

Protein-protein interactions: structure and systems approaches to analyze diverse genomic data

Anna R. Panchenko

National Center for Biotechnology Information, NIH, USA

Benjamin A. Shoemaker

National Center for Biotechnology Information, NIH, USA

Outline

It is now becoming clear that proteins interact with each other in a highly specific and regular manner determining the outcome of most cell processes, such as replication, transcription, translation, signal transduction and others. Distortion of normal protein-protein interfaces lead to the development of many known diseases. Therefore the study of protein-protein interactions is essential for understanding the mechanisms of biological processes, for elucidating the nature of various human diseases and can provide the clues to cure and prevent them. Protein-protein interactions are remarkably diverse making it very difficult to formulate general principles of formation and to develop reliable methods for prediction. In this review we attempt to classify and systemize the array of experimental and theoretical data on the identification and prediction of protein-protein interactions and their networks.

| | |
|---|----|
| Outline..... | 2 |
| Introduction..... | 4 |
| Experimental methods to identify protein-protein interactions | 5 |
| Yeast two-hybrid experiments (Y2H)..... | 5 |
| <i>Developments and variations of Y2H system</i> | 6 |
| <i>Disadvantages of the Y2H method</i> | 7 |
| <i>Advantages of Y2H</i> | 7 |
| Mass spectroscopy | 8 |
| TAP method of complex purification | 9 |
| Comparison between Y2H and TAP/MS..... | 10 |
| Correlation between gene expression and protein interactions..... | 10 |
| Verification of protein-protein interactions | 11 |
| Protein and domain interaction databases..... | 12 |
| Protein interaction databases..... | 12 |
| <i>Database of Interacting Proteins (DIP)</i> | 12 |
| <i>Biomolecular Interaction Network Database (BIND)</i> | 13 |
| <i>Munich MPact/MIPS database</i> | 13 |
| Domain interaction databases | 13 |
| <i>InterDom database</i> | 13 |
| <i>PIBASE database</i> | 14 |
| <i>3did database</i> | 14 |
| <i>Conserved Binding Mode (CBM) database</i> | 14 |
| <i>iPfam database</i> | 15 |
| <i>Domain Interaction Map (DIMA) database</i> | 15 |
| Methods of prediction of protein-protein interactions | 15 |
| Phylogenetic profile method | 15 |
| Rosetta Stone approach..... | 16 |
| Gene neighbor and gene cluster methods | 17 |
| Co-evolution of interacting proteins and correlated mutations methods | 17 |
| Classification methods | 19 |
| Predicting domain interactions from protein interactions..... | 20 |
| Association method..... | 20 |
| Maximum likelihood estimation method (MLE)..... | 21 |
| Domain pair exclusion analysis (DPEA) | 22 |
| Calculating P-values to predict domain interactions | 23 |
| Integrative method | 24 |
| Homology modeling | 24 |
| Automated complex modeling methods | 25 |
| CAPRI docking contest..... | 26 |
| Designed interfaces | 26 |
| Properties of protein interaction networks | 27 |
| Scale-free behavior of protein interaction networks..... | 27 |
| Conservation and alignment of protein interaction networks | 28 |
| Acknowledgements..... | 29 |
| References..... | 30 |

Introduction

Protein-protein interactions are remarkably diverse making it very difficult to formulate general principles of formation and to develop reliable methods for prediction. In this review we attempt to classify and systemize the array of experimental and theoretical data on the identification and prediction of protein-protein interactions and their networks.

Protein-protein recognition is determined by structural and physico-chemical properties of two interacting proteins and their interacting interfaces. It was reported earlier that the majority of protein complexes have a buried surface area of about $1600 \pm 400 \text{ \AA}^2$ (a “standard size” patch) and association does not involve large conformational changes of interacting proteins (Lo Conte et al. 1999). Large complexes with interfaces larger than the “standard size”, on the contrary, involve major conformational changes which are especially important in signal transduction. Moreover, in the following paper from the same group, the authors showed that large interfaces involve more than one interaction patch and multipatch interfaces usually contain two patches of at least “standard size” (Chakrabarti and Janin 2002). The authors (Bogan and Thorn 1998; Chakrabarti and Janin 2002) proposed a model of a protein recognition site which consists of a completely buried core and partially accessible rim. The amino acid composition of cores differs considerably from that of rims with some common features observed by various methods (Jones and Thornton 1997a; Bogan and Thorn 1998; Chakrabarti and Janin 2002). For example, amino acids Trp and Tyr (and also Met, Cys and Phe to much less extent) are abundant in the core, but Ser and Thr, Lys and Glu are particularly disfavored.

Protein-protein interactions can be categorized into different types depending on their strength (permanent and transient), the location of interacting partners within one or between two polypeptide chains and the similarity between interacting subunits (homo- and hetero-oligomers). It has been shown that different interface types are significantly different in amino acid composition so that it is possible to predict the type of interaction interface from amino acid composition alone with 63-100% accuracy (Ofra and Rost 2003). In addition to this, interactions formed by hydrophobic residues are more frequent in permanent interactions than in transient ones. Moreover, the same authors showed that disulfide bridges are observed very often on all types of interfaces, but salt bridges are not commonly found for homo-oligomers. Interestingly enough, the homo-oligomers have a significant excess of residue contacts involving identical residues which can be explained by the fact that

non-identical residue contacts would require two beneficial compensatory mutations to preserve the binding interface between the same chains, rather than just one mutation in the case of identical residue contacts (Ofra and Rost 2003).

Since proteins interact in a regular manner, there should be a certain degree of conservation in the interaction patterns between similar proteins and domains. Some previous studies pointed out that homodimer interface conservation is higher than expected by chance even for transient complexes (Valdar and Thornton 2001; Nooren and Thornton 2003), but nevertheless the conservation of protein interfaces is very weak compared to the rest of a protein (Grishin and Phillips 1994; Caffrey et al. 2004; Korkin et al. 2005). Poor conservation of interfaces can be the reason for low prediction accuracy of protein-protein interaction sites (Jones and Thornton 1997b; Panchenko et al. 2004).

A comprehensive analysis of interface conservation has been done on a test set including all protein domains from the Protein DataBank (PDB) (Aloy et al. 2003). The authors compared interactions by calculating root-mean-square-deviation between structure superpositions of two instances of domains on each other. They showed that if the measure of interaction similarity is plotted against the sequence identity between domains, the following pattern can be observed. Close homologs almost always interact the same way, while domains belonging to the same SCOP (Andreeva et al. 2004) fold but different superfamily categories have different interaction modes. In another study the authors examined conserved binding modes in pairs of interacting domains (Shoemaker et al. 2006) and found that interfaces between different functional subfamilies of the globin family are poorly conserved while interfaces within the same subfamily are well conserved and thereby can be used in homology modeling.

Experimental methods to identify protein-protein interactions

Yeast two-hybrid experiments (Y2H)

The yeast two-hybrid system was originally developed by Fields and Song (Fields and Song 1989) and later was advanced to analyse genome sequence data (Auerbach et al. 2002; Fields 2005). It is based on the fact that many eukaryotic transcription activators (ex: Gal4 eukaryotic transcription factor or bacterial repressor protein LexA) have at least two distinct domains, one that directs binding to a

promoter DNA sequence (BD) and another that activates transcription (AD). Fields and Song demonstrated that the DNA-binding domain can not activate transcription at a promoter unless physically (not necessarily covalently) associated with an activating domain. A protein of interest is fused to a DNA-binding domain (bait), this chimeric protein is cloned in an expression plasmid and then is transfected into a yeast cell. A similar procedure is performed to create a chimeric sequence of another protein which is fused to AD (prey). If two proteins physically interact, this causes the activation of the reporter gene in vivo.

One example of a Y2H system is the transcription activation system of the LacZ gene in yeast. Yeast promoters have TATA box regions and cis-regulatory elements (upstream activating sequences, UAS). UAS sequences are recognized by specific transcriptional activators, for example, by proteins GAL4. GAL4 proteins control in yeast the expression of proteins which participate in galactose metabolism, in particular, the expression of LacZ gene which codes for the beta-galactosidase. Target protein sequences are fused with the binding and activation domains of GAL4 proteins. If there is no galactose, GAL80 binds to GAL4 and blocks the transcription. When galactose is present GAL80 is removed from GAL4 activation domain and GAL4 can activate the transcription of beta-galactosidase. Expression is detected by turning cell colonies blue after exposing to 5-bromo-4-chloro-3-indolyl beta-D-galactoside. To avoid the interference by the natural GAL4 proteins, yeast host cells used in Y2H carry deletions of the GAL4 and GAL80 genes.

Developments and variations of Y2H system

- Yeast strains are developed to carry several reporter genes (lacZ, HIS3, LEU2 ...)
- Haploid yeast strains are developed with opposite mating type. Baits are transformed into yeast cells with one mating type, preys are transformed into another mating type, then the diploid cells are produced by mating these cells containing both baits and preys.
- One-hybrid system detects interactions between a prey protein and known DNA sequence (bait).
- RNA yeast three-hybrid system detects interactions between RNA and proteins. The bait RNA is a hybrid between the target RNA and MS2 RNA that can bind the MS2 coat protein. MS2 coat protein is fused into LexA BD.

- Protein yeast three-hybrid system detects the formation of complexes between several proteins.

Disadvantages of the Y2H method

- The interactions can not be tested if a protein under question can initiate transcription by itself.
- Fusion of a protein into another protein (chimeras) can change the structure of a test protein and effect its folding.
- Some cDNAs are fractional and do not represent the full length sequence of a target protein. In some cases a fragment of a protein might interact with another protein while the whole protein does not.
- Posttranslational modifications (formation of disulfide bridges, phosphorylation, glycosylation) which can alter interaction interfaces can occur differently in yeast and other organisms (but yeast is used as a host).
- Since two-hybrid reactions occur in the yeast nucleus, it is difficult to target extracellular proteins.
- A third protein can bridge the interactions between the bait and the prey.
- Proteins which can in general interact in Y2H experiments, may never interact in a cell due to different cell localizations or different expression times.

Advantages of Y2H

- This is an *in vivo* technique, so it is closer to the processes which occur in living cells of higher eukaryotes, compared to the techniques based on bacterial expression.
- Transient interactions between proteins can be detected due to the amplification of a signal by the reporter gene expression, Y2H can predict the affinity of an interaction.
- Fast and efficient.

Two approaches have been used for genome-wide analysis by Y2H: matrix-based and library-based:

- *matrix approach*: a matrix of prey clones is created, each yeast clone expressing each Y-AD protein in one well of a plate. Then this matrix of prey clones is added to the matrix of clones expressing a particular X-BD protein. Those diploids where X

and Y interact are selected based on the expression of a reporter gene (ex: producing blue color for beta-galactosidase).

- *library approach*: one bait X is screened against an entire library (library can contain random cDNA fragments or ORFs). Diploid positives are selected based on their ability to grow on specific substrates, sequences of interacting proteins are determined by DNA-sequencing. Since protein interactions very often can be detected by using protein fragments rather than the full-length proteins (if proteins are misfolded for example), the library-based approach is more sensitive than the matrix-based approach.

Two major genome-wide analyses of the yeast “interactome” revealed 692 and 841 putative interactions involving about 800 proteins (Uetz et al. 2000; Ito et al. 2001). The overlap between these two experimental studies was not very large, both methods shared only 141 interactions which constitutes about 20% of all interaction data (Ito et al. 2001).

Mass spectroscopy

Mass spectroscopy (MS) used in conjunction with complex purification is a powerful method to study macromolecular interactions. The principle of the MS method is to produce ions which then can be detected based on their mass-to-charge ratio thereby allowing the identification of polypeptide sequences (Causier 2004; Di Tullio et al. 2005). First, proteins are degraded enzymatically to peptides. The sample is evaporated into a vacuum, then an electron beam is used to fragment the sample into a set of pieces and those carrying a net charge are detected and separated based on their mass-to charge ratio. The detector measures the number of ions with a given mass-to-charge ratio. The fragmentations occur primarily at peptide bonds and the mass-to-charge ratios can be measured with an accuracy of less than 1 dalton per charge.

Electrospray ionization MS (ESI-MS) (Whitehouse et al. 1985) has been developed to produce isolated ions in the gas phase of large biomolecules. According to this technique protein molecules in an acidic solution are sprayed into a mass spectrometer under a strong electrical field; the solvent evaporates rapidly in a vacuum and protein molecules with a net positive charge become ionized in the gas phase. Integral net charges are assigned to different peaks of spectra. ESI-MS has proven to be very useful for detecting molecular recognition via noncovalent bonding

and therefore can be applied to analyze protein-protein interactions and large protein complexes. Another method of ionization called MALDI (Matrix Assisted Laser Desorption Ionization) uses proteins embedded on matrix which is bombarded by the laser to produce ions (Pieles et al. 1993).

Different algorithms have been developed to analyze a large number of peptide spectra to identify proteins by their sequence. Cross-correlation methods, for example, (Pevzner et al. 2000) find correlations between theoretical and experimental spectra while others using *de novo* algorithms infer peptide sequences from theoretical interpretation of the MS spectra (Taylor and Johnson 1997). Another group of algorithms for MS interpretation calculate the statistical significance of a match between the mass-to-charge ratios of experimentally produced peptides and the theoretical peptides produced by *in silico* digestion of a protein sequence library (Yates et al. 1995; Geer et al. 2004).

MS is a powerful method to decipher protein-protein interactions, but it has been shown that the limiting step in complex characterization is not in protein identification, but rather in protein complex purification. In this connection a tandem affinity purification method (TAP) has been developed.

TAP method of complex purification

A TAP tag consists of two IgG binding domains of Staphylococcus protein A and a calmodulin binding peptide, separated by the tobacco etch virus protease cleavage site (Rigaut et al. 1999). A target protein ORF is fused with the DNA sequences encoding TAP tag. The tagged ORFs are expressed in yeast cells and form native complexes with other proteins in a cell. At the first step of the TAP purification, protein A binds tightly to an IgG matrix and after washing out the contaminants protease cleaves the link between protein A and IgG matrix. The eluate of this first step is then incubated with calmodulin-coated beads in the presence of Ca. After washing, the target protein complex is released and the components of each complex are found by polyacrylamide gel electrophoresis. Protein bands are excised and corresponding proteins are cleaved by proteases. The resulting fragments are analyzed by MS and identified by bioinformatics methods.

In yeast several large-scale studies of protein complexes have been performed using TAP/MS methods (Gavin et al. 2002; Ho et al. 2002; Krogan et al. 2006). Gavin et al, for example, identified 1440 interacting proteins from 232

multiprotein complexes, and proposed new cellular roles for 344 proteins (Gavin et al. 2002). Ho et al identified 1578 interacting proteins (Ho et al. 2002). A more recent analysis showed that 7123 protein-protein interactions identified with high confidence can be clustered into 547 protein complexes, with about half of them absent from MIPS (Krogan et al. 2006). Comparative analysis of human and yeast complexes showed that orthologous proteins interact with complexes enriched by orthologs; essential gene products are more likely to interact with essential rather than nonessential proteins (Gavin et al. 2002).

Comparison between Y2H and TAP/MS

- Both methods generate a lot of false positives, both methods miss a lot of known interactions (false negatives).
- Y2H produces binary interactions, does not provide information about protein complexes, but can detect transient interactions.
- MS can detect large stable complexes and networks of interactions.

Correlation between gene expression and protein interactions

Since the function of a protein complex depends on the functionality of all subunits, the independent expression of each gene/subunit would not be efficient. Therefore, there should exist a relationship between gene expression levels of subunits in a complex. The large-scale study of whole-genome expression data in the context of protein-protein interactions has been performed (Jansen et al. 2002). The authors analyzed protein complexes from the MIPS catalog (Guldener et al. 2006) while expression profiles were taken from two different sources: cell cycle experiments and the Rosetta yeast compendium (Hughes et al. 2000). Cell cycle data comprised expression profiles obtained from synchronized cells in two cell cycles while Rosetta data contained expression ratios for the overall yeast genome for 300 stationary cell states.

The relationship between gene expressions was calculated as the difference between absolute expression levels as: $D = |E_i - E_j| / (E_i + E_j)$, where E_i and E_j are mRNA expression levels of protein subunits “i” and “j”. This quantity is calculated for all proteins in a complex and then the distribution of “D” is compared to the distribution of “D” for random gene/protein pairs. Another way to calculate the correlation between the expression profiles is to refer to their relative expression

levels rather than the absolute ones. In these cases the measure of similarity is chosen as a Pearson correlation coefficient between the two expression profiles.

The coexpression method was tested on specific complexes: ribosome, proteasome, RNA Polymerase II Holoenzyme and replication complex. It was found that the subunits from the same complex with the most obvious coexpression come from permanent complexes such as ribosome and proteasome. Some transient complexes can be subdivided into smaller permanent complexes, which show strong correlation with gene expression. It was also shown that for genome-wide Y2H data, there is only a weak correlation with the gene expression.

Verification of protein-protein interactions

Several methods have been proposed for verification of protein-protein interaction data (Deane et al. 2002; Sprinzak et al. 2003; Bader et al. 2004). Some of them are described here.

1. *Expression profile reliability method (EPR)* is based on the observation that interacting proteins are coexpressed. The distance between expression profiles of two proteins, A and B, can be calculated as:

$$d_{AB}^2 = \sum_i (\log(e_i^A / e_{ref}^A) - \log(e_i^B / e_{ref}^B))^2$$

Here each term in the sum is the log ratio of expression levels of a protein under condition “i”. Then the distributions of d^2 for non-interacting (ρ_n) and interacting proteins (ρ_i) are compared (reliable interactions taken from DIP-YEAST). Based on these two distributions one can define a parameter α which would characterize the accuracy of a given data set (for example Y2H data), or correspond to the fraction of false positives:

$$\rho_{exp}(d_{AB}^2) = \alpha \cdot \rho_i(d_{AB}^2) + (1 - \alpha) \cdot \rho_n(d_{AB}^2)$$

The parameter α can be obtained by fitting the expression protein interaction data distribution $P_{exp}(d^2)$.

2. *Paralogous verification method (PVM)* is based on the observation that if two proteins interact, their paralogs would most likely interact, it calculates the number of interactions between two families of paralogous proteins. This method identifies ~40% of true interactions at 1% error rate. Using PVM and EPR methods about 50% of DIP interactions can be considered reliable.

3. *Protein localization method.* True positives are defined as interacting proteins which are localized in the same cellular compartment and/or interacting proteins that are annotated to have a common cellular role (Sprinzak et al. 2003). The accuracy strongly depends on the method with up to 50% of true positives detected in Y2H experiments and up to 100% true positives detected in immunological experiments (coimmunoprecipitation is a method of detecting interacting proteins by removing them from solution after adding a specific antibody).

Comparing protein-protein interaction data is difficult as various techniques and methods have different goals, the data are obtained under different conditions and for different organisms. For example, none of the methods cover more than 60% of proteins in the yeast genome (von Mering et al. 2002). The low coverage can be explained by different factors:

- proteins form transient complexes in a cell which are difficult to identify;
- proteins behave differently in different parts of the cell, genome-scale cellular location assays provide data on the protein location;
- if two proteins separately interact on the same face of a third protein, the three proteins must not interact at the same time.
- Ancient, evolutionary conserved proteins have much better coverage than the proteins restricted to a certain organism.

Protein and domain interaction databases

Protein interaction databases

Database of Interacting Proteins (DIP)

DIP contains experimentally-determined protein-protein interactions and includes a core subset of interactions which have passed a quality assessment (Salwinski et al. 2004). Interaction data are obtained from literature; Protein

Databank (PDB); and high-throughput methods like Y2H, protein microarrays, and TAP / MS analysis of protein complexes. Several methods are employed to assess the quality of interaction data and are offered as a service for query interactions. DIP has links to a couple of related databases including LiveDIP, which records information about the state of a biological interaction, such as covalently modified, conformational or cellular location states (Duan et al. 2002). Another database related to DIP is Prolinks which brings together four methods of linking proteins: phylogenetic profiles, Rosetta Stone, gene neighbors and gene clusters (Bowers et al. 2004). The database includes a Proteome Navigator tool to browse the linkages and view accompanying data. DIP and related databases can be accessed at <http://dip.doe-mbi.ucla.edu>.

Biomolecular Interaction Network Database (BIND)

BIND includes high-throughput experimental datasets and protein complexes from PDB (Bader and Hogue 2000; Alfarano et al. 2005). It contains a large variety of experimental interaction data curated by an in-house team of curators. A generalized data specification was developed to handle not only various types of protein-protein interaction data, but also protein-small molecule interactions and protein-nucleic acid interactions. An Interaction Viewer is provided to browse the interaction space. BIND uses a grammar of unique icons to distinguish functional types of interactions in displays. Web access (with user registration) is at <http://www.bind.ca>.

Munich MPact/MIPS database

MPact is a resource to access MIPS, which contains a manually curated yeast protein-protein interaction dataset (Guldener et al. 2006). This set of 4,300 different interactions from 1,500 proteins has been collected by curators from the literature. The resource also includes high-throughput results for yeast, but keeps this data separate. Web-based analysis and visualization tools are available at <http://mips.gsf.de/services/ppi>.

Domain interaction databases

InterDom database

InterDom collects evidence for predicting protein domain interactions from a number of sources (Ng et al. 2003b). These sources include PDB, literature, protein

interactions stored in DIP and BIND as well as instances of domain fusion. The reliability of domain interactions is scored depending on the number/type of experimental evidence for each interaction. Web access can be found at <http://interdom.lit.org.sg>.

PIBASE database

PIBASE is a database of domain interactions from the protein structure data (Davis and Sali 2005). It uses SCOP and CATH domain definitions to find putative domain-domain interactions. Structural comparisons of interfaces are made for the same domain pair within one structure to remove redundancy. The database combines physicochemical properties of protein binding sites and has a link to MODBASE (Pieper et al. 2006) containing modeled three-dimensional structures which allows one to model putative interacting domain interfaces. Web access is at <http://alto.compbio.ucsf.edu/pibase>.

3did database

3did allows one to explore the details of domain interactions from protein structure data (Stein et al. 2005). For a particular domain an overview is given of all domain interactions, showing whether each occurs inter-chain, intra-chain, or both. A more detailed view is shown for a particular structure with lines connecting domains in different chains. Tables for a given domain list structures and domain information. In some cases dot plots of structural comparisons show the variance of the interactions between pairs of families. GO-based functional annotations and yeast interactions are also present in the database.

InterPreTS is a web-based service to predict domain interactions based on sequence homology of query proteins to a database of interacting domains (DBID) associated with the 3did database (Aloy and Russell 2003). Web access for 3did and InterPreTS can be found at <http://3did.embl.de>.

Conserved Binding Mode (CBM) database

The Conserved Binding Mode (CBM) database is a collection of domain-domain interactions from the structure data grouped by geometry into conserved interaction modes for each pair of domain families across all PDB structures (Shoemaker et al. 2006). Structural superpositions are used to infer CBMs from different members of interacting domain families docking in the same way. Such domain interactions with recurring structural themes have greater significance to be biologically relevant, unlike spurious crystal packing interactions. CBMs highlight the

commonalities and variation of a domain pair's interactions from all structural examples. Types of interacting domain pairs range from 1,000 (conserved) to 2,000 (all). Currently the CBM database is available by ftp download from the NCBI ftp site: <ftp://ftp.ncbi.nlm.nih.gov/pub/cbm>.

iPfam database

iPfam displays the interactions of Pfam domains from the PDB (Finn et al. 2005). The system is integrated into the Pfam website and allows for interactive browsing of all Pfam-Pfam domain interactions detected on PDB structures at the family and individual structure levels. Web access is at <http://www.sanger.ac.uk/Software/Pfam/iPfam>.

Domain Interaction Map (DIMA) database

DIMA is a domain interaction map derived from phylogenetic profiling Pfam domains (Pagel et al. 2006). Instead of looking at entire protein sequences, the algorithm compares the occurrences of domains across genomes and associates them for interaction with similar patterns of conservation. The method works well for domains with moderate information content which have distinct phylogenetic profiles. Web access is at <http://mips.gsf.de/genre/proj/dima/index.html>.

Methods of prediction of protein-protein interactions

Phylogenetic profile method

Pioneered by the work by (Pellegrini et al. 1999) the phylogenetic profile method is based on the hypothesis that functionally linked and possibly interacting proteins have orthologs in the same subset of fully sequenced organisms. Indeed, for many pathways and complexes all components should be present simultaneously in order to perform its function. A phylogenetic profile is constructed for each protein, using a vector of N elements, where N is the number of genomes. The presence/absence of a given protein in a given genome is indicated as “1” or “0” at each position of a profile. Proteins or their profiles are then clustered using bit-distance and those proteins from the same cluster are considered functionally related or interacting. One drawback of this method is that it is computationally expensive and ubiquitous proteins present in all genomes (profiles will have all “1”s) have very small distances between profiles which would result in a large number of false positives. The same is true for proteins which are specific to a given genome (the

profiles will have all but one “0”s). Function of genes and genetic map can be also identified by phylogenetic profiling of nonessential gene deletions. The method of synthetic lethality, for example, describes the genetic interaction when two non-lethal mutations results in lethality when combined at the same time (Bender and Pringle 1991; Brown et al. 2006; Ooi et al. 2006).

The idea of phylogenetic profiles can be applied to protein domains instead of entire proteins. In this case a profile is constructed for each domain (PFAM, SMART) and the presence/absence of this domain in different genomes is recorded (Pagel et al. 2004). This results in a domain interaction map (DIMA). This method can avoid computationally expensive all versus all sequence searches and can give information about domain-domain interactions. The method utilizes entropy filtering; and profiles with low information content are excluded. Performance is assessed by comparing the profile distance distribution for protein pairs known to interact to the distance distribution of random protein pairs. Limitation of DIMA is that domain databases are not complete and no predictions can be made for almost half of proteins. Another problem includes the presence of specialized domains which are found only in a few genomes. Major drawbacks of all phylogenetic profile approaches are that they can not make reliable predictions for low information profiles and that they rely on homology detection between distant organisms.

Rosetta Stone approach

The Rosetta Stone approach infers protein interactions from protein sequences in different genomes (Marcotte et al. 1999). It is based on the observation that some pairs of interacting domains have homologs which are fused into one protein chain, a so called Rosetta Stone protein. In *E.coli*, for example, this method found 6809 potentially interacting pairs of non-homologous proteins, both proteins from each pair had significant sequence similarity to a single protein from some other genome. Analysis of pairs found by this approach revealed that for more than half of them both members of a pair are functionally related and therefore this method can be used for inferring functional similarity. Comparison with the experimental data on protein-protein interactions from DIP database showed that about 6.4% of all experimental interactions can be linked by Rosetta Stone sequences.

Gene neighbor and gene cluster methods

Bacterial genes with closely related functions are often transcribed as a single unit – an operon. Different methods try to predict operon structures based on intergenic distances (Ermolaeva et al. 2001; Bowers et al. 2004). A systematic comparison of bacterial and archaeal genomes reveals some conservation of gene-order and operon structure (Dandekar et al. 1998; Overbeek et al. 1999; Galperin and Koonin 2000; Bowers et al. 2004). Gene pairs from conserved gene clusters appear to encode proteins which physically interact in a cell. It has been shown that gene order between the prokaryotic and archaeal species is conserved if the sequence identity shared by orthologs in two genomes is higher than 50%. Conservation of gene order can also be used to predict gene function by inferring its function from the functions of neighboring genes.

Co-evolution of interacting proteins and correlated mutations methods

Interacting protein or domain pairs very often coevolve and in these cases the phylogenetic trees of interacting partners show some degree of similarity. The similarity between phylogenetic trees can be quantified by calculating the correlation coefficient between the distance matrices used to construct the trees (Goh et al. 2000; Pazos and Valencia 2001). For example, the active site of phosphoglycerate kinase is formed by two domains and therefore the working enzyme required these two domains to coevolve. In other words, any changes in one domain which would lead to the loss of activity should be compensated by the correlated changes in another domain. To quantify co-evolution, first, the pairwise evolutionary distances between all members of each family of interacting proteins are calculated. For example, X_{ij} is a pairwise distance between sequences s_i and s_j from a family of one potentially interacting partner and Y_{ij} is the distance between sequences h_i and h_j of another interacting protein family, where sequences s_i and h_i are taken from the same species. Next, the correlation coefficient is calculated between two matrices X_{ij} and Y_{ij} and its large values indicate the coevolution between two proteins.

In order to compare phylogenetic trees one needs to know the corresponding branches of the two trees, but such information is not always available. Several computational methods have been developed to identify specific interaction partners between two interacting families (Gertz et al. 2003; Ramani and Marcotte 2003; Jothi et al. 2005). This is especially useful when both families contain paralogs with

different binding specificities. According to these methods, given a pair of proteins known to interact, their similarity matrices are aligned using simulated annealing algorithm to minimize the root mean square difference between the elements of two matrices. Then interactions are predicted between proteins corresponding to the aligned columns of two matrices. It has been shown that the prediction accuracy strongly depends on the phylogenetic tree complexity (measures how close is the tree to the radial one): as the tree complexity increases, the accuracy increases (Ramani and Marcotte 2003). A more formal measure of tree complexity was introduced in another paper (Jothi et al. 2005)

Gertz et al (Gertz et al. 2003) implemented similar Monte Carlo schemes to align two matrices with the preliminary clustering of proteins within the matrices. Protein clustering using the UPGMA method allowed to compare matrices with different dimensions and helped find biologically relevant one-to-many correspondence between proteins from two families. It should be mentioned, that all previously described methods can not perform the search successfully if the size of families is large (more than 30 proteins in a family as noted by Ramani and Marcotte). One way to reduce the search space is to use the information encoded in the phylogenetic tree (Jothi et al. 2005). In this case local minima can be avoided by swapping the whole isomorphic subtrees in a single move instead of a single column in the course of the Monte Carlo algorithm.

The similarity between two phylogenetic trees is influenced by the speciation process and therefore there is a certain “background” similarity between trees of any proteins, no matter if they interact or not. The following methods have been introduced to account for this background similarity (Pazos et al. 2005; Sato et al. 2005). According to the first method (Pazos et al. 2005) multiple alignments of orthologous sequences are constructed for all proteins under interest. At the next step the phylogenetic trees are made from the multiple sequence alignments and the evolutionary distances between the proteins in the alignment are calculated by summing up the branch lengths separating each pair of sequences. The “background” tree is constructed from the 16S rRNA sequences and is considered to be a canonical tree of life. The distance matrices used in the study are obtained by subtracting the normalized and rescaled rRNA based distances from the distances obtained for the original phylogenetic trees. Finally, the corrected distance matrices of two proteins are compared by calculating correlation coefficients or interaction scores. It has been shown that the method finds half of real interacting proteins at

6.4% false positive rate which is a higher accuracy than with the 16.5% false positive rate obtained using methods which compare phylogenetic trees without taking into account explicit evolutionary distances and “background” canonical tree (Goh et al. 2000; Pazos and Valencia 2001).

Classification methods

Different classification methods have been proposed for the prediction of protein interactions (Jansen et al. 2003; Chen and Liu 2005; Qi et al. 2005). These methods use different biological data sources including direct experimental data and indirect data (for example protein coexpression data) on protein interactions to train a classifier to distinguish between positive examples of truly interacting protein pairs from the negative examples of non-interacting pairs. Each protein or protein sequence is encoded as a feature vector where features may represent different information sources on protein-protein interactions such as gene coexpression of two proteins, domain-domain interactions and evidence coming from various experimental methods. As a result of a comparison of different classifiers, it has been shown that Random Forest classifiers outperform other methods with the Support Vector Machine being in second place (Qi et al. 2006). Moreover, by examining different feature combinations the same authors found that the importance of features in correct classification depends on the type of prediction problem. Namely, if it is a prediction of physically interacting proteins, co-complex interactions or pathway co-membership, gene expression was one of the most important features for all prediction tasks.

One of the Random Decision Forest methods introduced recently builds decision trees based on the domain composition of interacting and non-interacting proteins, explores all possible combinations of interacting domains and predicts at the end if a given pair of proteins interact (Chen and Liu 2005). Each protein pair is represented as a vector of length N , where N is the number of different domain types (features), where each feature can have values 2, 1 or 0 depending if this domain is found in both proteins, in one of them or not found in the protein pair. Given a training set of interacting protein pairs taken from the experimental data, the method constructs a decision tree (or many trees) which defines the best splitting feature at each node from a randomly selected feature subspace. The best feature is selected based on the measure of “goodness of fit” which estimates how well this feature can

discriminate between two classes of interacting and non-interacting pairs. The method stops growing the tree as soon as all pairs at a given node are well separated into two classes providing a classification for an unknown protein pair.

Predicting domain interactions from protein interactions

By far the most coverage of experimental data describing protein interaction networks comes from high-throughput experiments giving us the identity of protein pairs detected to interact. Unfortunately, these experiments reveal no structural details about the interaction interfaces and the formation of protein complexes. To deal with these limitations several approaches have been developed to predict domain-domain interactions given a set of experimental protein-protein interactions. The following section gives an overview of the approaches. Most methods begin with protein sequence searches of domains defined by Pfam, SCOP, CDD or other domain databases (Marchler-Bauer et al. 2002; Andreeva et al. 2004; Finn et al. 2006). The methods are trained on protein-protein interactions, typically high-throughput results from yeast or multi-genome data. Predicted domain-domain interactions are evaluated using structural data or by higher quality interaction sets such as MIPS (Guldener et al. 2006). Accounting for domains in proteins can also help in predicting protein interactions. For example it was shown that domain interactions in one organism can be successfully used to predict domain and protein interactions in another organism (Wojcik and Schachter 2001). Treating a protein as a collection of domains allows one to assign different probability values for different protein interactions depending on domain frequency and allows one to use such domain networks with weighted edges to predict protein interactions (Gomez and Rzhetsky 2002).

Association method

The association method was one of the first methods which looked for the characteristic sequence-signatures in a pair of interacting proteins (Sprinzak and Margalit 2001). Correlated sequence-signatures that are found together more often than expected by chance can be used as markers to indicate/predict a new type of protein-protein interaction. The authors used three sets of yeast protein-protein interaction data (including MIPS and DIP) to compute log-odds scores and to find correlated sequence-signatures. Sequence-signatures were defined using InterPro (we refer to them as “domains”). The log-odds score was computed as: $\log_2(P_{ij}/P_iP_j)$,

where P_{ij} is the observed frequency of domains i and j occurring in one protein pair; P_i and P_j are the background frequencies of domains i and j in the data. The average mutual information content calculated per domain was pretty high (2.48 bits) indicating a significant correlation between interacting proteins and predicted domain pairs. Domain-domain interactions were defined as those having positive log-odds scores (greater than 2) and having several instances of occurrence of a given domain pair in the database (more than 5 counts).

Maximum likelihood estimation method (MLE)

The association method proposed earlier considered each pair of interacting domains separately, ignoring other domains in a given pair of interacting proteins. Moreover, the association method did not incorporate the experimental errors of Y2H data into the scoring scheme. To account for this a new Maximum Likelihood Estimation method (MLE) has been developed (Deng et al. 2002). According to this method domain-domain interactions are considered to be independent and proteins interact using at least one pair of domains. The likelihood function is a function of parameters $\theta = (\lambda_{mn}, f_p, f_n)$, where λ_{mn} is the probability that domains m and n interact, f_p is the false positive rate and f_n is the false negative rate derived from the experimental Y2H data (which are fixed to 2.85E-4 and 0.64 respectively). It is difficult to maximize the likelihood function directly because of the large number of parameters (large number of different types of interacting domains). To solve this problem the iterative Expectation Maximization algorithm is used which finds estimates of unknown parameters θ using the complete data (the observed data together with the missing data). This procedure has two steps, expectation and maximization. The first step involves finding the expectation of the complete dataset, given the observed dataset and a set of parameters, θ . In the second step the maximum likelihood estimation of the parameters θ is obtained. Step one starts with initial parameter values and the two recursive steps are iterated until convergence.

The method has been analyzed in several indirect ways. First the accuracy of the method is assessed by predicting protein-protein interactions from the inferred domain-domain interactions and is compared with the experimental Y2H protein interaction data. Using two sets of Y2H data and excluding training data, the authors achieved accuracy with 42.5% specificity and 77.6% sensitivity. When the protein interaction predictions were compared with data derived from the MIPS database, the accuracy was reported to be nearly 100 times better than the accuracy of random predictions. The limitation, however, was that at this level of significance only 0.68%

of the MIPS interactions were predicted correctly, which is very low and might not be useful in practice. The predictions were also compared to pairwise correlation coefficients of gene expression profiles and it was found that the best predictions had higher correlation coefficients than random protein pairs.

Domain pair exclusion analysis (DPEA)

The DPEA method extends the previously described MLE method and can detect specific domain interactions which are hard to detect using MLE (Riley et al. 2005). MLE and other methods emphasize non-specific promiscuous domain interactions which are detected as those having large θ values. On the contrary, specific, rare interactions between certain members of two domain families would be neglected as they would have low values of θ . The DPEA method accounts for this by estimating an E score which is computed as a ratio of the probability that proteins m and n interact given that two domains i and j interact, and the probability that proteins m and n interact given that domains i and j do not interact. The Expectation Maximization procedure, similar to the one described in the previous section, is used to compute E-scores. The major difference between the two implementations of the EM algorithm is that in DPEA an additional step is performed when all instances of interacting domains i and j are excluded by fixing the interaction probability between domains i and j to zero and by allowing the competing domains to maximize θ_{ij} . The change in the likelihood (pointing to the confidence that domains i and j interact) is evaluated and expressed as an E-score.

A high E-score value shows the high propensity of two domains to interact while a low value indicates that competing domains from the same protein pair are more likely to be responsible for this interaction. Therefore, specific domain-domain interactions can be found by screening for low θ values and high E-scores. This model incorporates the protein interaction data from many organisms as present in DIP but does not account for false positives and negatives in the experimental data. The E-score is compared to a log-odds score and θ in terms of correct ranking/predicting physically interacting domains (PFAM-A) in PDB. It was shown that the E-score finds 71 times more true positive domain-domain interactions compared to the random assignments in 100 top predictions. When non-modular domains are excluded, E-scores considerably outperform other scores in predicting structurally interacting domains.

Calculating P-values to predict domain interactions

Another method has proposed a statistical framework to calculate p-values of domain pairs being responsible for protein-protein interactions (Nye et al. 2005). The authors test the null hypothesis that the presence of a particular domain pair in a protein pair has no effect on whether two proteins interact. To test this hypothesis a statistic is calculated for each domain pair which takes into account experimental error (fraction of false positives estimated for each experimental dataset) and incompleteness of the dataset (fraction of false negatives). The reference distribution is simulated by shuffling domains in proteins so that the network of protein interactions remains fixed. P-values show the reliability of domain-domain interactions given that two proteins interact and the domain pair with the lowest p-value is most likely to interact compared to other domain pairs within the interacting proteins. In this approach domains are defined using SCOP superfamily categories and the p-value simulation is performed on the three sets of yeast interaction data. Predictions are tested with domain interactions obtained from the Protein Quaternary Structure (PQS) database, which uses symmetry operations to make PDB protein assemblies more biologically meaningful.

The method has been compared to the Sprinzak association method, to the Deng MLE method and to random domain prediction. The results reveal that the method does better than the others when there are many domains found on a protein pair. Interestingly enough, for the majority of test cases, however, the random domain prediction outperforms all other methods, pointing to the low accuracy of all prediction methods of domain-domain interactions. The major limitations of these methods are:

- domains are assumed to interact independently, although their interactions can depend on other domains in a protein pair;
- if a protein contains several domains of the same type, the scoring schemes can not distinguish between their contacts;
- the gaps between domain assignments can contain another interacting domains and ignoring these gaps can lead to false positive and false negative predictions;
- many proteins can not be assigned any domains;
- methods are based on the assumption that the interacting domains make only one contact which is not true for many multidomain complexes;
- protein interaction data are not complete.

Integrative method

This method used by the web service Interdom combines the contribution of three different sources of data to rank the domain interaction predictions (Ng et al. 2003a). Domain interactions supported by different sources are more reliable and as a consequence should have a higher score compared to the domain interactions supported by one data source. An additive scoring scheme was used which integrated scores from three different data sources. The first score was calculated for domain-domain interactions derived from protein-protein interactions as defined by the DIP database. In this case a scoring scheme was based on odd-ratios and was calculated as a ratio of the observed weighted frequency of domain pairs and the background frequency of domain pair occurrence by chance. The second score was derived from protein complexes (Cellzome yeast protein complexes and PDB complexes) using a similar scoring scheme. The third source of data represented domain fusion events as found by searching SWISS-PROT for a pair of domains which are fused in one organism and are on separate chains in another organism. A probabilistic score could not be calculated in this case so a constant is assigned to the instances of fused domains.

The method was evaluated by looking at the number of protein-protein interactions matching predicted domain interactions in a 20-fold cross validation. It was found that the major improvement in the prediction was made when two sources were used (compared to the case when only protein interaction data were used), namely, the fraction of correctly predicted true positives increased from 39% to 58%, while the error rate did not change considerably (8% to 12%).

Homology modeling

Experimental techniques for protein structure determination have improved to the point that for single proteins, structures are solved quickly and decent coverage of major genomes can be expected in the near future (Aloy et al. 2005). Structure prediction can typically be handled by finding a homologous template to a query in the structure database and making a query model based on this template. The next challenge for protein structure prediction is the prediction of protein-protein interactions and making high-quality models of protein complexes with the ultimate goal of creating representative coverage of all genome protein-protein interactions. It has been estimated that roughly 2,000 out of 10,000 interaction types are known so

far from high-throughput methods (Aloy and Russell 2004). Unfortunately, there are a limited number of protein-protein complexes present in PDB and solving the structures of large complexes meets difficult technical challenges not readily overcome in general (Russell et al. 2004). The most likely path to protein interface characterization of large complexes would therefore involve multiple experimental methods together with homology modeling and docking of structural subunits.

To build a protein complex model, one can start with a set of protein interaction data from high-throughput identification methods such as yeast two-hybrid or affinity purification screens. Protein pairs tagged to interact are searched for homologous domains and evaluated for likely domain-domain interactions. For these proteins or more specifically for the predicted interacting domains, homology searches are made against structure data. In rare cases entire structural complexes of homologous proteins may be found, but sometimes only interacting domain dimers or, more often, single domains can be identified. At the next step, the pieces should be put together while avoiding steric hindrance and maximizing complementarity between interacting domains. In this case the docking potentials can be used to score different orientations between two interacting domains. The success of docking strongly depends on the similarity between the target protein and homologous proteins as well as on the presence of homologous multidomain complexes in the structure database.

Automated complex modeling methods

There are several automated methods available for modeling of protein-protein interactions between proteins X and Y.

- Interprets first matches Pfam domains to target sequences and constructs complexes from structures matching the same type of Pfam domains (Aloy and Russell 2002). The method uses empirical pair potentials from 3did to score putative interactions. It has been shown that the method yields good results for most classes of complexes, but poor predictions for peptidase/inhibitors.
- Multiprospector first threads sequences X and Y separately against a structure database of dimers to find single chains matching target protein sequences (Lu et al. 2003; Grimm et al. 2006). For those template structures which form a complex between X and Y, the method performs additional threading cycle for both proteins X and Y together by fixing the alignment of X to its single chain template and finding an optimal alignment of Y to its template in a complex and vice versa.

The fitness of target sequences into a protein complex is estimated using the conventional threading potential together with a separate score for the interfaces which is derived from the structural dimer database. This method was applied to the yeast genome and predicted 7,321 interactions from 304 complexes. The method was ranked third among large-scale methods of protein interaction prediction, and it has been found it did not bias towards abundant proteins while giving atomic detail of interaction surfaces.

CAPRI docking contest

The contest to critically assess protein interaction predictions (CAPRI) was designed in the spirit of CASP, the protein structure prediction contest, to make blind predictions before a crystal structure of the complex is released. In the CAPRI rounds, predictors build atomic models of complexes given structures of the unbound proteins. In some cases when two proteins bind, their conformations do not change and the prediction accuracy of the complex is very high as was shown in one of the CAPRI experiments (Mendez et al. 2005). However, the backbone of the bound form can significantly deviate from the unbound form and in this case it is difficult to make a correct prediction. For example, for homodimer docking of the PTS regulation domain from LicT (Wodak and Mendez 2004) the conformational changes upon binding two domains were as large as 12Å RMSD per domain.

The CAPRI experiment demonstrated that docking methods have a number of limitations (Wodak and Mendez 2004) which can also restrict the homology modeling methods described earlier:

- proteins can undergo significant conformational changes upon binding;
- docking potentials are not accurate enough;
- specific and non-specific types of protein interactions are not adequately distinguished between each other;

Designed interfaces

It should be mentioned that one area of research related to the prediction of protein interfaces is that of computational interface design. By modifying protein sequences, such as with point mutations and linkers, and subsequently expressing them researchers are able to explore a range of biological activity not found in

nature. For homology modeling of protein interfaces, additional information becomes available for refining potentials and for acceptable domain combinations.

One review of the research (Kortemme and Baker 2004) describes a range of examples from the alteration of oligomeric state in helical bundles (Harbury et al. 1998), to the creation of chimeric proteins through the linking of domains from different functional pathways (Howard et al. 2003). Design methods have increased the specificity of promiscuous domains (Shifman and Mayo 2002), have created novel interactions (Reina et al. 2002), and have automated the process (Havranek and Harbury 2003).

Properties of protein interaction networks

Scale-free behavior of protein interaction networks

For the past five years the scale-free behavior of complex networks has attracted a lot of attention. Many empirical studies indeed showed that the structure of metabolic and protein interaction networks can not be explained by the classical random network model (Barabasi and Albert 1999; Jeong et al. 2000; Wolf et al. 2002). According to the latter, the nodes are connected randomly, leading to the homogeneous network where most nodes have the same number of edges. The degree distribution or connectivity of such a network follows a Poisson distribution and the probability of finding a highly connected node decays exponentially. On the contrary, scale-free networks are highly heterogeneous with a few highly connected nodes (hubs) and a large number of poorly connected nodes. This structure can be explained by the preferential attachment of new vertices to the highly connected node in the network's expansion (Barabasi and Albert 1999). The degree distribution of these networks follows a power-law: $P(k) \sim e^{-k}$ reflecting their self-similarity under scale transformation. Other important properties of the scale-free networks include: small diameter (calculated as an average number of edges in the shortest path connecting two nodes), high tolerance to errors and high susceptibility to attacks. Random errors and removal of random nodes do not affect the diameter of a scale-free network, this property is very important for maintaining the integrity of biological networks upon external changes or errors. On the other hand, if the few, highly connected hubs are removed from the network, the network diameter increases

sharply which leads to the disruption of the network disintegrating it into many isolated clusters.

Indeed, mutagenesis experiments proved that yeast can tolerate mutations/deletions of a large number of proteins from its proteome (Winzeler et al. 1999; Jeong et al. 2001). This in turn implies that less connected proteins should be less essential for a cell compared to highly connected proteins. To answer this question, the yeast protein interaction network has been investigated and shown that proteins characterized by high connectivity are three times more likely to be essential than proteins with few connections (Jeong et al. 2001). Many models have been proposed describing the mechanisms reproducing scale-free protein interaction networks (Qian et al. 2001; Rzhetsky and Gomez 2001; Middendorf et al. 2005; Deeds et al. 2006). For example, according to the duplication-mutation-complementation model (DMC), gene duplication is followed by mutations and diversification, but gene functional complementarity is conserved (if one copy of a gene becomes dysfunctional, another copy can carry its function) (Middendorf et al. 2005). Another model emphasizes the role of desolvation in forming the protein-protein interaction interfaces and predicts the correlation between the number of interactions which a protein makes and the fraction of hydrophobic residues on its surface (Deeds et al. 2006).

Conservation and alignment of protein interaction networks

The fast development of experimental techniques for protein-protein interactions has enabled the construction and systematic analysis of interaction networks or maps of interacting proteins. Interaction maps obtained for one species can be used to predict interaction networks in other species, to predict function of unknown proteins and to get insight into the evolution of protein interaction patterns. The interaction map analyses and comparisons are based on the assumption/observation that many interactions are conserved among species and form so called “interologs” (Walhout et al. 2000). Sequence-based searches for conserved “interologs” were able to identify 16%-31% of true “interologs” (tested using two-hybrid system) even between remotely related species such as yeast and worm (Matthews et al. 2001). Analysis of gene-coexpression networks revealed 22,163 gene pairs coexpressed in humans, flies, worms and yeast (Stuart et al. 2003). The conservation of co-expression patterns among diverse organisms

suggests that these gene pairs correspond to the functionally related genes responsible for core biological processes. Moreover, a multiple-species network has been constructed by identifying pairs of genes with correlated expression in different organisms. It was shown that a multiple-species network performs better than a single-species network in linking together functionally related genes.

To measure the evolutionary distance at the level of network connectivity, a new algorithm of aligning two networks has been developed, called PATHBLAST (Kelley et al. 2003). The method searches for high-scoring pathway alignments between two networks, where proteins are paired based on their sequence similarity. Pathway alignments can allow gaps occurring when one path passes over a protein in another path and can accommodate misalignments occurring between two aligned proteins with low sequence similarity. The network alignment between worm, fly and yeast detected 71 network regions that were conserved between all three species (Sharan et al. 2005). Among these, 94% of the clusters contained at least 50% of proteins sharing the same annotation. Single network analysis of yeast resulted in much lower accuracy of 83%.

Instead of aligning two protein networks, the network topologies also can be compared by calculating the difference between the number of connections of identical proteins from two networks (Hoffmann and Valencia 2003). In this case the correlation coefficients between the protein connectivities of two networks is estimated which in turn quantifies the agreement between the networks obtained by different methods. Although the method can perform only pairwise comparisons, it is not restricted to only conserved interactions but rather can encompass all proteins covered by both methods. Applying this approach to networks obtained by different experimental and *in silico* methods showed that there exists statistically significant correlations between different experimental and theoretical methods, while the gene neighborhood method correlates with both experimental and *in silico* methods.

Acknowledgements

We thank Teresa Przytycka and Lewis Geer and Elena Zotenko for helpful suggestions and discussions.

References

- Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M et al. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* 33(Database issue): D418-424.
- Aloy P, Russell RB (2002) Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A* 99(9): 5896-5901.
- Aloy P, Russell RB (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics* 19(1): 161-162.
- Aloy P, Russell RB (2004) Ten thousand interactions for the molecular biologist. *Nat Biotechnol* 22(10): 1317-1321.
- Aloy P, Pichaud M, Russell RB (2005) Protein complexes: structure prediction challenges for the 21st century. *Curr Opin Struct Biol* 15(1): 15-22.
- Aloy P, Ceulemans H, Stark A, Russell RB (2003) The relationship between sequence and interaction divergence in proteins. *J Mol Biol* 332(5): 989-998.
- Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C et al. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 32(Database issue): D226-229.
- Auerbach D, Thaminy S, Hottiger MO, Stagljar I (2002) The post-genomic era of interactive proteomics: facts and perspectives. *Proteomics* 2(6): 611-623.
- Bader GD, Hogue CW (2000) BIND--a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* 16(5): 465-477.
- Bader JS, Chaudhuri A, Rothberg JM, Chant J (2004) Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol* 22(1): 78-85.
- Barabasi AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439): 509-512.
- Bender A, Pringle JR (1991) Use of a screen for synthetic lethal and multicopy suppressor mutants to identify two new genes involved in morphogenesis in *Saccharomyces cerevisiae*. *Mol Cell Biol* 11(3): 1295-1305.
- Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. *J Mol Biol* 280(1): 1-9.
- Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO et al. (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* 5(5): R35.
- Brown JA, Sherlock G, Myers CL, Burrows NM, Deng C et al. (2006) Global analysis of gene function in yeast by quantitative phenotypic profiling. *Mol Syst Biol* 2: 2006 0001.
- Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci* 13(1): 190-202.
- Causier B (2004) Studying the interactome with the yeast two-hybrid system and mass spectrometry. *Mass Spectrom Rev* 23(5): 350-367.
- Chakrabarti P, Janin J (2002) Dissecting protein-protein recognition sites. *Proteins* 47(3): 334-343.
- Chen XW, Liu M (2005) Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics* 21(24): 4394-4400.
- Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 23(9): 324-328.

- Davis FP, Sali A (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics* 21(9): 1901-1907.
- Deane CM, Salwinski L, Xenarios I, Eisenberg D (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics* 1(5): 349-356.
- Deeds EJ, Ashenberg O, Shakhnovich EI (2006) A simple physical model for scaling in protein-protein interaction networks. *Proc Natl Acad Sci U S A* 103(2): 311-316.
- Deng M, Mehta S, Sun F, Chen T (2002) Inferring domain-domain interactions from protein-protein interactions. *Genome Res* 12(10): 1540-1548.
- Di Tullio A, Reale S, De Angelis F (2005) Molecular recognition by mass spectrometry. *J Mass Spectrom* 40(7): 845-865.
- Duan XJ, Xenarios I, Eisenberg D (2002) Describing biological protein interactions in terms of protein states and state transitions: the LiveDIP database. *Mol Cell Proteomics* 1(2): 104-116.
- Ermolaeva MD, White O, Salzberg SL (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res* 29(5): 1216-1221.
- Fields S (2005) High-throughput two-hybrid analysis. The promise and the peril. *Febs J* 272(21): 5391-5399.
- Fields S, Song O (1989) A novel genetic system to detect protein-protein interactions. *Nature* 340(6230): 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(3): 410-412.
- Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V et al. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* 34(Database issue): D247-251.
- Galperin MY, Koonin EV (2000) Who's your neighbor? New computational approaches for functional genomics. *Nat Biotechnol* 18(6): 609-613.
- Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415(6868): 141-147.
- Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M et al. (2004) Open mass spectrometry search algorithm. *J Proteome Res* 3(5): 958-964.
- Gertz J, Elfond G, Shustrova A, Weisinger M, Pellegrini M et al. (2003) Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics* 19(16): 2039-2045.
- Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE (2000) Co-evolution of proteins with their interaction partners. *J Mol Biol* 299(2): 283-293.
- Gomez SM, Rzhetsky A (2002) Towards the prediction of complete protein-protein interaction networks. *Pac Symp Biocomput*: 413-424.
- Grimm V, Zhang Y, Skolnick J (2006) Benchmarking of dimeric threading and structure refinement. *Proteins* 63(3): 457-465.
- Grishin NV, Phillips MA (1994) The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *Protein Sci* 3(12): 2455-2458.
- Guldener U, Munsterkott M, Oesterheld M, Pagel P, Ruepp A et al. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res* 34(Database issue): D436-441.
- Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS (1998) High-resolution protein design with backbone freedom. *Science* 282(5393): 1462-1467.
- Havranek JJ, Harbury PB (2003) Automated design of specificity in molecular recognition. *Nat Struct Biol* 10(1): 45-52.

- Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L et al. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415(6868): 180-183.
- Hoffmann R, Valencia A (2003) Protein interaction: same network, different hubs. *Trends Genet* 19(12): 681-683.
- Howard PL, Chia MC, Del Rizzo S, Liu FF, Pawson T (2003) Redirecting tyrosine kinase signaling to an apoptotic caspase pathway through chimeric adaptor proteins. *Proc Natl Acad Sci U S A* 100(20): 11267-11272.
- Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R et al. (2000) Functional discovery via a compendium of expression profiles. *Cell* 102(1): 109-126.
- Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 98(8): 4569-4574.
- Jansen R, Greenbaum D, Gerstein M (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res* 12(1): 37-46.
- Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ et al. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302(5644): 449-453.
- Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* 411(6833): 41-42.
- Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL (2000) The large-scale organization of metabolic networks. *Nature* 407(6804): 651-654.
- Jones S, Thornton JM (1997a) Analysis of protein-protein interaction sites using surface patches. *J Mol Biol* 272(1): 121-132.
- Jones S, Thornton JM (1997b) Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol* 272(1): 133-143.
- Jothi R, Kann MG, Przytycka TM (2005) Predicting protein-protein interaction by searching evolutionary tree automorphism space. *Bioinformatics* 21 Suppl 1: i241-i250.
- Kelley BP, Sharan R, Karp RM, Sittler T, Root DE et al. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci U S A* 100(20): 11394-11399.
- Korkin D, Davis FP, Sali A (2005) Localization of protein-binding sites within families of proteins. *Protein Sci* 14(9): 2350-2360.
- Kortemme T, Baker D (2004) Computational design of protein-protein interactions. *Curr Opin Chem Biol* 8(1): 91-97.
- Krogan NJ, Cagney G, Yu H, Zhong G, Guo X et al. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440(7084): 637-643.
- Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein-protein recognition sites. *J Mol Biol* 285(5): 2177-2198.
- Lu L, Arakaki AK, Lu H, Skolnick J (2003) Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. *Genome Res* 13(6A): 1146-1154.
- Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY et al. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res* 30(1): 281-283.
- Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO et al. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science* 285(5428): 751-753.
- Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP et al. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res* 11(12): 2120-2126.

- Mendez R, Leplae R, Lensink MF, Wodak SJ (2005) Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins* 60(2): 150-169.
- Middendorf M, Ziv E, Wiggins CH (2005) Inferring network mechanisms: the *Drosophila melanogaster* protein interaction network. *Proc Natl Acad Sci U S A* 102(9): 3192-3197.
- Ng SK, Zhang Z, Tan SH (2003a) Integrative approach for computationally inferring protein domain interactions. *Bioinformatics* 19(8): 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(1): 251-254.
- Nooren IM, Thornton JM (2003) Structural characterisation and functional significance of transient protein-protein interactions. *J Mol Biol* 325(5): 991-1018.
- Nye TM, Berzuini C, Gilks WR, Babu MM, Teichmann SA (2005) Statistical analysis of domains in interacting protein pairs. *Bioinformatics* 21(7): 993-1001.
- Ofran Y, Rost B (2003) Analysing six types of protein-protein interfaces. *J Mol Biol* 325(2): 377-387.
- Ooi SL, Pan X, Peyser BD, Ye P, Meluh PB et al. (2006) Global synthetic-lethality analysis and yeast functional profiling. *Trends Genet* 22(1): 56-63.
- Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* 96(6): 2896-2901.
- Pagel P, Wong P, Frishman D (2004) A domain interaction map based on phylogenetic profiling. *J Mol Biol* 344(5): 1331-1346.
- Pagel P, Oesterheld M, Stumpflen V, Frishman D (2006) The DIMA web resource--exploring the protein domain network. *Bioinformatics* 22(8): 997-998.
- Panchenko AR, Kondrashov F, Bryant S (2004) Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci* 13(4): 884-892.
- Pazos F, Valencia A (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng* 14(9): 609-614.
- 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(4): 1002-1015.
- 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 U S A* 96(8): 4285-4288.
- Pevzner PA, Dancik V, Tang CL (2000) Mutation-tolerant protein identification by mass spectrometry. *J Comput Biol* 7(6): 777-787.
- Pieles U, Zurcher W, Schar M, Moser HE (1993) Matrix-assisted laser desorption ionization time-of-flight mass spectrometry: a powerful tool for the mass and sequence analysis of natural and modified oligonucleotides. *Nucleic Acids Res* 21(14): 3191-3196.
- Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS et al. (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* 34(Database issue): D291-295.
- Qi Y, Klein-Seetharaman J, Bar-Joseph Z (2005) Random forest similarity for protein-protein interaction prediction from multiple sources. *Pac Symp Biocomput*: 531-542.
- Qi Y, Bar-Joseph Z, Klein-Seetharaman J (2006) Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins* 63(3): 490-500.

- Qian J, Luscombe NM, Gerstein M (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J Mol Biol* 313(4): 673-681.
- Ramani AK, Marcotte EM (2003) Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J Mol Biol* 327(1): 273-284.
- Reina J, Lacroix E, Hobson SD, Fernandez-Ballester G, Rybin V et al. (2002) Computer-aided design of a PDZ domain to recognize new target sequences. *Nat Struct Biol* 9(8): 621-627.
- Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M et al. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol* 17(10): 1030-1032.
- Riley R, Lee C, Sabatti C, Eisenberg D (2005) Inferring protein domain interactions from databases of interacting proteins. *Genome Biol* 6(10): R89.
- Russell RB, Alber F, Aloy P, Davis FP, Korkin D et al. (2004) A structural perspective on protein-protein interactions. *Curr Opin Struct Biol* 14(3): 313-324.
- Rzhetsky A, Gomez SM (2001) Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics* 17(10): 988-996.
- Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU et al. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* 32(Database issue): D449-451.
- 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(17): 3482-3489.
- Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S et al. (2005) Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A* 102(6): 1974-1979.
- Shifman JM, Mayo SL (2002) Modulating calmodulin binding specificity through computational protein design. *J Mol Biol* 323(3): 417-423.
- Shoemaker BA, Panchenko AR, Bryant SH (2006) Finding biologically relevant protein domain interactions: conserved binding mode analysis. *Protein Sci* 15(2): 352-361.
- Sprinzak E, Margalit H (2001) Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol* 311(4): 681-692.
- Sprinzak E, Sattath S, Margalit H (2003) How reliable are experimental protein-protein interaction data? *J Mol Biol* 327(5): 919-923.
- Stein A, Russell RB, Aloy P (2005) 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res* 33(Database issue): D413-417.
- Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302(5643): 249-255.
- Taylor JA, Johnson RS (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 11(9): 1067-1075.
- Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS et al. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403(6770): 623-627.
- Valdar WS, Thornton JM (2001) Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* 42(1): 108-124.
- von Mering C, Krause R, Snel B, Cornell M, Oliver SG et al. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417(6887): 399-403.

- Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF et al. (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* 287(5450): 116-122.
- Whitehouse CM, Dreyer RN, Yamashita M, Fenn JB (1985) Electrospray interface for liquid chromatographs and mass spectrometers. *Anal Chem* 57(3): 675-679.
- Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K et al. (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285(5429): 901-906.
- Wodak SJ, Mendez R (2004) Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications. *Curr Opin Struct Biol* 14(2): 242-249.
- Wojcik J, Schachter V (2001) Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics* 17 Suppl 1: S296-305.
- Wolf YI, Karev G, Koonin EV (2002) Scale-free networks in biology: new insights into the fundamentals of evolution? *Bioessays* 24(2): 105-109.
- Yates JR, 3rd, Eng JK, McCormack AL, Schieltz D (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* 67(8): 1426-1436.



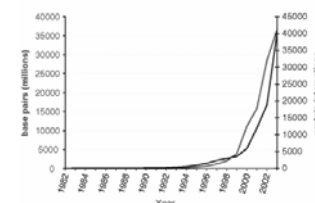
Protein-protein interactions: structure and systems approaches to analyze diverse genomic data

Anna R. Panchenko
Benjamin A. Shoemaker
NCBI / NIH

Importance of protein-protein interactions.

- Many cellular processes are regulated by multiprotein complexes.
- Distortions of protein interactions can cause diseases.
- Protein function can be predicted by knowing functions of interacting partners ("guilt by association").

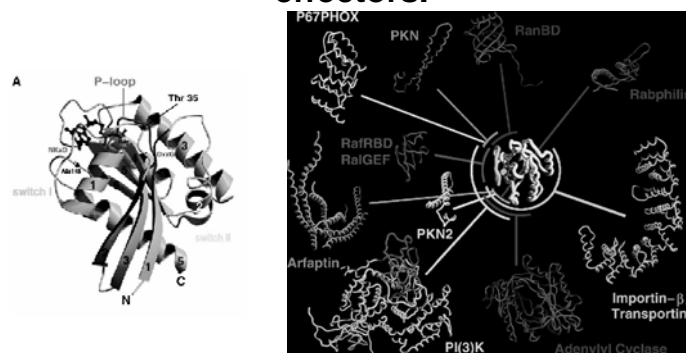
A comparison of sequence (GenBank) and protein-protein interaction data (DIP database)



Adapted from S. Fields, FEBS, 2005



Example: interaction of guanine- nucleotide binding domain with different effectors.

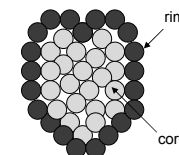


Adapted from Vetter & Wittinghofer, Science 2001

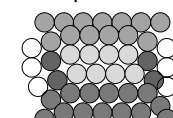


Common properties of protein- protein interactions.

- Majority of protein complexes have a buried surface area of about $1600 \pm 400 \text{ \AA}^2$ ("standard size" patch).
- Complexes of "standard size" do not involve large conformational changes of interacting proteins while large complexes do.
- Protein recognition site consists of completely buried core and a partially accessible rim.
- Trp and Tyr are abundant in the core, but Ser and Thr, Lys and Glu are particularly disfavored.



Top molecule



Bottom molecule

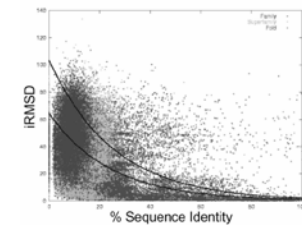


Different types of protein-protein interactions.

- Permanent and transient depending on their strength.
- External are between different chains; internal are within the same chain.
- Homo- and hetero-oligomers depending on the similarity between interacting subunits.
- Different interface types differ in amino acid composition; can predict interface type from amino acid composition with 63-100% accuracy (Ofra and Rost 2003).

Conservation of protein-protein interactions.

- Conservation of protein interfaces is weak compared to the rest of a protein → low accuracy of prediction of protein-protein interaction sites.
- Conservation of domain-domain interactions: at SCOP Family level (red) interactions are conserved, at Fold level (blue) are not conserved.



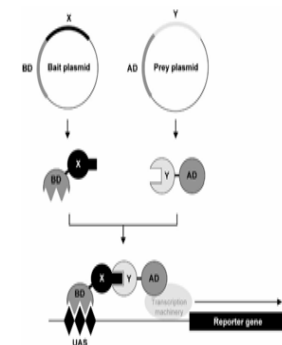
Adapted from Aloy et al., J. Mol. Biol., 2003

Experimental methods for identifying protein-protein interactions.

- Yeast two hybrid
- Mass spectroscopy
- TAP purification
- Gene expression

Yeast two-hybrid experiments.

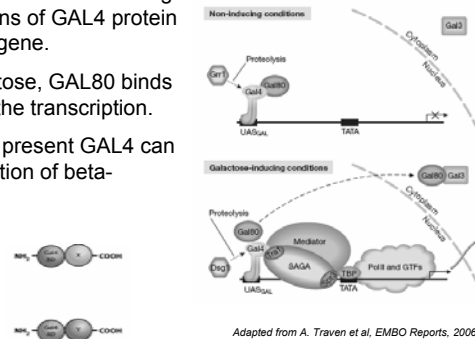
- Many transcription factors have two domains; one that binds to a promoter DNA sequence (BD) and another that activates transcription (AD).
- DNA-binding domain can not activate transcription at a promoter unless physically (not necessarily covalently) associated with an activating domain (Fields and Song, 1989).



Adapted from B. Causier, Mass Spectroscopy Reviews, 2004

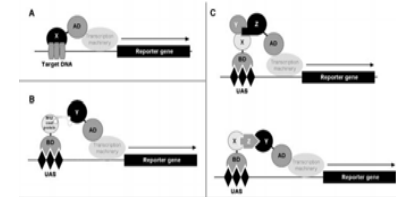
Gal4/LacZ Y2H system

- Target proteins are fused with binding and activation domains of GAL4 protein which activate LacZ gene.
- If there is no galactose, GAL80 binds to GAL4 and blocks the transcription.
- When galactose is present GAL4 can activate the transcription of beta-galactosidase.



Development and variations of Y2H system.

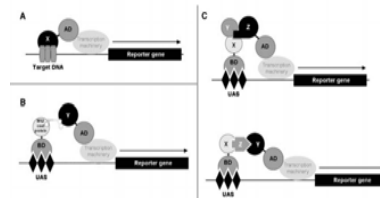
- Developing yeast strains that carry several reporter genes.
- Developing of haploid yeast strains of opposite mating type. Diploid cells are produced by mating containing both baits and preys.
- One-hybrid system detects interactions between a prey protein and a known DNA sequence (bait).



Adapted from B. Causier, Mass Spectroscopy Reviews, 2004

Development and variations of Y2H system.

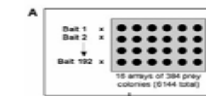
- RNA yeast three-hybrid system detects interactions between RNAs and proteins.
- Protein yeast three-hybrid system detects the formation of complexes between several proteins.



Adapted from B. Causier, Mass Spectroscopy Reviews, 2004

Genome-wide analysis by YTH.

- Matrix approach: a matrix of prey clones is added to the matrix of bait clones. Diploids where X and Y interact are selected based on the expression of a reporter gene.
- Library approach: one bait X is screened against an entire library. Positives are selected based on their ability to grow on specific substrates.



Interactions identified by HIS3 activity
Position on array identifies interactor



54 plates of 96 prey colonies
Positives selected based on growth

Uetz et al 2000, Ito et al 2001:
692-840 interactions detected using
library-based approach in yeast

Adapted from B. Causier, Mass Spectroscopy Reviews, 2004

Disadvantages of Y2H.

- The interactions can not be tested if a target protein can initiate transcription.
- Fusion of a protein into another domain (chimeras) can change the structure of a target protein.
- Protein interactions and posttranslational modifications can be different in yeast and other organisms.
- It is difficult to target extracellular proteins.
- Some cDNAs are fractional and do not represent the full length sequence of a target protein.
- Proteins which can in general interact in two-hybrid experiments, can never interact in a cell.

Advantages of Y2H.

- This is *in vivo* technique, good approximation of processes which occur in a living cells of higher eukaryotes.
- Transient interactions between proteins can be determined due to the amplification of a signal by the reporter gene expression, can predict the affinity of an interaction.
- Fast and efficient.

Mass spectroscopy.

1. Ionization (Ex: Electrospray ionization)

the solvent evaporates rapidly in a vacuum and protein molecules with a net positive charge become ionized;

- **Detection and recording of sample ions**
integral net charges are assigned to different peaks of spectra;
- **Analysis of MS spectra, protein identification**
search sequence database with mass fingerprint, find correlations between theoretical and experimental spectra.

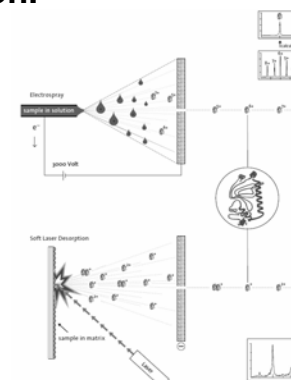
Ionization.

- Electrospray ionization, John Fenn, 2002)

The solvent evaporates rapidly in a vacuum and protein molecules with a net positive charge become ionized;

- Matrix Assisted Laser Desorption, K. Tanaka, 2002)

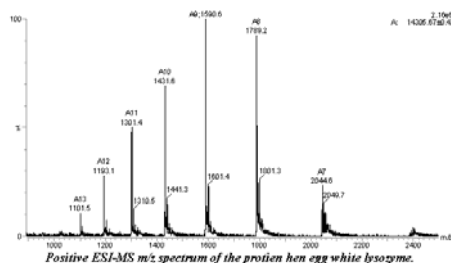
The laser ionizes protein molecules embedded on the matrix



From www.nobelprize.org

Detection.

- Peptide fragments are separated based on mass-to-charge ratios;
- Accuracy of **0.01%** of the total molecular mass of the sample *i.e.* within a 4 Daltons;



Differences and similarities between Y2H and MS.

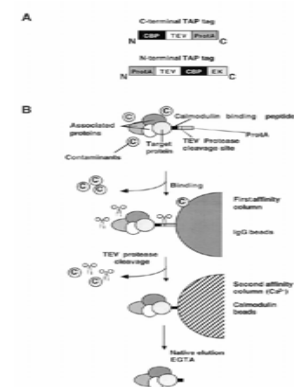
- Both methods generate a lot of false positives, only ~50% interactions are biologically significant. Both miss a lot of known interactions.
- Y2H produces binary interactions, lack of information about protein complexes, but can detect transient interactions.
- MS can detect large stable complexes and networks of interactions.

Tandem affinity purification method (TAP).

- Target protein ORF is fused with the DNA sequences encoding TAP tag;
- tagged ORFs are expressed in yeast cells and form native complexes;
- the complexes are purified by TAP method;
- components of each complex are found by gel electrophoresis, MS and bioinformatics methods.

Tandem affinity purification method (TAP).

TAP tag consists of two IgG binding domains of *Staphylococcus* protein A and calmodulin binding peptide;



O. Puig et al, Methods, 2001

Correlation between gene expression and protein interactions.

- There should exist a relationship between gene expression levels of subunits in a complex. → protein-protein interactions can be deduced from coexpression data.
- Methods are tested on specific protein complexes: ribosome, proteasome, RNA Polymerase II Holoenzyme and replication complexes.

Correlation between gene expression and protein interactions.

Jansen, Greenbaum & Gerstein, *Genome Research*, 2002

- Expression profiles were taken from two different sources: cell cycle experiments and expression ratios for overall yeast genome for 300 stationary cell states.
- Difference between absolute expression levels can be calculated as

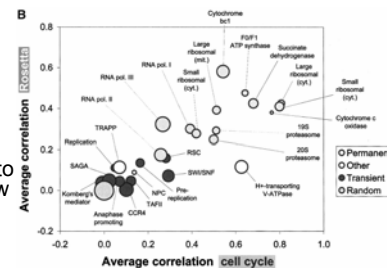
$$D = \frac{|E_i - E_j|}{(E_i + E_j)}$$

where E_i and E_j are mRNA expression levels of subunits "i" and "j".

Results of gene coexpression analysis.

Jansen, Greenbaum & Gerstein,
Genome Research, 2002

- Subunits from the same complex show coexpression, expression correlation is strong for permanent complexes.
- Transient complexes subdivided into smaller permanent complexes show strong correlation with gene expression.
- For genome-wide Y2H data, there is only a weak correlation with the gene expression.



Verification of experimental protein-protein interactions.

- Protein localization method.
- Expression profile reliability method.
- Paralogous verification method.

Protein localization method.

Sprinzak, Sattath, Margalit, *J Mol Biol*, 2003

A – A3: Y2H
B: physical methods
C: genetics
E: immunological

True positives:

- Proteins which are localized in the s. cellular compartment
- Proteins with a common cellular role

| Experimental method category | % TP by Localization, n=1 | % TP by Cellular Role, n=1 | % TP by Cellular Role, n=50,000 |
|------------------------------|---------------------------|----------------------------|---------------------------------|
| A3 | 7.8 | 6.5 | 5.9 |
| A1 | 26.7 | 27.4 | 27.4 |
| A2 | 43.5 | 43.5 | 43.5 |
| A3 | 43.5 | 43.5 | 43.5 |
| A0 | 43.5 | 43.5 | 43.5 |
| A | 43.5 | 43.5 | 43.5 |
| C | 76.3 | 76.3 | 76.3 |
| D1 | 76.3 | 76.3 | 76.3 |
| D2 | 76.3 | 76.3 | 76.3 |
| E1 | 88.0 | 88.0 | 88.0 |
| D4 | 88.0 | 88.0 | 88.0 |
| D2 | 88.0 | 88.0 | 88.0 |
| A4 | 93.7 | 93.7 | 93.7 |
| D | 93.7 | 93.7 | 93.7 |
| E2 | 93.7 | 93.7 | 93.7 |

Experimental method category

Fortaleza, Brazil
August 6-10, 2006

169

Expression profile reliability method.

The flowchart illustrates the Expression profile reliability method. It begins with a 'Putative protein interaction network' represented by a cluster of nodes and edges. An arrow labeled 'EPR' points to a box containing the text: 'Collect the mRNA expression levels of the interacting pairs under several conditions'. From this box, an arrow points to a series of four small graphs, each showing a different expression profile. An arrow then points to a larger graph labeled 'Create the distribution of distances for the network', which shows a distribution curve. This is followed by a box labeled 'Compare this distribution to those of standard interacting and non-interacting sets'. An arrow then points to a final graph labeled 'Percentage of true interaction in the network ~50%', which shows a distribution curve. The entire process is shown in a flowchart format with arrows indicating the sequence of steps.

Putative protein interaction network

EPR

Collect the mRNA expression levels of the interacting pairs under several conditions

Create the distribution of distances for the network

Compare this distribution to those of standard interacting and non-interacting sets

Percentage of true interaction in the network ~50%

Deane, C. M. (2002) Mol. Cell. Proteomics 1: 349-356

170


Expression profile reliability method.

Deane et al, *Molecular & Cellular Proteomics*, 2002

EPR method is based on observation that interacting proteins are coexpressed. The distance between expression profiles of two proteins:

$$d_{AB}^2 = \sum_i (\log(e_i^A / e_{ref}^A) - \log(e_i^B / e_{ref}^B))^2$$

Parameter α characterizes the accuracy of given data, or correspond to the fraction of false positives.

$$\rho(d_{AB}^2) = \alpha \cdot \rho_i(d_{AB}^2) + (1 - \alpha) \cdot \rho_n(d_{AB}^2)$$


171

Paralogous verification method.

The diagram illustrates the Paralogous Verification Method (PVM) process. It begins with a 'Putative protein interaction network' represented by a cluster of various shapes (circles, squares, hexagons) in different colors (grey, black, white). An arrow labeled 'PVM' points from this network to a single interaction, labeled 'Select an interaction'. This leads to a step labeled 'Collect the paralog of the interacting pair', which shows a central black square connected to three other shapes: a grey circle, a grey square, and a white square. These three shapes are grouped by a bracket labeled 'Paralog'. An arrow then points to a step labeled 'Count the number of paralogous interactions', which shows the same three paralogous shapes. Finally, an arrow points to the result: 'Score = 2' and 'P1 and P2 do interact'.

PVM method is based on observation that if two proteins interact, their paralogs would interact. Calculates the number of interactions between two families of paralogous proteins.

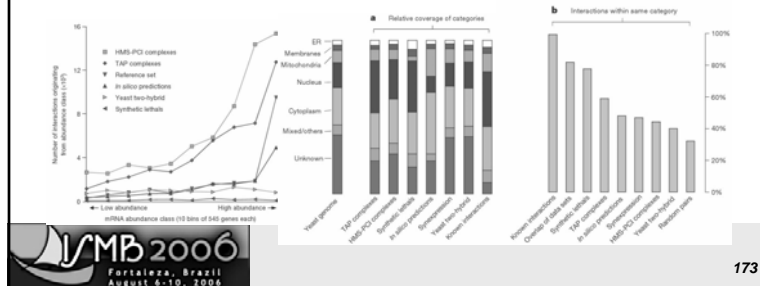
Deane, C. M. (2002) Mol. Cell. Proteomics 1: 349-356

172

Comparing large scale data of protein-protein interactions.

C. Von Mering et al, *Nature*, 2002:

- All methods except for Y2H and synthetic lethality technique are biased toward abundant proteins.
- PPI are biased toward certain cellular localizations.
- Evolutionary conserved proteins have much better coverage than the proteins restricted to a certain organism.

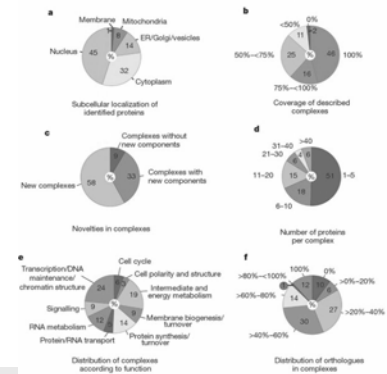


173

Functional organization of yeast proteome.

Gavin et al, *Nature*, 2002

- 589 protein assemblies,
- 232 multiprotein complexes,
- new cellular roles for 344 proteins.

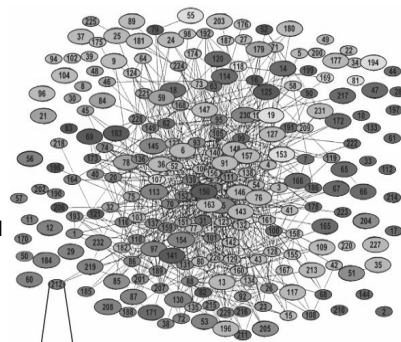


174

Functional organization of yeast proteome: network of protein complexes.

A. Gavin et al, *Nature*, 2002

- orthologous proteins interact with complexes enriched by orthologs;
- essential gene products are more likely to interact with essential rather than nonessential proteins



175

Interaction databases

- Protein-protein interactions from experiment (some pass quality assessment).
 - DIP (LiveDIP, ProLinks), BIND, MIPS
- Domain-domain interactions inferred from crystal structure data.
 - 3did, Pibase, CBM, iPFam

176

DIP database

- Documents protein-protein interactions from experiment
 - Y2H, protein microarrays, TAP/MS, PDB
- 55,733 interactions between 19,053 proteins from 110 organisms.

| Organisms | # proteins | # interactions |
|-------------------|------------|----------------|
| Fruit fly | 7052 | 20,988 |
| <i>H. pylori</i> | 710 | 1425 |
| Human | 916 | 1407 |
| <i>E. coli</i> | 1831 | 7408 |
| <i>C. elegans</i> | 2638 | 4030 |
| Yeast | 4921 | 18,225 |
| Others | 985 | 401 |

DIP database

Duan et al., *Mol Cell Proteomics*, 2002

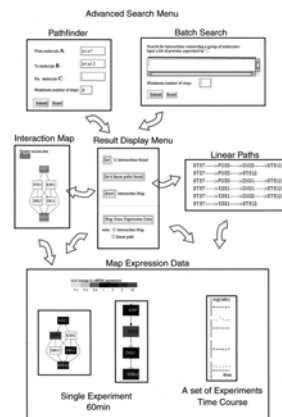
- Assess quality
 - Via proteins: PVM, EPR
 - Via domains: DPV
- Search by BLAST or identifiers / text



DIP database

Duan et al., *Mol Cell Proteomics*, 2002

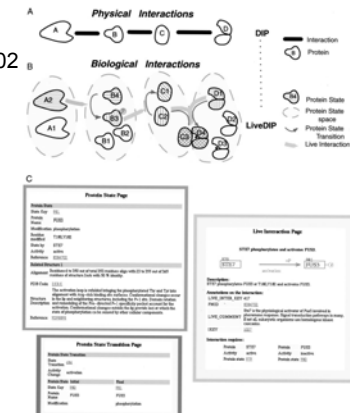
- Assess quality
 - Via proteins: PVM, EPR
 - Via domains: DPV
- Search by BLAST or identifiers / text
- Map expression data



LiveDIP

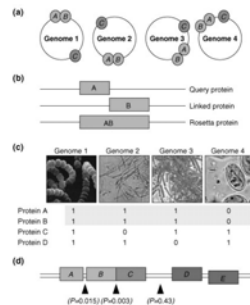
Duan et al., *Mol Cell Proteomics*, 2002

- Distinguish biological state
 - Covalently modified
 - Conformational
 - Cellular location



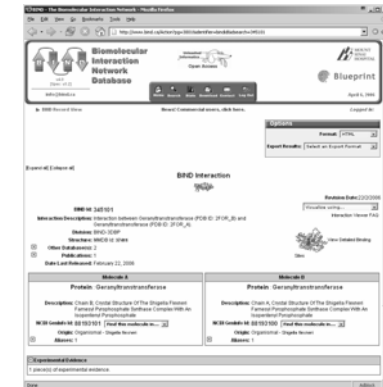
Prolinks database

- Gene neighbors
- Rosetta Stone
- Phylogenetic profiles
- Gene clusters



BIND database

- Contains experimental interaction data
- 83,517 protein-protein interactions
- Developed specification to handle diverse data



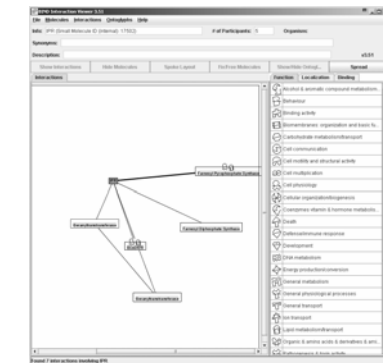
BIND

- 204,468 total interactions
- Includes small molecules, NAs, genes, complexes, photons



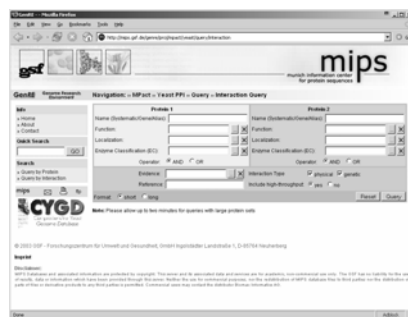
BIND

- Interaction Viewer
- Unique icons of functional classes



MPact/MIPS database

- Yeast protein-protein interactions
- Curated set:
 - 4,300 PPI
 - 1,500 proteins
- High-throughput available
- Web tools



InterDom database

- 30,037 predicted domain interactions from PPIs
 - Domain fusions
 - Protein interactions
 - Complexes
 - Literature
- Score interactions

The screenshot shows the InterDom database web interface. It features a search bar at the top with the InterDom logo. Below the search bar, there are several tabs: 'Search', 'Browse', 'Download', and 'Help'. The 'Search' tab is selected, and it shows a search form with fields for 'Domain 1', 'Domain 2', and 'Interaction'. There are also checkboxes for 'Include high-throughput' and 'Include low-throughput'. A 'Search' button is at the bottom right of the form. Below the search form, there is a table of results with columns for 'Domain 1', 'Domain 2', 'Interaction', and 'Score'. The table contains several rows of data, including domain names and interaction scores.

Pibase database

- Protein structures from PDB and PQS
- Domains defined with SCOP and CATH
- All inter-domain and inter-chain distances within 6.05 Å are considered interacting domains
- From interacting domain pairs, create list of interfaces with buried solvent accessible area > 300 Å²

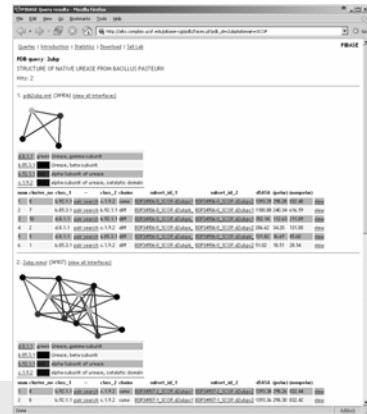
PIBASE

- Query by PDB, domain, interface
- 1,946 interacting SCOP domains
- 2,387 unique interaction types



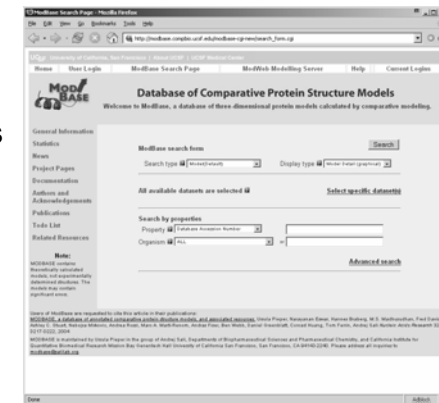
PIBASE

- Redundancy removed within a structure
- Properties listed



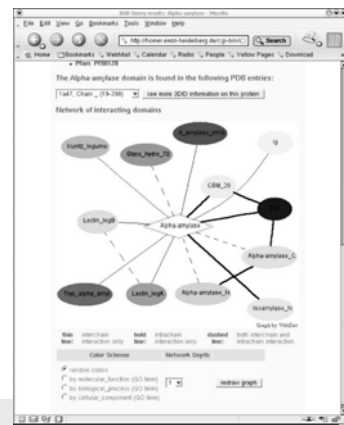
PIBASE/ModBase

- Protein structure models
- Predict interfaces with Pibase



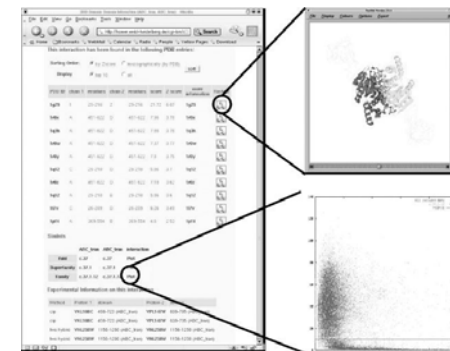
3did database

- Pfam domain-domain interactions
- Protein structure data
- 3,304 unique interaction types
- 2,247 interacting domains
- Display linkages and chain locations



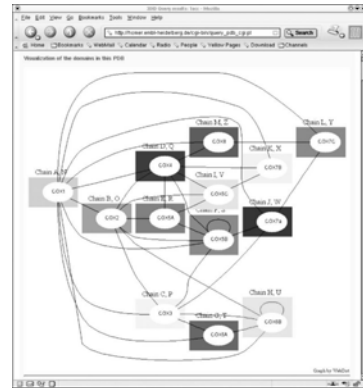
3did

- List structures
- Visualize interfaces
- View interface overlap distribution
- GO annotation



3did

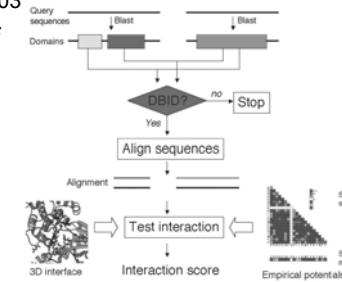
- Show domain linkages on a given structure



InterPReTS

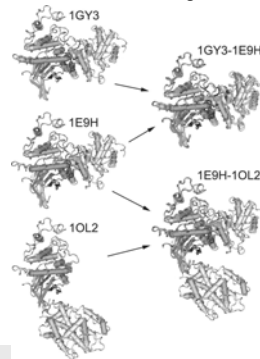
Aloy & Russell, *Bioinformatics*, 2003

- Structure prediction of interfaces
- Uses 3did



Protein-protein interactions available from structure data: NCBI CBM database

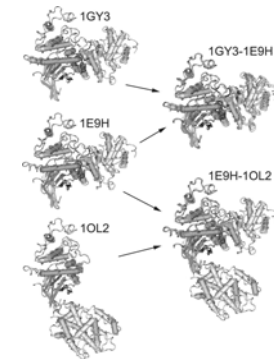
- CBM – database of interacting structural domains exhibiting Conserved Binding Modes
- To retrieve interactions:
 - Record interactions
 - Use VAST structural alignments to compare binding surfaces
 - Study recurring domain-domain interactions
- Currently available via FTP



Shoemaker et al., *Protein Sci*, 2006.

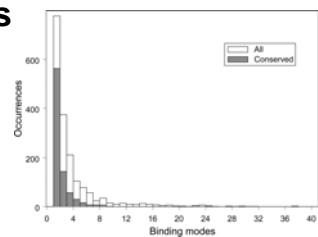
Definition of CBM

- Interacting domain pair – if at least 5 residue-residue contacts between domains (contacts – distance of less than 8 Å)
- Structure-structure alignments between all proteins corresponding to a given pair of interacting domains
- Clustering of interface similarity, those with >50% equivalently aligned positions are clustered together
- Clusters with more than 2 entries define conserved binding mode.



Number of interacting pairs and binding modes

- 833 conserved interaction types
- 1,798 total domain interaction types
- Up to 24 CBMs per interaction type



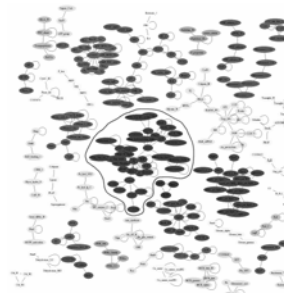
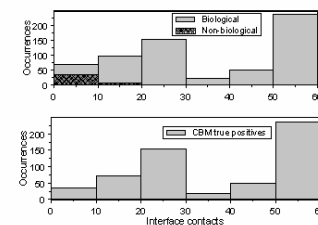
| CBM | Structures | Species |
|-----|------------|-------------------|
| 1 | 154 | Jawed vertebrates |
| 2 | 112 | Jawed vertebrates |
| 3 | 17 | Clam,earthworm |
| 4 | 4 | lamprey |
| 5 | 4 | V.stercoraria |
| 6 | 2 | Rice,soybeans |
| 7 | 2 | human |
| 8 | 2 | lamprey |

- Classify complicated domain pairs by CBMs
- Globin example:
 - 630 pairs
 - 2 CBMs account for majority

Shoemaker et al., *Protein Sci*, 2006.

CBMs distinguish biologically relevant interactions

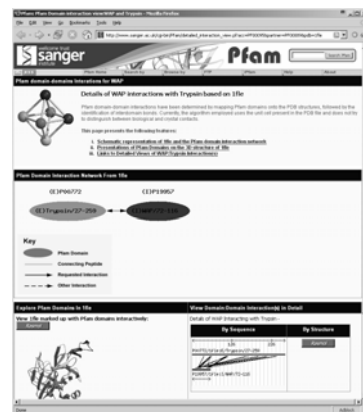
- Non-biological interactions (e.g. crystal packing) are not conserved among different structures.
- Interaction networks more clear



Shoemaker et al., *Protein Sci*, 2006.

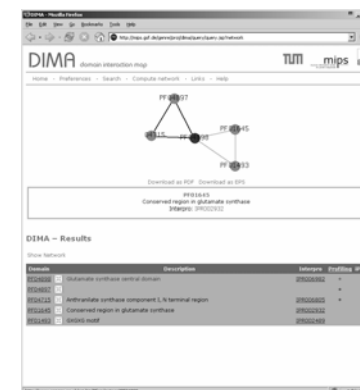
iPfam database

- View Pfam interactions on PDB structures
- View individual structures and sequence plots



DIMA database

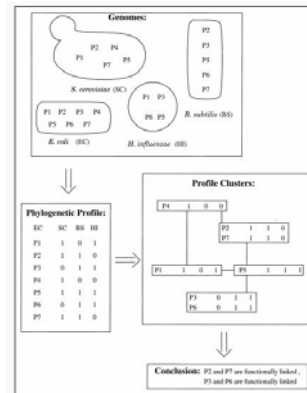
- Phylogenetic profiles of Pfam domain pairs
- Uses structural info from iPfam
- Works well for moderate information content



Phylogenetic profile method.

Pellegrini et al, *PNAS* 1999

Functionally linked and probably interacting proteins should have orthologs in the same subset of fully sequenced organisms

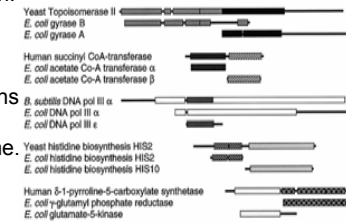


Rosetta Stone approach.

Marcotte et al, *Science*, 1999

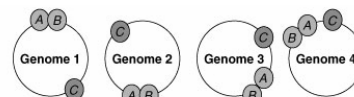
Some pairs of interacting domains have homologs which are fused into one protein chain – “Rosetta Stone” protein.

In *E. coli* method found 6809 pairs of non-homologous proteins, both proteins from each pair could be mapped to a single protein from some other genome.



Gene neighborhood method.

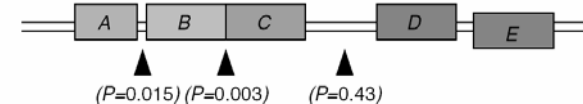
- Gene pairs from conserved gene clusters appear to encode proteins which are functionally related and possibly interact.
- Gene order between the prokaryotic and archaeal species is conserved if sequence identity shared by orthologs in two genomes > 50%.
- Conservation of gene order can be used to predict gene function.



Adapted from Bowers et al, *Genome Biology*, 2004

Gene cluster method.

- Bacterial genes of related function are often transcribed simultaneously – operon.
- Identification of operons is based on intergenic distances.

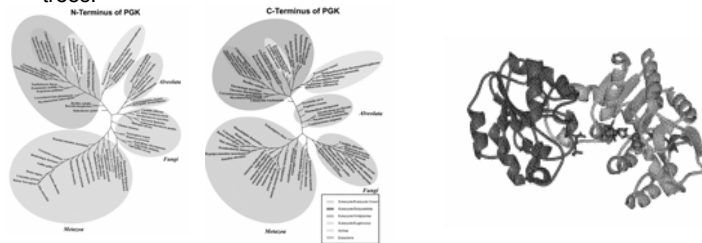


Adapted from Bowers et al, *Genome Biology*, 2004

Coevolution of interacting proteins/domains – “mirrortree” methods.

Goh et al. 2000; Pazos and Valencia 2001

- Interacting proteins very often co-evolve and their phylogenetic trees show some similarity.
- The similarity between phylogenetic trees can be quantified by correlation coefficient between distance matrices used to construct trees.



Adapted from Goh et al., J.Mol.Biol.,2000

Predicting interacting partners from two interacting protein families.

Problem:

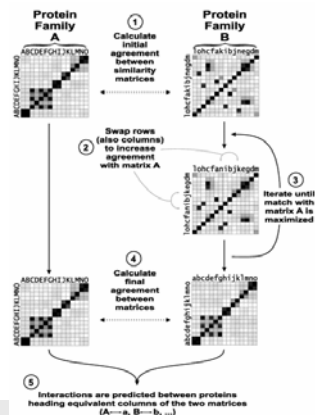
given interacting protein families $A \{a_1, \dots, a_n\}$ and $B \{b_1, \dots, b_m\}$:

- Find corresponding proteins a_i and b_i that interact.
- Predict interaction specificity of interaction, ex: families containing paralogs.
- Predict one-to-many correspondence between interacting partners.

Methods of predicting interacting partners.

Ramani & Marcotte, *J. Mol. Biol.*, 2003,
Gertz et al, *Bioinformatics*, 2003

- Proteins are clustered allowing to find one-to-many correspondence between proteins.
- Similarity matrices are aligned using simulated annealing, optimizing the root mean square difference/correlation coefficient between elements of two matrices.
- Interactions are predicted between proteins corresponding to the aligned columns of two matrices.



Adapted from Ramani & Marcotte, *J. Mol. Biol.*, 2003

Problems of matrix permutations methods:

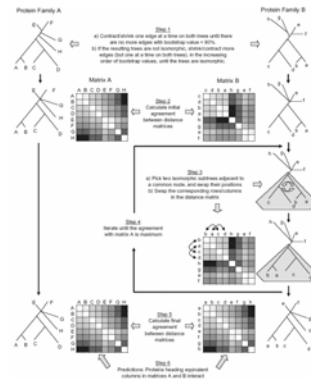
- $N!$ – permutations (N – number of proteins in a family) – search space is big!
- maximal agreement between similarity matrices does not mean correct pairing of proteins on phylogenetic tree.

Methods of predicting interacting partners.

MORPH method, Jothi et al, *Bioinformatics*, 2005

- To reduce the search space – by swapping whole isomorphic subtrees in a single move instead of a single column
- avoid local minima.

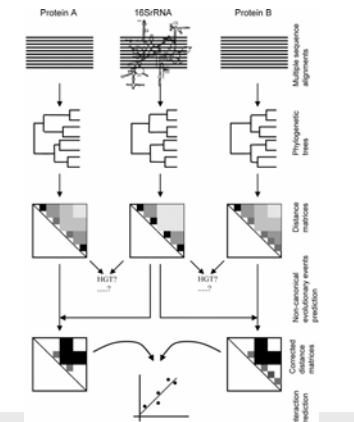
- Uses information encoded in the phylogenetic trees themselves.



Tree of life assists in prediction of protein-protein interactions.

Pazos et al, *J. Mol. Biol.*, 2005
Sato et al, *Bioinformatics*, 2005

- There exists certain “background” similarity between trees of any proteins, no matter if they interact or not.
- The “background” tree is constructed from 16S rRNA sequences.
- rRNA-based distances are subtracted from distances for the original phylogenetic tree.

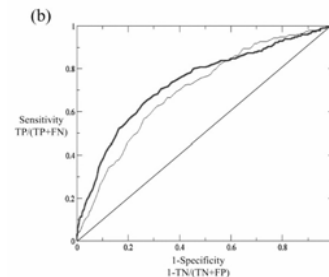


Adapted from Pazos et al, *J. Mol. Biol.*, 2005

Performance of “mirrortree” methods.

Pazos et al, *J. Mol. Biol.* 2005

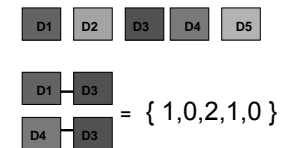
- Test set of 512 physically interacting proteins from *E. coli*
- “tol-mirrortree” method (blue) finds half of real interacting proteins at 6.4% false positive rate compared to 16.5% false positives rate with “mirrortree” method (black).



Adapted from Pazos et al, *J. Mol. Biol.*, 2005

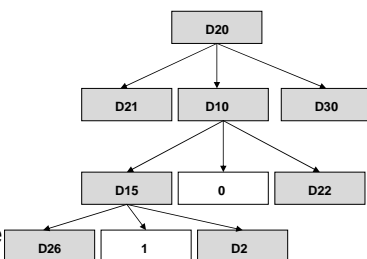
Classification methods: Random Decision Forest.

- Training set: interacting protein pairs + non-interacting pair;
- Each pair – vector of features (domain types) of dimension N.
- Values of vector:
0, if protein pair does not contain feature;
1, if at least one protein in a pair contains feature;
2, if two proteins contain the feature.



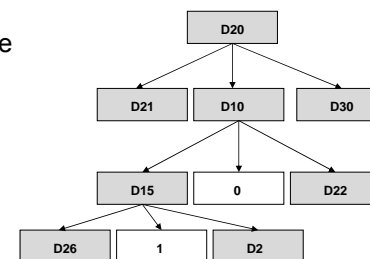
Classification methods: Random Decision Forest.

- Choose feature randomly (D20), get values of all pairs in a given position corresponding to this feature;
- Divide all pairs in three groups: those which both have this feature, only one protein has feature, no feature.



Classification methods: Random Decision Forest.

- Repeat splitting at next node and stop when node impurity is small.
- To classify a new protein pair – traverse along the tree



Node impurity = # interacting proteins / # non-interacting proteins

Predicting domain interactions from protein interactions

- Association method
- Maximum likelihood estimation method
- Domain Pair Exclusion Analysis
- Random decision forests
- Calculating P-values
- Integrative method

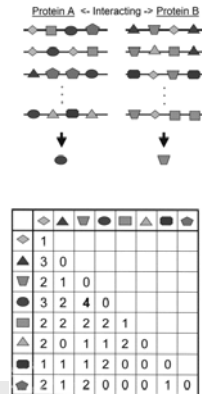
Predicting domain interactions from protein interactions

- Protein sequence search of Pfam, SCOP or CDD domains
- Train on high-throughput experimental data
- Evaluate with structures or MIPS
- Assign probabilities to protein interactions for further prediction

Association method

Sprinzak & Margalit, *J Mol Biol*, 2001

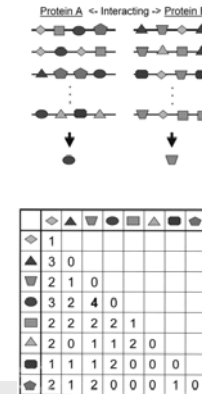
- Record domains
- List interacting protein pairs
- Tabulate domain pairs from protein pairs
- Compute log-odds values



Association method

Sprinzak & Margalit, *J Mol Biol*, 2001

- 2,286 domain pairs
- 1,141 pairs > 2 bits
- 40 pairs with > 2 bits & count of 5
- No experimental error



Association method

Sprinzak & Margalit, *J Mol Biol*, 2001

- Log-odds value: $\log_2(P_{ij}/P_iP_j)$
- P_i is the frequency of domain i in the data
- Average mutual information content per domain pretty high (2.48 bits) – significant correlation between interacting proteins and predicted domain pairs

Random decision forests

Chen and Liu, *Bioinformatics*, 2005

- Discussed earlier as protein interaction prediction method
- 3,000 domain pairs predicted
- No experimental error
- Doesn't assume independency
- Accounts for non-interactions
 - Riley et al. note that this makes it harder to find specific paralogous interactions

Expectation Maximization

Deng et al., *Genome Res*, 2002

1. Use initial parameters to get Z , expectation of complete dataset
2. Get maximum likelihood estimator of parameter set, Θ .
3. Iterate until convergence



221

Expectation Maximization

Deng et al., *Genome Res*, 2002

- $f_n = 0.64$, $f_p = 2.85E-4$
- 43% specificity, 78% sensitivity
- MIPS best predictors 100x > random
- But, only 0.68% predicted



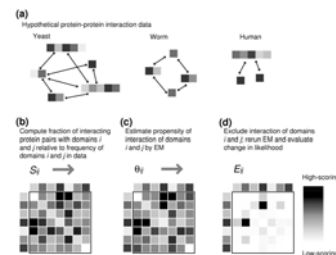
222

Domain Pair Exclusion Analysis

Extend MLE method to detect specific, rare interactions

Riley et al., *Genome Biol*, 2005.

1. S_{ij} frequencies
2. MLE of Θ_{ij} starting with S_{ij}
3. Recalculate Θ_{ij} with interaction probability ij fixed to zero. Get E_{ij} from difference.

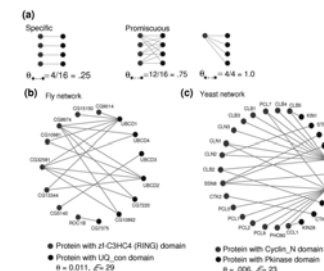


223

Promiscuous domains

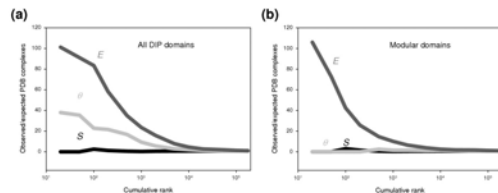
Riley et al., *Genome Biol*, 2005.

- High E-score: high propensity to interact
- Low E-score: competing domains more likely responsible for interaction
- Screen for low θ and high E to find specific domain-domain interactions



224

Domain Pair Exclusion Analysis



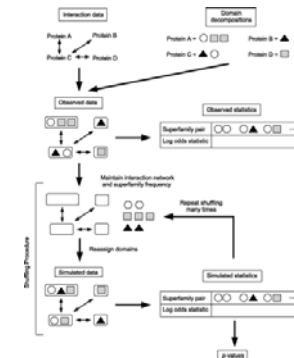
Riley et al., Genome Biol, 2005.

- E discriminates 100 top predictions 71x random
- Θ and S are ineffectual particularly with modular domains

Calculating P-values to predict domain interactions

Nye et al., Bioinformatics, 2005.

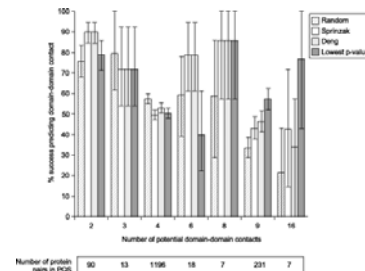
- SCOP superfamilies
- P-values for domain pairs
- Shuffle domains on sequences for null hypothesis
- Domain architectures considered



Calculating P-values to predict domain interactions

Nye et al., Bioinformatics, 2005.

- $f_n = 5.7E^{-4}$, $f_p = 0.1$
- Contrast to Deng (0.64, $2.85E^{-4}$)
- Predicts better at higher number of interacting partners
- Random wins in largest group



Integrative method

Ng et al., Bioinformatics, 2003

- Add scores from three sources:
 - DIP – odds ratio score
 - Protein complexes – odds ratio score
 - Domain fusions – simple constant
- 20-fold cross validation
 - Major change from DIP to DIP + complexes
 - TP: 39% to 58%, FPs: 8% to 12%

Limitations of domain interaction prediction methods

- Assume domain pairs interact independently
- Repeated domains not scored to distinguish contacts
- Missing domain assignments give false negatives and positives
- Many proteins have no assignments
- Assume domain pairs, though may require higher order assemblies



229

Homology modeling of protein interactions

- Comparison to modeling single proteins
- General procedure
- Automated methods
- CAPRI docking contest
- Designed interfaces



230

Homology modeling of single proteins

- Structures solved quickly with current techniques
- Decent coverage of major genomes expected
- Structure prediction:
 - Find homologous template to query
 - Make query model based on template



231

Homology modeling of protein interactions

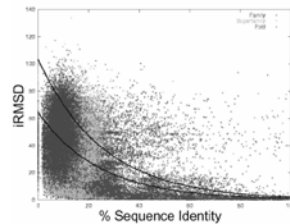
- Elucidate interaction networks: Roughly 2,000 out of 10,000 interaction types known
- Limited protein-protein complexes in PDB
- Large complex structure determination has technical challenges not readily overcome in general
- Likely path involves multiple experimental methods with homology modeling and docking of structural subunits



232

Support for modeling protein interactions

- Conservation of protein interfaces is weak compared to the rest of a protein → low accuracy of prediction of protein-protein interaction sites.
- Conservation of domain-domain interactions: at SCOP Family level (red) interactions are conserved, at Fold level (blue) are not conserved.



Adapted from Aloy et al., *J. Mol. Biol.*, 2003

Support for modeling protein interactions

Shoemaker et al., *Protein Sci*, 2006.

Globin example:

- Interfaces between different functional subfamilies poorly conserved
- Within the same subfamily well conserved
- Supports homology modeling of interaction interfaces



General procedure for homology modeling

- Start with high-throughput (Y2H, TAP/MS) protein interaction data
- Search proteins for homologous domains
- Evaluate likelihood of domain-domain interactions
- Search for homologous structures to query proteins/protein domains

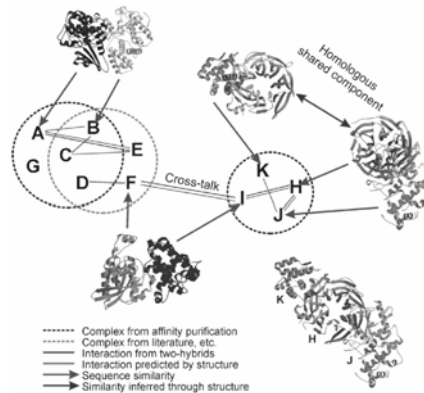
General procedure for homology modeling

- Homologous structures might be
 - Complete complexes (rare)
 - Interacting domain dimers (sometimes)
 - Single domains (most often)
- Put together structural pieces avoiding steric hindrance and maximize domain complementarity
- Docking potentials score orientations of two interacting domains
- Success depends strongly on similarity and completeness of homologous structures

Example: Modeling of yeast complexes

P. Aloy et al, Science, 2004

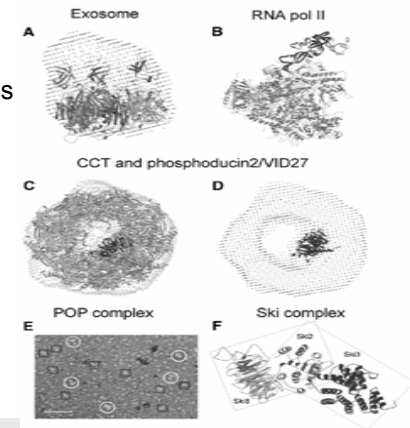
- Use data from:
 - Multiple experimental methods
 - Homology to structure
- Model interactions within and between complexes



Example: Modeling of yeast complexes

P. Aloy et al, Science, 2004

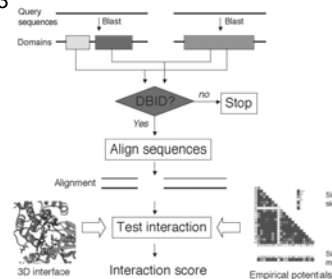
- Found 42 (out of 102) “nearly complete” models
- 12 partial models of interacting subunits
- Structures fit onto electron microscopy grids (A,C,D)
- Complexes assembled from multiple smaller complexes (F)



InterPRETS

Aloy & Russell, Bioinformatics, 2003

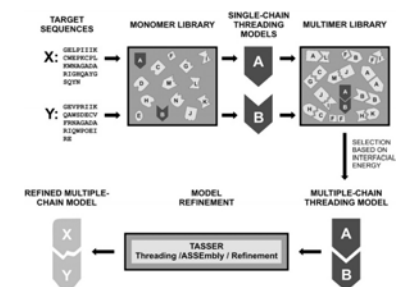
- Search for Pfam domains on target sequences
- Construct complexes matching the same Pfam types
- Score putative interactions with empirical pair potentials
- Good results except for peptidase / inhibitor class



Multiprospector

Grimm et al., Proteins, 2006.

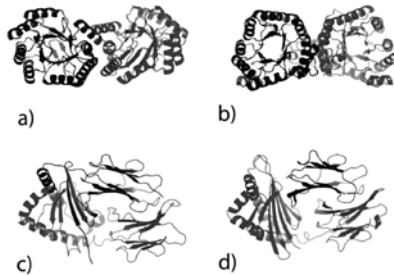
- Separately thread sequences X and Y against protein dimer database
- For X/Y matches to the same dimer, assess fitness by rethreading with an interface score derived from the dimer database



Multiprospector

Grimm et al., *Proteins*, 2006.

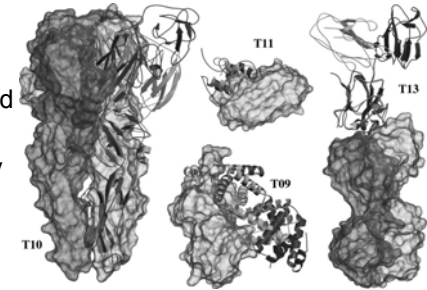
- On yeast genome, 7,321 interactions were predicted from 304 complexes
- Ranked 3rd amongst large-scale prediction methods
 - No bias towards abundant proteins
 - Provides atomic detail of interaction surfaces



CAPRI contest

Mendez et al., *Proteins*, 2005.

- Build atomic models of complexes given structures of the unbound proteins
- Bound/unbound differ by up to 12Å
- “Acceptable” to “highly accurate” predictions made



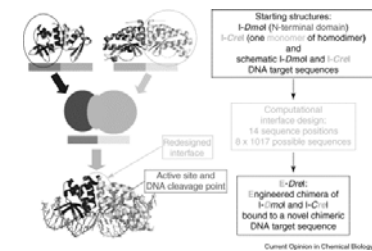
Limitations from CAPRI contest affecting homology modeling

- Proteins can undergo significant conformation changes upon binding
- Docking potentials require more accuracy
- Specific and non-specific protein interactions are not adequately distinguished

Interface design

Kortemme & Baker, *Curr Opin Chem Biol*, 2005

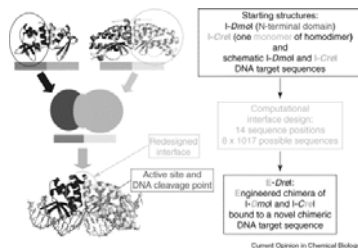
- Computationally alter interface to modify function
- Create useful complexes
- Better understand prediction



Interface design

Kortemme & Baker, *Curr Opin Chem Biol*, 2005

- Alter oligomeric state in helical bundle
- Increase specificity of promiscuous domains
- Novel interactions
- Automated the process



Basic notions of networks.

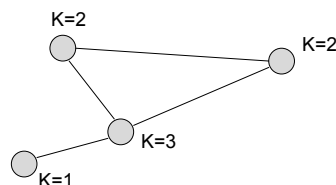
Network (graph) – a set of vertices connected via edges.

The degree of a vertex – the total number of connections of a vertex.

Random networks – networks with a disordered arrangement of edges.

Characteristics of networks: degree distribution.

$P(k, N)$ – degree distribution, k - degree of the vertex,
 N - number of vertices



If vertices are statistically independent and connections are random, the degree distribution completely determines the statistical properties of a network.

Different network models: Barabasi-Alberts.

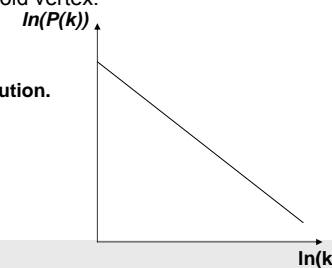
Barabasi & Albert, *Science*, 1999

Model of preferential attachment.

- At each step, a new vertex is added to the graph
- The new vertex is attached to one of old vertices with probability proportional to the degree of that old vertex.

Degree distribution – power law distribution.

$$p(k) \propto k^{-\gamma}$$

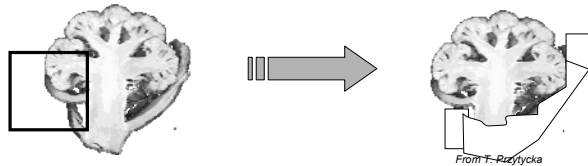


Power Law distribution

$$p(k) \sim k^{-\gamma}$$

Multiplying k by a constant, does not change the shape of the distribution – scale free distribution.

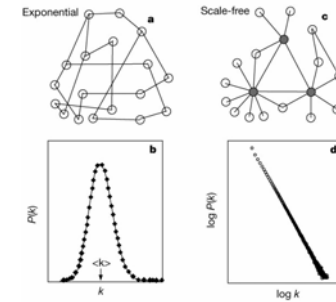
$$p(\alpha k) = (\alpha k)^{-\gamma} = \alpha^{-\gamma} p(k)$$



Difference between scale-free and random networks.

Random networks are homogeneous, most nodes have the same number of links.

Scale-free networks have a few highly connected vertices.



Adapted from Jeong et al, *Nature*, 2000

Multiple-species gene co-expression networks.

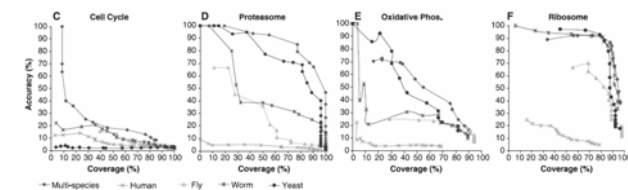
Stuart et al, *Science*, 2003

- Multiple-species network has been constructed by identifying pairs of genes with the correlated gene expression in different organisms.
- Multiple-species network performs better than single-species network in linking together functionally related genes.

Multiple-species gene co-expression networks.

Stuart et al, *Science*, 2003

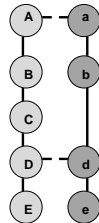
True positives – links from the same KEGG functional category;
accuracy - % links connecting two members of the same category;
coverage - % metagenes connected to at least one metagenes in the category.



Aligning protein interaction networks.

PATHBLAST (Kelley et al. , *PNAS*, 2003, Sharan et al, *PNAS*, 2005).

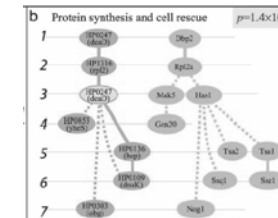
- The method searches for high-scoring pathway alignments between two networks, where proteins are paired based on their sequence similarity.



Aligning protein interaction networks.

PATHBLAST (Kelley et al. , *PNAS*, 2003, Sharan et al, *PNAS*, 2005).

- The network alignment between worm, yeast and fly detected 71 network regions that were conserved between all three species.



Comparing networks by their connectivities.

Hoffmann & Valencia, *TRENDS in genetics*, 2003

- Correlation coefficient between protein connectivities of two networks quantifies the agreement between the networks.
- Significant correlations between different experimental and theoretical methods: gene neighborhood method (GN) correlates with both experimental and in silico methods.

