Bioinformatics of Large-Scale Protein Interaction Networks

Vincent Schächter
Hybrigenics, Paris, France

Computational Proteomics Supplement 32:S16–S27 (March 2002)

ABSTRACT

We survey recent techniques for construction and prediction of large-scale protein interaction networks, focusing on computational processing steps. Special emphasis is placed on critical assessment of data completeness and reliability of the various approaches. Once built, protein interaction networks can be used for functional annotation or to generate higher-level biological hypotheses on pathways.

INTRODUCTION

With the completion of the full genome sequence for several model organisms, new approaches are emerging to comprehensively characterize the function of gene products. In so-called "functional proteomics" approaches, large-scale assays on the complete set of proteins of a given organism—the proteome—enable the study of the function of proteins in their context, rather than individually, through systematic identification of physical interactions between proteins. Indeed, most properties that are now grouped under the generic term "function " (of a protein) can be characterized more precisely through knowledge of protein interaction patterns. Moreover, networks of interacting proteins also extend this purely local view of function by providing a first level of understanding of cellular mechanisms (see Reference 43 for review).

While two promising approaches to the systematic analysis of protein complexes using mass spectrometry were published recently (see Note added in proofs), high-throughput techniques to construct protein interaction networks are mostly derived from the yeast two-hybrid system (9). Because of their underlying mechanism—the measuring of interactions between two chimeric and heterologous proteins in a yeast cell nucleus—two-hybrid assays cannot detect all protein-protein interactions and exhibit a certain proportion of false-positive and false-negative results. Existing studies rely on one of two main categories of yeast two-hybrid techniques: the matrix or the whole library approach. These techniques differ considerably not only in scale, but also by the nature and reliability of the results they yield.

While experimentally derived interaction data have started accumulating, so far they cover only a very small fraction of sequenced proteomes. Increasing recognition that knowledge of protein-protein interactions is key to the understanding of protein function motivated the design of predictive algorithms that generate hypotheses about interactions. These computational methods have so far been based mainly on sequence information (from the potential interacting partners as well as from their neighbors or orthologs in the case of methods based on "genomic context"), which set a priori limits on their predictive power (7).

Very recently, to capitalize on the availability of high-quality experimental interaction maps that include interaction domain information and cover a significant fraction of their underlying proteome (34), the first predictive methods based on a reference interaction dataset were proposed (44).

Even more than experimental techniques, predictive methods call for validation of specific results and for assessment of method sensitivity and selectivity. Proper validation methodologies, resting on a rigorous definition of what, precisely, is being assessed, as well as on reliable and reasonably complete reference datasets, have yet to be designed.

In the first part of this article, we briefly survey experimental approaches to protein interaction map construction, with a particular emphasis on the coverage and reliability of each type of approach. Computational methods for protein interaction map construction are then described, starting with sequence-based approaches, moving on to methods based on an interaction reference dataset, and concluding with a critical review of existing assessment and validation techniques. Finally, we discuss the uses of protein interaction networks, primarily for functional annotation, but also for more theoretical analyses of cellular network topology.

CONSTRUCTION OF PROTEIN INTERACTION NETWORKS FROM EXPERIMENTAL DATA

Experimental Technologies

While low-throughput technologies (co-immunoprecipitations, far-Western blots, "pull-downs", etc.; see Reference 33 for review) are commonly used for interaction studies on individual proteins, the study of interactions at the proteome level calls for high-throughput assays. The term functional proteomics is often used to refer to the corresponding technologies.

Two-hybrid in yeast. The yeast two-hybrid system (9) can detect interactions between two known proteins or polypeptides and can also search for unknown partners (preys) of a given protein (bait) (for review, see Reference 41). Yeast two-hybrid assays

are the main technology for large-scale interaction network construction. Two strategies, namely the matrix approach and the library screening approach, have been tested to find the most efficient way to explore interactions within the proteome (Figure 1).

The so-called "matrix approach" relies on a collection of predefined open reading frames (ORFs), usually full-length proteins, as both bait and prey for interaction assays. The experimental approach consists of amplifying ORFs by PCR, cloning them into two-hybrid vectors (both bait and prey), and expressing the fusion proteins individually in yeast cells of opposite mating type. Combinations of bait and prey can be assessed individually or after pooling cells expressing different bait or prey proteins. This strategy is intrinsically limited to the testing of predefined proteins. It was first used to explore interactions among *Drosophila* proteins involved in the control of cell cycle (10). Several studies have now been published for the vaccinia virus (26) and for the yeast proteome, comprehensive (15,16, 40) or using only a subset of specific baits (29).

The alternative yeast two-hybrid assay strategy uses exhaustive libraries to screen for the identification of new protein interacting partners. Repeating such screening experiments with a series of proteins involved in the same biochemical process led to the concept of specific functional protein interaction maps that could identify other previously uncharacterized proteins involved in the same pathway. This strategy can be extended to whole-cell interactomes. Moreover, as a single experiment simultaneously tests for interaction of a bait protein with a large number of randomly distributed fragments, upper-approximations of interacting domains can be computed as the common sequences shared by selected overlapping prey fragments (38). This approach was first applied to determine protein networks for the T7 phage proteome which contains 55 proteins (3) and later applied to the yeast proteome focused on the RNA metabolism (13), to hepatitis C virus (HCV) polypeptide interactions (11), and to the human gastric pathogen *Helicobacter pylori* (34).

The matrix and library strategies are depicted in Figure 1. Table 1 summarizes the results of major two-hybrid large-scale assays performed so far. For a more detailed comparative study of these technologies, see reviews (21, 36).

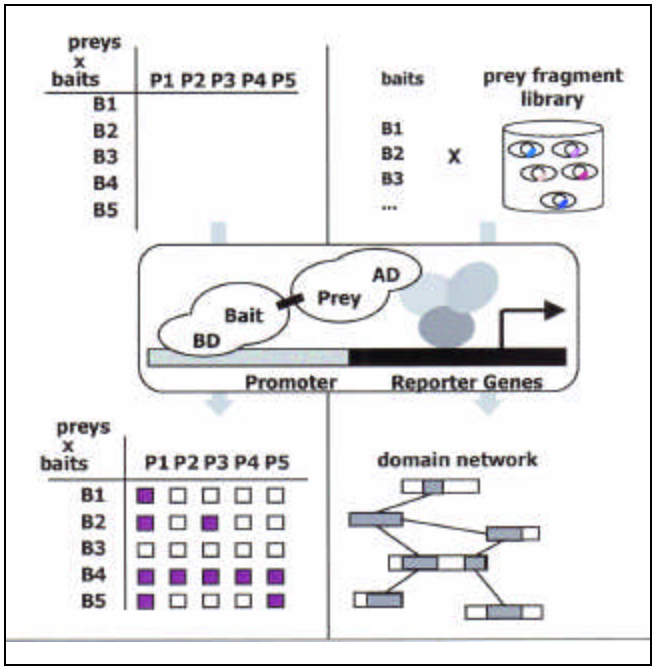

Figure 1. Yeast two-hybrid strategies. The central box illustrates the principle of the yeast two-hybrid assay: a protein domain that binds specifically to DNA sequences (BD) is fused to a "bait" polypeptide, and a domain that recruits the transcription machinery (AD) is fused to a "prey" polypeptide. Transcription of a reporter gene will occur if and only if the bait and the prey polypeptides interact together. The matrix approach (first column) uses the same collection of proteins as baits (B1-B5) and preys (P1-P5). The results can be represented in a matrix where bait auto-activators (B4 for example) and "sticky" prey proteins (P1, for example, interacts with many proteins) are identified and discarded. The final result can be summarized as a list of interactions. The library screening approach identifies for a given bait the set of its interacting prey partners, from which one can deduce the interaction domains on prey proteins. Sticky prey proteins are identified as fragments of proteins that are often selected regardless of the bait protein. An autoactivator bait can be used in the screening process with more stringent selective conditions.

From Raw Experimental Results to Protein Interaction Networks

The raw output of experimental methods is unusable for all but the smallest-scale studies. Transforming raw experimental data into understandable, publishable, and computationally tractable protein interaction networks requires bioinformatics support for a number of key processing steps.

From a computational biology perspective, a protein network can be seen as a graph with proteins (bait or prey) as vertices and interactions as edges. Accumulation of experimental data results in incremental construction of the graph. When two protein partners are identified experimentally, an edge is added to the graph, and possibly a vertex as well if one of the partners was not included in the preexisting graph; if the experiment is asymmetrical, as in two-hybrid techniques, the edge may be directed.

This step is trivial when the two partners are known beforehand, for example, in the two-hybrid matrix approach. When a partner is screened against a library and selects a prey, however, post-processing is required: the prey gene must be sequenced and identified in sequence databases using tools such as BLAST (2). Moreover, in the case of a two-hybrid strategy using fragment libraries, an upper approximation of the interacting domains can be mapped on proteins: the common sequence shared by the selected overlapping prey fragments defines the smallest selected docking site of the bait (38). A more accurate representation of the experimental results is thus a graph where vertices represent protein domains instead of full-length proteins.

To enable high-throughput production of protein interaction information, these computational steps must be integrated into a processing pipeline. For instance, the strategy that was used to construct the *H. pylori* interaction map (34) (see Figure 2) was supported by a dedicated integrated laboratory production management system, the PIM Builder®. The PIM Builder tracks all biotechnological or bioinformatics operations performed during the production processes, stores information about all biological objects produced during experiments, and interfaces with robots and bioinformatics modules. It also implements the processing steps necessary to construct interaction maps from raw experimental data, including its two major computational steps, identification of interaction domains and scoring of interactions.

Table 1. Key Figures in Large-Scale Datasets for Protein-Protein Interaction Maps

Organism	Technology	Number of Assays Baits × Preys	Number of Interactions	Reference
vaccinia virus	Protein array	proteome × proteome	37	(26)
Saccharomyces cerevisiae	Protein array Pools of preys	192 × proteome proteome × proteome	281 692	(40)
S. cerevisiae	Pools of baits and preys	430 assays of pools (96 × 96)	175	(16)
S. cerevisiae	Pools of baits and preys	3.844 assays of pools (96 × 96)	841*	(15)
S. cerevisiae	Protein array	162 × 162	213	(29)
Caenorhabditis elegans	Protein array Library screening	29 × 29 27 × proteome	8 124	(42)
HCV	Protein array Library screening	10 × proteome 22 fragments × proteome	0 5	(11)
S. cerevisiae	Library screening	15 × proteome	170	(13)
S. cerevisiae	Library screening	11 × proteome	113	(12)
H. pylori	Library screening	261 × proteome	1524	(34)

*This number corresponds to highly significant interactions (more than three hits, see Reference 15)

Several additional light processing steps are implemented to facilitate experiment design and execution, including a "bait program" that automatically designed oligonucleotides for PCR amplification and sequencing of bait constructs, a "prey program" that determined the position of each fragment in the genome and its coding capacity (intergene, antisense, nucleotide position in an ORF, coding frame, etc.).

Coverage and Reliability

One major issue with the high-throughput experimental technologies described above is the generation of potential false negatives and false positives.

False-negative interactions are biological interactions that are missed, because of incorrect folding, inadequate subcellular localization, lack of specific post-translational modifications, and the like. The yeast two-hybrid matrix approach is likely to generate a high level of false negatives (see Table 2), because only two assays are performed for each pair of proteins (bait versus prey, and vice versa), whereas the fragment library approach tests for millions of potential interactions simultaneously and is therefore more likely to capture a fragment that exhibits the appropriate folding if such a fragment exists. For instance, the two exhaustive studies of the yeast proteome (15,40) have failed to recapitulate as much as 90% of interactions previously described in the literature (15). Intrinsic limitations of the matrix approach regarding the choice of selective conditions can also explain this high rate of false negatives (for review see Reference 21).

On the other hand, searching for many potential interactions, especially when screening a random fragment library, increases the likelihood of selecting biologically nonsignificant interacting polypeptides, thus leading to false positives. First, some bait proteins might have a predisposition to activate the transcription of reporter genes without specific interaction with any prey protein. These auto-activator bait proteins may interact with a large random set of prey proteins. Second, some chimeric prey proteins, dubbed sticky proteins, may similarly be nonspecifically selected by many independent bait proteins. Discarding auto-activator bait proteins (that select many prey proteins in one screen) or sticky prey proteins (that are selected in many screens) significantly reduces the rate of false positives, although it may also slightly increase the rate of false negatives (15,16). The technology described above for the construction of the *H. pylori* interaction map (34) was designed to specifically address major known causes for false-negative and false-positive results in two-hybrid assays. Parallel screening against highly complex libraries of fragments greatly increased the number of able two-hybrid candidates, and bait selectivity fine-tuning and quality-control steps were implemented to tackle toxicity and auto-activation issues. In combination, these measures considerably reduced the rate of false negatives that arose with the matrix approach. On the false positives front, quality-control measures, elimination of auto-activating baits, and use of adapted reporter systems ensured that the two-hybrid technology realized its full potential. More importantly, the statistical nature of the experimental procedure (i.e., sets of screens against fragment libraries) allowed the detection of nonspecific partners (sticky SIDs) through a scoring scheme that computes an E-value for each bait-prey interaction. The score was obtained by comparing the observed pattern of selected prey fragments with the theoretical pattern that would be obtained by randomly picking fragments in the library. Global connectivity was also taken into

account; the score was computed incrementally over the whole network, and its discriminatory power increased as screening results accumulate. A specific score category was added to distinguish interactions involving only highly connected prey domains (SIDs which were found as prey with frequency greater than a fixed threshold). As a result, each interaction was tagged with one of several discrete reliability values. These values can be used to filter or prioritize the interactions that are generated by two-hybrid data.

COMPUTATIONAL PREDICTION OF PROTEIN INTERACTION NETWORKS

Prediction of Functional Links by Comparative Genomics Techniques

Several approaches to computational prediction of protein networks have been explored over the last two years. The majority of these attempt to predict functional links "ab initio," on the sole basis of sequence data from completely sequenced genomes. The underlying algorithms are inspired by comparative genomics techniques, i.e., identification of (sets of) ortholog genes in two or more related organisms (see Reference 7 for a detailed review).

- The gene-fusion event (i.e., Rosetta Stone) method (8,22) is based on an evolutionary interaction hypothesis: if two genes x and y are separate in a given organism and exist as fused as a single gene z in an ancestor organism, a functional link is inferred.
- The gene neighborhood approach (5,31) rests on the more general hypothesis that interaction of encoded proteins is one of the reasons for conservation of gene proximity and order. In Reference 31, functional links are inferred between genes x and y if these are neighbors (with respect to chromosomal location) in organism A and have orthologs in organism B that are neighbors as well.
- The phylogenetic profiles method (32) deduces functional links between genes that have similar occurrence patterns of orthologs in a set of reference genomes, since these genes are then assumed to have co-evolved.

Each of the above approaches exhibits an a priori bias corresponding to the biological hypothesis underlying the prediction algorithm. Comparison with experimental data confirms this bias (23). One way to reduce the bias and to minimize the rate of false-positive predictions is to combine several approaches, seen as providing independent sets of clues that hint at the existence of functional links. Marcotte et al. (23) show that Rosetta Stone and phylogenetic profile predictions are combined with metabolic pathways information on *E. coli* from EcoCyc (20), yeast two-hybrid interaction data from DIP (45), and links inferred from yeast cell cycle expression data (39) to yield higher-confidence predictions.

Another limitation of these comparative genomics methods is that the exact biological nature of the predicted functional links—participation in the same structural complex, in the same biological pathway, the same biological process, or, in some cases, existence of a physical interaction—cannot be specified without additional information.

Prediction of Protein Interactions Across Organisms

Once a protein network is built in a given organism (by experimental or predictive methods), a natural question that arises is how much information can be deduced from that network about the interactions taking place in another organism? The straightforward "comparative genomics" answer would involve two major steps:

1. Establishment of a correspondence between proteomes, classically by identifying orthologs between organisms by sequence comparison.
2. Transport of links in the source protein network to the target proteome along this correspondence.

The accuracy of this type of approach is obviously highly dependent on the criteria chosen for orthology and its relationship

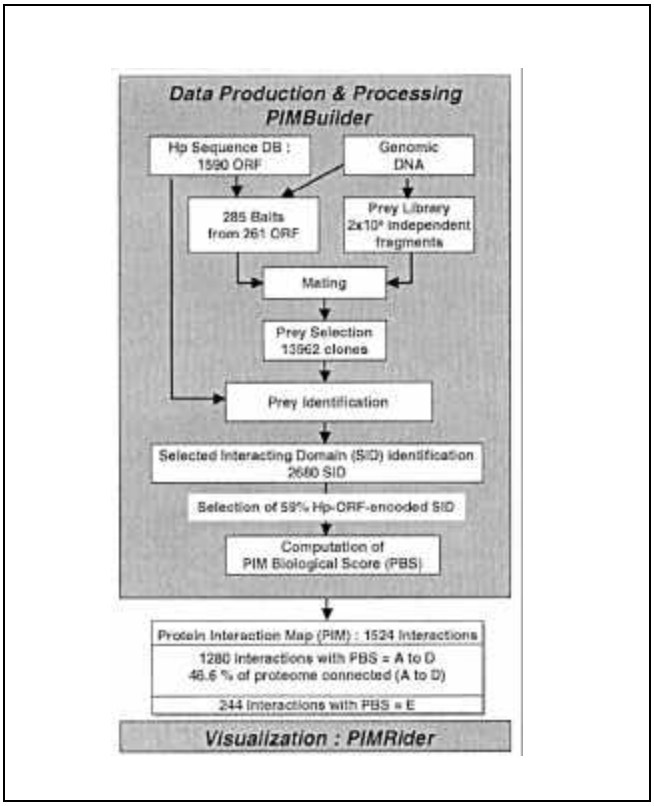

Figure 2. General outline of the strategy for building *H. pylori* (Hp) proteome-wide interaction map. Bait constructs were specifically adapted for interaction screens, and the selective pressure was adjusted for each construct according to results obtained from a preliminary small-scale screening experiment. Each bait was screened against a library of *H. pylori* random genomic fragments. The PIM Builder LIMS was populated with raw data from screening experiments. After identification of almost all positive clones, overlapping prey fragments were clustered into families to define the selected interacting domains (SIDs). Those families that had no biological coding capability (antisense or intergenic region, out-of-frame fragments occurring in a single frame) were discarded. The PIM Biological Score (PBS) was then computed for *H. pylori* ORF-encoded SIDs. Interactions were grouped into categories A to D (from high to low heuristic values). The global connectivity of the protein interaction map (PIM) was also analyzed to detect highly connected prey polypeptides. Those interactions were grouped within the E category. Processing of data and visualization of interactions were performed by an in-house built bioinformatic platform (the PIM Rider®).

to the nature of the link to be predicted. For instance, in the hypothetical extreme case where links in the source protein network are completely independent from sequence features and orthology is simply defined by sequence similarity, the inference is obviously meaningless.

To better address the issue of prediction of physical interactions, a technique was introduced (44) to predict protein-protein interaction maps across organisms using experimental interaction data as input, the interaction-domain profile-pair (IDPP) method (see Figure 3). This technique was designed to fully exploit the properties of the richest experimental interaction maps now available, namely the existence of domain information for each interaction and the fact that for a given ID domain d, a large-scale map will typically provide several instances of domains interacting with d. The algorithm combines sequence similarity searches with clustering based on interaction patterns and interaction domain information. The source map is first transformed into an abstract interaction map connecting clusters of interaction domains. A correspondence is then built between this abstract interaction map and the target proteome, and the interactions are inferred along this correspondence.

Wojcik and Schächter (44) applied IDPP to the prediction of an interaction map of *E. coli* using the *H. pylori* interaction map as reference, together with an inference method following the straightforward comparative genomics approach described above (correspondence according to sequence similarity on full-length sequences), referred to as the "naive" method.

From the 1524 interactions of the original *H. pylori* network, the IDPP method led to 881 interaction predictions, connecting 412 proteins of *E. coli* (9.6%). Compared to the naive method, the IDPP method yielded 35 additional, highly domain-specific, predicted interactions. The use of sequence similarity searches restricted to interacting domains rather than full-length proteins increases the sensitivity of the method, while the use of interacting domain clusters instead of single interacting domain sequences allowed the detection of homologies at lower levels of sequence similarity. For instance, *H. pylori* protein HP1411 has no homolog in *E. coli*, whether one considers its full-length sequence or a sub-sequence corresponding to an interaction domain. Nevertheless, because HP1411 interacts with the gyrA *H. pylori* protein and also shares a sequence similarity with gyrA, a profile merging gyrA and HP1411 sequences was built and succeeded in selecting the homologous *E. coli* gyrA protein. A gyrA homodimer was thus predicted in *E. coli* (Figure 4). This prediction was confirmed by SWISS-PROT annotations, according to which gyrA forms an A2-B2 complex with gyrB.

Six hundred and fifty-one interactions were predicted by the naive method but not by the IDPP method. Two hundred and fifty-two of these 651 interactions were demonstrated to be false positives of the naive method, since the prediction was achieved through sequence similarity of a region that does not contain the interacting domain. The 399 remaining interactions were obtained through sequence similarity that was significant when

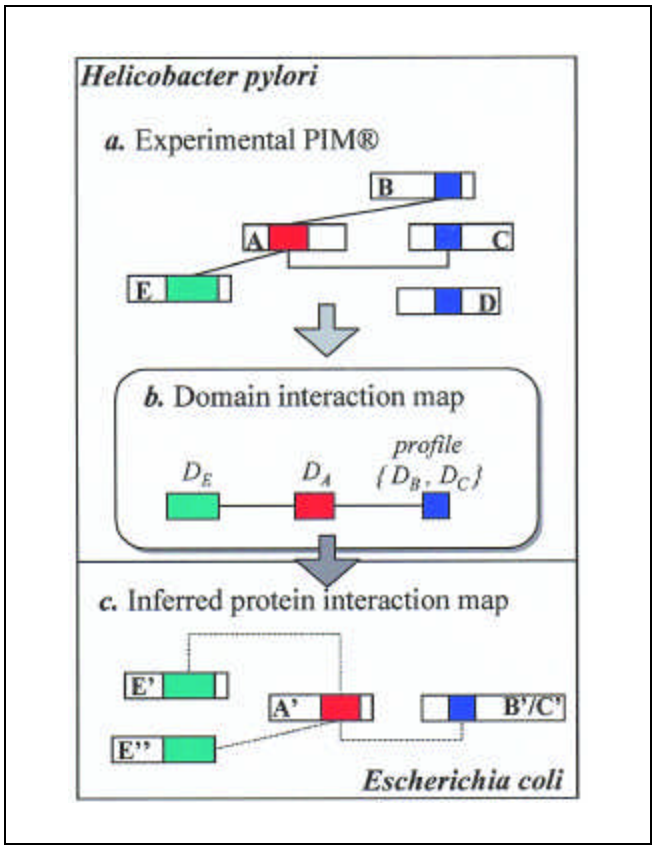

Figure 3. The Interacting Domain Profile Pair method. From the initial protein interaction map of *H. pylori* (a), an abstract domain cluster interaction map is derived (b). Domains are clustered together if (*i*) they share a significant sequence similarity and (*ii*) they share a common interaction property with a third partner (e.g., interacting domains of proteins B and C both interact with A). Each domain or profile of domains is then used as a probe to screen a library of *E. coli* protein sequences and domain cluster interactions are transferred (c).

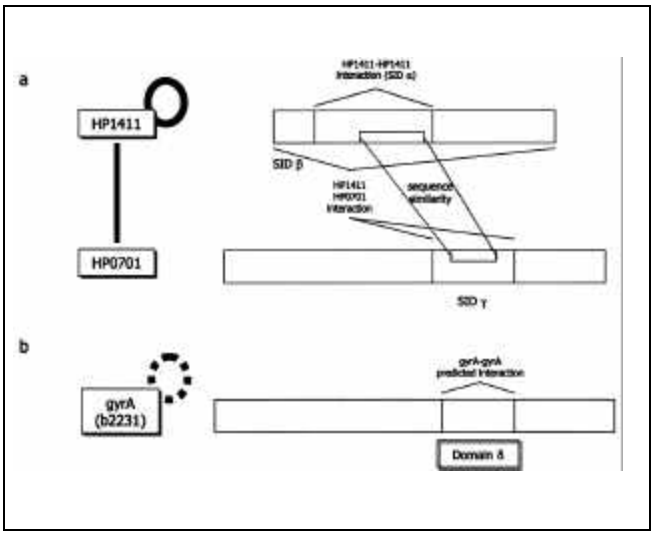

Figure 4. Prediction of gyrA homodimerization in *E. coli*. In the reference Hp protein interaction map, ID β of HP1411 interacts with ID γ of HP0701 and HP1411 interacts with itself through ID α (a). When the IDPP method is applied, ID α and ID γ are clustered in the same domain cluster since they both interact with the same region of HP1411 and the 197-332 region of HP1411 is similar to the 498-627 region of HP0701. The resulting abstract domain cluster is a "homodimer." When used as a probe to screen an *E. coli* protein sequence library, the profile of this cluster selected a 172 amino acid long domain δ on the gyrA protein, and gyrA was predicted to interact with itself through this domain (b).

considering the whole protein but not when considering the shorter included interacting region.

The domain-based method was thus shown to eliminate a significant amount of false positives of the naive method that are the consequences of multi-domain proteins and increase the sensitivity compared to the naive method by identifying new potential interactions.

Assessment and Validation of Predictions

Each of the predictive algorithms described above is based on specific biological hypotheses, including several empirically chosen quantitative parameters and applied to an input (or training) dataset (e.g., sequence data, expression profiles, a reference interaction map, etc.). Their predictions can in theory be validated by comparing them to accepted biological knowledge. The actual scientific goal, however, is generally to validate the method itself, whereas accuracy of specific predictions is also a function of parameter choice and the quality of the input dataset—one consequence being the notorious "garbage-in, garbage-out" problem. It is especially difficult to distinguish the respective roles in the overall reliability of results when the input dataset is already indirect, deduced information, such as interpretation of nonstandardized experiments.

In practice, even validating predictions is a tricky proposition; reference information is scarce, and accessing this information, especially with automated procedures, may also be problematic. In addition, reference status is often subjective, as is the interpretation of all but the most finalized functional information. We distinguish below between automated validation methods, which are somewhat more objective and can be applied at high-throughput but often yield weak biological confirmation, and manual expert validation methods, which typically tap a wider body of knowledge but are labor-intensive and difficult to reproduce or compare.

Automated validation. Predicted protein-protein links can be evaluated by directly checking their existence in dedicated databases, such as MIPS (27), DIP (45), or OMIM (14). For instance, this approach was used to validate literature networks (17). Predictions can also be confronted to other types of data, for example, gene clusters from microarray data (17). In both cases, the significance of the predictions is evaluated by calculating the fold improvement over a virtual random experiment and/or the correlation between the two datasets. However, interaction informa-

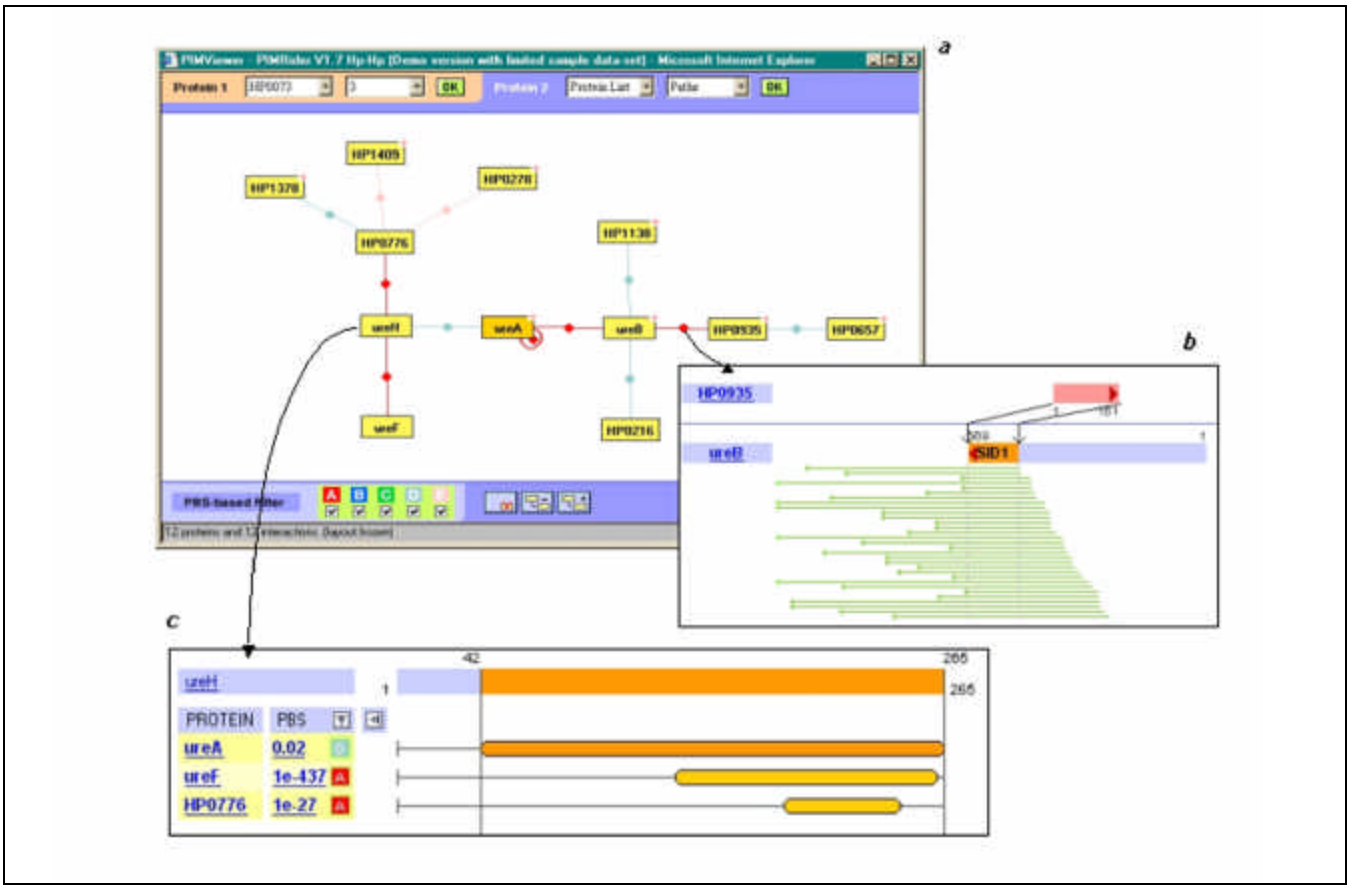

Figure 5. The PIM Rider: an integrated exploration platform for protein interaction networks. The main window (a) displays the protein interaction map as a graph and allows the biologist to navigate through the network, filter interactions and proteins on criteria such as interaction reliability (PBS), and focus on a particular pathway. Clicking on a specific interaction gives access to a summary of the supporting primary two-hybrid data (b), where interacting fragments and the computed selected interacting domain (SID) are positioned relative to the coding sequence of the two proteins. All interacting domains of a given protein (c) and paths between two proteins can also be queried.

tion directly available in tractable format from public databases is scarce, so it is often necessary to resort to indirect methods.

The most widely used validation method is known as the "keyword retrieval" technique. The principle is simple: if two proteins are linked together in the network, one compares the sets of functional keywords associated to these proteins in a given database; if the pair shares similar keywords, the weight of the link is reinforced. The percentage of shared keywords at the network level can be compared to a theoretical background noise to evaluate the global validity of the prediction. For instance, the keywords can be SWISS-PROT annotation keywords or functional categories (17,23,44). This validation method, however, rests heavily on the quantity, quality, homogeneity, and comparability of database annotations. For example, Marcotte et al. state that "even truly related proteins show only a partial [SWISS-PROT] keyword overlap, for example 35%" (23). Thus, the method, while adequate for confirming that a prediction algorithm performs better than random on a given space, yields weak biological validation in general.

Another approach to validation circumvents the reference dataset availability problem by cross-validating predictions with predictions resulting from other, ideally independent, experimental methods. The underlying principle is that a prediction confirmed by two or more independent functional clues is significantly more reliable. It has been used to define high-confidence links in protein networks (23) and to assess interaction predictions against physical location of genes in prokaryotic genome (44). One problem in that scenario is to assess the real independence of prediction methods, the majority of them being based on sequence data.

Validation by manual expert analysis. In this approach, each predicted link between a pair of proteins of the network is assessed by manually comparing the annotations in public databases and checking reference literature and original literature of each protein partner. This low-throughput method obviously leaves a wider margin to individual interpretation but significantly increases the quantity of accessible reference knowledge.

This method has been applied to the assessment of protein interactions inferred from *H. pylori* to *E. coli* (J. Wojcik, J.G. Boneca, and P. Legrain, personal communication). The inference process is based on clustering and a definition of orthology restricted to interaction protein domains (44). The true-positive prediction rate was evaluated to be at least 12%, i.e., at least 12% of the 1280 predicted interactions makes biological sense according to biological curators. Three main causes were identified to explain predictions that were not confirmed by the literature: (*i*) one of the gene functions in the source interaction was completely lost during evolution (this gene has only paralogs in *E. coli*); (*ii*) the source interaction is a false positive of the two-hybrid system; or (*iii*) the predictions are real true positives but are not yet referenced in the literature. Comparison of these precise but statistically nonsignificant results with those obtained by automated validation by keyword retrieval (44), that were significantly better than random, points at the need to have real and exhaustive reference datasets in order to validate predictions.

Literature Mining

Literature mining, sometimes called "information retrieval," can be viewed both as an assessment method for predicted protein networks and as a prediction method in its own right. Assuming that the greater part of current biological knowledge is to be found in scientific literature, the parsing and analysis of titles, headings, abstracts, and/or full texts of articles should enable us to extract links between genes or proteins and then build networks. Several techniques exist to perform this extraction, including syntactic and semantic analysis methods imported from the computational linguistics community (e.g., 30) and statisti-

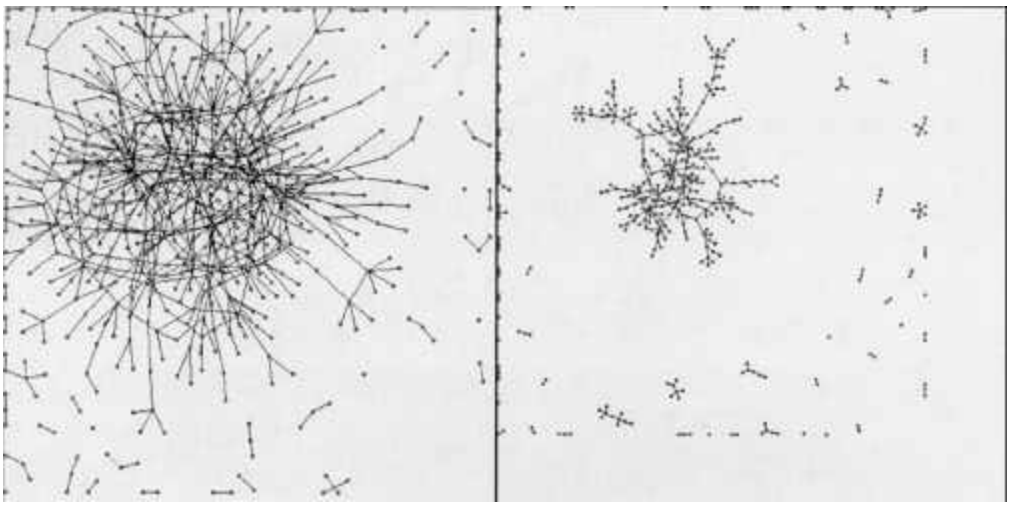

Figure 6. Possible influence of data reliability on network topology of large-scale protein interaction networks. The interaction map on the left includes *H. pylori* interactions (34) in the 3 best score categories out of 5, while the map on the right represents the same dataset, but includes only interactions in the best score category.

Table 2. Main Protein-Protein Interaction Databases

Database	URL	Reference
BIND	www.bind.ca	
Cellzome	yeast.cellzome.com	
CuraGen portal	portal.curagen.com	(40)
DIP	dip.doe-mbi.ucla.edu	(45)
FlyNets	gifts.univ-mrs.fr/FlyNets/	(35)
Interact	bioinf.man.ac.uk/interactso.htm	(6)
MIPS	www.mips.biochem.mpg.de	(27)
PIM Rider	pim.hybrigenics.fr	(34)
ProNet	pronet.doubletwist.com	

cal methods that estimate discriminating word distributions (e.g., 24). One major issue in these studies is the establishment of an unambiguous nomenclature for gene or gene product names. Gene name dictionaries can be built from various nomenclature databases such as HUGO, LocusLink, or GENATLAS, but problems remain because of insufficient synonym definitions, synonym variations, and gene families with fuzzy naming conventions.

A recent work aimed to analyze over 10 million MEDLINE records to detect and count human gene symbol or names co-occurring in titles or abstracts. This resulted in a protein interaction network containing about 140 000 edges connecting 7512 human genes (17), the largest protein network predicted from literature mining so far. Until these literature mining techniques mature, they can also be used with less ambitious but perhaps more directly efficient goals, such as helping experts filter or preselect scientific articles. The Database of Interacting Protein (DIP) (24) is one example of such computer-assisted curation.

EXPLOITATION OF PROTEIN INTERACTION NETWORKS

Experimental or predicted protein networks constitute large bodies of information on molecular mechanisms. The first step toward fruitful utilization of these networks is visual exploration. More sophisticated approaches involving additional computational analyses can be grouped in two categories: assignment of functional attributes to proteins in context and analysis of global network structure, for example in an attempt to uncover or confirm evolutionary hypotheses.

Visualization and Exploration Tools

The first interaction databases available on the Internet provided a basic display of the alphabetical interaction list: an interaction is represented by its two protein partners, sometimes accompanied by basic annotations or cross-references to other protein databases. A few Web sites also propose packages for interaction networks visualization (28). The main protein-protein interaction sources are listed in Table 2.

Simple lists of interactions are hardly sufficient to evaluate the reliability of results or compare with other information. Access to primary data is especially important to evaluate false positives and reproducibility. For example, interactions listed at the MIPS (27) only provide a brief indication of the experimental source, like "two-hybrid" or "co-immunoprecipitation", without any assessment of quality nor reference to the source experiment or laboratory. More recent bioinformatics tools, such as the PIM Rider (34), give access to primary data (Figure 5) and offer functionalities beyond basic visualization to help the biologist in the discovery process searching for paths between two given proteins, filtering of interactions on the basis of their reliability value, or simultaneous displaying of all interacting domains identified in one specific protein.

Functional Annotations: The "Guilt-by-Association" Rule

First attempts to assign function on the basis of interaction information are applications of the guilt-by-association principle: a protein is annotated using some lowest common denominator of the annotations of its interacting partners, or, more generally, of proteins belonging to a given cluster of interactants (25).

For instance, a set of yeast protein interactions described in the literature or revealed by large-scale two-hybrid screens was clustered using cellular role and subcellular localization annotations (37) from the Yeast Proteome Database (4). Functions were assigned to uncharacterized protein on the basis of the known functions of its interacting partners. Twenty-nine proteins had two or more interacting proteins with at least one common function and were assigned a function by this approach.

Guilt-by-association methods should be used with caution. First, predictions are highly dependent on the annotations available from a given database, which are often vague (e.g., one keyword) and sometimes false. Poorly defined annotations often group different levels or types of descriptions and induce clustering that is devoid of biological significance. The quality of the source protein network is also a key factor: false negatives (missing connections) will result in missed predictions, false positives in false predictions. In the case of two-hybrid interaction data, for which highly connected nodes in the network may be false positives, reliability is even more critical.

Finally, a major difficulty with this kind of automated function annotation method, yet common to all bioinformatics prediction algorithms, is the absence of an independent reference dataset and validation methods. For instance, the 29 function assignments made in the former study were assessed by comparison with the corresponding high confidence links of the study of Marcotte et al. (23), which were themselves partly predicted from interactions listed in MIPS, one of the yeast protein interaction databases used in the original study (37). This exemplifies the fact that predictions must be used with caution; oversight of the initial hypotheses and lack of independent data sources could lead to biased conclusions.

Annotating proteins using their interaction partners' annotations is still a promising technique that will become more and more useful as the interaction data accumulate and their quality improves. Meanwhile, confidence in such functional annotations can be increased by aggregating conclusions based on interaction data with functional clues of clearly independent origin.

Analysis of the Global Structure of Protein Networks

Several authors have proposed analyses of the topology of large-scale interaction networks, to gain insight into global cellular mechanisms within an evolutionary perspective.

Jeong and co-workers published such an analysis of the public yeast protein interaction map (18), showing that it forms a scale-free network. In such networks, the probability that a given protein interacts with k partners follows a power law; functionality (i.e., lines of communication) is mostly preserved when the network is attacked randomly, yet is fragile when certain key vertices are disrupted (1). Scale-free networks can be generated by random mechanisms that add connections to a graph with a positive bias for already well-connected components of the network, which is consistent with current hypotheses on the evolution of pathways, favoring robustness in terms of resistance to random mutations. Indeed, similar network structures were also evidenced for metabolic networks (19) and another protein interaction map in bacteria (34). The authors establish a positive correlation between connectivity and lethality: highly connected proteins are three times more likely to be essential (i.e., the yeast cell dies if the corresponding gene is deleted).

While the existence of such a correlation makes biological sense, the relative weight of technological bias in the argument should not be underestimated. Jeong's conclusions rest mainly on interaction data produced by a given high-throughput two-hybrid system in yeast, which is known to induce high rates of false negatives and false positives, as exemplified in a more recent similar study (15). The corresponding network of 1870 proteins (31% of the yeast proteome), is far from complete, and the shape of the actual interaction network could be quite different (see Figure 5). For instance, proteins that exhibit few interacting partners in this network could actually represent highly connected nodes. Conversely, false positives of the two-hybrid system are likely to result in highly connected nodes of the network (so-called "sticky prey" proteins).

In conclusion, while these approaches will certainly be refined in years to come, for the time being both local guilt-by-association functional assignment rules and global network analysis methods are fragile against poor interaction data quality or incompleteness. Consequently, they probably should not be used as stand-alone sources of local conclusions, at least without prior assessment of the technological bias. Rather, conclusions should be seen as hypotheses asking for confirmation by other means.

CONCLUSION

Large-scale protein interaction maps are, with gene expression profiles, among the first examples of datasets generated without prior knowledge on functions of genes. These technology-driven experiments are valuable tools for protein function prediction, despite the occurrence of typical artifacts. A number of technological variants of the yeast-two hybrid system have been tested during the last two years. Bioinformatics tools enable high-throughput production of experimental results with quality control, transformation of these results into protein networks, and exploitation through visualization and analysis tools. Recent improvements on throughput of the experimental method, as well on reliability and level of detail of the resulting protein interaction networks, have led to proteome-wide datasets that can be used as the basis for predictive methods, a welcome complement to algorithms that predict functional links from sequence information using comparative genomics techniques.

Perspectives opened by such predictive methods include ab initio prediction of virtual protein interaction maps on a variety of organisms related to existing experimental protein interaction maps, combined use of prediction and experimental work to accelerate the construction of new interaction maps, or the identification of new shared interaction domains.

While protein networks represent a new and potentially very rich type of functional information, their use for functional annotation should always be accompanied by a critical assessment of the intrinsic limitations and biases of their construction methodology, as well as by explicit distinction of different levels of confidence on the scale that stretches from "likely physical interaction in artificial context" to "validated interaction with biological meaning." The fact that reference datasets are still almost nonexistent should always be kept in mind.

Ultimately, combination of functional clues from different experimental origins—yeast two-hybrid assays, microarrays, mass spectrometry—with sequence data and predictive methods, structured by the appropriate knowledge management tools, should facilitate the assignment of functional annotations. One step further, combination with other sources of functional information (queried from structured databases or extracted from literature) can help transform protein networks into detailed descriptions of cellular pathways, enabling a shift in our view of function, from "local property of a protein" to "the role of that protein in one or several processes." Such knowledge integration is a major challenge for computational biology, requiring the development of pathway models that can bridge the gap between different representations adapted to specific experimental data types and enable in silico hypotheses testing on incomplete information sets.

ACKNOWLEDGMENTS

We are grateful to Alexandre Hamburger and Jerome Wojcik for sharing their views on computational methods related to protein interactions, and to Pierre Legrain for many stimulating discussions on protein interaction network construction techniques and on biological data interpretation and reliability issues. Our thanks also go to Pierre Legrain and Donny Strosberg for a critical reading of the manuscript.

REFERENCES

1. **Albert, R., H. Jeong, and A.L. Barabasi.** 2000. Error and attack tolerance of complex networks. Nature *406*:378-382.
2. **Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*:3389-3402.
3. **Bartel, P.L., J.A. Roecklein, D. SenGupta, and S. Fields.** 1996. A protein linkage map of Escherichia coli bacteriophage T7. Nat. Genet. *12*:72-77.
4. **Costanzo, M.C., J.D. Hogan, M.E. Cusick, B.P. Davis, A.M. Fancher, P.E. Hodges, P. Kondu, C. Lengieza, et al.** 2000. The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): compre-

hensive resources for the organization and comparison of model organism protein information. Nucleic Acids Res. *28*:73-76.
5. **Dandekar, T., B. Snel, M. Huynen, and P. Bork.** 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. *23*:324-328.
6. **Eilbeck, K., A. Brass, N. Paton, and C. Hodgman.** 1999. INTERACT: an object-oriented protein-protein interaction database, p. 87-94. *In* Seventh International Conference of Intelligent Systems for Molecular Biology.
7. **Eisenberg, D., E.M. Marcotte, I. Xenarios, and T.O. Yeates.** 2000. Protein function in the post-genomic era. Nature *405*:823-826.
8. **Enright, A.J., I. Iliopoulos, N.C. Kyrpides, and C.A. Ouzounis.** 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature *402*:86-90.
9. **Fields, S. and O. Song.** 1989. A novel genetic system to detect protein-protein interactions. Nature *340*:245-246.
10. **Finley, R.L., Jr., and R. Brent.** 1994. Interaction mating reveals binary and ternary connections between Drosophila cell cycle regulators. Proc. Natl. Acad. Sci. USA *91*:12980-12984.
11. **Flajolet, M., G. Rotondo, L. Daviet, F. Bergametti, G. Inchauspé, P. Tiollais, C. Transy, and P. Legrain.** 2000. A genomic approach of the hepatitis C virus generates a protein interaction map. Gene *242*:369-379.
12. **Fromont-Racine, M., A.E. Mayes, A. Brunet-Simon, J.C. Rain, A. Colley, I. Dix, L. Decourty, N. Joly, F. Ricard, J.D. Beggs, and P. Legrain.** 2000. Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. Yeast *17*:95-110.
13. **Fromont-Racine, M., J.C. Rain, and P. Legrain.** 1997. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens [see comments]. Nat. Genet. *16*:277-282.
14. **Hamosh, A., A.F. Scott, J. Amberger, D. Valle, and V.A. McKusick.** 2000. Online Mendelian Inheritance in Man (OMIM). Hum. Mutat. *15*:57-61.
15. **Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki.** 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA *98*:4569-4574.
16. **Ito, T., K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki.** 2000. Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc. Natl. Acad. Sci. USA *97*:1143-1147.
17. **Jenssen, T.K., A. Laegreid, J. Komorowski, and E. Hovig.** 2001. A literature network of human genes for high-throughput analysis of gene expression. Nat. Genet. *28*:21-28.
18. **Jeong, H., S.P. Mason, A.-L. Barabasi, and Z.N. Oltvai.** 2001. Lethality and centrality in protein networks. Nature *411*:41-42.
19. **Jeong, H., B. Tombor, R. Albert, Z.N. Oltvai, and A.-L. Barabasi.** 2000. The large-scale organization of metabolic networks. Nature *407*:651-654.
20. **Karp, P.D., M. Riley, M. Saier, I.T. Paulsen, S.M. Paley, and A. Pellegrini-Toole.** 2000. The EcoCyc and MetaCyc databases. Nucleic Acids Res. *28*:56-59.
21. **Legrain, P., J. Wojcik, and J.M. Gauthier.** 2001. Protein—protein interaction maps: a lead towards cellular functions. Trends Genet. *17*:346-352.
22. **Marcotte, E.M., M. Pellegrini, H.L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg.** 1999. Detecting protein function and protein-protein interactions from genome sequences. Science *285*:751-753.
23. **Marcotte, E.M., M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg.** 1999. A combined algorithm for genome-wide prediction of protein function [see comments]. Nature *402*:83-86.
24. **Marcotte, E.M., I. Xenarios, and D. Eisenberg.** 2001. Mining literature for protein-protein interactions. Bioinformatics *17*:359-363.
25. **Mayer, M.L. and P. Hieter.** 2000. Protein networks—built by association. Nat. Biotechnol. *18*:1242-1243.
26. **McCraith, S., T. Holtzman, B. Moss, and S. Fields.** 2000. Genome-wide analysis of vaccinia virus protein-protein interactions. Proc. Natl. Acad. Sci. USA *97*:4879-4884.
27. **Mewes, H.W., D. Frishman, C. Gruber, B. Geier, D. Haase, A. Kaps, K. Lemcke, G. Mannhaupt, F. Pfeiffer, C. Schuller, S. Stocker, and B. Weil.** 2000. MIPS: a database for genomes and protein sequences. Nucleic Acids Res. *28*:37-40.
28. **Mrowka, R.** 2001. A Java applet for visualizing protein-protein interaction. Bioinformatics *17*:669-671.
29. **Newman, J.R., E. Wolf, and P.S. Kim.** 2000. A computationally directed screen identifying interacting coiled coils from Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA *97*:13203-13208.
30. **Ono, T., H. Hishigaki, A. Tanigami, and T. Takagi.** 2001. Automated extraction of information on protein-protein interactions from the biological literature. Bioinformatics *17*:155-161.
31. **Overbeek, R., M. Fonstein, M. D'Souza, G.D. Pusch, and N. Maltsev.** 1999. The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. USA *96*:2896-2901.
32. **Pellegrini, M., E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates.** 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA *96*:4285-4288.
33. **Phizicky, E.M. and S. Fields.** 1995. Protein-protein interactions: methods for detection and analysis. Microbiol. Rev. *59*:94-123.
34. **Rain, J.C., L. Selig, H. de Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, et al.** 2001. The protein-protein interaction map of Helicobacter pylori. Nature *409*:211-216.
35. **Sanchez, C., C. Lachaize, F. Janody, B. Bellon, L. Roder, J. Euzenat, F. Rechenmann, and B. Jacq.** 1999. Grasping at molecular interactions and genetic networks in Drosophila melanogaster using FlyNets, an Internet database. Nucleic Acids Res. *27*:89-94.
36. **Schächter, V.** 2002. Construction and prediction of protein-protein interaction maps. Ernst Schering Research Foundation, Vol. 38. Springer-Verlag, New York.
37. **Schwikowski, B., P. Uetz, and S. Fields.** 2000. A network of protein-protein interactions in yeast. Nat. Biotechnol. *18*:1257-1261.
38. **Siomi, M.C., M. Fromont, J.C. Rain, L. Wan, F. Wang, P. Legrain, and G. Dreyfuss.** 1998. Functional conservation of the transportin nuclear import pathway in divergent organisms. Mol. Cell. Biol. *18*:4141-4148.
39. **Spellman, P.T., G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher.** 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell *9*:3273-3297.
40. **Uetz, P., L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, et al.** 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature *403*:623-627.
41. **Vidal, M. and P. Legrain.** 1999. Yeast forward and reverse 'n'-hybrid systems. Nucleic Acids Res. *27*:919-929.
42. **Walhout, A.J., R. Sordella, X. Lu, J.L. Hartley, G.F. Temple, M.A. Brasch, N. Thierry-Mieg, and M. Vidal.** 2000. Protein interaction mapping in C. elegans using proteins involved in vulval development [see comments]. Science *287*:116-122.
43. **Walhout, A.J.M. and M. Vidal.** 2001. Protein interaction maps for model organisms. Nat. Rev. Mol. Cell. Biol. *2*:55-62.
44. **Wojcik, J. and V. Schächter.** 2001. Protein-protein interaction map inference using interacting domain profile pairs. Bioinformatics *17*:S296-S305.
45. **Xenarios, I., D.W. Rice, L. Salwinski, M.K. Baron, E.M. Marcotte, and D. Eisenberg.** 2000. DIP: the database of interacting proteins. Nucleic Acids Res. *28*:289-291.

Note added in proofs: Two recently published articles should be cited:

1. **Gavin, A. et al.** 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*:141-147.
2. **Ho, Y. et al.** 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature *415*:180-183.

Address all correspondence to:

Vincent Schächter
Hybrigenics
3-5 impasse Reille
Paris 75014, France
e-mail: vschachter@hybrigenics.fr