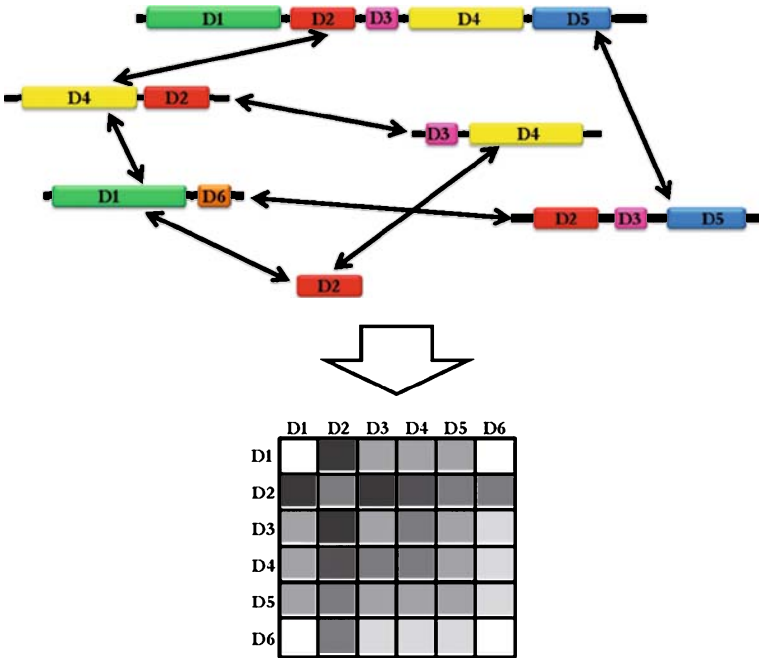


Fig. 15 Deriving the likelihood of domain interactions from experimental data of protein interactions. Six different proteins are shown containing domains D1., D2., . . . , D6 in different combinations. Known interactions between these proteins are shown as black arrows. The matrix in the bottom part of the figure shows the likelihood for each pair of domains to interact – from low (white) to high (dark)

can be estimated as the fraction of protein pairs known to interact among all proteins in the dataset containing this domain pair.

This basic idea has been improved upon by using a maximum-likelihood (ML) approach based on the expectation-maximization (EM) algorithm. This method finds the maximum likelihood estimator of the observed protein–protein interactions by an iterative cycle of computing the expected likelihood (E-step) and maximizing the unobserved parameters (domain interaction propensities) in the M-step. When the algorithm converges (i.e. the total likelihood cannot be further improved by the algorithm), the ML estimate for the likelihood of the unobserved domain interactions is found (Deng et al. 2002; Riley et al. 2005). Riley and colleagues further improved this method by excluding each potentially interacting domain pair from the dataset and re-computing the ML-estimate to obtain an additional confidence value for the respective domain–domain interaction. This domain pair exclusion (DPEA) method measures the contribution of each domain pair to the overall likelihood of the protein interaction network based on domain–domain interactions. In particular, this approach enables the prediction of specific domain–domain interactions between selected proteins which would have been missed by the basic ML method. Another ML-based algorithm is InSite which takes differences in the reliability of the protein–protein interaction data into account (Wang et al. 2007a). It also integrates external evidence such as functional annotation or domain fusion events.

An alternative method for deriving domain interactions is through co-evolutionary analysis that exploits the notion that mutations of residue pairs at the interaction interfaces are correlated to preserve favorable physico-chemical properties of the binding surface (Jothi et al. 2006). The pair of domains mediating interactions between two proteins P1 and P2 may therefore be expected to display a higher similarity of their phylogenetic trees than other, non-interacting domains (Fig. 16). The degree of agreement between the evolutionary history of two domains, D_i and D_j, can be computed by the Pearson's correlation coefficient r_{ij} between the similarity matrices of the domain sequences in different organisms:

$$r_{ij} = \frac{\sum_{p=1}^{n-1}\sum_{q=p+1}^{n}(M_{pq}^{i} - \overline{M}^{i})(M_{pq}^{j} - \overline{M}^{j})}{\sqrt{\sum_{q=p+1}^{n-1}\sum_{q=p+1}^{n}((M_{pq}^{i} - \overline{M}^{i}))\sum_{p=1}^{n-1}\sum_{q=p+1}^{n}((M_{pq}^{j} - \overline{M}^{j}))}},$$

where n is the number of species, M_{pq}^{i} and M_{pq}^{j} are the evolutionary distances between species, and $\overline{M}^{i}$ and $\overline{M}^{j}$ are the mean values of the matrices, respectively. In Figure 16 the evolutionary tree of the domain D2 is most similar to those of D5 and D6, corroborating the actual binding region.

A well-known limitation of the correlated mutation analysis is that it is very difficult to decide whether residue co-variation happens as a result of functional co-evolution

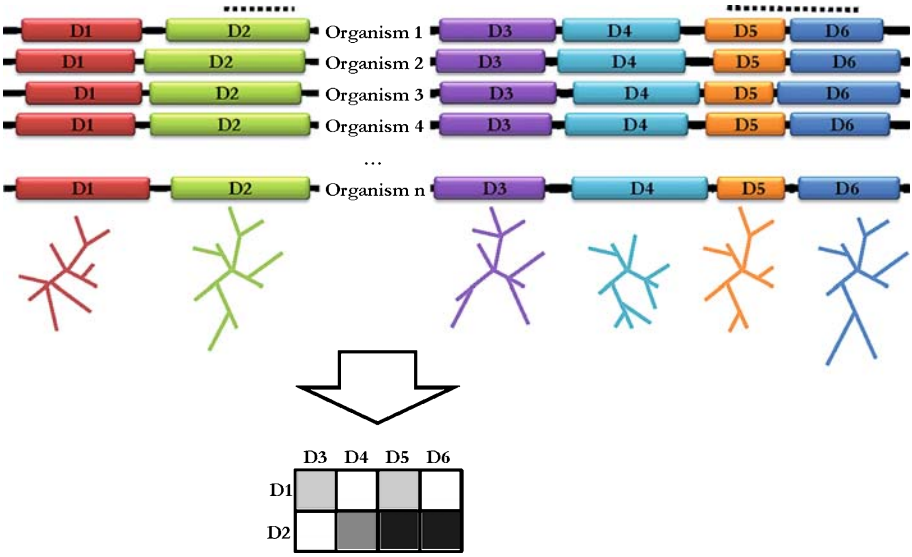


Fig. 16 Co-evolutionary analysis of domain interactions. Two orthologous proteins from different organisms known to interact with each other are shown. The first protein consists of two domains, D1 and D2, while the second protein includes the domains D3, D4, D5, and D6. Evolutionary trees for each domain are shown, their similarity serves as an indication of interaction likelihood that is encoded in the interaction matrix

directed at preserving interaction sites, or because of sequence divergence due to speciation. To address this problem, (Kann et al. 2007) suggested to distinguish the relative contribution of conserved and more variable regions in aligned sequences to the co-evolution signal based on the hypothesis that functional co-evolution is more prominent in conserved regions.

Finally, interacting domains can be identified by phylogenetic profiling, as described above for full-chain proteins. As in the case of complete protein chains, the similarity of evolutionary patterns shared by two domains may indicate that they interact with each other directly or at least share a common functional role (Pagel et al. 2004). As illustrated in Fig. 17, clustering protein domains with similar phylogenetic profiles allows researchers to build domain interaction networks which provide clues for describing molecular complexes. Similarly, the DomainTeam method (Pasek et al. 2005) considers chromosomal neighborhoods at the level of conserved domain groups.

A number of resources provide and combine experimentally derived and predicted domain interaction data. InterDom (http://interdom.i2r.a-star.edu.sg/) integrates domain-interaction predictions based on known protein interactions and complexes with domain fusion events (Ng et al. 2003b). DIMA (http://mips.gsf.de/genre/proj/dima2) is another database of domain interactions, which integrates experimentally demon-

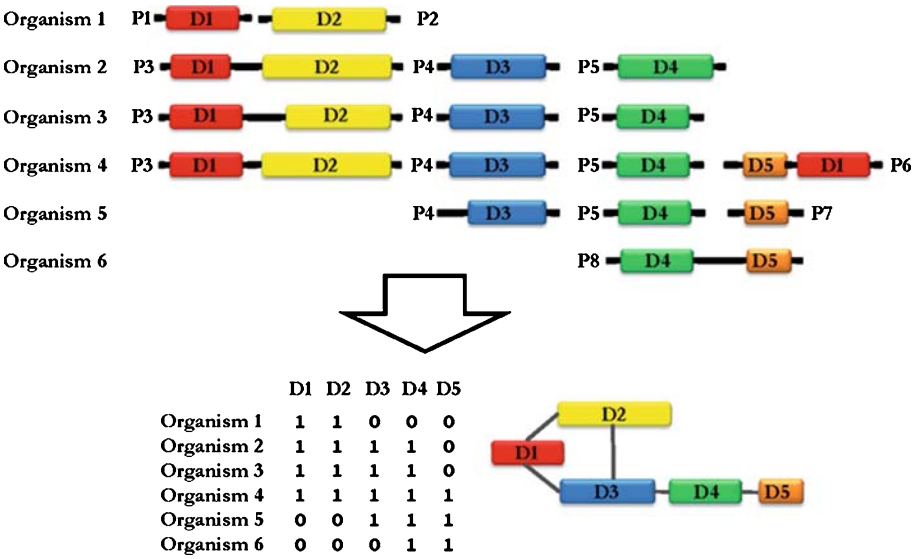


Fig. 17 Similarity of domain phylogenetic profiles can be used to build a domain interaction network

strated domain interactions from iPfam and 3did with predictions based on the DPEA algorithm and phylogenetic domain profiling (Pagel et al. 2007). Recently, two new comprehensive resources, DOMINE (http://domine.utdallas.edu) (Raghavachari et al. 2008) and DASMI (http://www.dasmi.de) (Blankenburg et al. 2008, submitted), were introduced and are available online. These resources contain iPfam and 3did data and predicted domain interactions taken from several other publications. Predictions are based on several methods for deriving domain interactions from protein interaction data, phylogenetic domain profiling data and domain coevolution. With the availability of an increasing number of predictions the task of method weighting and quality assessment becomes crucial. A thorough analysis of the quality of domain interaction data can be found in Schlicker et al. (2007).

Beyond domain–domain contacts, an alternative mechanism of mediating molecular recognition is through binding of protein domains to short sequence regions (Santonico et al. 2005), typically from three to eight residues in length (Zarrinpar et al. 2003; Neduva et al. 2005). Such linear recognition motifs can be discovered from protein interaction data by identifying amino acid sequence patterns overrepresented in proteins that do not possess significant sequence similarity, but share the same interacting partner (Yaffe 2006). Web services like EML (http://elm.eu.org (Puntervoll et al. 2003)), support the identification of linear motifs in protein sequences.

As described above, specific adapter domains can mediate protein–protein interactions. While some of these interaction domains recognize small target peptides, others

are involved in domain–domain interactions. As short binding motifs have a rather high probability of being found by chance and the exact mechanisms of binding specificity for this mode of interaction are not understood completely, predictions of protein–protein interactions based on binding domains is currently limited to domain–domain interactions for which reliable data is available.

Predicting PPIs from domain interactions may simply be achieved by reversing the ideas discussed above, that is, by using the domain composition of proteins to evaluate the interaction likelihood of proteins (Bock and Gough 2001; Sprinzak and Margalit 2001; Wojcik and Schachter 2001). In a naive approach, domain interactions are treated as independent, and all protein pairs with a matching pair of interacting domains are predicted to engage in an interaction. Given that protein interactions may also be mediated by several domain interactions simultaneously, more advanced statistical methods take into account dependencies between domains and exploit domain combinations (Han et al. 2004) and multiple interacting domain pairs (Chen and Liu 2005).

Exercising and validating these prediction approaches revealed that the most influential factor for PPI prediction is the quality of the underlying data. This suggests that, as for most biological predictions in other fields, the future of prediction methods for protein and domain interactions may lie in the integration of different sources of evidence and weighting the individual contributions based on calibration to gold-standard data. Further methodological improvements may include the explicit consideration of cooperative domains, that is, domain pairs that jointly interact with other domains (Wang et al. 2007b).

15 Biomolecular docking

Basic interactions between two or up to a few biomolecules are the basic elements of the complex molecular interaction networks that enable the processes of life and, when thrown out of their intended equilibrium, manifest the molecular basis of diseases. Such interactions are at the basis of the formation of metabolic, regulatory or signal transduction pathways. Furthermore the search for drugs boils down to analyzing the interactions between the drug molecule and the molecular target to which it binds, which is often a protein.

For the analysis of a single molecular interaction, we do not need complex biological screening data. Thus it is not surprising that the analysis of the interactions between two molecules, one of them being a protein, has the longest tradition in computational biology of all problems involving molecular interactions, dating back over three decades. The basis for such analysis is the knowledge of the three-dimensional structure of the involved molecules. To date, such knowledge is based almost exclusively on experimental measurements, such as X-ray diffraction data or NMR spectra. There are

also a few reported cases in which the analysis of molecular interactions based on structural models of protein has led to successes.

The analysis of the interaction of two molecules based on their three-dimensional structure is called molecular docking. The input is composed of the three-dimensional structures of the participating molecules. (If the involved molecule is very flexible one admissible structure is provided.) The output consists of the three-dimensional structure of the molecular complex formed by the two molecules binding to each other. Furthermore, usually an estimate of the differential free energy of binding is given, that is, the energy difference ΔG between the bound and the unbound conformation. For the binding event to be favorable that difference has to be negative.

15.1 Protein-ligand docking

This slight misnomer describes the binding between a protein molecule and a small molecule. The small molecule can be a natural substrate such as a metabolite or a molecule to be designed to bind tightly to the protein such as a drug molecule. Protein-ligand docking is the most relevant version of the docking problem because it is a useful help in searching for new drugs. Also, the problem lends itself especially well to computational analysis, because in pharmaceutical applications one is looking for small molecules that are binding very tightly to the target protein, and that do so in a conformation that is also a low-energy conformation in the unbound state. Thus, subtle energy differences between competing ligands or binding modes are not of prime interest. For these reasons there is a developed commercial market for protein-ligand docking software.

Usually the small molecule has a molecular weight of up to several hundred Daltons and can be quite flexible. Typically, the small molecule is given by its 2D structure formula, e.g., in the form of a SMILES string (Weininger 1988). If a starting 3D conformation is needed there is special software for generating such a conformation (see, e.g. (Pearlman 1987; Sadowski et al. 1994)).

Challenges of the protein ligand problem are (i) finding the correct conformation of the usually highly flexible ligand in the binding site of the protein, (ii) determining the subtle conformational changes in the binding site of the protein upon binding of the ligand, which are termed induced fit, (iii) producing an accurate estimate of the differential energy of binding or at least ranking different conformations of the same ligand and conformations of different ligands correctly by their differential energy of binding. Methods tackling problem (ii) can also be used to rectify smaller errors in structural models of proteins whose structure has not been resolved experimentally. The solution of problem (iii) provides the essential selection criterion for preferred ligands and binding modes, namely those with lowest differential energy of binding.

Challenge (i) has basically been conquered in the last decade as a number of docking programs have been developed that can efficiently sample the conformational space of

the ligand and produce correct binding modes of the ligand within the protein, assuming that the protein is given in the correct structure for binding the ligand. Several methods are applied here. The most brute-force method is to just try different (rigid) conformations of the ligand one after the other. If the program is fast enough one can run through a sizeable number of conformations per ligand (McGann et al. 2003). A more algorithmic and quite successful method is to build up the ligand from its molecular fragments inside the binding pocket of the protein (Rarey et al. 1996). Yet another class of methods sample ligand conformations inside the protein binding pocket by methods such as local search heuristics, Monte Carlo sampling or genetic algorithms (Abagyan et al. 1994; Jones et al. 1997; Morris et al. 1998). There are also programs exercising combinations of different methods (Friesner et al. 2004). The reported methods usually can compute the binding mode of a ligand inside a protein within fractions of a minute to several minutes. The resulting programs can be applied to screening through large databases of ligands involving hundreds of thousands to millions of compounds and are routinely used in pharmaceutical industry in the early stages of drug design and selection. They are also repeatedly compared on benchmark datasets (Kellenberger et al. 2004; Chen et al. 2006; Englebienne et al. 2007). More complex methods from computational biophysics, such as molecular dynamics (MD) simulations that compute a trajectory of the molecular movement based on the forces exerted on the molecules take hours on a single problem instance and can only be used for final refinement of the complex.

Challenges (ii) and (iii) have not been solved yet. Concerning problem (ii), structural changes in the protein can involve redirections of side chains in or close to the binding pocket and more substantial changes involving backbone movement. While recently methods have been developed to optimize side-chain placement upon ligand binding (Claußen et al. 2001; Sherman et al. 2006), the problem of finding the correct structural change upon binding involving backbone and side-chain movement is open (Carlson 2002). Concerning problem (iii), there are no scoring functions to date that are able to sufficiently accurately estimate the differential energy of binding on a diverse set of protein-ligand complexes (Wang et al. 2003; Huang and Zou 2006). This is especially unfortunate as an inaccurate estimate of the binding energy causes the docking program to disregard correct complex structures even though they have been sampled by the docking program because they are labeled with incorrect energies. This is the major problem in docking which limits the accuracy of the predictions. Recent reviews on protein-ligand docking have been published in Sousa et al. (2006) and Rarey et al. (2007).

One restriction with protein-ligand docking as it applies to drug design and selection is that the three-dimensional structure of the target protein needs to be known. Many pharmaceutical targets are membrane-standing proteins for which we do not have the three-dimensional structure. For such proteins there is a version of drug screening that can be viewed as the negative imprint of docking: Instead of docking the

drug candidate into the binding site of the protein – which is not available – we superpose the drug candidate (which is here called the test molecule) onto another small molecule which is known to bind to the binding site of the protein. Such a molecule can be the natural substrate for the target protein or another drug targeting that protein. Let us call this small molecule the reference molecule. The suitability of the new drug candidate is then assessed on the basis of its structural and chemical similarity with the reference molecule. One problem is that now both the test molecule and the reference molecule can be highly flexible. But in many cases largely rigid reference molecules can be found, and in other cases it suffices to superpose the test moelcule onto any low-energy conformation of the reference molecule. There are several classes of drug screening programs based on this molecular comparison, ranging from (i) programs that perform a detailed analysis of the three-dimensional structures of the molecules to be compared (e.g. (Lemmen et al. 1998; Krämer et al. 2003)) across (ii) programs that perform a topological analysis of the two molecules (Rarey and Dixon 1998; Gillet et al. 2003) to (iii) programs that represent both molecules by binary or numerical property vectors which are compared with string methods (McGregor and Muskal 1999; Xue et al. 2000). The first class of programs require fractions of seconds to fractions of a minute for a single comparison, the second can perform hundreds comparisons per second, the third up to several ten thousand comparisons per second. Reviews of methods for drug screening based on ligand comparison are given in (Lengauer et al. 2004; Kämper et al. 2007).

15.2 Protein–protein docking

Here both binding partners are proteins. Since drugs tend to be small molecules this version of the docking problem is not of prime interest in drug design. Also, the energy balance of protein–protein binding is much more involved that for protein-ligand binding. Optimal binding modes tend not to form troughs in the energy landscape that are as pronounced as for protein-ligand docking. The binding mode is determined by subtle side-chain rearrangements of both binding partners that implement the induced fit along typically quite large binding interfaces. The energy balance is dominated by difficult to analyze entropic terms involving the desolvation of water within the binding interface. For these reasons, the software landscape for protein–protein docking is not as well developed as for protein-ligand docking and there is no commercial market for protein–protein docking software.

Protein–protein docking approaches are based either on conformational sampling and MD – which can naturally incorporated molecular flexibility but suffers from very high computing demands – or on combinatorial sampling with both proteins considered rigid in which case handling of protein flexibility has to be incorporated with methodical extensions. For space reasons we do not detail methods for protein–protein docking. A recent review on the subject can be found in Hildebrandt et al. (2007).

A variant of protein–protein docking is protein-DNA docking. This problem shares with protein–protein docking the character that both binding partners are macromolecules. However, entropic aspects of the energy balance are even more dominant in protein-DNA docking than in protein–protein docking. Furthermore DNA can assume nonstandard shapes when binding to proteins which deviate much more from the known double helix than we are used to when considering induced fit phenomena.

References

Abagyan R, Totrov M, Kuznetsov D (1994) ICM-a method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comput Chem 15: 488–506

Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y (2006) Gene prioritization through genomic data fusion. Nat Biotechnol 24: 537–544

Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118: 4947–4957

Aloy P, Ceulemans H, Stark A, Russell RB (2003) The relationship between sequence and interaction divergence in proteins. J Mol Biol 332: 989–998

Aloy P, Russell RB (2004) Ten thousand interactions for the molecular biologist. Nat Biotechnol 22: 1317–1321

Aloy P, Russell RB (2006) Structural systems biology: modelling protein interactions. Nat Rev Mol Cell Biol 7: 188–197

Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32: D226–D229

Apic G, Gough J, Teichmann SA (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol 310: 311–325

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29

Assenov Y, Ramirez F, Schelhorn SE, Lengauer T, Albrecht M (2008) Computing topological parameters of biological networks. Bioinformatics 24: 282–284

Bader GD, Betel D, Hogue CW (2003) BIND: the Biomolecular interaction network database. Nucleic Acids Res 31: 248–250

Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101–113

Barker D, Meade A, Pagel M (2007) Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. Bioinformatics 23: 14–20

Barker D, Pagel M (2005) Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput Biol 1: e3

Barrios-Rodiles M, Brown KR, Ozdamar B, Bose R, Liu Z, Donovan RS, Shinjo F, Liu Y, Dembowy J, Taylor IW, Luga V, Przulj N, Robinson M, Suzuki H, Hayashizaki Y, Jurisica I, Wrana JL (2005) High-throughput mapping of a dynamic signaling network in mammalian cells. Science 307: 1621–1625

Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35: D301–D303

Bock JR, Gough DA (2001) Predicting protein–protein interactions from primary structure. Bioinformatics 17: 455–460

Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32: D267–D270

Bouwmeester T, Bauch A, Ruffner H, Angrand PO, Bergamini G, Croughton K, Cruciat C, Eberhard D, Gagneur J, Ghidelli S, Hopf C, Huhse B, Mangano R, Michon AM, Schirle M, Schlegl J, Schwab M, Stein MA, Bauer A, Casari G, Drewes G, Gavin AC, Jackson DB, Joberty G, Neubauer G, Rick J, Kuster B, Superti-Furga G (2004) A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. Nat Cell Biol 6: 97–105

Bowers PM, Cokus SJ, Eisenberg D, Yeates TO (2004) Use of logic relationships to decipher protein network organization. Science 306: 2246–2249

Breitkreutz BJ, Stark C, Tyers M (2003) The GRID: the General Repository for Interaction Datasets. Genome Biol 4: R23

Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N, Davey M, Parkinson J, Greenblatt J, Emili A (2005) Interaction network containing conserved and essential protein complexes in *Escherichia coli.* Nature 433: 531–537

Calderwood MA, Venkatesan K, Xing L, Chase MR, Vazquez A, Holthaus AM, Ewence AE, Li N, Hirozane-Kishikawa T, Hill DE, Vidal M, Kieff E, Johannsen E (2007) Epstein-Barr virus and virus human protein interaction maps. Proc Natl Acad Sci USA 104: 7606–7611

Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, Cho RJ, Chen RO, Brownstein BH, Cobb JP, Tschoeke SK, Miller-Graziano C, Moldawer LL, Mindrinos MN, Davis RW, Tompkins RG, Lowry SF (2005) A network-based analysis of systemic inflammation in humans. Nature 437: 1032–1037

Camargo LM, Collura V, Rain JC, Mizuguchi K, Hermjakob H, Kerrien S, Bonnert TP, Whiting PJ, Brandon NJ (2006) Disrupted in Schizophrenia 1 Interactome: evidence for the close connectivity of risk genes and a potential synaptic basis for schizophrenia. Mol Psychiatry 12: 74–86

Carlson HA (2002) Protein flexibility is an important component of structure-based drug discovery. Curr Pharm Des 8: 1571–1578

Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (2007) MINT: the Molecular INTeraction database. Nucleic Acids Res 35: D572–D574

Chen H, Lyne PD, Giordanetto F, Lovell T, Li J (2006) On evaluating molecular-docking methods for pose prediction and enrichment factors. J Chem Inf Model 46: 401–415

Chen XW, Liu M (2005) Prediction of protein–protein interactions using random decision forest framework. Bioinformatics 21: 4394–4400

Chen Y, Dokholyan NV (2006) The coordinated evolution of yeast proteins is constrained by functional modularity. Trends Genet 22: 416–419

Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3: 140

Claußen H, Buning C, Rarey M, Lengauer T (2001) FlexE: efficient molecular docking considering protein structure variations. J Mol Biol 308: 377–395

Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD (2007) Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2: 2366–2382

Colland F, Daviet L (2004) Integrating a functional proteomic approach into the target discovery process. Biochimie 86: 625–632

Colland F, Rain JC, Gounon P, Labigne A, Legrain P, De Reuse H (2001) Identification of the Helicobacter pylori anti-sigma28 factor. Mol Microbiol 41: 477–487

Cui Q, Ma Y, Jaramillo M, Bari H, Awan A, Yang S, Zhang S, Liu L, Lu M, O'Connor-McCourt M, Purisima EO, Wang E (2007) A map of human cancer signaling. Mol Syst Biol 3: 152

Cusick ME, Klitgord N, Vidal M, Hill DE (2005) Interactome: gateway into systems biology. Hum Mol Genet 14 Spec No. 2: R171–R181

Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23: 324–328

Date SV, Marcotte EM (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. Nat Biotechnol 21: 1055–1062

Deng M, Mehta S, Sun F, Chen T (2002) Inferring domain–domain interactions from protein–protein interactions. Genome Res 12: 1540–1548

Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. Febs J 272: 5129–5148

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755–763

Englebienne P, Fiaux H, Kuntz DA, Corbeil CR, Gerber-Lemaire S, Rose DR, Moitessier N (2007) Evaluation of docking programs for predicting binding of Golgi alpha-mannosidase II inhibitors: a comparison with crystallography. Proteins 69: 160–176

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 86–90

Fields S, Song O (1989) A novel genetic system to detect protein–protein interactions. Nature 340: 245–246

Finn RD, Marshall M, Bateman A (2005) iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions. Bioinformatics 21: 410–412

Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34: D247–D251

Fishman MC, Porter JA (2005) Pharmaceuticals: a new grammar for drug discovery. Nature 437: 491–493

Flajolet M, Rotondo G, Daviet L, Bergametti F, Inchauspe G, Tiollais P, Transy C, Legrain P (2000) A genomic approach of the hepatitis C virus generates a protein interaction map. Gene 242: 369–379

Franke L, Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 78: 1011–1025

Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47: 1739–1749

Fryxell KJ (1996) The coevolution of gene family trees. Trends Genet 12: 364–369

Gaasterland T, Ragan MA (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. Microb Comp Genomics 3: 199–217

Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, Mishra G, Nandakumar K, Shen B, Deshpande N, Nayak R, Sarker M, Boeke JD, Parmigiani G, Schultz J, Bader JS, Pandey A (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet 38: 285–293

Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631–636

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B,

Kuster B, Neubauer G, Superti-Furga G (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141–147

Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D'Souza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabasi AL, Oltvai ZN, Osterman AL (2003) Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. J Bacteriol 185: 5673–5684

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Guldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. Nature 418: 387–391

Giallourakis C, Henson C, Reich M, Xie X, Mootha VK (2005) Disease gene discovery through integrative genomics. Annu Rev Genomics Hum Genet 6: 381–406

Gillet VJ, Willett P, Bradshaw J (2003) Similarity searching using reduced graphs. J Chem Inf Comput Sci 43: 338–345

Giot L (2003) A protein interaction map of Drosophila melanogaster. Science 302: 1727–1736

Goehler H, Lalowski M, Stelzl U, Waelter S, Stroedicke M, Worm U, Droege A, Lindenberg KS, Knoblich M, Haenig C, Herbst M, Suopanki J, Scherzinger E, Abraham C, Bauer B, Hasenbank R, Fritzsche A, Ludewig AH, Bussow K, Coleman SH, Gutekunst CA, Landwehrmeyer BG, Lehrach H, Wanker EE (2004) A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. Mol Cell 15: 853–865

Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE (2000) Co-evolution of proteins with their interaction partners. J Mol Biol 299: 283–293

Goll J, Uetz P (2007) Analyzing Protein Interaction Networks. In: Lengauer T (ed) Bioinformatics – from genomes to therapies. Wiley-VCH, Weinheim, pp 1121–1179

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL (2007) The human disease network. Proc Natl Acad Sci USA 104: 8685–8690

Graham DL, Lowe PN, Grime GW, Marsh M, Rittinger K, Smerdon SJ, Gamblin SJ, Eccleston JF (2002) MgF(3)(-) as a transition state analog of phosphoryl transfer. Chem Biol 9: 375–381

Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35: D291–D297

Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V (2006) MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res 34: D436–D441

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33: D514–D517

Han DS, Kim HS, Jang WH, Lee SD, Suh JK (2004) PreSPI: a domain combination based prediction system for protein–protein interaction. Nucleic Acids Res 32: 6312–6320

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach

B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32 (Database issue): D258–D261

Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402: C47–C52

He X, Zhang J (2006) Why do hubs tend to be essential in protein networks? PLoS Genet 2: e88

Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SG, Sander C, Bork P, Zhu W, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe B, Hogue C, Apweiler R (2004) The HUPO PSI's molecular interaction format – a community standard for the representation of protein interaction data. Nat Biotechnol 22: 177–183

Hildebrandt A, Kohlbacher O, Lenhof H-P (2007) Modeling protein–protein and protein-DNA docking. In: Lengauer T (ed) Bioinformatics – from genomes to therapies. Wiley-VCH, Weinheim, pp 601–650

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415: 180–183

Hu Z, Mellor J, Wu J, Kanehisa M, Stuart JM, DeLisi C (2007) Towards zoomable multidimensional maps of the cell. Nat Biotechnol 25: 547–554

Hu Z, Mellor J, Wu J, Yamada T, Holloway D, Delisi C (2005) VisANT: data-integrating visual framework for biological networks and modules. Nucleic Acids Res 33: W352–W357

Huang SY, Zou X (2006) An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. J Comput Chem 27: 1876–1882

Huynen MA, Bork P (1998) Measuring genome evolution. Proc Natl Acad Sci USA 95: 5849–5856

Itzhaki Z, Akiva E, Altuvia Y, Margalit H (2006) Evolutionary conservation of domain–domain interactions. Genome Biol 7: R125

Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. Science 302: 449–453

Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411: 41–42

Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267: 727–748

Jones RB, Gordus A, Krall JA, MacBeath G (2006) A quantitative protein interaction network for the ErbB receptors using protein microarrays. Nature 439: 168–174

Jonsson PF, Bates PA (2006) Global topological features of cancer proteins in the human interactome. Bioinformatics 22: 2291–2297

Jothi R, Cherukuri PF, Tasneem A, Przytycka TM (2006) Co-evolutionary analysis of domains in interacting proteins reveals insights into domain–domain interactions mediating protein–protein interactions. J Mol Biol 362: 861–875

Juan D, Pazos F, Valencia A (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proc Natl Acad Sci USA 105: 934–939

Kaltenbach LS, Romero E, Becklin RR, Chettier R, Bell R, Phansalkar A, Strand A, Torcassi C, Savage J, Hurlburt A, Cha GH, Ukani L, Chepanoske CL, Zhen Y, Sahasrabudhe S, Olson J, Kurschner C, Ellerby LM, Peltier JM, Botas J, Hughes RE (2007) Huntingtin interacting proteins are genetic modifiers of neurodegeneration. PLoS Genet 3: e82

Kämper A, Rognan D, Lengauer T (2007) Lead Identification by virtual screaning. In: Lengauer T (ed) Bioinformatics – from genomes to therapies. Wiley-VCH, Weinheim, pp 651–704

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36: D480–D484

Kann MG (2007) Protein interactions and disease: computational approaches to uncover the etiology of diseases. Brief Bioinform 8: 333–346

Kann MG, Jothi R, Cherukuri PF, Przytycka TM (2007) Predicting protein domain interactions from coevolution of conserved regions. Proteins 67: 811–820

Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK (2007) Relating protein pharmacology by ligand chemistry. Nat Biotechnol 25: 197–206

Kellenberger E, Rodrigo J, Muller P, Rognan D (2004) Comparative evaluation of eight docking tools for docking and virtual screening accuracy. Proteins 57: 225–242

Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H (2007a) IntAct–open source resource for molecular interaction data. Nucleic Acids Res 35: D561–D565

Kerrien S, Orchard S, Montecchi-Palazzi L, Aranda B, Quinn AF, Vinod N, Bader GD, Xenarios I, Wojcik J, Sherman D, Tyers M, Salama JJ, Moore S, Ceol A, Chatr-Aryamontri A, Oesterheld M, Stumpflen V, Salwinski L, Nerothin J, Cerami E, Cusick ME, Vidal M, Gilson M, Armstrong J, Woollard P, Hogue C, Eisenberg D, Cesareni G, Apweiler R, Hermjakob H (2007b) Broadening the Horizon – Level 2.5 of the HUPO-PSI Format for Molecular Interactions. BMC Biol 5: 44

Kitano H (2007) A robustness-based approach to systems-oriented drug design. Nat Rev Drug Discov 6: 202–210

Korbel JO, Doerks T, Jensen LJ, Perez-Iratxeta C, Kaczanowski S, Hooper SD, Andrade MA, Bork P (2005) Systematic association of genes to phenotypes by genome and literature mining. PLoS Biol 3: e134

Korbel JO, Jensen LJ, von Mering C, Bork P (2004) Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nat Biotechnol 22: 911–917

Krämer A, Horn HW, Rice JE (2003) Fast 3D molecular superposition and similarity search in databases of flexible molecules. J Comput Aided Mol Des 17: 13–18

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae.* Nature 440: 637–643

Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P (2008) STITCH: interaction networks of chemicals and proteins. Nucleic Acids Res 36: D684–D688

LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth JR, Schoenfeld LW, Ota I, Sahasrabudhe S, Kurschner C, Fields S, Hughes RE (2005) A protein interaction network of the malaria parasite Plasmodium falciparum. Nature 438: 103–107

Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, Moreau Y, Brunak S (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 25: 309–316

Legrain P, Selig L (2000) Genome-wide protein interaction maps using two-hybrid systems. FEBS Lett 480: 32–36

Lemmen C, Lengauer T, Klebe G (1998) FLEXS: a method for fast flexible ligand superposition. J Med Chem 41: 4502–4520

Lengauer T, Lemmen C, Rarey M, Zimmermann M (2004) Novel technologies for virtual screening. Drug Discov Today 9: 27–34

Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P (2006) SMART 5: domains in the context of genomes and networks. Nucleic Acids Res 34: D257–D260

Lim J, Hao T, Shaw C, Patel AJ, Szabo G, Rual JF, Fisk CJ, Li N, Smolyar A, Hill DE, Barabasi AL, Vidal M, Zoghbi HY (2006) A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell 125: 801–814

Lin J, Gan CM, Zhang X, Jones S, Sjoblom T, Wood LD, Parsons DW, Papadopoulos N, Kinzler KW, Vogelstein B, Parmigiani G, Velculescu VE (2007) A multidimensional analysis of genes mutated in breast and colorectal cancers. Genome Res 17: 1304–1318

Liu M, Liberzon A, Kong SW, Lai WR, Park PJ, Kohane IS, Kasif S (2007) Network-based analysis of affected biological processes in type 2 diabetes models. PLoS Genet 3: e96

Loscalzo J, Kohane I, Barabasi AL (2007) Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Mol Syst Biol 3: 124

Lu X, Jain VV, Finn PW, Perkins DL (2007) Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. Mol Syst Biol 3: 98

Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH (2007) CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res 35: D237–D240

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D (1999a) Detecting protein function and protein–protein interactions from genome sequences. Science 285: 751–753

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999b) A combined algorithm for genome-wide prediction of protein function. Nature 402: 83–86

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999c) A combined algorithm for genome-wide prediction of protein function [see comments]. Nature 402: 83–86

McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK (2003) Gaussian docking functions. Biopolymers 68: 76–90

McGregor MJ, Muskal SM (1999) Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. J Chem Inf Comput Sci 39: 569–574

Mintseris J, Weng Z (2005) Structure, function, and evolution of transient and obligate protein–protein interactions. Proc Natl Acad Sci USA 102: 10930–10935

Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavnath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, Kumar HG, Nagini M, Kumar GS, Jose R, Deepthi P, Mohan SS, Gandhi TK, Harsha HC, Deshpande KS, Sarker M, Prasad TS, Pandey A (2006) Human protein reference database–2006 update. Nucleic Acids Res 34: D411–D414

Morett E, Korbel JO, Rajan E, Saab-Rincon G, Olvera L, Olvera M, Schmidt S, Snel B, Bork P (2003) Systematic discovery of analogous enzymes in thiamin biosynthesis. Nat Biotechnol 21: 790–795

Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. J Compu Chem 19: 1639–1662

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2007) New developments in the InterPro database. Nucleic Acids Res 35: D224–D228

Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, Gibson TJ, Lewis J, Serrano L, Russell RB (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol 3: e405

Ng SK, Zhang Z, Tan SH (2003a) Integrative approach for computationally inferring protein domain interactions. Bioinformatics 19: 923–929

Ng SK, Zhang Z, Tan SH, Lin K (2003b) InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. Nucleic Acids Res 31: 251–254

Nicastro G, Menon RP, Masino L, Knowles PP, McDonald NQ, Pastore A (2005) The solution structure of the Josephin domain of ataxin-3: structural determinants for molecular recognition. Proc Natl Acad Sci USA 102: 10493–10498

Noirot P, Noirot-Gros MF (2004) Protein interaction networks in bacteria. Curr Opin Microbiol 7: 505–512

Nooren IM, Thornton JM (2003) Diversity of protein–protein interactions. EMBO J 22: 3486–3492

Oda K, Kitano H (2006) A comprehensive map of the toll-like receptor signaling network. Mol Syst Biol 2: 2006 0015

Oda K, Matsuoka Y, Funahashi A, Kitano H (2005) A comprehensive pathway map of epidermal growth factor receptor signaling. Mol Syst Biol 1: 2005 0010

Orchard S, Kerrien S, Jones P, Ceol A, Chatr-Aryamontri A, Salwinski L, Nerothin J, Hermjakob H (2007a) Submit Your Interaction Data the IMEx Way: a Step by Step Guide to Trouble-free Deposition. Proteomics: 28–34

Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stumpflen V, Ceol A, Chatr-aryamontri A, Armstrong J, Woollard P, Salama JJ, Moore S, Wojcik J, Bader GD, Vidal M, Cusick ME, Gerstein M, Gavin AC, Superti-Furga G, Greenblatt J, Bader J, Uetz P, Tyers M, Legrain P, Fields S, Mulder N, Gilson M, Niepmann M, Burgoon L, De Las Rivas J, Prieto C, Perreau VM, Hogue C, Mewes HW, Apweiler R, Xenarios I, Eisenberg D, Cesareni G, Hermjakob H (2007b) The minimum information required for reporting a molecular interaction experiment (MIMIx). Nat Biotechnol 25: 894–898

Orengo CA, Thornton JM (2005) Protein families and their evolution-a structural perspective. Annu Rev Biochem 74: 867–900

Oti M, Brunner HG (2007) The modular nature of genetic diseases. Clin Genet 71: 1–11

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) Use of contiguity on the chromosome to predict functional coupling. In Silico Biol 1: 93–108

Pacifico S, Liu G, Guest S, Parrish JR, Fotouhi F, Finley RL Jr (2006) A database and tool, IM Browser, for exploring and integrating emerging gene and protein interaction data for Drosophila. BMC Bioinformatics 7: 195

Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stumpflen V, Mewes HW, Ruepp A, Frishman D (2005) The MIPS mammalian protein–protein interaction database. Bioinformatics 21: 832–834

Pagel P, Oesterheld M, Tovstukhina O, Strack N, Stumpflen V, Frishman D (2007) DIMA 2.0 predicted and known domain interactions. Nucleic Acids Res 36: D651–D655

Pagel P, Wong P, Frishman D (2004) A domain interaction map based on phylogenetic profiling. J Mol Biol 344: 1331–1346

Pages S, Belaich A, Belaich JP, Morag E, Lamed R, Shoham Y, Bayer EA (1997) Species-specificity of the cohesin-dockerin interaction between Clostridium thermocellum and Clostridium cellulolyticum: prediction of specificity determinants of the dockerin domain. Proteins 29: 517–527

Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. Nat Biotechnol 24: 805–815

Park J, Lappe M, Teichmann SA (2001) Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. J Mol Biol 307: 929–938

Parrish JR, Yu J, Liu G, Hines JA, Chan JE, Mangiola BA, Zhang H, Pacifico S, Fotouhi F, Dirita VJ, Ideker T, Andrews P, Finley RL Jr (2007) A proteome-wide protein interaction map for Campylobacter jejuni. Genome Biol 8: R130

Pasek S, Bergeron A, Risler JL, Louis A, Ollivier E, Raffinot M (2005) Identification of genomic features using microsyntenies of domains: domain teams. Genome Res 15: 867–874

Pawson T, Nash P (2003) Assembly of cell regulatory systems through protein interaction domains. Science 300: 445–52

Pazos F, Ranea JA, Juan D, Sternberg MJ (2005) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J Mol Biol 352: 1002–1015

Pazos F, Valencia A (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Eng 14: 609–614

Pearlman RS (1987) Rapid generation of high quality approximate 2-dimension molecular structures. Chem Des Auto News 2: 1–6

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 96: 4285–4288

Pujana MA, Han JD, Starita LM, Stevens KN, Tewari M, Ahn JS, Rennert G, Moreno V, Kirchhoff T, Gold B, Assmann V, Elshamy WM, Rual JF, Levine D, Rozek LS, Gelman RS, Gunsalus KC, Greenberg RA, Sobhian B, Bertin N, Venkatesan K, Ayivi-Guedehoussou N, Sole X, Hernandez P, Lazaro C, Nathanson KL, Weber BL, Cusick ME, Hill DE, Offit K, Livingston DM, Gruber SB, Parvin JD, Vidal M (2007) Network modeling links breast cancer susceptibility and centrosome dysfunction. Nat Genet 39: 1338–1349

Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 31: 3625–3630

Raghavachari B, Tasneem A, Przytycka TM, Jothi R (2008) DOMINE: a database of protein domain interactions. Nucleic Acids Res 36: D656–D661

Ralser M, Albrecht M, Nonhoff U, Lengauer T, Lehrach H, Krobitsch S (2005) An integrative approach to gain insights into the cellular function of human ataxin-2. J Mol Biol 346: 203–214

Ramírez F, Schlicker A, Assenov Y, Lengauer T, Albrecht M (2007) Computational analysis of human protein interaction networks. Proteomics 7: 2541–2552

Rarey M, Degen J, Reulecke I (2007) Docking and scoring for structure-based drug design. In: Lengauer T (ed) Bioinformatics – from genomes to therapies. Wiley-VCH, Weinheim, pp 541–600

Rarey M, Dixon JS (1998) Feature trees: a new molecular similarity measure based on tree matching. J Comput Aided Mol Des 12: 471–490

Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261: 470–489

Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B (1999) A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotechnol 17: 1030–1032

Riley R, Lee C, Sabatti C, Eisenberg D (2005) Inferring protein domain interactions from databases of interacting proteins. Genome Biol 6: R89

Ruepp A, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Stransky M, Waegele B, Schmidt T, Doudieu ON, Stumpflen V, Mewes HW (2008) CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res 36: D646–D650

Ruffner H, Bauer A, Bouwmeester T (2007) Human protein–protein interaction networks and the value for drug discovery. Drug Discov Today 12: 709–716

Sadowski J, Gasteiger J, Klebe G (1994) Comparison of automatic three-dimensional models builders using 639 X-ray structures. J Chem Inf Comput Sci 34: 1000–1008

Saito R, Suzuki H, Hayashizaki Y (2002) Interaction generality, a measurement to assess the reliability of a protein–protein interaction. Nucleic Acids Res 30: 1163–1168

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32: D449–D451

Santonico E, Castagnoli L, Cesareni G (2005) Methods to reveal domain networks. Drug Discov Today 10: 1111–1117

Sato T, Yamanishi Y, Horimoto K, Kanehisa M, Toh H (2006) Partial correlation coefficient between distance matrices as a new indicator of protein–protein interactions. Bioinformatics 22: 2488–2492

Sato T, Yamanishi Y, Kanehisa M, Toh H (2005) The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 21: 3482–3489

Schlicker A, Huthmacher C, Ramirez F, Lengauer T, Albrecht M (2007) Functional evaluation of domain–domain interactions and human protein interaction networks. Bioinformatics 23: 859–865

Schuster-Bockler B, Bateman A (2007) Reuse of structural domain–domain interactions in protein networks. BMC Bioinformatics 8: 259

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504

Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T (2005) Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci U S A 102: 1974–1979

Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. Mol Syst Biol 3: 88

Sherman W, Day T, Jacobson MP, Friesner RA, Farid R (2006) Novel procedure for modeling ligand/receptor induced fit effects. J Med Chem 49: 534–553

Snel B, Huynen MA (2004) Quantifying modularity in the evolution of biomolecular systems. Genome Res 14: 391–397

Sousa SF, Fernandes PA, Ramos MJ (2006) Protein-ligand docking: Current status and future challenges. Proteins 65: 15–26

Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci USA 100: 12123–12128

Sprinzak E, Margalit H (2001) Correlated sequence-signatures as markers of protein–protein interaction. J Mol Biol 311: 681–692

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34: D535–D539

Stein A, Russell RB, Aloy P (2005) 3did: interacting protein domains of known three-dimensional structure. Nucleic Acids Res 33: D413–D417

Stelzl U, Wanker EE (2006) The value of high quality protein–protein interaction networks for systems biology. Curr Opin Chem Biol 10: 551–558

Suderman M, Hallett M (2007) Tools for visually exploring biological networks. Bioinformatics 23: 2651–2659

Tewari M, Hu PJ, Ahn JS, Ayivi-Guedehoussou N, Vidalain PO, Li S, Milstein S, Armstrong CM, Boxem M, Butler MD, Busiguina S, Rual JF, Ibarrola N, Chaklos ST, Bertin N, Vaglio P, Edgley ML, King KV, Albert PS, Vandenhaute J, Pandey A, Riddle DL, Ruvkun G, Vidal M (2004) Systematic interactome mapping and genetic perturbation analysis of a *C. elegans* TGF-β signaling network. Mol Cell 13: 469–482

Uetz P (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. Nature 403: 623–627

Uetz P, Dong YA, Zeretzke C, Atzler C, Baiker A, Berger B, Rajagopala SV, Roupelieva M, Rose D, Fossum E, Haas J (2006) Herpesviral protein networks and their interaction with the human proteome. Science 311: 239–242

Uetz P, Rajagopala SV, Dong YA, Haas J (2004) From ORFeomes to protein interaction maps in viruses. Genome Res 14: 2029–2033

Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L (2007) Reactome: a knowledge base of biologic pathways and processes. Genome Biol 8: R39

Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA (2004) Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol 14: 208–216

von Brunn A, Teepe C, Simpson JC, Pepperkok R, Friedel CC, Zimmer R, Roberts R, Baric R, Haas J (2007) Analysis of intraviral protein–protein interactions of the SARS coronavirus ORFeome. PLoS ONE 2: e459

von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P (2007) STRING 7 – recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 35: D358–D362

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Comparative assessment of large-scale data sets of protein–protein interactions. Nature 417: 399–403

Wang H, Segal E, Ben-Hur A, Li QR, Vidal M, Koller D (2007a) InSite: a computational method for identifying protein–protein interaction binding sites on a proteome-wide scale. Genome Biol 8: R192

Wang R, Lu Y, Wang S (2003) Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem 46: 2287–2303

Wang RS, Wang Y, Wu LY, Zhang XS, Chen L (2007b) Analysis on multi-domain cooperation for predicting protein–protein interactions. BMC Bioinformatics 8: 391

Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction and encoding rules. J Chem Inf Comput Sci 28: 31–36

Wells JA, McClendon CL (2007) Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. Nature 450: 1001–1009

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35: D5–D12

Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Liebundguth N, Lockhart DJ, Lucau-Danila A, Lussier M, M'Rabet N, Menard P, Mittmann M, Pai C, Rebischung C, Revuelta JL, Riles L, Roberts CJ, Ross-MacDonald P, Scherens B, Snyder M, Sookhai-Mahadeo S, Storms RK, Veronneau S, Voet M, Volckaert G, Ward TR, Wysocki R, Yen GS, Yu K, Zimmermann K, Philippsen P, Johnston M, Davis RW (1999) Functional char-

acterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science 285: 901–906

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

Xia Y, Yu H, Jansen R, Seringhaus M, Baxter S, Greenbaum D, Zhao H, Gerstein M (2004) Analyzing cellular biochemistry in terms of molecular networks. Annu Rev Biochem 73: 1051–1087

Xu J, Li Y (2006) Discovering disease-genes by topological features in human protein–protein interaction network. Bioinformatics 22: 2800–2805

Xue H, Xian B, Dong D, Xia K, Zhu S, Zhang Z, Hou L, Zhang Q, Zhang Y, Han JD (2007) A modular network model of aging. Mol Syst Biol 3: 147

Xue L, Godden JW, Bajorath J (2000) Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. J Chem Inf Comput Sci 40: 1227–1234

Yaffe MB (2006) "Bits" and pieces. Sci STKE 2006: pe28

Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M (2007) Drug-target network. Nat Biotechnol 25: 1119–1126

Yook SH, Oltvai ZN, Barabasi AL (2004) Functional and topological characterization of protein interaction networks. Proteomics 4: 928–942

Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput Biol 3: e59

Zarrinpar A, Bhattacharyya RP, Lim WA (2003) The structure and function of proline recognition domains. Sci STKE 2003: RE8