Computational Analyses of High-Throughput Protein-Protein Interaction Data

Yu Chen[1, 2] and Dong Xu[1, 2]*

[1]*Protein Informatics Group, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA,* [2]*UT-ORNL Graduate School of Genome Science and Technology, Oak Ridge, TN, 37830, USA*

Abstract: Protein-protein interactions play important roles in nearly all events that take place in a cell. High-throughput experimental techniques enable the study of protein-protein interactions at the proteome scale through systematic identification of physical interactions among all proteins in an organism. High-throughput protein-protein interaction data, with ever-increasing volume, are becoming the foundation for new biological discoveries. A great challenge to bioinformatics is to manage, analyze, and model these data. In this review, we describe several databases that store, query, and visualize protein-protein interaction data. Comparison between experimental techniques shows that each high-throughput technique such as yeast two-hybrid assay or protein complex identification through mass spectrometry has its limitations in detecting certain types of interactions and they are complementary to each other. *In silico* methods using protein/DNA sequences, domain and structure information to predict protein-protein interaction can expand the scope of experimental data and increase the confidence of certain protein-protein interaction pairs. Protein-protein interaction data correlate with other types of data, including protein function, subcellular location, and gene expression profile. Highly connected proteins are more likely to be essential based on the analyses of the global architecture of large-scale interaction network in yeast. Use of protein-protein interaction networks, preferably in conjunction with other types of data, allows assignment of cellular functions to novel proteins and derivation of new biological pathways. As demonstrated in our study on the yeast signal transduction pathway for amino acid transport, integration of high-throughput data with traditional biology resources can transform the protein-protein interaction data from noisy information into knowledge of cellular mechanisms.

Keywords: protein-protein interaction, high-throughput data, yeast two hybrid, protein complex, proteome, bioinformatics.

1. INTRODUCTION

1.1. Protein-Protein Interaction in A Proteome

Protein-protein interactions are at the heart of biological activities [1-3]. They play a critical role in most cellular processes and form the basis of biological mechanisms such as DNA replication and transcription, enzyme-mediated metabolism, signal transduction, and cell cycle control [4,5]. Protein-protein interactions give the information about the biological context in which an individual protein plays its cellular role. Knowing the interactions that an uncharacterized protein has can provide a clue about its biological function. To fully understand a biological machinery of a cell or a biological pathway, it is also essential to know how the involved proteins directly interact with each other.

The advent of genome sequencing projects makes it possible to analyze protein-protein interactions at the genome scale. There are now more than half a million nonredundant sequences deposited on Genebank [6]. The complete genomes of more than 50 bacteria have been sequenced. Several eukaryotes have been sequenced at the genome scale as well, including fly (*Drosophilla melanogaster*) [7], worm (*Caenorhabditis elegans*) [8], yeast (*Saccharomyces cerevisiae*) [9] and *Arabidopsis thaliana* [10]. The human genome sequence is almost completed and the draft of mouse genome sequence has been finished [11,12]. The whole genome sequence provides the information about all the proteins in genome, i.e., the so-called "proteome". Such information allows us, for the first time, to characterize all the protein-protein interactions in an organism, which is referred to as the protein interaction map or "interactome" [13,14]. Several high-throughput technologies have been developed to characterize the protein-protein interaction map. This is in contrast to the traditional biology approach, where protein-protein interaction is determined and studied one at a time. In the protein interaction map, life reveals itself not as a mere collection of proteins, but rather as a sophisticated network. In other words, we can see not only the tree but also the forest. The protein interaction map provides a unique approach to address challenges in this post genome era, especially for understanding the functions of many newly discovered genes whose functions have not been characterized. For example, only one-third of all 6200 predicted yeast genes were functionally characterized when the complete sequence of yeast genome became available [15]. At present, 3800 yeast genes have been characterized by genetic or biochemical techniques and an addition of 600 genes have been identified based on homologs of known functions in other organisms. This leaves about 1800 genes with unknown functions [16]. Another challenge in the post

*Address correspondence to this author at the Protein Informatics Group, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA; Tel: 865-574-8934; Fax: 865-547-8934; Email: xud@ornl.gov

1389-2037/03 $41.00+.00

genome era is to understand how the proteins coded in the genome interact with each other to perform cellular mechanism [17]. The protein interaction map can provide essential information to address this challenge.

1.2. Physical Interaction and Genetic Interaction

The protein-protein interactions that we address in this review are direct or indirect, stable or transient physical interactions. The proteins involved are physically in contact through the binary interaction or the formation of a protein complex. This is in contrast with the genetic interactions, where the change of one gene affects the expression of another gene, or mutations of two genes at the same time can produce a novel phenotype that is not displayed by either mutation alone. Some of the genetic interaction screens are based on either loss or gain of viability for a phenotype. Several classical approaches were developed to identify genetic interaction. The synthetic lethality screen in yeast is a very powerful method for finding interactions between gene products [18]. It identifies non-allelic and non-lethal mutations that are lethal in combination with a non-lethal mutation in a gene of interest. A systematic genetic interaction analysis in yeast was developed to enable high-throughput synthetic lethal analysis by using an ordered array of about 4700 viable mutants [19]. It is possible that a pair of proteins can have both genetic and physical interactions.

1.3. Complexity of Protein Interaction

Protein-protein interactions within proteome (the complete set of proteins in a given organism) are of a dynamic nature. They change during different development stages or in response to different environmental stimuli. Furthermore, proteins interact with others and form a large interaction network, in which they regulate and support each other. Protein-protein interactions are inherently complex. Some interactions are transient, which are temporal and specific to a certain condition or a subset of cellular states, while others are stable, which are maintained throughout most cellular conditions. Moreover, post-translation modifications may change the interaction partners and patterns. Some proteins may have different subcellular localizations and they can interact with other proteins through translocalization into a specific cellular compartment upon receiving signals.

1.3.1. Transient interaction vs. Stable Complex

Protein expressions and protein-protein interaction patterns can change during the development or morphogenesis or in response to many different environmental conditions. There exist different interaction types, for example, transient interactions and stable complexes. Some interactions are transient, which are induced in response to a specific cellular event and quickly released after triggering a reaction. On the other hand, some interacting proteins form stable complexes to perform biological roles together and such complexes can last a long time in cell. In particular, some proteins cannot even fold into stable structures by themselves. They can only have stable structures and perform their function in a complex.

1.3.2. Post-translational Modification Effect

Post-translational modification (PTM) is very important for protein formation, regulation, and interaction. Many proteins, especially in eukaryotes, are modified after their synthesis by adding sugars (glycosylation), phosphate (phosphorylation), sulfate and some other chemicals. Such modifications often play an important role in modulating the function carried out by the protein. For example, some proteins can switch between active and inactive forms by such modifications. In other cases, a newly synthesized protein coming off the ribosome is often an inactive precursor protein, then it is cleaved into smaller proteins, which interact with other proteins and perform the biological function. Therefore, the protein-protein interaction patterns and partners are dynamic and highly dependent on PTMs. When the biological condition is changed, a protein can undergo PTMs and has a new modulating function with new interacting partners.

1.3.3. Multi-body Effects

Sometimes, due to "multi-body" effects, protein-protein interactions in a complex may not be decomposed into a set of independent binary protein-protein interactions. For example, two proteins may interact in a protein complex that has multiple components, but they do not interact with each other without the presence of the other components in the complex, since the two proteins alone cannot form a stable complex. More interestingly, whether two proteins interact with each other may depend on the presence of a small molecule, i.e., the so-called allosteric effects. For example, for a signaling protein built from multiple modular domains, a specific ligand can robustly activate or deactivate the interactions between these domains. There are many cases that have been studied thoroughly, including cAMP-mediated allosteric control over cAMP receptor protein (CRP) conformation and activity [20]. The allosteric effects can generate cooperative repression or reciprocally cooperative activation using multiple weak interactions, displaying higher specificity and sensitivity for signaling switches [21]. These multi-body effects contribute to the dynamic nature of the protein-protein interaction map.

1.4. Types of Protein-protein Interaction

Protein-protein interactions are also complex from structure perspective. The structural interface between two interacting proteins can be of three different types: (a) coiled-coil interaction, (b) rigid-body protein binding, and (c) flexible binding.

a. Coiled-coil Interaction

Coiled-coil conformation contains twisted α-helices, which are characterized by a repeating sequence of seven amino acids, $(abcdefg)_n$, in which the a- and d-position residues are hydrophobic, while the e- and g-position residues are usually polar or charged [22]. Coiled coils

mediate protein-protein interactions or oligomerization through intertwining two coiled coils together.

b. Rigid-body Binding

"Rigid-body binding" [23] means that each polypeptide component in a protein complex has a stable structure by itself and the structure of a component in a protein complex closely resembles its structure in its free, native state. This does not exclude some small conformational changes, in particularly on the side chains of residues buried at the binding interface.

c. Induced Binding

"Induced binding" refers to the case where the backbone conformation is significantly changed upon protein binding [24]. Sometimes, in solvent a polypeptide component in a protein complex does not have a stable structure by itself.

1.5. Size of Interactome

The total number of interactions between all proteins in an organism, or the size of interactome (*Nint*) can be estimated based on current experimental data and the size of proteome (i.e., the number of proteins in a genome) [25,26]. *Nint* depends on the number of predicted ORFs (*N*), the average number of interactions per protein observed in experiments (*a*), the percentage of questionable interactions or false positives (*b*, typically 10-20%), and the number of ORFs with unknown function (*c*). It is found that ORFs with unknown function tend to have only half as many interactions as known proteins. Therefore, estimated *Nint* is:

$$Nint = (a*N/2) - (a*b*N/2) - (a*c/4) = N*a*[(1 - b) - c/2] /2$$

For example, in yeast: *a=10, N=6400, b=10%, c=2000.* Hence, *Nint = [6400*(1-0.1)-2000/2]*10/2 = 23,800*. *Nint* in other genomes can be estimated similarly. The values of *a* and *b* are expected to be similar in all genomes.

1.6. Importance of Bioinformatics in Analyzing Protein-protein Interaction Data

High-throughput protein-protein interaction data are generated from technology-driven experiments, which provide rich information with the ever-increasing volume. However, the information explosion does not mean biological knowledge explosion. Understanding biological meaning from the raw outputs of experimental techniques is becoming the bottleneck in the application of high-throughput protein-protein interaction data. These challenges require bioinformatics in a number of aspects. First, with more and more data accumulating, the databases are needed to store, document, and describe the protein-protein interactions and visualization tools are needed to display and navigate the interaction network. It is indisputable that publicly available databanks play a fundamental role in disseminating the data to the biological community. Second, inherent to the high-throughput nature of the experimental techniques is heterogeneity in data quality with the false positives and false negatives. For the efficient use of data, the computational and statistical models are needed to deal with data quality control such as reliability assessment and validation. Finally, new computational tools are in demand to infer new biological discoveries and validate those hypotheses based on high-throughput protein-protein interaction data. The computational approaches towards the fruitful utilization of protein-protein interaction data will not only provide tools for experimental biologists but also result in important scientific insights into the cellular mechanisms.

To our knowledge, there has not been any review paper to comprehensively address the computational analysis of high-throughput protein-protein interaction data, although some topics have been discussed in several review papers [27,28]. In this paper, we will provide a systematic and comprehensive survey for computational analyses of high-throughput protein-protein interaction data, including their databases, assessment, prediction, analyses, and biological inferences.

2. EXPERIMENTAL METHODS FOR PROTEIN-PROTEIN INTERACTION IDENTIFICATION

There are many experimental methods for the identification of protein-protein interactions and characterization of their biological importance [29]. Traditionally, protein-protein interactions have been studied on the individual basis by low-throughput technologies (immuno-precipitations [30], pull-down [31], etc). In the so-called "proteomics" approaches, several techniques are applied for studying protein-protein interactions in a proteome scale [32,33]. These techniques are summarized in Table **1**. To effectively analyze high-throughput protein-protein interaction data, it is important to know the source of the data together with the strength and limitation of the associated experimental technique.

Different experimental methods can generate different types of protein-protein interactions. Some technologies such as yeast two hybrid, protein chip, and phage display can detect the binary interactions while the others can identify protein complexes. In a protein interaction graph, a binary interaction can be represented as an edge with the two interacting proteins as vertices. A protein complex can be regarded as a connected graph. However, the topology of the graph, i.e., which pairs of the proteins within a complex physically interact with each other is unknown. Therefore, we cannot get the exact binary information from a protein complex. A more complicated issue is that due to multi-body effects as discussed in section 1.3.3, a true protein-protein interaction within a complex may not be detectable in a binary interaction.

One major issue with the high-throughput experimental technologies is the generation of false negatives and false positives. Proteins interact with one another with a wide-range of affinities and timescales. Detection of such interactions is often at the margin of observation, and non-physiological interactions result in noise. The different techniques have different noise level since each technique has its own strengths and weaknesses in detecting certain types of interaction. One should take into account the

Table 1. Current Major Technologies in Studying Protein-protein Interaction

Method	Experimental condition	Binary interaction *vs.* complex	High-throughput	Noise level
Two-hybrid system	*In vivo*	Binary	Yes	High
Immuno-precipitations	*In vitro*	Complex	No	Low
Pull-down	*In vivo*	Complex	No	Low
Mass spectrometry	*In vitro*	Complex	Yes	High
Protein chip	*In vitro*	Binary	Yes	High
Phage display	*In vitro*	Binary	Yes	High

technique bias and limitation for computational analyses of the data. For example, most current protein-protein interaction experimental techniques are not effective to characterize protein interactions involving integral membrane proteins. To overcome this shortcoming, some genetic screening systems have been developed for assaying membrane protein interactions, such as the Ras recruitment system [34], the G protein based screening system [35], the split-ubiquitin system [36,37], etc.

The genome-wide protein-protein interaction studies have been carried out in many organisms such as bacteriophage T7 [38], Hepatitis C virus (HCV) [39], *Helicobacter pylori* [40], *Caenorhabditis elegans* [41,42], *Saccharomyces cerevisiae* [43-46] and mouse [47]. While our review addresses general issues in the analyses of protein-protein interaction data, the examples used will focus on yeast *S. cerevisiae*, which is not only a good model organism for eukaryotes but also contains the most protein-protein interaction data generated so far.

2.1. Yeast Two-Hybrid System

Yeast two-hybrid system is the most widely used method for detecting protein-protein interactions, since its original description in 1989 [48]. Initially it was designed as a test to identify an interaction between two known proteins, and then it was rapidly developed as a screening assay to find partners for a protein at the high-throughput mode [49]. The yeast two hybrid technique carries out two fusions: a bait protein fused to the DNA-binding domain of a transcription factor and potential interacting partners fused to a transcriptional activation domain. An interaction between the bait and an interacting partner (prey) results in the formation of a functional transcription factor that induces the expression of a specific reporter gene, thereby, allowing such interactions to be detected. It should be noted that this approach forces the protein-protein interaction between the bait and prey to occur in nuclei, and some errors of measuring protein interactions may result from this restriction. In the recent years, several variations of the two-hybrid system have been developed [50], and the methods also extend to organisms others than yeast, including bacteria and viruses [51].

Many protein-protein interaction data have been generated using two-hybrid system. In a proteome-wide study on yeast by Uetz *et al*. two designed experiments were used, i.e., one with a low-throughput protein array and one with a high-throughput array. In the low-throughput array, 192 bait proteins were tested against a completed set of 6000 prey proteins, a total of 281 binary interactions were identified. The high-throughput approach used the complete set of 6000 yeast proteins as baits against the completed set of 6000 prey proteins. This second approach identified 692 interacting protein pairs involving 817 unique proteins as either bait or prey proteins. An independent, large-scale project by Ito *et al*. was also conducted for the whole yeast proteome. This study detected 3278 proteins involved in 4589 putative protein-protein interactions.

2.2. Mass Spectrometry

A typical approach to identify proteins in a protein complex is done by the separation of the various proteins of an extract by gel electrohporesis followed by mass spectrometric analysis of the protein gel spot. The precise identification of polypeptides can be done by searching the molecular weights against a protein database. High throughput is achieved by MALDI(automated matrix-assisted laser desorption/ionization), providing a list of masses of the fragmented peptides. Matching this list against a list of pre-calculated peptide masses from an appropriate protein sequence database can characterize the isolated protein.

Recently, Gavin *et al.* and Ho *et al.* took a new approach to screen protein-protein interaction in the proteome-wide scale. This method is particularly effective for identifying protein complexes that contain three or more components. First, the authors attached amino acid tags to hundreds of proteins, thus, creating bait proteins. Then they encoded these proteins into yeast cells, allowing the modified proteins to be expressed in the cells and to form physiological complexes with other proteins. Then, by using the tag, each bait protein was pulled out, and usually it fished out the entire complex. The proteins extracted with the tagged bait were identified using MALDI method. This approach for characterization of protein complexes in a large scale was named TAP (tandem affinity purification). Notably using the tags may perturb some protein interactions and result in errors.

By using TAP, Gavin *et al.* have identified 1440 distinct proteins within 232 multi-protein complexes in yeast after processing 1739 genes as baits. 91% of these complexes contain at least one protein of unknown function. Ho *et al.* reported another application example for yeast using the same general approach, which they termed HMS-PCI (high-throughput mass spectrometric protein complex identification) methods. Ho *et al.* constructed an initial set of 725 bait proteins, from which they identified 3617 associated proteins, covering about 25% of the yeast proteome.

2.3. Protein Chip

Another approach to generate the protein-protein interaction map is protein chip technology [52]. In this approach, proteins are expressed, purified and screened in a high-throughput scale so that a large number of proteins can be attached to a planar substrate (chip) as discrete spots at known locations, where the proteins keep their folded conformation and their ability to interact specifically with other proteins. A solution containing labeled protein(s) to be tested is then incubated with the chip, and then the chip is washed. Specific interactions between proteins on the chip and protein(s) from the solution are indicated by the position of the label. In addition to the rapid simultaneous measurement of large number of samples, protein chip technology has substantial advantages over conventional methods, especially the high signal-to-noise ratio, small amount of sample needed, and high sensitivity. On the other hand, attachment of proteins to chip can disrupt some protein interactions as well.

Recently Zhu *et al.* [53] identified many new calmodulin and phospholipid interacting proteins by application of this technique. They first fused 4800 yeast ORFs to glutathione-S-transferase (GST) and expressed the fused proteins in yeast. Subsequently, they printed the purified proteins onto glass slides, thus generating a matrix that was then screened for finding the interacting proteins and phospholipids. Zhu *et al.* also developed protein chips to conduct high-throughput biochemical assays of 119 protein kinases for 17 different substrates.

2.4. Phage Display

Phage display is another method for studying protein-protein interactions [54]. It is based on the ability of bacteriophage to express engineered proteins on their surface coat. Diverse libraries such as peptides, antibodies and protein domains corresponding to gene fragments can be displayed on the coat through an artificially inserted DNA sequence. By immobilizing a protein on an affinity support, phages with display proteins binding to the immobilized protein can be selected from the library. These phages selected on the basis of their interaction with the immobilized protein can be enriched, and the protein on the phage that interacts with the immobilized protein can be identified. Bartel *et al.* screened a library of random bacteriophage T7 protein fragments against random libraries of T7 activation domains. The authors found 25 interactions among the 55 phage proteins. Since the displayed protein is expressed artificially rather than in its native cellular environment, inevitably some error can occur when using this method to detect protein-protein interaction.

3. PROTEIN-PROTEIN INTERACTION DATABASES AND VISUALIZATION

Data management is critical for using high-throughput biological data, including protein-protein interaction data. The massive amount of protein-protein interaction data that have been generated are impossible to handle systematically without a computer database, let alone many more such data being obtained daily. To collect, retrieve, and describe protein-protein interactions, several databases have been established. Protein-protein interaction information retrieved from the literature can also be added to the databases [55,56]. These databases can be accessed through the Internet and they typically have user-friendly interfaces. Most of them also provide good search capacity. One can search the interactions that a particular protein involves by querying its ORF name or gene name. The functional annotation, when available, is usually given for a protein participating in an interaction. The experimental source and reference are also provided for a particular interaction in some databases. As we will discuss in section 4, such information can help evaluate errors and validate protein-protein interactions. In addition, protein-protein interaction data in a centralized database provide a starting point (input) for computer programs that analyze protein-protein interaction at the proteome scale.

Most protein-protein interaction databases also provide visualization tools, where a protein interaction network is represented as a graph with proteins as vertices and interactions as edges. In such a graph, all the interacting partners of a specific protein can be displayed and the paths between two given proteins can be easily identified. With more and more protein-protein interaction data collected into databases, the text listing of interactions are hardly sufficient to evaluate and compare such huge amount of information. The visualization tools can help researchers validate interactions from different experimental sources, make sense of interaction paths, and construct hypotheses for biological pathways.

To fully utilize protein interaction network, integrated tools implemented in the database enable the combinational analysis of various types of biological data on proteins and interactions, which will help researchers in biological discoveries. For example, PIMRider provides an integrated platform for the exploration of protein interaction maps and other genomic/proteomic information [57].

In this section, we will review nine widely used protein-protein interaction databases, as summarized in Table **2**. We will also provide an example of visualization for protein-protein interactions.

DIP

The DIP [62] provides an integrated tool for browsing and extracting information about protein-protein interactions, which are either generated from various high-throughput

Table 2. Online Protein-protein Interaction Databases

Database name	Acronym	URL	Database size		Visu.	Acad.	Com
			Binary	Compl.			
Database of Interacting Proteins	DIP [58]	http://dip.doe-mbi.ucla.edu	18,000		Yes	Yes	No
Biomolecular Interaction network Database	BIND[59]	http://binddb.org	6171	851	Yes	Yes	Yes
Munich Information Center for Protein Sequences	MIPS [60]	http://mips.gsf.de/proj/yeast/ CYGD/db/	11,200	1050	No	Yes	Yes
Molecular Interaction Database	MINT [61]	http://cbm.bio.uniroma2.it/ mint/	3786	782	Yes	Yes	Yes
Biomolecular Relations in Information Transmission and Expression	BRITE	http://www.genome.ad.jp/ brite/	5506		No	Yes	Yes
Pathcalling Yeast Interaction database	PathCalling	http://genome.c.kanazawa-u.ac.jp/Y2H/	957		Yes	Yes	No
A Protein-Protein Interaction database	Interact	http://www.bioinf.man.ac.uk/ resources/interactpr.shtml	1000	200	Yes	Yes	Yes
Hybrigenics	PIMRider	http://pim.hybrigenics.com/	1400		Yes	Yes	No
The General Repository for Interaction Datasets.	GRID	http://biodata.mshri.on.ca/grid/	14,318		Yes	Yes	Yes

The table shows, in different columns, the name of the protein-protein interaction database, its acronym, its Web address, the size of the database as of August 2002 (for number of binary protein-protein interactions and number of protein complexes), and whether the database has visualization tool (Visu.), free academic use (Acad.), and free commercial use (Com.).

experiments or collected from literature search [63]. The vast majority of data are from yeast, *Helicobacter pylori* and human. The DIP allows the visual representation and navigation of protein-protein interaction networks. The reproducibility of a given interaction can be assessed visually by the thickness of the line between two proteins [64]. A related tool LiveDIP also integrates protein-protein interactions network with large-scale gene expression data [65].

BIND

The BIND database stores various interactions between molecular compounds including protein-protein, protein-RNA, protein-DNA and protein-ligand interactions. Description of an interaction includes subcellular localizations of the proteins involved in an interaction and experimental conditions used to observe the interaction. This database also contains the information of molecular complexes and pathways. BIND can be visually navigated using a Java applet. Currently 11,171 various interactions and 851 protein complexes are represented. BIND also provides a framework for users to build their own protein-protein interaction databases.

MIPS

The MIPS Comprehensive Yeast Genome Database (CYGD) [66] provides the protein-proteins interactions together with sequence and function information for all the genes in the budding yeast *Saccharomyces cerevisiae*. All the protein-protein interaction data are available to download in a text file. In addition, the database contains other compiled yeast data for download, such as functional classification category, subcellular localization category, EC number category, etc.

MINT

The MINT database stores data on functional interactions between proteins, which are extracted from the scientific literature. MINT also includes the information about enzymatic modifications of one of the partners. The interaction data can be extracted and visualized graphically. Presently MINT contains 4568 interactions, 782 of which are indirect or genetic interactions.

BRITE

BRITE [67] is a database of binary protein-protein interactions retrieved from literature, high-throughput data based on yeast two-hybrid system of *S. cerevisiae*, and yeast two-hybrid interactions of *H. pylori* proteins.

Path Calling

The Pathcalling database contains yeast protein-protein interaction data from high-throughput yeast two-hybrid experiment. Data are available at the Curagen website

(http://www.curagen.com) for free academic use. Visualization tools are available for the protein interaction network. Fig. **1** shows an example using the graph to map interactions around the protein Nup100p. In this graph, each node represents one protein and each edge marks an interaction. All the immediate neighbors and the next immediate neighbors in the protein interaction map for Nup100p are displayed. By clicking the node, one can navigate its neighbors and the detail description of gene, including protein sequence and function.

Interact

Interact is a database for protein-protein interactions constructed with object oriented technology that provides a means to fully accommodate and query the data associated with protein interactions. Unified Modelling Language (UML) 6 was used to model the database. In this database 3D visualization of protein cluster is available.

PIMRider

PIMRider contains protein-protein interaction data for *Helicobacter pylori*, which has been studied using genome-wide two-hybrid assay [68]. 1273 protein-protein interactions are viewed as a graph and assigned with PIM Biological Score (PBS® score) that quantifies the reliability of each interaction and allows the filtering of interactions based on their reliability. The PBS® score takes into account the characteristics of the libraries screened and the target organisms as well as the results of the screens. The PBS® score ranges from 0 (the best) to 1 (the worst). In PIMRider, tools are also developed to identify the specific protein domain involved in a given interaction and to query pathways between two proteins.

GRID

The GRID is a database of genetic and physical interactions. It includes 14,138 unique protein-protein interactions at present, including the data from MIPS and BIND. Osprey Network Visualization System (a graphical visualization tool at http://biodata.mshri.on.ca/osprey/index.html) is integrated into the database to let users visualize searched results. Users can also upload their own datasets and visualize the interaction maps.

4. ASSESSMENT OF PROTEIN-PROTEIN INTERACTION DATA

A general strategy for high-throughput experimental technologies in detecting protein-protein interactions is to be selective enough to minimize the report of false interactions yet sensitive enough to maximize the detection of all biologically true interactions. However, currently this goal is far from being achieved. In fact, one major issue with the high-throughput protein-protein interaction data is the high error rate, compared with the data generated from traditional low-throughput methods. To use high-throughput protein-protein interaction data for biological inference effectively, it

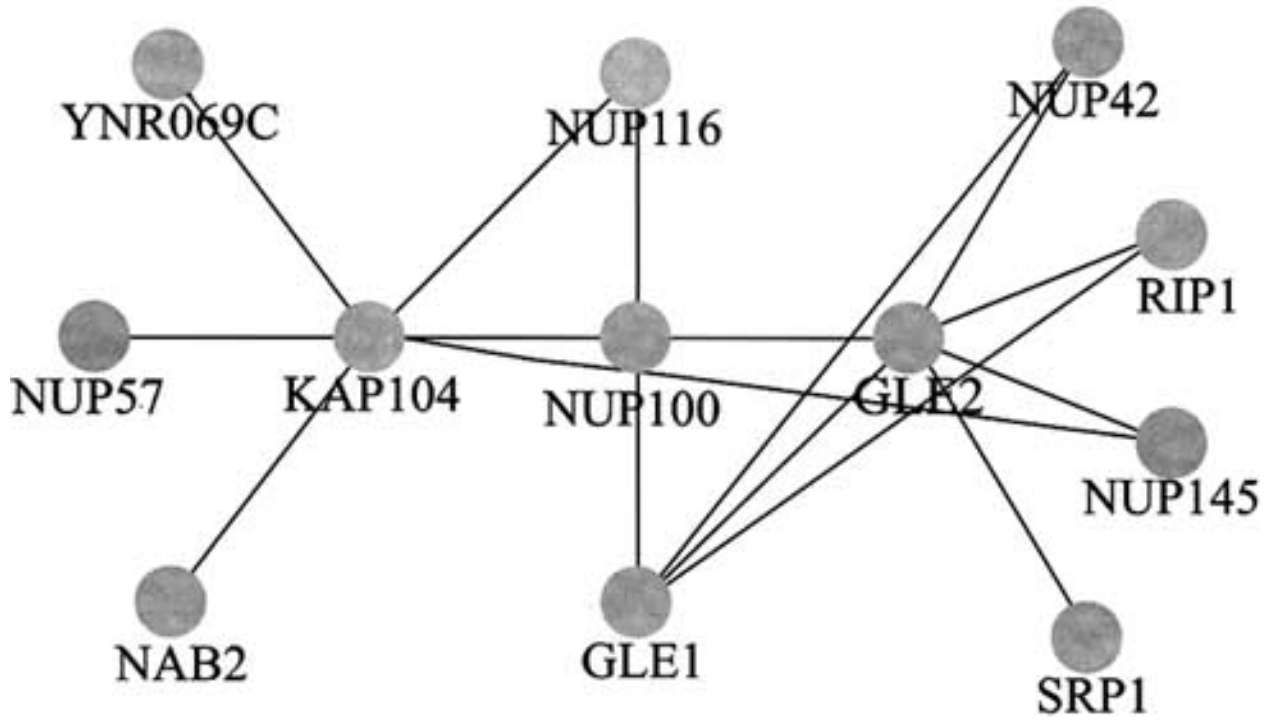

Fig. (1). The protein interaction map around Nup100p from PathCalling. A gene is represented as a vertex and a protein-protein interaction is indicated as an edge.

is essential to evaluate the coverage and reliability of the data. In this section, we will discuss the origin of errors and provide examples to show the characteristics of the errors. We will also address how to assess the reliability of protein-protein interaction data using computational methods.

4.1. False Negatives and False Positives

The difference between actual biological protein-protein interactions and measured protein-protein interactions may arise from three factors. (1) **The dynamic nature of protein interaction map.** Protein expressions and interaction patterns are changing under different biological conditions. Proteins interact with one another with a wide-range of affinities and time scales. Consequently, detection of such interactions is often at the margin of observation and each measurement of protein-protein interactions can only capture a snapshot of the dynamic protein interaction map under a specific condition. (2) **The limitation of the technologies.** As discussed in Section 2, any high-throughput protein-protein interaction technology creates a substantial disruption of normal cellular function, which can make the protein interaction pattern deviate from the one under the native biological condition. For example, mass spectrometry might fail to uncover transient or weak interactions while yeast two-hybrid assay might not detect interactions that are dependent on PTMs or interactions having the "multi-body" effects. (3) **The errors during the measurement.** In this case, the technology is capable of identifying an interaction correctly. But due to operation problems during the experiment, the interaction is not identified correctly. These three factors make the protein-protein interaction maps different with the use of different technologies and in different labs using the same technology. Here we focus on the second and the third factors, i.e., errors caused by the technology drawbacks and measurements, including both false negatives and false positives.

False negatives are the biological interactions that are not detected by the experiments. For example, in yeast two-hybrid assay, which relies on the transcriptional activation of the reporter gene, the incorrect folding, inappropriate subcelluar localization, and absence of certain necessary post-translational modifications can cause the false negatives. For protein complex mass spectrometry identification methods, it is also likely to generate the false negatives. For example, it may not detect some transient interactions and it may miss some complexes that are not presented under the given experimental conditions. Moreover, the loosely associated components in a complex may be washed off during the purification process.

False positives are generated by experiments that are not true biological interactions. In two-hybrid assay, false positives arise when the expression of the reporter gene occurs under conditions that are not dependent on bait/prey protein-protein interactions. For example, bait proteins may activate the transcriptional of reported genes above a threshold level by themselves in the actual physiological conditions. Two-hybrid assay can also produce some non-specific interactions that are not biologically relevant, especially between proteins normally existing in different subcellular location or different tissues. Large-scale protein complex identification approaches can also generate false positives. When the bait protein is used to fish out the entire complex components, some other unrelated proteins (e.g., proteins in different compartments of a cell) may attach with the complex and be pulled out together. Even within a true complex, it is challenging to distinguish the true binary interactions between the component proteins. If we assign binary interactions between all proteins in a complex, it can generate false positives.

4.2. Overlap and Complementation Analysis of Protein-protein Interaction Data

Until now, there are 5125 publicly available binary interactions identified from yeast two-hybrid experiments in high-throughput assays or low-throughput assays. In addition, 49,094 binary interactions can be assigned for the protein complexes identified by TAP (tandem affinity purification) and HMS-PCI (high-throughput mass spectrometric protein complex identification) methods, assuming any two components in a protein complex interact with each other. However, our analysis shows that strikingly few interactions (55 interactions) are commonly represented in yeast two-hybrid, HAP and HMS-PCI. There are only 1920 interactions supported by at least two out of the three technologies.

Unexpectedly, not only the data produced by different technologies do not overlap significantly, the data produced at different labs using the same technology differ substantially. For yeast two-hybrid data, only 141 interactions were common in both data sets from Uetz *et al.* and Ito *et al.* Interestingly, neither of those two studies identified more than 15% of previous published interactions [69], suggesting that coverage of protein interaction map is very sparse and the map in a simple organism like yeast may be more complex than expected. The approaches taken by Gavin *et al.* and Ho. *et al.* are clearly powerful, but they also have limitations. Both groups found a significant number of false-positive interactions with failure to identify many known associations. Gavin *et al.* estimated that the probability of detecting the same protein in two different purifications from the same entry point is about 70% by purifying 13 large complexes at least twice. We also studied the overlap and coverage using the datasets from Uetz *et al.* (yeast two-hybrid) and Ho *et al.* (mass spectrometric protein complex identification) and compared the binary interactions involved (see Table **3**). We found that the common interactions detected by both yeast two-hybrid assay and mass spectrometric protein complex identification are only 4.4%.

In fact, not only the coverage of different techniques is different, the protein-protein interaction data generated by each technique have unique characteristics. Mering *et al.* [70] comparatively assessed the high-throughput protein-protein data generated from different sources in yeast such as yeast two-hybrid assay, mass spectrometry of purified complexes, correlated mRNA expression, genetic interactions and *in silico* predictions through genome analysis and found that data generated from different methods have different distributions with respect to functional categories of interacting proteins, thus indicating

Table 3. The Coverage Comparison between Protein-protein Interaction Data Generated from Yeast Two-hybrid Array (Uetz *et al.*) and Protein Complex Mass Spectrometry Identification Method (Ho *et al.*). The known Interactions come from 2301 Annotated Binary Protein-protein Interactions Maintained at MIPS[60] which we used as Reference Dataset

Experimental method	Baits having interaction	Identified interactions	Known interactions
Yeast two-hybrid assay	71	224	10 (4.5%)
Protein complexes mass spectrometry identification	121	1182	53 (4.5%)

that those methods have specific strengths and weaknesses. The lack of overlap between datasets demonstrates that the current data is far from saturating, which suggests that high-throughput technologies might provide complementarities to each other. Therefore, the combination of protein-protein interaction data from different resources will substantially expand our knowledge of protein-protein interaction network.

The described systematic protein-protein interaction assay methods clearly show that single screens rarely capture all proteins capable of interacting with the given bait. In the yeast two-hybrid array, even multiple screens with different two-hybrid variants usually produce only partial overlaps. This again shows that several different methods are needed to complement each other in order to identify as many true interactions as possible. To illustrate the complementarily between the two-hybrid and mass spectrometry methods, we considers two examples. The first example is the interactions with Rrn10p (YBL025W), RNA polymerase Ⅰ-specific transcription initiation factor. In Ho *et al.*'s dataset there was no detected protein complex when using Rrn10p as bait protein. Yeast two-hybrid assay detected the interaction between Rrn10p and Rrn9p, which was also validated by experiments. RRN10 mRNA abundance is 0.7 copy per cell based on the genome-wide analysis of mRNA abundance in yeast while the average number is 2.8 copies per cell [71]. Thus, it is likely that the protein complex may not be formed due to the low abundance of the bait protein. It is also possible that this pair of proteins has only a transient interaction rather than forming a stable complex. On the other hand, yeast two hybrid is independent of the protein concentration level and capable of detecting the transient and weak interactions.

A converse example is Cmd1p, a small ubiquitous Ca^{2+}-binding protein regulating a wide variety of proteins and processes in all eukaryotes [72]. For this protein, yeast two hybrid cannot detect the interactions while mass spectrometry using different baits can discover several complementary protein complexes (see Fig. **2**). In response to a Ca^{2+} signal, Cmd1p binds Ca^{2+} and consequently undergoes a conformational change that allows it to bind and activate a host of target proteins. Probably due to the absence of native physiological condition, yeast two-hybrid assay cannot detect such a protein-protein interaction.

4.3. Reliability

Currently there is no systematic statistical method developed to assess the confidence level of an interaction accurately. However, several heuristic approaches have been used for this purpose. These methods can provide the side evidence for an interaction, and as a result increase the

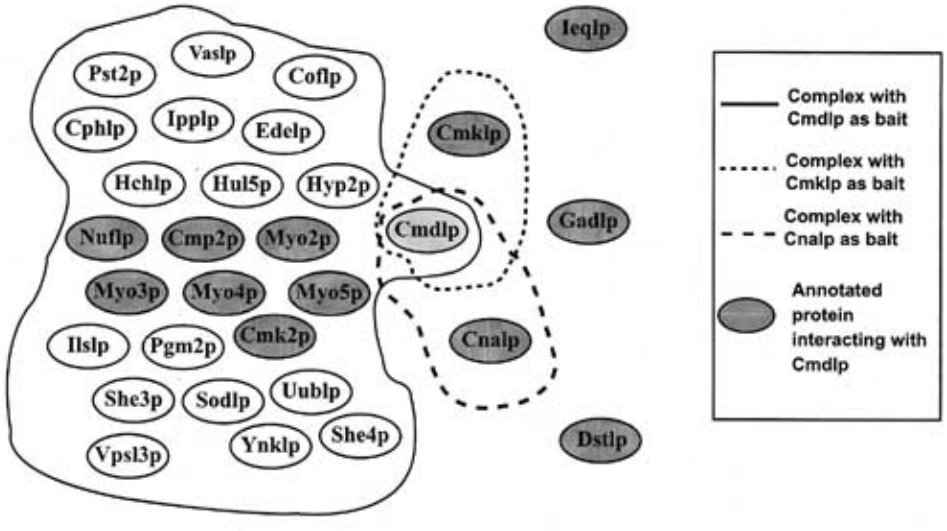

Fig. (2). Three Cmd1p related protein complexes identified by Mass spectrometry. The annotated Cmd1p interacting partners, shown in dark color, come from known experimental results reviewed in the paper from Cyert .

confidence level of interactions measured from high-throughput techniques.

Reliability of a reported interaction is increased by the observations of the same interaction using different methods. For example, if an interaction is detected by two distinct experiments, the joint observations enhance the confidence level for this particular interaction. Large-scale two-hybrid screens can identify some classes of systematic false positives using multiple, independent screen under standardized conditions. For example, some false positives related to particular proteins tend to appear repeatedly in screens with unrelated baits. In a yeast study if prey proteins are selected with more than three unrelated bait proteins from a pool of 100 bait proteins, they will be discarded. The array screens by Uetz *et al.* used reproducibility to estimate the reliability by testing each individual two-hybrid pair twice in a highly standardized and parallel fashion. False positives are often generated by mutations in the baits, prey plasmids, or reporter genes. When screens are done in duplicate, such mutations are unlikely to occur simultaneously.

Literature is a valuable resource to validate the protein-protein interaction generated by high-throughput techniques. The idea is that if two protein names appear in the same article, they have a chance to interact with each other. Such information about interacting proteins, albeit unreliable, can validate protein-protein interactions or at least provide clues for judging an interaction. Based on the literature mining method, a recent work created gene-to-gene co-citation network for 13,712 named human genes from analyzing over 10 million MEDLINE records [73].

Computational approaches can also be used to assess the reliability of the observations of high-throughput protein-protein interactions. To verify protein-protein interaction data, Deane *et al*. [74] developed two methods, i.e., expression profile reliability index and paralogous verification method (PVM). By comparing gene expression profiles of the proteins involved in an interaction, expression profile reliability index estimates the likelihood of the interaction to be biologically meaningful. The idea is that proteins with higher correlated expression pattern are more likely to interact with each other. Paralogous verification method is based on the observation that if two proteins are paralogs, the proteins that they interact with tend to be paralogs as well. PVM evaluated 8000 pairwise protein interactions in yeast and 3003 interactions were confidently identified. Some other computational methods that we will address in the following sections can also help assess the confidence level of a protein interaction, including using the relationship between protein-protein interaction data and other types of data (Section 5), as well as *in silico* prediction (Section 7) to help validate protein interactions.

5. RELATIONSHIP BETWEEN PROTEIN-PROTEIN INTERACTION DATA AND OTHER BIOLOGICAL DATA

Inherent in the growing collections of protein function and subcellular localization data, protein structure, gene expression data and protein-protein interaction data is the internal relationships between different aspects of the same set of genes/functions. These relationships provide a basis for cross-validating the data and offering more information than what a single source of data can achieve. For example, the protein functional role and subcellular localization information can be used to validate protein-protein interaction data. Given relatively low reliability of these computational relationships, one can use them to increase the confidence of a protein-protein interaction, but it is hard to reject a protein-protein interaction just because its pattern deviates from the general relationships. The correlation study between protein-protein interaction and gene expression can be used to formulate more meaningful biological hypotheses by improving hypotheses generated from either approach individually. Protein structure provides rich information about how proteins interact with each other at the atomic details. Therefore, the integration analysis of biological data from different sources offers a deepened knowledge exploration for understanding cellular mechanisms. In this section, we will discuss the relationship between protein-protein interaction data and other types of data, including subcellular location, function category, gene expression profiles, and protein structures.

5.1. Subcellular Localization

The subcellular distribution of proteins within a proteome is useful and important to a global understanding of the molecular mechanisms of a cell. Protein localization can be seen as an indicator of its function. Localization data can be used as a means of evaluating protein information inferred from other resources. Furthermore, the subcellular localization of a protein often reveals its activity mechanism. In a physical protein-protein interaction, the two proteins involved should be localized at the same subcellular compartment. If an interaction between two proteins that are known to have the same subcellular localization, the confidence level for the interaction increases. Therefore, the study of relationship between protein-protein interactions and the partners' subcellular localizations can provide an evaluation method for validating protein-protein interaction data generated from high-throughput experiments. On the other hand, a protein may have several subcellular localizations. For example, a translocalization of NK-κB can move the protein from cytoplasm to nuclei [75]. In this case, we can find pairs of interacting proteins have different subcellular localizations, when the alternative subcellular localizations of the proteins are not recorded in the database.

We assembled 2301 annotated binary protein-protein interactions maintained at MIPS, a manually curated database and took them as the trusted true interactions. We also derived protein subcellular localization information from MIPS. In yeast 2358 ORFs have been known their subcellular localizations, among which 169 ORFs can be localized in more than one subcellular compartments. For all 2301 interactions, the localizations of both partners in each interaction are known. We found that there are 2124 interactions (92 %) whose partners have the same subcellular localizations. The data set is biased towards particular cellular localizations of interacting proteins (Fig **3**), for

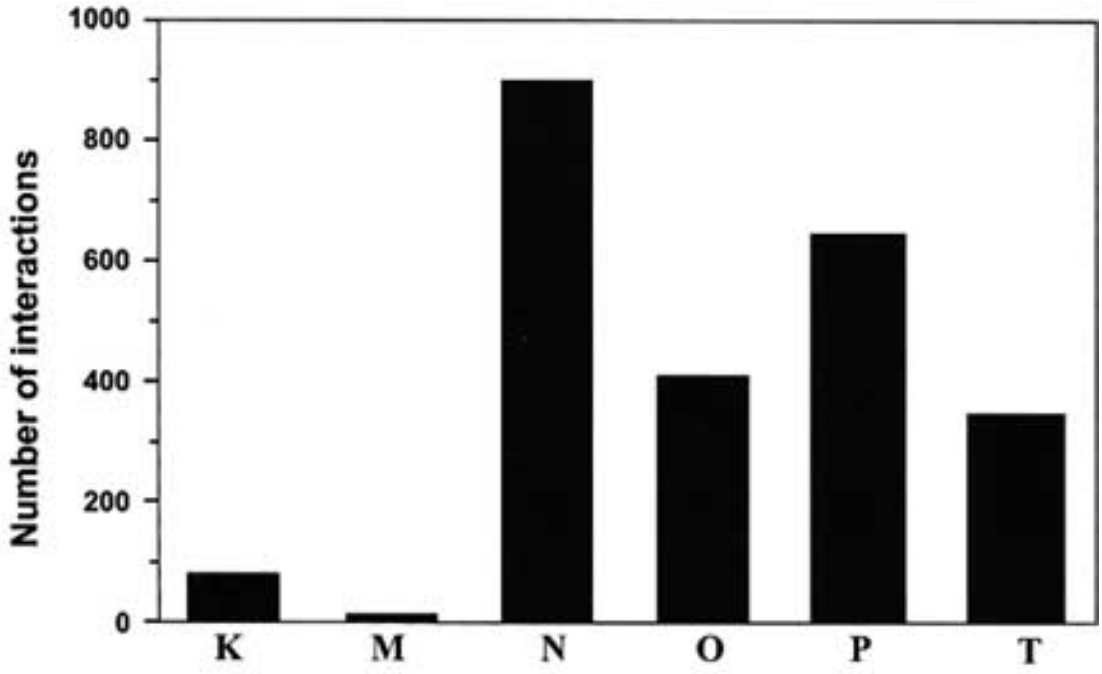

Fig. (3). The distribution of protein-protein interactions whose partners have the same subcellular localizations. The abbreviations of localizations are: N--nucleus; P--cytoplasm; K--cytoskeleton; M--plasma membrane; T--mitochondria, O--cell organelles (ER, Golgi, transport vesicles, peroxisome, endosome, vacuole, microsomes, and lipid particles).

example, the number of interactions involving plasma membrane proteins is very small, showing the technological limitation in detecting such interactions.

Table **4** shows a comparison between the observed number of protein-protein interaction pairs for a given combination of subcellular localizations and the expected number calculated from random combination of proteins involved in the interactions. Clearly the observed number of protein-protein interactions belonging to the same subcellular localization is much greater than expected. Conversely the observed number of protein-protein interactions belonging to different subcellular localizations is much less than expected, except for the interactions between nuclear proteins and proteins in cell organisms such as ER, golgi, transport vesicles, peroxisome, endosome, vacuole, microsomes, lipid particles. This may be because proteins can move between the two compartments (protein translocalization), and some nuclear proteins require modification and sorting in those cell organisms.

5.2. Function Catalogue

A protein interaction is often associated with a particular biological pathway. Hence, it is not surprising to see a pair of interacting proteins to have the same cellular role. To further assess the relationship between the cellular roles of a pair of interacting proteins, we used 3936 yeast ORFs' cellular functions that have been hierarchically classified at MIPS. We clustered those cellular functions into 11 broad functional categories using the same classification method proposed by Mering *et al.*, as shown in Table **5**. For 2301 well-annotated protein-protein interactions at MIPS, each ORF can be assigned into a known function category and both proteins participating in an interaction belong to the same function category for all the cases. It is likely that any interaction involving two proteins of different cellular roles was removed from this data set since the interaction is considered unreliable. Fig. **4** showed the distribution of protein-protein interactions for different function categories, indicating that the distribution is biased. However, the biased distribution may be caused by the small size of dataset, which is far from saturating currently.

5.3. Gene Expression Data

Analysis of gene expression data is currently one of the most exciting research fields in genomics. Computationally clustering individual gene expression measurements provides a new way to exploit and infer the information in

Table 4. The Observed Number and Expected Number of the Protein-protein Pairs belonging to the Same or Different Subcellular Localizations. "Ob." Means the observed number and "Pr." Means the Expected Number Calculated from the Probability Distribution, based on the Assumption that Two Proteins involving in an Interaction have Independent Probabilities of Subcellular Localization Distribution. The Expected Number of Interactions between Proteins in X Subcellular Localization and Proteins in Y Subcellular Localization Equals the Total Number of Interactions Multiplied by the Probability of X and the Probability of Y and 2, where the Probability of X and the Probability of Y are Calculated from 2358 yeast ORFs with known Subcellular Localizations. The Notations for Subcellular Localization are: N--nucleus; P--cytoplasm; K--cytoskeleton; M--plasma membrane; T--mitochondria, and O--cell organelles (ER, golgi, transport vesicles, peroxisome, endosome, vacuole, microsomes, lipid particles)

Sub. Loc.	K		M		N		O		P		T	
	Ob.	Pr.	Ob.	Pr.	Ob.	Pr.	Ob.	Pr.	Ob.	Pr.	Ob.	Pr.
K	**79**	**6**	0	1	8	60	0	26	9	41	0	22
M	0	1	14	0	2	14	0	6	10	9	0	5
N	8	60	2	14	898	640	227	275	14	436	0	232
O	0	26	0	6	227	275	406	118	6	187	0	99
P	9	41	10	9	1 4	436	6	187	622	297	1	158
T	0	22	0	5	0	232	0	99	1	158	347	84

order to characterize biological processes. For example, clustering analysis of gene expression results in hypotheses of function based on the assumption that groups of genes that are co-expressed are likely to mediate related function [76]. So, what is the general relationship between protein-protein interaction and gene expression?

Grigoriev's study showed that protein pairs encoded by co-expressed genes interact with each other more frequently than random protein pairs based on an analysis of bacteriophage T7 and yeast [77]. In the paper of Ge *et al.* [78], the authors gave a global evidence that genes with similar expression profiles are more likely to encode interacting proteins by using the transcriptome-interactome correlation mapping strategy to compare the interactions between proteins encoded by genes that belongs to the same expression clusters (intra-clusters) with those between proteins encoded by genes that belong to different clusters (inter-clusters). From 1709 protein–protein interactions in yeast, 347 interacting pairs could be assigned to one of the 30 clusters of gene expression data. The average intra-cluster protein interaction density is 5.1 times that of the inter-cluster interaction. Jansen *et al*. investigated the relationship between known protein complexes and mMRA expression level of genes that encode these proteins. They found that subunits of the same protein complex are significantly co-expressed in the absolute mRNA level [79].

Table 5. Yeast Protein Function Categories Retrieved from MIPS

Category	Description
A	Transport and sensing
B	Transcription control
C	Cellular fate/organization
D	Genome maintenance
E	Energy metabolism
F	Protein fate (folding, modification, destination)
G	Amino acid metabolism
M	All the other metabolism categories
O	Cellular transport and transport control
P	Protein synthesis
T	Transcription

It seems that interacting proteins should be simultaneously represented in cell. However, the relationship between protein-protein interaction and gene expression can be complicated. The gene expression level does not necessarily represent its true protein abundance. Furthermore, protein-protein interactions are in a complex and dynamic manner. Moreover, due to the high noisy level of high-throughput data, the information derived from these data might not be exact enough. Nevertheless, the cross-correlation study between gene expression and protein-protein interaction reveals the general trend inside the data. Therefore, it is important and useful to formulate more meaningful hypotheses by integrating the gene expression data and protein-protein interaction data. For example, gene expression profiles were used to verify protein-protein interactions by quantifying the reliability of each interaction.

6. TOPOLOGY OF PROTEIN INTERACTION NETWORK

A protein-protein interaction network (map) can be viewed as a graph, where proteins are nodes and interactions between proteins are edges. Analyses of the global architecture of this large-scale interaction networks can give

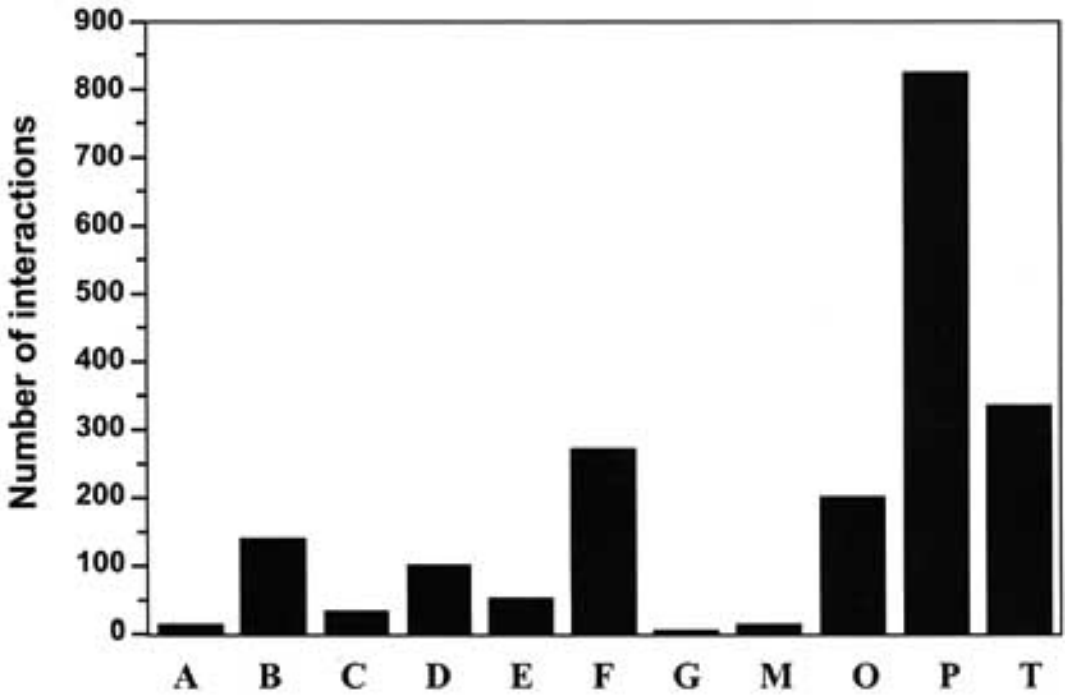

Fig. (4). The distribution of protein-protein interactions vs. function category.

us insights in the evolution of general cellular mechanisms. Jeong *et al.* [80] published such an analysis of the yeast interaction map, which showed that the map forms a highly heterogeneous scale-free network, not an inherently uniform exponential topology [81].

In a scale-free network, the probability for a given protein to interact with *k* partners follows an inverse power law as a function of *k*. In this case, majority of proteins in the network have a small number of interactions while a few proteins interact with many other proteins. Scale-free networks can be generated by randomly adding edges of a node to the existing nodes in a graph with a positive bias for already well-connected nodes in the network. This is consistent with current hypothesis of evolution. Such a type of network architecture is also common to other complex systems such as metabolic network [82], and it is error-tolerant and robust to random mutations. The authors also established a positive correlation between the connectivity and lethality; in particular, highly connected proteins are three times more likely to be essential (i.e., lethal upon deletion). This result is not surprising, since a mutation of a highly connected protein tends to affect more significantly the protein-protein interaction network, and hence, it is more likely to be lethal.

While the positive correlation between the connectivity and lethality makes biological sense, it is worthwhile further checking the argument because of the error associated with high-throughput protein-protein interaction data. The study of Jeong *et al.* was based on a small data set that had 2240 interactions involving 1870 proteins, and the data were obtained mostly from yeast two-hybrid assay, which is known to have high false negatives and false positives. The shape of the actual interaction network might be quite different due to false negatives and false positives. For example, proteins that exhibit few interactions in this network could actually represent highly connected nodes due to false negatives. Conversely, false positives of two-hybrid system may generate many artificially interactions for a particular protein.

To further explore the relationship between connectivity and lethality of a protein, we studied the difference between lethal proteins and viable proteins in their connectivity distributions by using correspondence analysis [83]. We used a core dataset from DIP, which are mostly obtained from small-scale experiments with high confidence. The dataset contains 3003 interactions involving 1020 proteins. We also used the annotation about viability in MIPS, where 878 genes were assigned essential roles for viability of cells and 3158 genes were found to be non-essential based on the literature reports (http://mips.gsf.de/proj/yeast/CYGD/db/index.html). Table **6** shows the distribution of essential ORFs and non-essential ORFs having different numbers of protein interaction partners in yeast. Fig. **5** plots the deviation of the observed connectivity frequency from the expected frequency. The plot shows that essential ORFs

Table 6. The Distribution of Essential ORFs and Non-essential ORFs in Yeast According to the Number of Interactions in the DIP Core Protein-protein Interaction Dataset

ORF classification	Number of interactions				
	>20	10-20	5-10	1-5	0
Essential ORFs	11	31	95	182	558
Non-essential ORFs	8	43	129	412	2564

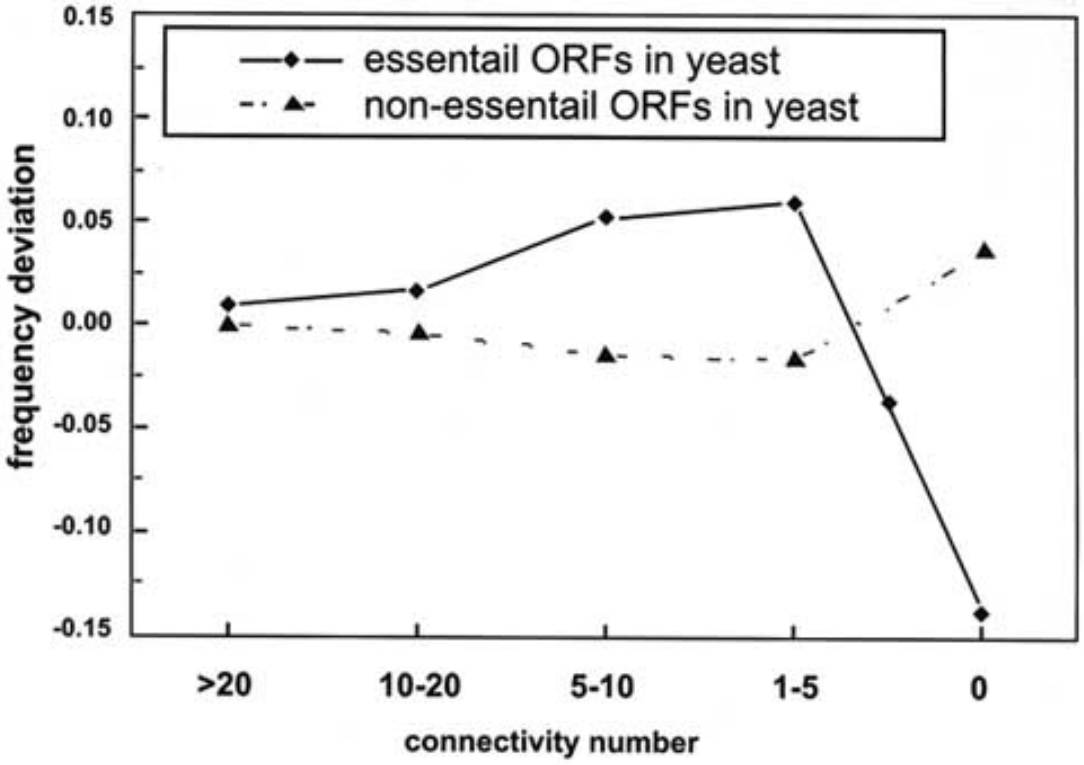

Fig. (5). Yeast essential ORFs and non-essential ORFs connectivity distribution. The Y-axis represents the frequency of profile deviated from the average of row frequency. It showed that yeast essential ORFs connectivity distribution profile deviates greatly from the expected value. Correspondence analysis was used to study the preference of essential ORFs and non-essential ORFs as a function of connectivity number. For a m×n contingency table with the cell frequencies N_{ij}, $i=1, 2, \ldots, m, j=1, 2, \ldots, n$, the i-th row profile is determined by $\frac{N_{ii}}{\sum_{i=1}^{n} N_{ii}}$.

have more interactions in protein-protein interaction network than expected, and *vice versa* for non-essential ORFs. Our result further supports the early studies.

The large-scale protein-protein interaction data allows us to study the general relationship between protein-protein interaction and evolution, especially the effect of interaction network topology on protein evolution. Fraser *et al.* [84] studied the correlation between the number of interactions of a protein and its evolutionary rate in yeast. The authors compiled a list of 3541 interactions between 2445 different proteins. The well-conserved orthologs between *Saccharomyces cerevisiae* and *Caenorhaditis elegans* were selected. Among 164 yeast proteins having well-conserved orthologs in the nematode, there is a negative correlation between the number of interaction of a protein and its evolutionary rate. This correlation does not depend on the evolutionary fitness of the protein itself. The authors suggested that a protein with more interactors tends to evolve slowly because a greater proportion of the protein is directly involved in its function given that different interactions to the same protein may depend on different sites of the protein. Using high-throughput yeast two-hybrid data, Wanger's study showed that protein-protein interaction network resembles a random graph, where it consists of many small subsets and one large connected subset [85]. The relationship between gene duplication rate and interaction showed that after gene duplication, the likelihood of loosing an interaction exceeds 2.2×10^{-3} /Myr, i.e., for every 300 million years, as many as half of all interactions may be replaced by new interactions.

Recently, to address the feature of molecular networks operating in living cell, Maslov and Sneppen [86] analyzed

the topological properties of protein-protein interaction and gene regulatory networks in yeast *Saccharomyces cerevisiae*. Correlations between these two networks in their connectivities of interacting nodes were calculated and compared with a null mode of a network. For both protein interaction network and regulatory network, connections between highly connected proteins are systematically suppressed, suggesting that the propagation of deleterious perturbations over the network is repressed. This indicates that the organized interaction pattern of molecular networks is robust and specific.

7. *IN SILICO* PREDICTION OF PROTEIN-PROTEIN INTERACTIONS

In addition to systematic analyses of protein interactions by high-throughput experiments, a number of computational methods have been developed for the prediction of protein-protein interactions based on protein/DNA sequence information. The predicted interactions can be found in databases like Predictome (http://predictome.bu.edu/) [87], which stores predicted interactions between the proteins in 44 genomes based on three computational methods, i.e., chromosomal proximity, phylogenetic profiling and domain fusion [88,89]. The predicted protein-protein interactions are less reliable than those generated from high-throughput data. However, they expand the score of experimental data and are useful to assess the protein-protein interactions generated from high-throughput experiments. Notably, the combination of experimental approaches and computational analyses has advantages in validating protein-protein interaction network, which is particularly effective in reducing noise. For example, Tong *et al.* [90] developed a strategy that combines large-scale yeast two-hybrid data with the computational prediction of protein-protein interactions from preferred ligands consensus sequences generated by phage display.

In this section, we will describe five major computational techniques for prediction of protein-protein interactions, i.e., gene fusion, conserved genetic neighborhood, co-occurrence of genes in genomes, predictions from domain interactions, and predictions based on structural information.

7.1. Gene Fusion

The Gene fusion or "Rosetta stone" method [91] for predicting protein-protein interaction is based on the observation that some pairs of interacting proteins whose homologs are fused into a single protein chain in another organism. For example, two separate proteins A and B in organism X are expressed as a fusion protein in an another organism Y. When expressed as a fused protein in Y organism, A and B as protein domains generally interact with each other physically, and this implies that A and B as separate proteins in X organism probably interact too. Thus, a successful search through genome sequences for the corresponding fused protein is powerful evidence that A and B physically interact and are functionally linked. This method, although limited by the relative infrequency of fusion events, is highly sensitive with low false positive rate. Searching sequences from many genomes revealed 6809 such putative protein-protein interactions in *Escherichia coli* and 45,502 interactions in yeast.

7.2. Conservation of Gene Neighborhood

Proteins with conserved genetic neighborhood in bacteria, i.e., a group of genes arranged in tandem in one genome and also appeared in a similar fashion in its related genomes, tend to interact with each other to form complexes [92-94]. Operons represent one such conserved gene context. Identification of operons or "conserved gene contexts" can provide clues about which set of proteins may form a complex. This can be done through discovering the sequential arrangement of genes in a microbial genomic sequence and conserved gene context across multiple related microbial genomes. Typically the intergenic distance between neighboring genes in an operon is short (less than 100 bases). Using such information, there are a number of existing algorithms for identification of operons [95]. One of the main limitations of this method is that it is only directly applicable to bacteria.

7.3. Co-occurrence of Genes in Genome

The phylogenetic profiling approach [96,97] is based on the assumption that proteins functioning together in a complex are likely to evolve in a correlated fashion. During evolution, all such functionally linked proteins tend to be either all preserved or completely eliminated in the new species. Such information can be represented by a phylogenetic profile that records the presence or absence of a protein in every known genome in a phylogenetic tree. It is shown that proteins having similar profiles tend to have physical interactions and to be functionally linked, for example, insulin and its receptors [98] and dockerins and cohexins [99]. The method of phylogenetic profiling can be used to establish the probability of two proteins interacting with each other. In a more quantitative approach, Pazos and Valencia [100] proposed to calculate the phylogenetic profile based on the evolutionary distances between the sequences in the associated protein family for a protein. To demonstrate the capacity of the method for large-scale predictions of protein-protein interactions, the authors applied it to a collection of more than 67 000 pairs of *E. coli* proteins, and they predicted 2742 pairs belonging to interacting proteins.

7.4. Prediction from Domain Interactions

Protein domain, as a unit of structure, function, and evolution, is also a unit for protein-protein interactions [101,102]. Many physical interactions between domains are preserved regardless of which proteins contain these domains [103]. Therefore, understanding the protein-protein interaction at the domain level can give a global view of the protein-protein interaction network and can be used to expand the knowledge of protein-protein interactions. One can derive the rules underlying protein recognition mediated by a small number of protein modules based on protein-protein interactions and homologous domain repertories [104]. These rules in turn can be used to predict protein-

protein interactions from the domain-domain interactions. Sprinzak and Margalit [105] analyzed the distribution of well-characterized sequence domain of interacting protein pairs. This information was further used to search for putative new interacting pairs that contain an interacting domain pair. Wojcik *et al.* [106] developed an approach named "Interacting Domain Profile Pairs" based on a combination of homologous domain searching and clustering method to infer a protein-protein interaction map of *Escherichia coli* from a *Helicobacter pylori* reference interaction map. Deng *et al.* [107] formulated this problem in a more systematic way by taking into account the errors of false negatives and false positives. They used a Maximum Likelihood approach to infer domain-domain interaction using 5719 protein-protein interactions in yeast. Their results from the inferred domain-domain interactions performed well on an independent test set of known protein-protein interactions and they also predicted novel protein-protein interactions.

7.5. Prediction from Structure Information

Like protein folding, where the folded structure is solely dependent on the protein sequence [108,109], a protein-protein interaction is also just dependent on the sequences of the two interacting proteins. This suggests a possibility of protein-protein interaction prediction directly from the involved proteins' sequences and their characteristics. Some attempts have been made, although more studies are needed to be done to evaluate the applicability of such an approach given the weak detectable signal in protein sequences for interaction. Given a database of known protein-protein interaction pairs, Support Vector Machine (SVM), a machine learning system, was trained to recognize and predict interaction based solely on protein sequences and their associated physicochemical properties [110] through recognition of correlated pattern between protein sequences and their interactions. Another direction for predicting interaction based on protein sequences is to explore the information on the evolution of these sequences. Protein-protein interaction sites are evolutionarily conserved and they can be detected from the sequence traces, especially the correlated pairs between monomers tend to occur at the contact interface [11,112]. The use of the correlation information may detect interacting protein pairs and their contact regions. Using such an approach, Pazos *et al.* [113] proposed a "*in silico* two hybrid system" to predict protein interactions and the most likely sequence regions involved in the interactions. They applied this system to predict protein-protein interactions in *E. coli*.

8. BIOLOGICAL INFERENCE THROUGH PROTEIN-INTERACTION DATA

Protein-protein interaction network contains the information of individual proteins, including their partners, functions and interactive complexes, as well as the information on biological pathways, which are often the results of several directly physical protein-protein interactions. Thus, protein-protein interaction data are useful to assign function to the uncharacterized gene product, and protein-protein interactions are a new and rich source to construct biological pathways, in particular the signal transduction pathways. However, biological inference from protein-protein interaction data is not trivial, given the complexity and the errors of the protein interaction map. When the protein-protein interaction information is insufficient, it may be important to use other valuable sources of data, including genomic sequence and gene expression data to refine biological hypotheses generated from protein-protein interactions. The integration analysis of protein-protein interaction, gene expression and protein/DNA sequence can be a powerful means to infer cellular functions and pathways, and it represents a grant challenge for bioinformatics.

8.1. Protein Function Prediction

A protein often performs its function through interacting with other proteins of the same cellular function. This is reflected in the statistical study that two interacting proteins often share the same function category, as shown in Section 5.2. Hence, one can use the protein-protein interaction information to assign putative function for a hypothetical protein [114] based on "guilt by association"[115]. For example, if protein X (uncharacterized) is found to interact with protein Y and protein Z, and both Y and Z are components of the DNA transcription processing machinery, then it is likely that protein X is also involved in this process, perhaps as part of the complex containing Y and Z. Based on such an approach, high-throughput protein-protein interaction data provide a good coverage for many novel proteins whose functions cannot be assigned by sequence comparison. Schwikowski *et al.* collected 2709 published protein-protein interactions in yeast and clustered them based on cellular role and subcellular localization annotated in the Yeast Proteome Database (YPD at http://www.proteome.com/YPDhome.html). They compiled a list of about 370 proteins with unknown function that interact with at least one protein with known function. Among them, 29 proteins have two or more interacting partners with the common function. To assign protein function by using protein-protein interaction data in a more systematic and rigorous way, Deng *et al.* [116] developed a mathematical model based on the theory of Markov random fields. Instead of searching for the simple consensus among the functions of the interacting partners, the method assigns a probability (with a confidence level) for a hypothetical protein to have the annotated function using Bayesian approaches.

Annotating proteins using their interaction partners' information is a promising technique and such an approach will become more and more useful as the protein-protein interaction data accumulate and their quality improves. This approach can be used in conjunction with other methods. For example, one can also use computational strategies to assign functions based on the co-evolution of proteins [117], etc. as described in Section 6. It is also possible to integrate gene expression data for this purpose.

8.2. Biological Pathway Construction

Protein-protein interaction networks not only allow the assignment of cellular functions to novel proteins but also

provide the resource for constructing the biological pathways. The study of biological pathways is a challenging research topic. The pathway network is a complex and synergistic system. There are rich interaction networks among the constituents, including non-protein elements such as ligands. These interactions are numerous and have nonlinear characteristics. For instance, even a small change in the expression of some components might cause the system to respond in an entirely new fashion. Moreover, there are cross-talks among multiple pathways. Despite the complexity, physical protein-protein interactions generally provide the backbones for biological pathways. For example, in the signal transduction pathway, protein-protein interactions often play the main role in the signal transduction cascade. Therefore, high-throughput protein-protein interaction data can help construct biological pathways *in silico*, at least partially.

Fig. **6** illustrated a validation of this approach. As a reference, we chose the well-known MAPK signal transduction pathway of filamentation from KEGG [118,119] (see Fig. **6A**). This pathway can be rebuilt based solely on the protein-protein interaction mapping as shown in (Fig. **6B**). To build the pathway, it is important to integrate the protein interaction data from different sources such as yeast two hybrid, protein complexes etc. This also indicates the complementarily of yeast two-hybrid method and protein-complex identification method in the case of their coverage as we discussed in section 4.2.

Protein interaction maps may also be used to construct new biological pathways, especially to constitute the part of the pathway that involves protein-protein interaction. For example, in a signal transduction cascade involving protein-protein interactions, if we know the sensor protein of the signal and the terminal transcription factor that the signal affects, we can construct a protein interaction pathway between the two ends. However, protein-protein interaction network is very complex. As we see in Fig. **6C**, one protein has many potential interaction partners and typically multiple paths exist between two proteins. In addition, the protein-protein interaction map contains false paths because of the error in the data. To assess a potential pathway, some priori knowledge or other type of data may be needed to perform a comprehensive evaluation for the pathway based on protein function (cellular role), gene expression pattern, and subcellular localization. Two proteins with true physical interaction generally have the similar cellular role, correlated gene expression pattern, and the same subcellular localization.

We have used protein-protein interaction data to study the signal transduction pathway in yeast amino acid transport (Chen *et al*., paper in preparation). The ability of yeast cells to rapidly respond and adapt to changing environmental conditions is essential for viability. A prerequisite for the generation of a proper physiological response is the ability to sense the extracellular nutrient environment and to regulate gene expression through signal transduction pathways. Previous genetic and biochemical experimental studies have shown that the SPS (Ssy1p-Ptr3p-Ssy5p) amino acid sensor system is required for amino acid-induced transcription of amino acid transporter genes (*e.g.*, *AGP1*, *BAP2*, *TAT1*,

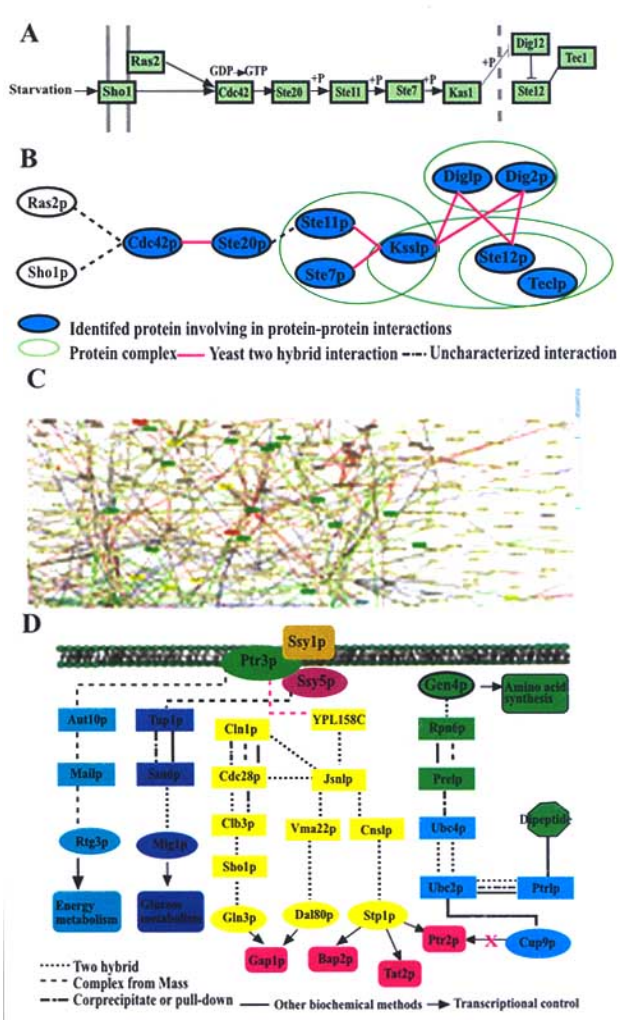

Fig. (6). Relationship between biological pathway and protein-protein interaction data. A: The MAPK signaling pathway for filamentation taken from KEGG. B: A MAPK signal transduction pathway constructed from protein-protein interaction data. C: A snapshot of protein-protein interaction map of yeast proteome taken from the Web site at http://depts.washington.edu/sfields. For color-coded lines, red means that the cellular roles and subcellular localizations of interacting proteins are identical; blue indicates only the subcellular localizations are identical; green indicates that only the cellular roles are identical; black means both the cellular roles and localizations are different or the information is unknown for at least one protein involved in the interaction. D. The signal transduction pathway for amino acid transport in yeast constructed from the high-throughput protein-protein interaction data, where the different colors indicate different pathways. The different shapes of lines represent protein-protein interactions detected by different experimental techniques. The ovals represent the transcription factors and the boxes represent the intermediate proteins between the SPS sensor and transcription factors. Among those pathways, the pathways of Ssy5p-Tup1p-Ssn6p-Mig1p and Dipeptide-Ptr1p-Ubc2p-Cup9p-Ptr2p are already known in literature, while the others are not characterized from experiments yet.

TAT2)[120] and the di-/tripeptide transporter, Ptr2p[121]. A functional SPS sensor is also required for the regulation of *GAP1*, a nitrogen-regulated amino acid transporter gene[122]. Experimental evidence has also suggested that *PTR2* is suppressed by the binding of Cup9p (a transcriptional repressor). This regulation is mediated by the ubiquitination of Cup9p by the Ptr1p-Ubc2p complex[123]. However, it is not clear at all what the other proteins are involved in this signal transduction pathway and how they form the signal transduction cascade, and how the cross-talks between related pathways take place.

We constructed a model of protein-protein interaction pathways from amino acid sensors to energy metabolism, glucose pathway, and transporter regulators (see Fig. **6D**). Based mainly on protein-protein interaction maps, we constructed the following signal transduction pathways between the amino acid sensor SPS and the related transcription factors for amino acid transport:

(1) Amino acid transporter gene expression regulation pathways:

Ptr3p-YPL158C-Jsn1p-Csn1p-Stp1p;

Ptr3p-YPL158C-Jsn1p-Cln1-Cdc28p-Clb3-Sho1p-Gln3p;

Ptr3p-YPL158C-Jsn1p-Vma22p-Dal80p.

(2) A feedback pathway:

Dipeptide-Ptr1p-Ubc2p-Cup9p-Ptr2p

(3) Other related pathways:

Ptr3p-Aut10p-Mai1p-Rtg3p;

Ssy5p-Tup1p-Ssn6p-Mig1p;

Dipeptide-Ptr1p-Ubc2p-Ubc4p-Pre1p-Rpn6p-Gcn4p.

The protein function and subcellular localization information was used to select the most probable pathways, i.e., all the constituents of a path with more reasonably related functions and subcellular localizations have better chance to be in the correct pathway. Regulatory region analysis and gene expression analysis validate the pathway model by showing how the selected transcription factors control the amino acid transporters and how the cross-talks between the amino acid transport pathway and the other related pathways are achieved. Although the pathway model contains some local information known before, it is the first global pathway hypothesis for the amino acid transport regulation process.

Our study shows that to better use protein-protein interaction for pathway studies, it is crucial to incorporate other bioinformatics techniques, including protein structure prediction, gene expression analysis, and regulatory region analysis. This type of integrative study will probably be more and more common in the future. A study by Ideker *et al.*[124] provided an another example. The authors used DNA microaray, mass spectrometry and protein-DNA interactions to analyze the galactose metabolism pathway in yeast. The combination of bioinformatics sources will help transform the protein interaction network from the local property description to the general understanding of cellular pathways.

9. DISCUSSION

As the highlight in biology research is moving from genome to proteome, it is expected that in the near future the function-based cell map will become a new hot spot. Analysis of protein-protein interactions will play a more and more important role, given the fact that protein-protein interactions occur in nearly all events that take place in a cell, and most cellular processes are regulated by multi-protein complexes. Identification of protein interaction network is also of great interest for drug screening and design. High-throughput protein-protein interaction data are very useful for providing the valuable global information compared with the traditional biology. The major efforts of traditional biology focus on identification of genes and proteins responsible for specific phenomena and on investigation of how particular genes and proteins function. These studies have been conducted typically in some *ad hoc* manners and on a small scale. The results, although accurate, are often only pieces of local information. The high-throughput data, given their genome-wide coverage, provide an opportunity to merge pieces of local information together for a global view about how a cell works at the molecular level.

With the growth of experimental data by an unprecedented amount, computational analysis plays a vital role in managing and deciphering protein-protein interaction data. Many useful researches and developments have been done along the line, as reviewed in the paper. On the other hand there are many more challenges. While several well-structured databases and visual tools have been implemented, a challenge is how to systematically integrate protein-protein interaction data with information from other various sources such as literature information (based on traditional biochemical/genetic approaches) and other types of high-throughput experimental data (genomic sequence, gene expression profile, etc.). The integration requires a uniform database to store and query heterogeneous information as well as a data-mining framework to automatically construct biological hypothesis. Furthermore, a rigorous statistical model is needed to be developed for data quality control before computational analyses of protein-protein interaction network. The model should allow automatic assessment and validation so as to combine protein-protein interaction data sets from different sources. An open question in this aspect is how to use protein complex information for assessing binary interaction effectively. This will also help function prediction and pathway construction as described in section 8, since currently only the binary information is used for this purpose. Another very difficult aspect for analyzing the protein-protein interaction map is its dynamic nature of the map, as well as the roles of non-proteins (e.g., ligands) in protein-protein interaction. Such a difficult has not been solved so far at all, and some new paradigms may be needed. Finally, the major challenge is to infer new biological

discoveries and validate hypotheses from protein-protein interaction data. Integrated analyses that combine information from various sources as we discussed above provide a promising trend. How to make the integration automated to fully utilize the underlying information in the data is far from being solved. Although the information contained in high-throughput data is rich, deriving biological knowledge/hypothesis from the information is very challenging, given the complexity of biological pathways and the signal-to-noise ratio in high-throughput protein-protein interaction data (i.e., the false positives and false negatives). Clearly, it is valuable to integrate the traditional approaches with the high-throughput protein-protein interaction data. Other types of high-throughput data such as genome sequence and gene expression data are also valuable resources for generating hypotheses about genes involved. For example, clustering analysis of gene expression data can be used to predict functions of unknown proteins and identify the proteins that may be involved in a pathway[125,126]. The sequence comparison as well as structural information can be modeled to reveal protein functions[127,128]. A consensus approach, or better yet, an expert system using a variety of methods to analyze and integrate high-throughput data will give us a global and comprehensive understanding of the related biological processes. An attempt has been made along the line to verify protein-protein interaction data with microarray data and then to further annotate function by the consensus analysis of the genome-wide high-throughput data[129]. Nevertheless, much more work is needed to be done.

With improved high-throughput experimental techniques, more and more high quality protein-protein interaction data will be generated. Moreover, we believe that well developed databases and integrated computational tools will enable biologists to easily navigate protein-protein interaction network and explore new biological discoveries. It may not be far before computational analysis of protein-protein interaction data becomes a necessary protocol in many biological laboratories, just like sequence homology search being carried out today.

ACKNOWLEDGEMENTS

We would like to thank Drs. Ying Xu, Victor Olman, and Jeffrey M. Becker for helpful discussions. We also thank Trupti Joshi for a critical reading of this manuscript. The work was supported by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory and by the Office of Biological and Environmental Research, U.S. Department of Energy, under Contract DE-AC05-00OR22725, managed by UT-Battelle, LLC.

REFERENCES

[1] Rudert, F., Ge, L. and Hag, L. (**2000**) *Biotech. Ann. Rev., 5*, 45-86.

[2] Eisenberg, D., Marcotte, E. M., Xenarios, I. and Yeates, T. O. (**2000**) *Nature, 405*, 823-826.

[3] Auerbach, D., Thaminy, S., Hottiger, M. and Stagljar, I. (**2002**) *Proteomics, 2*, 611-623.

[4] Kone, B. C. (**2000**) *Acta Physiol. Scand., 168*, 27-31.

[5] Wang, J. **(2002)** *Trends Biochem. Sci., 27*, 122-126.

[6] Eisen, J.A. (**2002**) *Nature, 415*, 845-848.

[7] Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C. A., Gocayne J. D., Amanatides, P. G. and Scherer S.E. **(2000)** *Science, 287*, 2185-2195.

[8] The *C. elegans* Sequencing Consortium, (**1998**) *Science, 282*, 2012-2018.

[9] Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S.G. (**1996**) *Science, 274*, 563-567

[10] The Arabidopsis Genome Initiative, (**2000**) *Nature, 408*, 796-815

[11] Birnery, E., Bateman, A., Clamp, M. and Hubbard, T. (**2001**) *Nature, 409*, 827-828.

[12] Battey, J., Jordan, E., Cox, D. and Dove, W. (**1999**) *Nature Genetics, 21*, 73-75.

[13] Legrain, P., Wojcik, J. and Gauthier, J. M. (**2001**) *Trends Gene, 7*, 346-352.

[14] Gerstein, M., Lan, N. and Jansen, R. (**2002**) *Science, 11*, 284-285.

[15] Goffeau, A., Barrell, B.G., Bussey, H. and Davis R.W. (**1996**) *Science, 546*, 563-567.

[16] Kumar, A. and Snyder, M. (**2001**) *Nat. Rev. Genet., 2*, 302-312.

[17] Garrels, J. I. (**2002**) *Funct. Integr. Genomics, 2*, 212-237.

[18] Simons, A. H., Dafni, N., Dotan, I., Oron, Y. and Canaani, D. (**2001**) *Nucleic Acids Res., 29*, 100-106.

[19] Tong, A. H. Y., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C. W., Bussey, H., Andrews, B., Tyers, M. and Boone, C. (**2001**) *Science, 294*, 2364-2368.

[20] Harman, J. G. (**2001**) *Biochim. Biophys. Acta, 1547*, 1-17.

[21] Lim, W. A. **(2002)** *Curr. Opin. Struct. Biol., 12*, 61-68.

[22] Wagschal, K., Tripet, B., Lavigne, P., Mant, C. and Hodges, R. S. (**1999**) *Protein Sci., 8*, 2312-2329.

[23] Jackson, R. M., Gabb, H. A. and Sternberg, M. J. (**1998**) *J. Mol. Biol., 276*, 265-285.

[24] Betts, M. J. and Sternberg, M. J. (**1999**) *Protein Eng., 12*, 271-283.

[25] Sali, A. and Blundell, T. L. (**1993**) *J. Mol. Biol., 234*, 779-815.

[26] Tucker, C. L., Gera, J. F. and Uetz, P. (**2001**) *Trends Cell Biol., 11*, 102-6.

[27] Valencia, A. and Pazos, F. (**2002**) *Curr. Opion. in Struct. Biol., 12*, 368-373.

[28] Schachter, V. (2002) (**2002**) *Biotechniques, Suppl., 16-8, 20*-4, 26-7.

[29] Yarmush, M. L. and Jayaraman, A. (**2002**) *Annu. Rev. Biomed. Eng., 4,* 349-373.

[30] Phizicky, E. M. and Field, S. (**1995**) *Microbiol. Rev., 59,* 94-123.

[31] Martzen, M. R., McCraith, S. M., Spinelli, S.L., Torres, F. M., Fields, S., Grayhack, E. J. and Phizicky, E. M. (**1999**) *Science, 286,* 1153-1155.

[32] Pelletier, J. and Sidhu, S. **(2001)** *Curr. Opin. Biotechnol., 12,* 340-347.

[33] Chen, Z. and Han, M. (**2000**) *Bioessays, 22,* 503-506.

[34] Broder, Y. C., Katz, S., Aronheim, A. (**1998**) *Curr. Biol., 8,* 1121-1124.

[35] Ehrhard, K. N., Jacoby, J J., Fu, X.Y., Jahn, R. and Dohlman, H. G. (**2000**) *Nat. Biotechnol., 18,* 1075-1079.

[36] Johnsson, N., Varshavsky, A. (**1994**) *Proc. Natl. Acad. Sci. USA, 91,* 10340-10344.

[37] Stagljar, I., Korostensky, C., Johnsson, N. and te Heesen, S. (**1998**) *Proc. Natl. Acad. Sci. USA, 95,* 5187-5192.

[38] Bartel, P. L., Roecklein, J. A., SenGupta, D. and Fields, S. (**1996**) *Nat Genet., 12,*72-77.

[39] Flajolet, M., Rotondo, G., Daviet, L., Bergametti, F., Inchauspe, G., Tiollais, P., Transy, C. and Legrain, P. (**2000**) *Gene, 241,* 369-379.

[40] Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A. and Legrain, P. (**2001**) *Nature, 409,* 211-215.

[41] Walhout, A. J., Soredella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M. A., Thierry-Mieg, N. and Vidal, M. (**2000**) *Science, 287,* 116-122.

[42] Davy, A., Bello, P., Thierry-Mieg, N., Vaglio, P., Hitti, J., Doucette-Stamm, L., Thierry-Mieg, D., Reboul, J., Boulton, S., Walhout, A. J., Coux, O. and Vidal, M. (**2001**) *EMBO Rep., 2,* 821-828.

[43] Uetz, P., Giot, l., Cagney, G., Mansfield T. A., Judson R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J. M. (**2000**) *Nature, 403,* 623-627.

[44] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (**2001**) *Proc. Natl. Acad. Sci. USA, 98,* 4569-4574.

[45] Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. and Superti-Furga, G. (**2002**) *Nature, 415,* 141-147.

[46] Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J. R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D. and Tyers, M. (**2002**) *Nature, 415,* 180-183.

[47] Suzuki, H., Fukunishi, Y., Kagawa, I., Saito, R., Oda, H., Endo, T., Kondo, S., Bono, H., Okazaki, Y. and Hayashizaki, Y. (**2001**) *Genome Res., 11,* 1758-1765.

[48] Field, S. and Song, O. (**1989**) *Nature, 340,* 245-246.

[49] Chien, C. T. Bartel, P. L., Sternglanz, R. and Fields, S. (**1991**) *Proc. Natl. Acad. Sci. USA, 88,* 9578-9582.

[50] Drees, B. L. (**1999**) *Curr . Opin. Chem. Biol., 3,* 64-70

[51] Uetz, P. and Hughes, R. E. (**2000**) *Curr. Opin. Microbiol., 3,* 303-308.

[52] Zhu, H., Klemic, J. F., Chang, S., Bertone, P., Casamayor, A., Klemic, K. G., Smith, D., Gerstein, M., Reed, M.A. and Snyder, M. (**2000**) *Nat. Gene., 26,* 283-289.

[53] Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R. A., Gerstein, M. and Snyder, M. (**2001**) *Science, 293,* 2101-2105

[54] Mullaney, B. P. and Pallavicini, M. G. **(2001)** *Exp. Hematol., 29,* 1136-1146.

[55] Ono, T., Hishigaki, H., Tanigami, A. and Takagi, T. (**2001**) *Bioinformatics., 17,* 155-161.

[56] Thomas, J., Milward, D., Ouzounis, C., Paulman, S. and Carroll, M. (**2000**) *Pac. Symp. Biocomput., 541*-542.

[57] Schachter, V. (**2002**) *Drug Discov. Today, 7,* S48-S54.

[58] Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S., Eisenberg, D. (**2002**) *Nucleic Acids Res., 30,* 303-305.

[59] Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T., Hogue, C. W. (**2001**) *Nucleic Acids Res., 29,* 242-245.

[60] Mewes, H. W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schuller, C., Stocker, S. and Weil, B. (**2000**) *Nucleic Acids Research, 28,* 37-40.

[61] Zanzoni, A., Montechi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M. and Cesareni, G. (**2002**) *FEBS Letters, 513,* 135-140.

[62] Xenarios, I., Rice, D.W., Salwinski, L., Baron, M. K., Marcotte, E.M. and Eisenberg, D. (**2000**) *Nucleic Acids Research, 28,* 289-291.

[63] Marcotte, E. M., Xenarios, I. and Eisenberg, D. (**2001**) *Bioinformatics., 17,* 359-63.

[64] Xenarios, I. and Eisenberg, D. (**2001**) *Curr. Opin. Biotechnol, 12,* 334-339.

[65] Duan, X. J., Xenarios, I. and Eisenberg, D. (**2002**) *Mol. Cell Proteomics, 1,* 104-116.

[66] Mewes, H. W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. and Weil, B. (**2002**) *Nucleic Acids Research, 30,* 31-34.

[67] Kanehisa, M., Goto, S., Kawashima, S., Nakaya, A. (**2002**) *Nucleic Acids Res., 30,* 42-46.

[68] Rain, J. C., Selig, L., Reuse, H. D., Battaglia, V., Reverdy, C., Simon S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A. and Legrain, P. (**2001**) *Nature, 409,* 211-216.

[69] Hazbun, T. R. and Field, S. (**2001**) *Proc. Natl. Acad. Sci. USA, 98,* 4277-4278.

[70] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. (**2002**) *Nature, 417,* 399-403

[71] Holstege, F. C., Jennings, E.G., Wyrick. J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S., Young, R. A. (**1999**) *Cell, 95,* 717-728.

[72] Cyert, M. S. (**2001**) *Annu. Rev. Genet, 35,* 647-672.

[73] Jenssen, T. K., Laegreid, A., Komorowski, J. and Hovig, E. (**2001**) *Nat. Genet., 28,* 21-28.

[74] Deane, C.M., Salwinski, L., Xenarios, I. and Eisenberg, D. (**2002**) *Mol. Cell Prot., 1,* 349-356.

[75] Shimohashi, N., Nakamuta, M., Uchimura, K., Sugimoto, R., Iwamoto, H., Enjoji, M. and Nawata, H. *(***2000**) *J. Cell Biochem., 78,* 595-606.

[76] Eisen, M. B., Spellman, P. L., Brown, P. O. and Botstein, D. (**1998**) *Proc. Natl. Acad. Sci. USA, 95,* 14863-14868.

[77] Grigoriev, A. (**2001**) *Nucleic Acids Res., 29,* 3513-3519.

[78] Ge, H., Liu, Z., Church, G. and Vidal, M. (**2001**) *Nat. Genet., 29,* 482-486.

[79] Jansen, R., Greenbaum, D. and Gerstein, M. (**2002**) *Genome Res., 12,* 37-46.

[80] Jeong, H., Mason, S. P., Barabasi, A. L. and Oltvai, Z. N. (**2001**) *Nature, 411,* 41-42.

[81] Albert, R., Jeong, H. and Barabasi, A. L. (**2000**) *Nature, 406,* 378-381.

[82] Jeong, H., Tombor, B., Albert, R., Oltvai, Z.H. and Barabasi, A.L. (**2000**) *Nature, 407,* 651-654.

[83] Jobson, J. D. (**1992**) Applied Multivariate Data Analysis, pp436-444, Springer-Verlag, New York.

[84] Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. and Feldman, M. W. (**2002**) *Science, 296,* 750-752.

[85] Wagner, A. (**2001**) *Mol. Boil. Evol., 18,* 1283-1292.

[86] Maslov, S. and Sneppen, K. (**2002**) *Science, 296,* 910-913.

[87] Mellor, J. C., Yanai, I., Clodfeter, K. H., Mintseris, J. and DeLisi, C. (**2002**) *Nucleic Acids Res., 30,* 306-309.

[88] Galperin, M. Y. and Koonin, E. V. (**2000**) *Nat. Biotechol., 18,* 609-613.

[89] Huynen, M., Snel, B., Lathe, W. and Brol, P. (**2000**) *Genome Res., 10,*1204-1210.

[90] Tong, A. H., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C.W., Fields, S., Boone, C. and Cesareni (**2002**) *Science, 295,* 321-324.

[91] Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. and Eisenberg, D. (**1999**) *Science, 285,* 751-753.

[92] Tamames, J., Casari, G., Ouzounis, C. and Valencia, A. (**1997**) *J. Mol. Evol., 44,* 66-73.

[93] Dandekar, T., Snel, B., Huuynen, M. and Bork, P. (**1998**) *Trends biochem. Sci., 23,* 324-328.

[94] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. and Maltsev, N. *(***1999**) *Proc. Natl. Acad. Sci. USA, 96,* 2896-2901.

[95] Lathe, W. C., 3rd, Snel, B. and Bork, P. (**2000**) *Trends Biochem. Sci., 25,* 474-479.

[96] Gaasterland, T. and Ragan, M. (**1998**) *Journal of Microbial and Comparative Genomics, 3,* 199-217.

[97] Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. and Yeates, T. O. (**1999**) *Proc. Natl. Acad. Sci. USA, 96,* 4285-4288.

[98] Fryxell, K. J. (**1996**) *Trends Genet., 23,* 364-369.

[99] Pages, S., Belaich, A., Belaich, J. P., Morag, E., Lamed, R., Shoham, Y. and Bayer, E. A. (**1997**) *Protiens, 29,* 517-527.

[100] Pazos, F. and Valencia, A. (**2001**) *Protein Eng., 14,* 609-614.

[101] Holm, L. and Sander, C. **(1994)** *Proteins, 19,* 256-268.

[102] Ponting, C. P. and Russell, R. R. (**2002)** *Annu. Rev. Biophys. Biomol. Struct., 31,* 45-71.

[103] Fanning, A. S. and Anderson, J. M. (**1996**) *Curr. Biol., 6,* 1385-1388.

[104] Zucconi, A., Panni, S., Paoluzi, S. (**2000**) *FEBS letters, 480,* 49-54.

[105] Sprinzak, E. and Margalit, H. (**2001**) *J. Mol. Biol, 311,* 681-692

[106] Wojcik, J. and Schachter, V. (**2001**) *Bioinformatics, 17,* Suppl. 1, 296-305.

[107] Deng, M., Mehta, S., Sun, F. and Chen, T. (**2002**) *RECOMB'02,* 117-126.

[108] Walls, P. H. and Sternberg, M. J. (**1992**) *J. Mol. Biol., 228,* 277-297.

[109] Young, L., Jernigan, R. L. and Covell, D. G. (**1994**) *Protein Sci., 3,* 717-729.

[110] Bock, J. R. and Gough, D. A. (**2001**) *Bioinformatics, 17,* 455-460.

[111] Pazos, F., Helmer-Citterich, M., Ausiello, G. and Valencia, A. (**1997**) *271,* 511-523.

[112] Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D. and Cohen, F. E. (**2000**) *J. Mol. Biol, 299*, 283-293.

[113] Pazos, F. and Valencia, A. (**2002**) *Proteins, 47*, 219-227.

[114] Schwikowski, B., Uetz, P. and Fields, S. (**2000**) *Nat. Biotechnol., 18*, 1257-1261.

[115] Olover, S. (**2000**) *Nature, 403*, 601-603

[116] Deng, M., Zhang, K., Mehta, S., Chen, T. and Sun, F. (**2002**) In Proceedings of the IEEE Computer Society Bioinformatics Conference (CSB2002). IEEE Computer Society, LosAlamitos, California. Pages 197-206.

[117] Pellegrini, M., Marcocptte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (**1999**) *Proc. Natl. Acad. Sci. USA, 96*, 4285-4288.

[118] Kanehisa, M. and Goto, S. (**2000**) *Nucleic Acids Res., 28*, 27-30.

[119] Kanehisa, M., Goto, S., Kawashima, S. and Nakaya, A. (**2002**) *Nucleic Acids Res., 30*, 42-46.

[120] Iraqui, I., Vissers, S., Andre, B., Urrestarazu, A.I. **(1999)** *Mol. Cell Biol., 19*, 989-1001.

[121] De Boer, M., Bebelman, J. P., Goncalves, P.M., Maat, J., Van Heerikhuizen, H. and Planta, R. J. (**1998)** *Mol. Microbiol., 29*, 297-310.

[122] Klasson, H., Fink, G. R. and Ljungdahl, P. O. (**1999**) *Mol. Cell Biol., 19*, 5404-5416.

[123] Turner, G. C., Du, F., Varshavsky, A. (**2000**) *Nature, 405***,** 579-583.

[124] Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R. and Hood, L. (**2001**) *Science, 292*, 929-934.

[125] Brown, M., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. Jr., Haussler, D. (**2000**) *Proc. Natl. Acad. Sci. USA, 97*, 262-267.

[126] Eisen, M., Spellman, P. T., Brown, P. O. and Botstein, D. (**1998**) *Proc. Natl. Acad. Sci. USA, 95*, 14863-14868.

[127] Kell, D. B. and King, R. D. (**2000**) *Trends Biotechnol., 18*, 93-98.

[128] King, R. D., Karwath, A., Clare, A. and Dehaspe, L. (**2001**) *Bioinfrmatics, 17*, 445-454.

[129] Kemmeren, P., Berkum, N. L., Vilo, J., Bijma, T., Donders, R., Brazma, A. and Holstege, F. C. (**2002**) *Mol. Cell, 9*, 1133-1143.