Protein-interaction networks: from experiments to analysis

Vincent Schächter

Functional proteomics approaches aim to characterize comprehensively the function of gene products, and provide a first-level understanding of cellular mechanisms. Here, we review recent techniques for the construction and prediction of large-scale protein-interaction networks, with a particular emphasis on computational processing steps and comparative assessment of the reliability and completeness of the various approaches. We also discuss the use of protein-interaction network information in functional annotation and in the generation of higher-level biological hypotheses on pathways.

Vincent Schächter
Hybrigenics
3–5 impasse Reille
Paris 75014
France
tel: +33 1 58103808
fax: +33 1 58103849
e-mail: vschachter@hybrigenics.fr

▼ Large-scale assays enable the study of the function of proteins in a general context (rather than individually) through identification of physical interactions between proteins. Networks of interacting proteins thus build on the previously confined view of function, providing a first-level understanding of cellular mechanisms (see [1] for review).

Apart from two approaches published very recently that use mass spectrometry for systematic analysis of protein complexes [2,3], high-throughput techniques to construct protein-interaction maps are mostly derived from the yeast two-hybrid system [4]. Existing studies rely on one of two main categories of yeast two-hybrid techniques: the matrix approach and the fragment-library approach. These techniques differ considerably not only in scale but also by the nature and reliability of the results they yield.

Over the past two years, several computational methods have been proposed to generate critical masses of interaction data faster. These use sequence data from sets of genomes ('genomic context' methods) [5] or experimentally produced interaction data [6], to predict novel interactions.

In the first part of this article, we briefly review experimental approaches to protein-interaction map construction, with a particular emphasis on the coverage and reliability of each type of approach. Next, we describe predictive methods and techniques to validate their accuracy. Finally, we discuss possible exploitation of protein-interaction networks; from visual exploration of potential pathways to functional annotation and target discovery.

Construction of protein-interaction networks from experimental data

Although low-throughput technologies (e.g. co-immunoprecipitations, far-western blots and 'pull-downs'; see [7] for review) are commonly used for interaction studies on individual proteins, the study of interactions at the proteome level requires higher-throughput technologies.

High-throughput two-hybrid assays

The yeast two-hybrid system [4] can measure interactions between two known proteins or polypeptides and can also search for unknown partners (preys) of a given protein (bait) (see [8] for review). To date, two types of assay have been used for proteome-wide screening: the matrix approach and the library screening approach.

The so-called matrix approach relies on a collection of pre-defined open reading frames (ORFs), usually full-length proteins, as both bait and prey for interaction assays. This strategy is intrinsically limited to the testing of pre-defined proteins. Several studies have been published, ranging from the vaccinia virus [9] to the yeast proteome, using either the complete proteome [10–12] or only a subset, as a set of baits [13].

The second yeast two-hybrid assay strategy uses exhaustive libraries to screen for the identification of new protein-interacting partners. Screens can be carried out on a series of proteins involved in the same biochemical process,

1359-6446/02/$ – see front matter PII: S1359-6446(02)02281-X

enabling the identification of other previously uncharacterized proteins involved in the same pathway, or can be applied systematically on a whole-cell scale. A single screen typically tests for interactions of a bait protein with millions of protein fragments, randomly distributed on the genome. Upper-approximations of interacting domains can be computed from the results, giving an indication of selected overlapping prey fragments [14]. Fragment-library approaches have been applied to a variety of organisms, including T7 phage [15], yeast [16], HCV [17] and *Helicobacter pylori* [18].

The matrix and library strategies are shown in Fig. 1. Table 1 summarizes the results of major two-hybrid large-scale assays performed to date. For a more detailed comparative study of these technologies, see [19,20].

From raw experimental data to protein-interaction networks

Transforming raw experimental data into understandable, publishable and computationally tractable protein-interaction networks requires bioinformatics support for several key processing steps. From a computational biology perspective, a protein network can be visualized as a graph, with proteins (bait or prey) as vertices, and interactions (directed or undirected) as edges. Accumulation of experimental data results in incremental construction of the graph, with a new edge added to the graph after experimental identification of each interaction.

This step is trivial in the matrix approach, where the two partners are known beforehand. However, post-processing is required when a partner is screened against a fragment library: selected prey fragments must be sequenced and identified using homology search strategies against sequence databases that range from the very simple (e.g. identifying perfect similarity with a known prokaryotic gene) to the fairly complex (e.g. use of EST databases and alternative splicing models).

These computational steps must be integrated into a processing pipeline to enable high-throughput production of protein-interaction information. For example, the strategy used to construct the *Helicobacter pylori* interaction map [18] (Fig. 2) was supported by PIM Builder®, which is a dedicated, integrated laboratory production management system that implements the processing steps necessary to construct interaction maps from raw experimental data.

Coverage and reliability

Assessment and minimization of false-negatives and -positives is a key issue in interpreting the results of the high-throughput experimental technologies described previously.

False-negatives are biological interactions that are missed because of incorrect folding, inadequate subcellular localization, lack of specific post-translational modifications, and so on. Several factors increase the probability of matrix approaches generating high levels of false-negatives, most notably the fact that only two assays are performed for each pair of proteins (bait versus prey, and vice versa), and that the choice of selective conditions is standardized (see [19] for review).

By contrast, the fragment library approach simultaneously tests for millions of potential interactions with able two-hybrid candidates, and is therefore more likely to capture a fragment that exhibits the appropriate folding, should such a fragment exist [18].

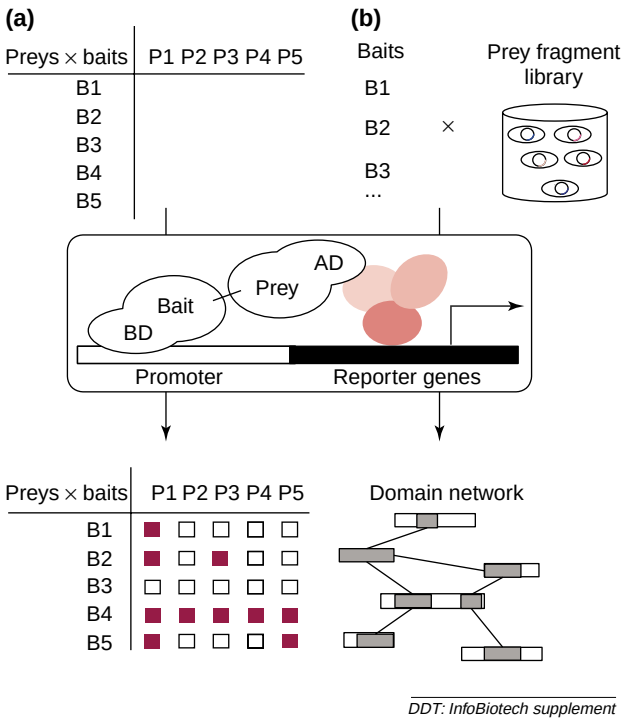

Figure 1. Yeast two-hybrid strategies. The central box shows the principle of the yeast two-hybrid assay: a protein domain that binds specifically to DNA sequences (BD) is fused to a 'bait' polypeptide, and a domain that recruits the transcription machinery (AD) is fused to a 'prey' polypeptide. Transcription of a reporter gene occurs if and only if the bait and the prey polypeptides interact. (a) The matrix approach uses the same collection of proteins as baits (B1–B5) and preys (P1–P5). Results can be represented in a matrix where bait autoactivators (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. (b) The fragment library screening approach identifies, for a given bait, a set of interacting prey partners, from which the interaction domains on prey proteins can be identified. Sticky prey proteins are identified as fragments of proteins that are often selected regardless of the bait protein. The statistical nature of the experimental procedure (that is, sets of screens against fragment libraries) enables the computation of a score for each interaction. This is an expected value (E-value), 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.

Table 1. Key figures in large-scale datasets for protein–protein interaction maps

Organism	Technology	Number of assays Baits × preys	No. of interactions	Refs
Vaccinia virus	Protein array	Proteome × proteome	37	[9]
Saccharomyces cerevisiae	Protein array	192 × proteome	281	[12]
	Pools of preys	Proteome × proteome	692	
S. cerevisiae	Pools of baits and preys	430 assays of pools (96 × 96)	175	[10]
S. cerevisiae	Pools of baits and preys	3844 assays of pools (96 × 96)	841[a]	[11]
S. cerevisiae	Protein array	162 × 162	213	[13]
Caenorhabditis elegans	Protein array	29 × 29	8	[35]
	Library screening	27 × proteome	124	
Hepatitis C virus	Protein array	10 × proteome	0	[17]
	Library screening	22 fragments × proteome	5	
S. cerevisiae	Library screening	15 × proteome	170	[16]
S. cerevisiae	Library screening	11 × proteome	113	[36]
Helicobacter pylori	Library screening	261 × proteome	1524	[18]

[a]This number corresponds to highly significant interactions (that is, more than three hits [11]).

Searching for many potential interactions increases the probability of selecting interacting polypeptides of no biological significance (that is, false-positives). Some bait proteins, termed auto-activators, might directly activate the transcription of reporter genes and thus appear to interact with a large random set of prey proteins. Similarly, certain chimeric prey fragments (sticky domains) might be non-specifically selected by many independent bait proteins. Fine-tuning of bait selectivity and quality control steps can be implemented to tackle toxicity and auto-activation issues [10,11,18]. In the fragment-library approach, sticky domains can be detected by the statistical nature of the experimental procedure (i.e. sets of screens against fragment libraries), enabling the computation of a scoring scheme that computes an expected value (E-value) for each interaction on the basis of results from multiple screens [18].

Computational prediction of protein-interaction networks

Although high-throughput assays yield increasingly large volumes of interaction information, data produced to date cover only a very small fraction of sequenced proteomes.

Over the past two years, increasing recognition that knowledge of protein–protein interactions is key to the understanding of protein function has motivated the design of predictive algorithms that generate hypotheses about interactions. These approaches typically attempt to predict functional links *ab initio*, using sequence data from completely sequenced genomes as input. Underlying algorithms are inspired by comparative genomics techniques (see [5] for a detailed review).

For example, the gene-fusion event (or 'Rosetta stone') method [21,22] is based on the following evolutionary interaction hypothesis: if two genes, x and y, exist separately in a given genome, but are fused as a single gene z in an ancestor genome, these two genes must be functionally linked. The gene neighborhood approach [23,24] rests on the more general hypothesis that one of the reasons for conservation of gene proximity and order is that the corresponding proteins interact. The phylogenetic profile method [25] assumes that genes with similar occurrence patterns of orthologs in a set of reference genomes must have co-evolved and therefore must be functionally linked.

As confirmed by comparisons with experimental data [26], each of these approaches is biased by the biological hypothesis behind the prediction. Another limitation is that the exact biological nature of the predicted functional links cannot be specified without additional information, that is, participation in the same structural complex, in the same biological pathway, in the same biological process, or, in some cases, existence of a physical interaction.

Aside from sequence-based approaches, a new method to predict a protein network in one organism using a large-scale reference protein network in another organism, was recently proposed [6]. It was designed to exploit fully the properties of the richest experimental interaction networks now available, namely the existence of domain information for each interaction, and large-scale maps that typically provide several instances of domains interacting with a given domain. The algorithm combines sequence similarity searches with clustering based on interaction patterns and interaction domain information, and reasons on an abstract interaction map connecting clusters of domains.

It is worth noting that the accuracy of prediction methods that rely on genome comparison is obviously highly dependent both on the criteria chosen for orthology and their relationship to the nature of the link to be predicted.

Predictions generated by each of these methods can, in theory, be validated by comparing them to accepted biological knowledge. These resources include interaction information stored in protein databases, keyword retrieval, and manual assessment of each interaction by an expert with access to the literature (see Box 1). One major issue with these techniques is the lack of sufficiently complete and reliable reference datasets, forcing the use of weaker indirect tests of predictions against other predictions.

Exploitation of protein-interaction networks

The first step towards fruitful use of protein networks is to store them in adequately structured databases, on top of which visual exploration tools can be implemented. Through additional computational analyses and integration with other data types, interaction information can then be used for functional annotation of proteins, leading to downstream applications such as target discovery.

Databases, visualization and exploration tools

Initially, interaction information available on the Internet was stored in protein-centric databases [27,28] as attributes of individual proteins. Then came lists of interactions, where an interaction is represented by its two protein partners, sometimes accompanied by basic annotations or cross-references to other databases [29]. To enable true exploitation, however, interaction-network information needs to be stored within dedicated databases: only data models designed to handle graph-like structures can search efficiently for immediate interaction partners (neighbors), paths between two proteins, or clusters of connected proteins [18,30]. The main protein–protein-interaction sources are listed in Table 2.

As highlighted previously, interpretation of interaction data requires a good understanding of the conditions of their creation; simple interaction lists are therefore clearly insufficient. Although some databases such as MIPS (Munich Information Centre for Protein Sequences; http://mips.gsf.de/mips/) [31] provide a brief indication of the experimental source (e.g. two-hybrid or co-immunoprecipitation), newer data-models are designed to store detailed information on the nature, laboratory and conditions of the experiment, in addition to quality assessment scores [18,30]. For experimentally generated networks, access to primary data is especially important to evaluate false-positives and reproducibility. Integrated tools such as PIM Rider® [18] provide such access (see Fig. 3), and offer several exploration functionalities to help the researcher during the discovery process.

As interaction data models grow more complex, encompassing more biological background information on proteins and interactions, and supporting integration with other types of experimental results, such as mRNA expression levels, constraints on their design get closer to those of general pathways-orientated

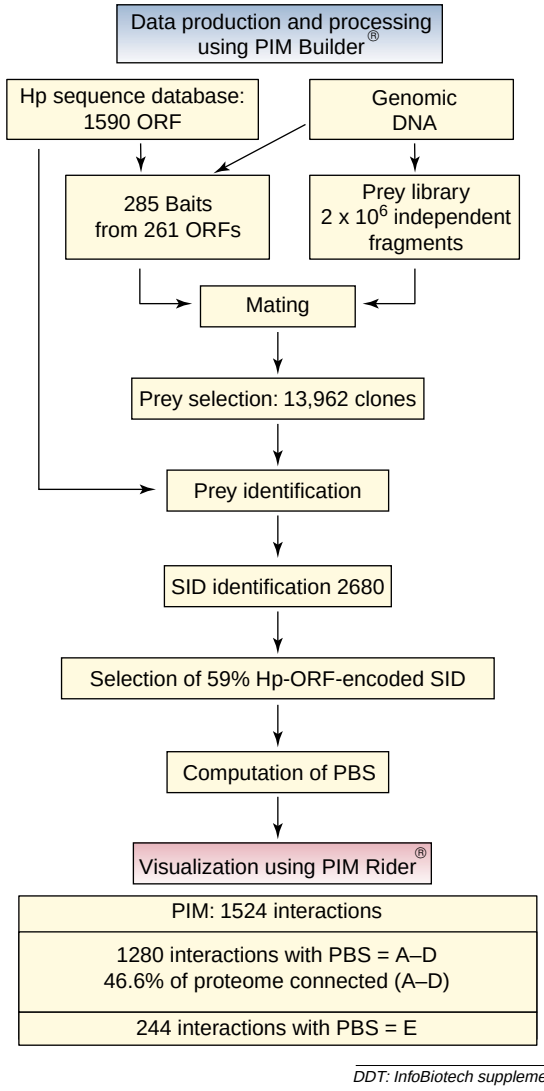

Figure 2. General outline of the strategy for building a *Helicobacter pylori* (Hp) proteome-wide interaction map. Each of the 285 bait proteins was screened against a library of *H. pylori* random genomic fragments. Bait constructs were specifically adapted for interaction screens, and the selective pressure was adjusted for each construct using results from a preliminary small-scale screening experiment. The PIM Builder® Laboratory Information System tracked all biotechnological or bioinformatics operations performed during the production processes and all interfaces with robots and external bioinformatics modules. It also implemented the processing steps necessary to construct interaction maps from raw experimental data, including the identification of interaction domains and scoring of interactions. 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 regions) were discarded. The interaction score was then computed, and interactions were grouped into categories A–D (from high to low heuristic values). The global connectivity of the protein-interaction map (PIM) was also analyzed to detect and tag highly connected prey polypeptides. Processing of data and visualization of interactions was performed by an in-house-built bioinformatics platform (the PIM Rider®).

Box 1. Approaches to validation of interaction-prediction methods

Protein–protein interactions can be evaluated by checking their existence in dedicated databases, such as MIPS (Munich Information Centre for Protein Sequences; http://mips.gsf.de/mips/) [a], DIP (Database of Interacting Proteins; http://dip.doe-mbi.ucla.edu/) [b], or OMIM (Online Mendelian Inheritance in Man; http://www3.ncbi.nlm.nih.gov/omim) [c]. However, the lack of of interaction information that can be directly extracted from public databases makes it necessary to resort to indirect methods such as the 'keyword retrieval' technique. For example, two proteins linked together in a network are compared by their associated set of functional keywords from a given database (e.g. SWISSPROT; http://www.expasy.ch/sprot/): the more similar the two sets of keywords, the higher the computed probability that the interaction exists [d–f].

Automated validation methods rely heavily on the quantity, quality, homogeneity and comparability of database annotations. Although they are adequate for confirming that a prediction algorithm performs better than random on a given space, they generally yield weak biological conclusions. These can be improved by cross-validation with predictions from other, ideally independent, experimental methods [d,e].

The approach that probably yields the most valuable biological conclusions, is the manual assessment of predicted interactions by an expert using information from public databases and the literature (J. Wojcik, *et al*. unpublished). Although obviously time-consuming, and leaving a wider margin for individual interpretation, it significantly increases the quantity of accessible reference knowledge. Emerging text-mining techniques could accelerate that process by pre-selecting relevant fragments, or even by automatically extracting relevant information from abstracts or full texts [g–i].

References

a Mewes, H.W. *et al.* (2000) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 28, 37–40

b Xenarios, I. *et al.* (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.* 28, 289–291

c Hamosh, A. *et al.* (2000) Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.* 15, 57–61

d Wojcik, J. and Schächter, V. (2001) Protein–protein interaction map inference using interacting domain profile pairs. *Bioinformatics* 17 (Suppl.), S296–S305

e Marcotte, E.M. *et al.* (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* 402, 83–86

f Jenssen, T.K. *et al.* (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 28, 21–28

g Marcotte, E.M. *et al.* (2001) Mining literature for protein–protein interactions. *Bioinformatics* 17, 359–363

h Blaschke, C. *et al.* (1999) Automatic extraction of biological information from scientific text: protein–protein interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 60–67

i Ono, T. *et al.* (2001) Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics* 17, 155–161

models. This is an active area of research in the computational biology community [32].

Functional annotations

The straightforward approach to functional annotation based interaction information, is to apply the 'guilt-by-association' principle: a protein is annotated using the lowest common denominator of the annotations of either its interacting partners, or, more generally, of proteins belonging to a given cluster of interactants [33,34].

For example, a set of yeast protein interactions was clustered according to cellular role and subcellular localization annotations [34] from the Yeast Proteome Database (http://www.incyte.com/sequence/proteome/databases/YPD.shtml) [28]. Functions were assigned to uncharacterized proteins based on annotations of their interacting partners, yielding assignments for 29 proteins that had two or more partners with at least one common function.

Guilt-by-association methods should be used carefully, however, as their predictions are very sensitive to the quality and coverage of both functional annotations and the protein-interaction network. False or vague annotations, and false-positives or false-negatives of the map could result in incorrect or missing predictions (for example, see [26]). Until quality and quantity improve, confidence in functional annotations can be increased by integrating conclusions based on interaction data with functional clues of a different nature.

Target discovery

Protein-interaction networks provide an excellent foundation on which a variety of target discovery strategies can be implemented [19]. Proteins that are found to interact with members of a known pathogenic process, and meet a more-or-less stringent set of drugability criteria, provide good target candidates. As do novel proteins discovered as preys in a fragment library two-hybrid assay. Generally speaking, interaction mapping can lead to expansion of the initial mechanism-of-action into a more complex set of interconnected pathways involved in pathogenicity, yielding a variety of new target candidates and enabling fine-tuning of specificity.

To turn target candidates selected through these strategies into highly reliable, validated targets, a carefully designed validation phase must follow the interaction-mapping phase.

Although the optimal design of the corresponding validation strategy is beyond the scope of this review, suffice it to say that validation platforms that are high-throughput and generic enough to match the interaction data production rate are yet to come. Until then, validation must be a careful mix of integrating knowledge from pre-existing heterogeneous sets of experimental results (*in silico* validation) and development of specific low-throughput functional assays.

Table 2. Main protein–protein interaction databases

Database	URL	Refs
BIND	http://www.bind.ca	[30]
Cellzome	http://www.yeast.cellzome.com	[2]
CuraGen portal	http://portal.curagen.com	[12]
DIP	http://dip.doe-mbi.ucla.edu	[29]
FlyNets	http://www.gifts.univ-mrs.fr/FlyNets/	[37]
Interact	http://www.bioinf.man.ac.uk/interact.shtml	[38]
MIPS	http://www.mips.biochem.mpg.de	[31]
PIM Rider®	http://pim.hybrigenics.fr	[18]
ProNet	http://pronet.doubletwist.com	

Concluding remarks

Large-scale protein-interaction networks are, together with gene expression profiles, among the first examples of datasets generated without previous knowledge of gene function. Bioinformatics tools enable high-throughput production of experimental results with: (1) quality control, (2) transformation of these results into protein networks, and (3) exploitation through visualization and analysis tools. Recent improvements in the throughput of experimental methods, as well as in the 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.

Although protein networks hold a wealth 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, which can be partly summarized by a confidence level. However, the fact that reference datasets are still almost non-existent should always be kept in mind.

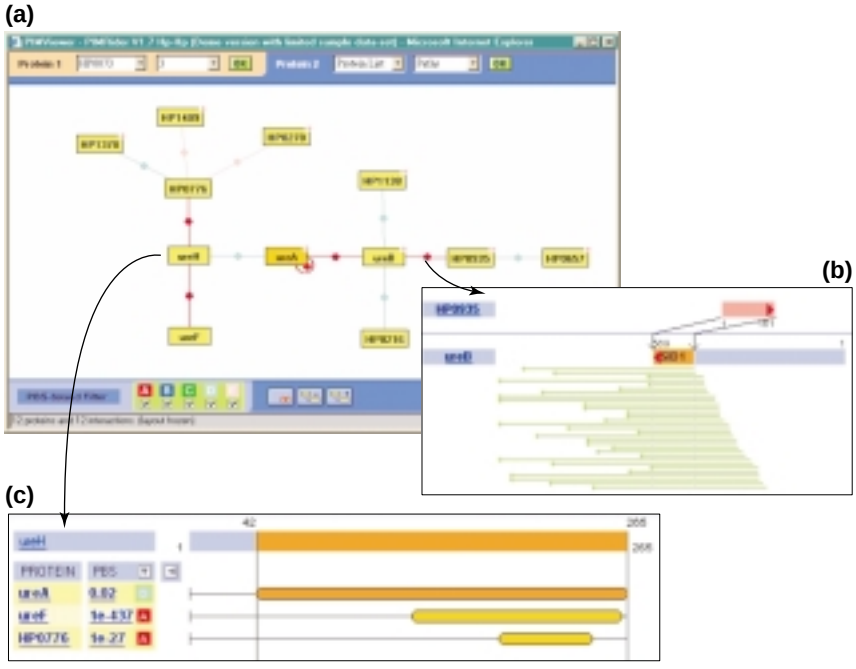

Figure 3. The PIM Rider® is an integrated exploration platform for protein-interaction networks. (a) The main window displays the protein-interaction map as a graph and enables the biologist to navigate through the network, filter interactions and proteins on criteria such as interaction reliability (PBS), and focus on a particular pathway. (b) Clicking on a specific interaction gives access to a summary of the supporting primary two-hybrid data, where interacting fragments and the computed selected interacting domain (SID) are positioned relative to the coding sequence of the two proteins. (c) The interacting domains of a given protein, and pathways between two proteins, can also be queried.

Ultimately, integration of functional clues from different experimental origins (e.g. yeast two-hybrid assays, microarrays and mass spectrometry) with sequence data and predictive methods structured by the appropriate knowledge management tools, should greatly enhance functional annotation. Indeed, combination with other sources of functional information can help transform protein networks into detailed descriptions of cellular pathways, enabling a shift in our view of function from 'the specific properties of a protein' to 'the role of that protein in one or several processes'. Such knowledge integration is a major challenge for computational biology. It requires the development of pathway models in a way that both bridges the gap between the different representations adapted to specific experimental data types, and enables *in silico*

hypothesis testing on incomplete information sets. In a more applied perspective, it should also enable actual interaction-based high-throughput target discovery by ensuring seamless progression of the discovery pipeline from interaction mapping to late-stage target validation.

Acknowledgements

I would like to thank Pierre Legrain for many stimulating discussions on protein-interaction network construction techniques and biological data interpretation and reliability issues. I would also like to thank Jean-Michel Gauthier for his input on the use of interaction networks for target discovery. Finally, I would like to acknowledge the unfailing support of Hybrigenics.

References

1 Walhout, A.J.M. and Vidal, M. (2001) Protein interaction maps for model organisms. *Nat. Rev. Mol. Cell Biol.* 2, 55–62

2 Gavin, A. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147

3 Ho, Y. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183

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

5 Eisenberg, D. *et al.* (2000) Protein function in the post-genomic era. *Nature* 405, 823–826

6 Wojcik, J. and Schächter, V. (2001) Protein–protein interaction map inference using interacting domain profile pairs. *Bioinformatics* 17 (Suppl.), S296–S305

7 Phizicky, E.M. and Fields, S. (1995) Protein–protein interactions: methods for detection and analysis. *Microbiol. Rev.* 59, 94–123

8 Vidal, M. and Legrain, P. (1999) Yeast forward and reverse 'n'-hybrid systems. *Nucleic Acids Res.* 27, 919–929

9 McCraith, S. *et al.* (2000) Genome-wide analysis of vaccinia virus protein–protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* 97, 4879–4884

10 Ito, T. *et al.* (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. U. S. A.* 97, 1143–1147

11 Ito, T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4569–4574

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

13 Newman, J.R. *et al.* (2000) A computationally directed screen identifying interacting coiled coils from *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* 97, 13203–13208

14 Siomi, M.C. *et al.* (1998) Functional conservation of the transportin nuclear import pathway in divergent organisms. *Mol. Cell Biol.* 18, 4141–4148

15 Bartel, P.L. *et al.* (1996) A protein linkage map of *Escherichia coli* bacteriophage T7. *Nat. Genet.* 12, 72–77

16 Fromont-Racine, M. *et al.* (1997) Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.* 16, 277–282

17 Flajolet, M. *et al.* (2000) A genomic approach of the hepatitis C virus generates a protein interaction map. *Gene* 242, 369–379

18 Rain, J.C. *et al.* (2001) The protein–protein interaction map of *Helicobacter pylori*. *Nature* 409, 211–216

19 Legrain, P. *et al.* (2001) Protein–protein interaction maps: a lead towards cellular functions. *Trends Genet.* 17, 346–352

20 Schachter, V. (2002) Construction and prediction of protein–protein interaction maps. In *Ernst Schering Research Foundation Workshop 38 (Bioinformatics and Genome Analysis)*, Springer Verlag

21 Enright, A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90

22 Marcotte, E.M. *et al.* (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science* 285, 751–753

23 Dandekar, T. *et al.* (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* 23, 324–328

24 Overbeek, R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2896–2901

25 Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4285–4288

26 Marcotte, E.M. *et al.* (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* 402, 83–86

27 Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45–48

28 Costanzo, M.C. *et al.* (2000) The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res.* 28, 73–76

29 Xenarios, I. *et al.* (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.* 28, 289–291

30 Bader, G.D. *et al.* (2001) BIND – the Biomolecular Interaction Network Database. *Nucleic Acids Res.* 29, 242–245

31 Mewes, H.W. *et al.* (2000) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 28, 37–40

32 Neumann, E.K. and Schächter, V. (2002) The informatics for making sense of the genome: a progress report from the biopathways consortium. *Comp. Funct. Genomics* (March)

33 Mayer, M.L. and Hieter, P. (2000) Protein networks – built by association. *Nat. Biotechnol.* 18, 1242–1243

34 Schwikowski, B. *et al.* (2000) A network of protein–protein interactions in yeast. *Nat. Biotechnol.* 18, 1257–1261

35 Walhout, A.J. *et al.* (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* 287, 116–122

36 Fromont-Racine, M. *et al.* (2000) Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. *Yeast* 17, 95–110

37 Sanchez, C. *et al.* (1999) Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an Internet database. *Nucleic Acids Res.* 27, 89–94

38 Eilbeck, K. *et al.* (1999) INTERACT: an object-oriented protein–protein interaction database. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 87–94