REVIEW

Experimental and bioinformatic approaches for interrogating protein–protein interactions to determine protein function

Arnaud Droit, Guy G Poirier and Joanna M Hunter

Health and Environment Unit, Faculty of Medicine, Eastern Quebec Proteomics Centre, 2705 Laurier Blvd, Ste-Foy, Quebec, Canada G1V 4G2

(Requests for offprints should be addressed to Guy G Poirier; Email: guy.poirier@crchul.ulaval.ca)

Abstract

An ambitious goal of proteomics is to elucidate the structure, interactions and functions of all proteins within cells and organisms. One strategy to determine protein function is to identify the protein–protein interactions. The increasing use of high-throughput and large-scale bioinformatics-based studies has generated a massive amount of data stored in a number of different databases. A challenge for bioinformatics is to explore this disparate data and to uncover biologically relevant interactions and pathways. In parallel, there is clearly a need for the development of approaches that can predict novel protein–protein interaction networks *in silico*. Here, we present an overview of different experimental and bioinformatic methods to elucidate protein–protein interactions.

Journal of Molecular Endocrinology (2005) **34**, 263–280

Introduction

Protein complexes are dynamic structures that assemble, store and transduce biological information. A major post-genomic scientific and technological pursuit is to describe the functions performed by the proteins encoded by the genome. Within the cell, proteins assemble into complexes and dynamic macromolecular structures. It is primarily as components of complexes that proteins perform cellular functions. These macromolecular structures engage in tasks critical for cell survival, such as regulation of metabolic pathways, and control DNA replication and progression through the cell cycle, as well as a myriad of minor but important functions.

The implementation of large high-throughput sequencing projects constitutes a revolution in our approach to human genomics. The human genome project has provided the sequence for approximately 30 000 genes. This number does not differ substantially from the number of genes of the nematode worm *Caenorhabditis elegans*, suggesting that genomic complexity may partly rely on the contextual combination of gene products. Moreover, protein splice variants and post-translational modifications complicate matters by transforming these human genes into millions of proteins. Fortunately, proteomics (the study of the protein complement of the genome) offers opportunities to understand protein expression and protein function. While the genome is fixed, the proteome is much more

dynamic. It changes during cellular development and in response to external stimuli. To fully understand the cellular machinery, simply identifying the proteins present is not enough. All of the interactions between them must also be delineated. The characterization of protein–protein interactions is essential to the understanding of the molecular role of the cell in the execution of various biological functions. These interactions form the basis of phenomena such as DNA replication and transcription, metabolic pathways, signaling pathways and cell cycle control. The association of more than two partners in a single complex introduces levels of complexity and regulation beyond binary interactions. The networks of protein interactions described recently (e.g. (Uetz *et al.* 2000, Ito *et al.* 2001, Gavin *et al.* 2002, Ho *et al.* 2002)) represent a higher level of proteome organization that goes beyond simple representations of protein networks. They represent a first draft of the molecular integration/regulation of the activities of cellular machineries.

The challenge of mapping protein interactions is vast, and many novel approaches have recently been developed for this task in the fields of molecular biology, proteomics and bioinformatics.

The charting of genome-scale protein–interaction maps is a first step forward addressing this challenge. A eukaryotic protein–protein interaction, or interactome, mapping effort has been initiated for *Saccharomyces cerevisiae*. However, many of the protein–protein

interactions that are relevant for understanding human biology, diseases and development only take place in multicellular organisms. Recently, the interactome maps of multicellular model organisms have emerged (Walhout *et al.* 2000, Giot *et al.* 2003).

The term proteomics was introduced in 1995 (Wasinger *et al.* 1995). This domain has seen a tremendous growth over the last nine years, as illustrated by the number of publications related to proteomics. The major goal of proteomics is to make an inventory of all proteins encoded in the genome and to analyze protein properties such as expression level, post-translational modifications and interactions. A number of recently described technologies have provided ways to approach these problems. The most common technologies used in proteomics today are two-dimensional sodium dodecylsulfate polyacrylamide gel electrophoresis (2D SDS-PAGE) for protein separation, mass spectrometry (MS) and protein identification through manual interpretation or database correlation of mass spectra. Integration of these steps is essential for a successful proteome experiment yet it relies on accurate knowledge of the parameters influencing each step. The improvement of these techniques has led to large-scale research in proteomics (Fleischmann *et al.* 1995, Anderson *et al.* 2000). It is now possible to identify a large fraction of proteins of a given proteome.

The final step in the characterization of proteins requires the application of bioinformatics tools to process existing experimental information. Bioinformatics tools provide sophisticated methods to answer the questions of biological interest. A number of different bioinformatics strategies have been proposed to predict protein–protein interactions. These include the use of information derived from reference maps of interacting domain profile pairs (Wojcik & Schachter 2001), conserved gene-pairs and correlated prokaryotic interacting gene products (Dandekar *et al.* 1998), clusters of orthologous proteins (Tatusov *et al.* 1997), phylogenetic profile (Pellegrini *et al.* 1999) or tree similarity (Pazos *et al.* 1997), gene fusion events (Marcotte *et al.* 1999), location within a functional cluster map (Schwikowski *et al.* 2000) and others.

Bioinformatics therefore has a critical role in the analysis of protein–protein interaction. Several databases that accumulate these data are currently available. These databases play an essential role in visualizing and integrating their own experimental data with the information about protein–protein interactions available in the Database of Interacting Proteins (DIP) (Xenarios *et al.* 2002), the General Repository for Interaction Datasets (GRID) (Breitkreutz *et al.* 2003a), the Biomolecular Interaction Network Database (BIND) (Figeys 2003) and the Human Protein Reference Database (HPRD) (Peri *et al.* 2004).

Figure 1 Representation of methods employing mass spectrometric identification of proteins.

The focus of this paper is to describe experimental and bioinformatics approaches for determining protein–protein interactions (Fig. 1). We also discuss the interpretation of protein–protein interaction information to elucidate protein function.

Experimental approaches

I: Molecular biology based methods

Traditionally, protein interactions have been studied by genetic, biochemical and biophysical techniques such as protein–protein affinity chromatography, immunoprecipitation, sedimentation and gel-filtration (Phizicky & Fields 1995). However, the speed with which new proteins are being discovered and predicted has created a need for high-throughput interaction detection methods. Consequently, in the last few years, methods based on molecular and cellular biology have been developed for the discovery of protein–protein interactions (Uetz *et al.* 2000, Ito *et al.* 2001, Gavin *et al.* 2002, Ho *et al.* 2002).

The best known method for screening binary protein–protein interactions is the yeast two-hybrid system, first introduced in 1989 (Fields & Song 1989, Ito *et al.* 2000). Two-hybrid procedures are typically carried out by screening a protein of interest against a random library of potential protein partners. Plasmid DNA is recovered from cells expressing interacting proteins, and the genes are identified by DNA sequencing. Thus, this approach provides information about the interaction between two proteins. It can also be used to identify unknown proteins. Additionally, it has been applied in high-throughput mode across the entire proteome of an organism to produce a comprehensive protein–protein interaction map (Uetz *et al.* 2000, Ito *et al.* 2001).

The yeast two-hybrid system is based on the fact that many eukaryotic transcription factors have discrete and separable DNA-binding and transcriptional activation domains. In this system, protein–protein interactions are tested by fusing one protein to the DNA-binding domain of the yeast GAL4 transcription factor, and the second protein to the GAL4 activation domain. The GAL4 transcription factor consists of separable domains responsible for DNA-binding and transcriptional activation (Keegan *et al.* 1986). Plasmids encoding two-hybrid proteins, one consisting of the GAL4 DNA-binding domain fused to protein X (protein of interest, called the bait) and the other consisting of the GAL4 activation domain fused to protein Y (often called the prey) are transfected into yeast. Interaction between proteins X and Y leads to the transcriptional activation of a reporter gene containing a binding site for GAL4. The interaction is detected by assaying for expression of a responsive reporter gene.

Although the yeast two-hybrid system is a powerful method, it has several limitations. First, it cannot detect interactions requiring three or more proteins or those depending on post-translational modifications (Ito *et al.* 2002). Also, variations in the state of protein–protein interactions under different cellular conditions (e.g. normal versus apoptotic cell) cannot be confirmed. Moreover, it is not applicable to studies of the kinetics of protein–protein interactions. Finally, a major disadvantage is the significant number of false positives (interaction that is not of biological origin) (Uetz 2002). The use of two or more reporter genes to assay for an interaction can diminish this number, but true interactions must generally be confirmed by an alternative method. One reason for the high number of false positives is that detected interactions could be unrelated to the physiological setting. This may also generate false negative results for proteins not native to the nucleus. The two-hybrid protein interactions take place in the nucleus, and so many proteins are not in their native compartment. For example, detection of the endogenous interactions of compartmentalized proteins such as cell surface receptors is not possible (Wehrman

et al. 2002). Finally, in some cases, the protein of interest may not be endogenous to yeast, making it possible to identify proteins that do not normally interact under physiological conditions.

High-throughput yeast two-hybrid methods have been used to detect protein interactions in yeast (Uetz *et al.* 2000, Ito *et al.* 2001). The array method was used by Uetz *et al.* to screen 192 yeast bait proteins against nearly all of the 6000 predicted yeast proteins. From this, 281 discrete interacting protein pairs from 87 bait proteins were found. A total of 5341 yeast bait proteins (87% of total yeast open reading frame (ORFs)) were screened by Ito *et al.* against the yeast prey library. In their study, 692 discrete interacting proteins pairs from 817 bait proteins were identified. A most interesting result of these large-scale studies is the fact that the vast majority of protein–protein interactions reveal novel potential interacting proteins.

To use these datasets to build robust and statistically significant protein–protein interaction networks, it will be essential to quantify the intrinsic error rates in these interactions. This will be challenging, because it is difficult to define a basis protein–protein interaction dataset with which to validate the experimentally identified interactions. Yet it is critical because in yeast, for example, there are a possible 18 million protein interactions (given 6000 proteins), of which only a small fraction is relevant biologically.

In an attempt to evaluate the quality of these interaction datasets, the results of these two large-scale studies of interacting proteins have been analyzed in detail (von Mering *et al.* 2002). The conclusion is that the interaction datasets contain a considerable number of false positives and are missing many true protein interactions. However, comparisons of interaction data are difficult because datasets are often derived under different conditions, come in different formats and need to be benchmarked against a trusted reference set. There are some plausible reasons for this small overlap. The methods may not have reached saturation, are known to produce a significant fraction of false positives and may have difficulty with certain types of interactions (Uetz *et al.* 2000, Ito *et al.* 2001).

Despite these drawbacks, the yeast two-hybrid system is established as a standard technique in molecular biology and serves as an appropriate method for proteomics analyses (Pandey & Mann 2000). High-throughput yeast two-hybrid screens have been performed in *Escherichia coli* (Bartel *et al.* 1996), hepatitis C virus (Flajolet *et al.* 2000), *Vaccinia* virus (McCraith *et al.* 2000), *S. cerevisiae* (Ito *et al.* 2000, Uetz *et al.* 2000), *Helicobacter pylori* (Rain *et al.* 2001) and *C. elegans* (Walhout *et al.* 2000). One benefit of the method is that it is independent of endogenous protein expression. In addition, this *in vivo* technique is often more sensitive than *in vitro* techniques, and thus may be more suited for the detection of weak or transient and unstable

interactions. Although yeast cells offer a convenient system for these types of interaction studies, the two-hybrid system has also been adapted to function in bacterial and mammalian cells (Figeys 2002). This system has undergone numerous modifications and has been adapted for the study of not only protein–protein interactions but also DNA–protein interactions and RNA–protein interactions (Causier & Davies 2002).

An important contribution of the two-hybrid method in the field of nuclear receptors is shown by numerous publications on steroid receptors and thyroid hormone receptors. It helped to determine the role of coactivators and corepressors in the mechanism of steroid/thyroid receptor action (Shibata *et al.* 1997). Giovannone *et al.* (2003) worked with Grb10, a protein that binds to the intracellular domain of activated tyrosine kinase receptors, including insulin-like growth factor (IGF-I) and insulin receptors. Using yeast two-hybrid screening, the N-terminus of Grb10 was shown to interact with two novel proteins designated GIGYF1 (Grb10 interacting GYF protein 1) and GIGYF2. The cloning of the vitamin D receptor partner steroid receptor coactivator-1 (SRC-1) was carried out by a two-hybrid study (Gill *et al.* 1998). Convinced by the usefulness of this technique in the field of signal transduction, Pirson *et al.* (1999) applied this two-hybrid system to isolate partners of new proteins implicated in signal transduction in the thyroid. Two proteins were isolated by differential screening as specifically regulated by the cAMP pathway, the other one is involved in the inositol-phosphates cascade.

Immunoprecipitation seems to be the simplest method to test interaction between two proteins. For example, the *in vitro* avid binding of glutathione (GTH) to glutathione S-transferase (GST) is used in such a way that one protein is fused with GST (carried out using recombinant techniques (Smith & Johnson 1988, Kaelin *et al.* 1991)), and the fused protein is bound, by way of GST, to GTH previously attached to sepharose beads. The second protein is labeled (often with sulfur 35). If the two proteins interact with each other, a complex forms on the sepharose beads. The complex of two proteins with GST can be eluted by competition with an excess of free GTH (Smith & Johnson 1988) or by boiling in the presence of SDS (Kaelin *et al.* 1991). It is then subjected to SDS electrophoresis (Porter *et al.* 1997). If, in the course of the preparation of the fused protein, a sequence is incorporated for a site-specific protease (e.g. thrombin), the GST-protein bound can be cleaved and the two-protein complexes isolated (Smith & Johnson 1988). When radioactive labeling is not used, the two proteins–GST complex can be detected with fluorography (Ing *et al.* 1992).

The pull-down assay has been used in a number of studies. Examples of interactions investigated are as follows: estrogen receptor (ER with and without activation by estradiol) with SRC-1 (Mak *et al.* 1999, Nishikawa *et al.*

1999, Tremblay *et al.* 1999), hormone-binding domain of ER with a potential coactivator SPT6 (Banahmad *et al.* 1995), ER- α with Transcription Intermediary Factor (TIF) proteins or with Gustatory Factor (GR) interact proteins 1, ER- α and ER- β with TRAP220 (NR-binding subunit of the mammalian mediator complex), progesterone receptor (PR) (activated by progesterone) with SRC-1, the orphan member of the steroid hormone receptor superfamily chicken ovalbumin upstream promoter-transcription factor (COUP-TFC) with Transcription Factor IIB (TFIIB) or ER.

Protein–protein interaction assays based on the fusion of N- and C-terminal to interacting proteins that regenerate enzymatic activity after dimerization are particularly well suited for monitoring inducible protein interactions. Wehrman *et al.* (2002) described a method to monitor protein–protein interactions in mammalian cells using the β -lactamase enzyme. In this approach, they utilized the complementary α and ω fragments of β -lactamase to build a protein interaction reporting assay. Fusion proteins are constructed with the α and ω fragments, and the interaction is tested by co-transfection in mammalian cells. If the α and ω domains are brought into close proximity, the activity of β -lactamase is recovered, allowing growth in ampicillin-containing media. When the two proteins cannot interact to form a complex, the formation of active β -lactamase is not favored (Fig. 2).

These systems have important advantages including low-level expression of the test proteins, generation of signal as a direct result of the interaction, and enzymatic amplification. Consequently, they are highly sensitive and physiologically relevant assays. Additionally, assays based on enzyme complementation can be performed in any cell type of interest or in diverse cellular compartments such as the nucleus, secretory vesicles or plasma membrane. Although several systems have been developed that use chimeras of proteins of interest and enzyme fragments to access protein interactions, each has its limitations (Wehrman *et al.* 2002). Clearly, the β -lactamase approach has great potential for determining protein interactions in mammalian cells.

Another approach called ubiquitin-based split-protein sensor (USPS) makes it possible to examine kinetic and equilibrium aspects of protein–protein interaction at its natural sites in a living cell (Johnsson & Varshavsky 1994). Ubiquitin (ub) is a small and conserved protein of 76 amino acids found in all eukaryotes. It is usually coupled to the N-terminus of proteins as a signal for their degradation by ubiquitin specific proteases (UBP). The enzymatic mechanisms of ub-protein conjugation and the role of ubiquitination in protein degradation have been largely established through studies using purified components from cell-free extracts (see Rechsteiner 1988).

Figure 2 Mapping protein interactions using the β -lactamase assay. (A) A ribbon diagram of β -lactamase. (B) A construct is made to create a fusion protein of a protein-bait and the ω fragment of β -lactamase. A second construct is designed to create a fusion protein between a protein-prey and the α fragment of β -lactamase. (C) If the bait protein binds to the prey protein, β -lactamase activity is observed.

The ub can be rationally dissected into two fragments and the fragments are fused to two test proteins that are thought to bind to each other. Folding of the ub from its fragments is catalyzed by the binding of the test proteins to each other and is detected with gene reporter. When a C-terminal fragment of ubiquitin is expressed as a fusion to a reporter protein, the fusion is cleaved only if an N-terminal fragment of ubiquitin is also expressed in the same cell. The reconstitution of the native ubiquitin from its fragment is recognized by a UBP, resulting in the cleavage of the attached protein. The cleavage can be visualized with a stable reporter protein attached to the C terminus of ubiquitin.

In 1948, Förster formulated the principle of Fluorescence Resonance Energy Transfer (FRET), a phenomenon that occurs when two different chromophores (donor and acceptor) with overlapping emission/absorption spectra are separated by a suitable orientation and a distance in the range 10–80 Å. Recently, however, the introduction of the green fluorescent protein (GFP) to FRET-based imaging microscopy gave new life to its use as a sensitive probe of protein-protein interactions and protein conformational changes *in vivo*. This was the beginning of real-time *in vivo* imaging of dynamic molecular events. Intermolecular and intramolecular FRET between two spectrally overlapping green fluorescent protein variants fused to two different host proteins offer a unique opportunity to monitor real-time protein-protein interactions or protein conformational changes. Intermolecular FRET can occur

between one molecule (protein A) fused to the donor and another molecule (protein B) fused to the acceptor. When the two proteins bind to each other, FRET produces a characteristic shift in the emission spectra (Truong & Ikura 2001). When they dissociate, FRET diminishes.

By using fluorescence digital imaging microscopy, one can visualize the location of GFPs within a living cell and follow the time course of the changes in FRET corresponding to cellular events at a millisecond time resolution. The observation of such dynamic molecular events *in vivo* provides vital insight into the action of biological molecules.

Limitations of FRET are the inter- and intramolecular spatial constraints and the sensitivity of detection, because there is no amplification of the signal.

Finally, another technology for monitoring protein-protein interaction is bioluminescence resonance energy-transfer (BRET), which relies on the same principles as FRET, except that the donor is a bioluminescence macromolecule (*Renilla* luciferase) that acquires luminescence when activated by a cofactor (Xu *et al.* 1999). Angers *et al.* (2000) demonstrated that BRET could be used as a tool to study constitutive protein-protein interactions *in vivo* between β 2 adrenergic receptor (β 2AR) fused to *Renilla* luciferase and another fused to yellow fluorescent protein (YFP). β 2AR is a G-protein-coupled receptor that forms a constitutive dimer. One disadvantage of BRET is its substantial overlap with the YFP emission, which contributes to a low signal-to-noise

ratio for the system (Xu *et al.* 1999). Therefore, BRET does not require an external excitation source, as each fluorophore is independent.

These two methods produce a significant number of false negatives, but they are useful because they are *in vivo* techniques.

We note that each of the techniques described above produces a unique distribution of interactions with respect to functional categories of interacting proteins. These differences in proteome coverage suggest that the various methods have specific strengths and weaknesses. The datasets based on purified complexes, for example, predict relatively few interactions for proteins involved in transport and signaling (possibly because they are enriched in transmembrane proteins, which are more difficult to purify). Similarly, interactions detected by the yeast two-hybrid technology largely fail to cover certain categories, such as proteins involved in translation (von Mering *et al.* 2002).

II: Mass spectrometry based methods

The high-throughput identification of proteins necessary for profiling experiments has become feasible because of advancements in mass spectrometry (MS) as well as bioinformatics (Fenyo 2000, Pandey & Mann 2000, Figeys 2003). Although protein interaction studies in yeast are fairly routine, the limiting factor in functional proteomics studies of human protein–protein interactions has been the number of cells required for the analysis. Fortunately, over the last several years, the sensitivity of mass spectrometry has continuously improved, thus significantly reducing the requirement for cellular material. Indeed, it is now possible to study protein–protein interactions by MS using a single dish of human cell lines (Figеys 2002).

Briefly, MS-based approaches generally consist of the selective purification and enrichment of a bait protein and its interactors from a cell lysate. The isolated proteins are digested into peptides using a protease such as trypsin. The peptide mixtures are then analyzed by mass spectrometry, and protein interactors are identified by database searching. Immunological methods such as immunoprecipitations (*in vivo*) or purifications are commonly used to isolate protein complexes. These methods often utilize tagged target proteins. A variety of expression vectors with different tag sequences have been designed for fusion to almost any target protein that can be cloned and expressed in a microbial host. A cDNA clone coding for a tagged bait protein is engineered, cells are transfected with the clone, and the expressed protein and its interacting protein partners are purified using the tag present on the bait. The properties of the additional tag provide a handle for purification of the fusion protein by affinity chromatography. The size of these tags can range from only one or a few amino acids to complete proteins

and can be attached to either the N- or the C-terminus of the desired protein or, in some applications, to both termini. The biological activity of the fusion proteins depends on the location and on the amino acid composition of the tag (Bucher *et al.* 2002).

Two common affinity tags are GST and FLAG. Immunoprecipitation approaches using antibodies against these tags (Einhauser & Jungbauer 2001) are often used for the purification/enrichment of protein complexes. These affinity tag systems share the following features: (a) one step adsorption purification; (b) a minimal effect on tertiary structure and biological activity; (c) simple and accurate assay of recombinant protein during purification; (d) applicability to a number of different proteins. Furthermore, the protein of interest may be expressed in a relevant cell line. Unfortunately, because of the need for transfection, this technique is not useful for mapping interactions in tissues.

Rigaut *et al.* (1999) have described a technique that might more efficiently isolate low abundance complexes. A fusion cassette encoding calmodulin-binding protein (CBP), a TEV cleavage site, and ProtA was constructed and named the tandem affinity purification (TAP) tag. The cDNA of a bait protein is cloned with this cassette followed by the introduction into the host cell or organism. The fusion protein and associated components are recovered from cell extracts by affinity selection on an IgG matrix. After washing, the TEV protease is added to release the bound proteins. In a second affinity purification step, the eluate is incubated with calmodulin-coated beads in the presence of calcium. The two purification steps help to reduce the number of non-specifically bound proteins. However, weakly bound or transient interactions may not be recovered.

Simple generic affinity purification methodologies have increasingly been applied for the purposes of large-scale proteomic studies, particularly in yeast. However, these often use reagents (e.g. antibodies) of variable affinities that increase background and/or intermediate steps (e.g. purification or affinity tag removal) that restrict their simple application in more complex protein sources such as mammalian cells. de Boer *et al.* (2003) presented an attractive approach for protein complex purification based on the very high affinity of avidin/streptavidin for biotinylated templates. They coexpressed bacterial BirA biotin ligase and hematopoietic transcription factor tagged by an N-terminal fusion of small artificial peptide. These tags have been successfully biotinylated in bacterial, yeast, insect and mammalian cells (Cronan 1990, Cronan & Reed 2000, Parrott & Barry 2000, 2001). Biotinylation can occur either by endogenous protein-biotin ligases or through the coexpression of exogenous biotin ligase, in most cases that of the bacterial BirA enzyme. The advantage of this technique is that tagged proteins can be efficiently and specifically biotinylated in mammalian

Figure 3 Experimental approach to determine protein–protein interactions. (A) Sequence and structure of the TAP tag. The various domains constituting the tag are indicated. (B) The purification protein assemblies are separated by denaturing gel electrophoresis. The separated proteins are then digested by trypsin, and the resulting peptides are analyzed by mass spectrometry. (C) Using bioinformatics methods, interacting proteins can be characterized.

cells and transgenic mice and can be efficiently purified in a single step by binding to streptavidin beads.

Large scale projects

Two recent large-scale analyses of protein complex composition in *S. cerevisiae* by Gavin *et al.* (2002) and Ho *et al.* (2002) have generated an unprecedented amount of protein interaction information. Similar strategies were adopted in both studies. Tagged proteins were used to capture complexes whose protein components are subsequently identified using mass spectrometry (Fig. 3).

In the study by Gavin *et al.* (2002), bait proteins were generated by inserting, using homologous recombination, a gene-specific cassette containing a TAP tag at the 3' end of the target genes. In this study, 1739 genes were tagged, and 1167 genes were expressed in yeast. The interacting proteins were purified by the high-affinity (TAP) purification method of Rigaut *et al.* (1999) and then separated by gel electrophoresis. At this stage, only 589 of the 1739 attempted baits were successfully

purified. Of these, roughly 130 baits did not provide interactors. In total, 1440 yeast proteins were identified, representing approximately 25% of the yeast genome. Simultaneously, Ho *et al.* (2002) used recombinant based cloning to add a FLAG epitope tag to 725 yeast genes. These clones were then transfected in yeast. Once expressed, the baits and their interactors were purified and separated by gel electrophoresis. Interactions were discovered for roughly 70% of the clones for a total of 1578 different interacting proteins, again representing approximately 25% of the yeast genome.

Thus, in spite of the differences in tagging methods, the two studies reached a similar conclusion. A wide majority of the targets were found to be associated with other proteins, confirming that most proteins exist as complexes rather than as monomeric entities. Although large in scope, these projects provide only a partial analysis of the yeast protein–protein interaction space, together providing information on 2283 yeast proteins covering about 25% of the yeast proteome.

Although a large number of protein interactions were identified in these two studies, these strategies have some

drawbacks. Because the tagged protein is overexpressed, its stoichiometry is not the same as that of its partners in the complex, and therefore it must compete for complex assembly with the nontagged version of the protein encoded in the genome. In addition, the single step purification may be insufficient for recovering low abundance complexes above the background of the overexpressed protein. The more stringent two-step TAP purification utilized by Gavin *et al.* permits the recovery of complexes present at levels as low as 15 copies per cell (Puig *et al.* 2001, Gavin *et al.* 2002). However, the method of homologous recombination cloning is not readily compatible with mapping interactions in human proteins. Conversely, the method used by Ho *et al.* (2002) is readily applicable to the mapping of protein–protein interactions in human.

There are clear advantages and disadvantages to each of the TAP and FLAG approaches, as noted in a recent study by von Mering *et al.* (2002). The advantages are that the interactions are formed in relevant cell lines and that the tag is small and has limited interference with the protein function or its localization (Figeys 2003). The principal disadvantage is a high false positive rate, particularly as the immunoprecipitations were only performed once for the high-throughput studies in yeast (Gavin *et al.* 2002, Ho *et al.* 2002). Differentiating non-specific binding could be an issue when limited repetitions of experiments are performed. Also, the increased level of expression of the bait protein could increase the occurrence of non-specific interactions. Furthermore, these approaches might miss some complexes that are not formed under the given conditions, either because tagging may disturb complex formation, or because loosely associated components may be washed off during purification.

Targeted approaches

The last section described large scale studies for detecting protein interactions, but these same techniques have been applied in more targeted studies.

The application of the TAP approach in mammalian cells has not been fully explored (Bauer & Kuster 2003). In yeast, it is relatively easy, using homologous recombination, to replace the gene encoding the endogenous protein by a recombinant tagged version. The expression levels of the tagged protein often track the endogenous protein. However, this has recently been accomplished in cultured mammalian cells via retroviral gene transfer (Knuesel *et al.* 2003). The TAP tag retroviral expression vector was fused to the N-terminus of the human SMAD3 and SMAD4 to allow stable expression of the TAP-tagged proteins at close to the physiological levels in mammalian cells. Retroviral gene transfer allows stable expression of exogenous proteins in a variety of mammalian cell lines within two days of

infection. This publication demonstrates the power of combining epitope tagging into a bicistronic retroviral vector with mass spectrometry to identify proteins that form complexes with endogenous components. The authors have identified HSP70 as a specific interacting protein of SMAD3 *in vivo*. It is conceivable that rapid purification of tagged exogenous proteins from cultured mammalian cells will facilitate analysis and identification of expressed proteins in various physiological states.

There are very few publications relating to the analysis of endocrine-related protein–protein interactions with mass spectrometry. However, this number is expected to rise as endocrinologists utilize these techniques more frequently.

Ranish *et al.* (2003) have published a method for the mapping of protein–protein interactions that minimizes the number of false positives. In this study, stable isotope tagging with isotope-coded affinity tag (ICAT) reagents and mass spectrometry were used to determine the specific composition and changes in the relative abundance of components of protein complexes. Proteins isolated from the two specific purifications were labeled with either heavy or normal stable isotope tags (Gygi *et al.* 1999, Cagney & Emili 2002, Zhou *et al.* 2002). The commercially available ICAT reagent utilizes a reactive group that is specific for cysteine thiol residues in proteins and peptides. After proteolysis, sample complexity is reduced in three sequential chromatographic steps. The samples are then analyzed by nano-liquid chromatography mass spectrometry/mass spectrometry (nano-LC-MS/MS), where the peptide pairs labeled with heavy and light tag are quantified by measuring their peak ratios as they co-elute from a reversed phase HPLC column into the mass spectrometer. The resulting MS/MS spectra are used to search protein databases using SEQUEST (Yates *et al.* 1995a,b, Perkins *et al.* 1999) to sequence the peptides and thus identify the proteins from which they originated. The relative quantification can be used to distinguish specific complex components from co-purifying proteins or to detect changes in the abundance and composition of complexes isolated from cells in different states.

This quantitative proteomics strategy provides a way of reliably distinguishing specific complex components from co-purifying proteins, even when the specific components are more than 20 times less abundant than the co-purifying proteins. Thus, the ICAT technology offers an alternative approach for analysis of complexes isolated by simple one-step affinity purifications. Protein losses resulting from multiple purification steps are avoided, and the potential to identify weakly associated proteins is increased. There are, however, some limitations to this isotopic labeling approach. The ICAT method requires proteins to contain cysteine residues flanked by appropriately spaced protease cleavage sites. Because cysteine residues are relatively rare in proteins, sequence coverage is quite poor, and some proteins may not be identified. In

addition, the ICAT method is not well suited to the detection of protein post-translational modifications or protein isoforms generated by alternative mRNA splicing (Patton *et al.* 2002).

Comparison of methods

Deng *et al.* (2003) compared the data from all the large-scale yeast interaction screens. They developed a maximum likelihood estimation method to assess the reliability of the interaction data, and found that the Uetz data were more reliable than the Ito data, and that the Gavin data were more reliable than the Ho data. In addition, they suggested that the MS-based analysis of protein complexes performed better in function predictions than the two-hybrid data, thus validating the theory that each component of a complex can be assigned a function based on that of the whole complex.

Protein complex formation is seen as more than the sum of its binary interactions (Gavin *et al.* 2002). Because the two-hybrid system is suited to the characterization of binary interactions, it may not be adequate for the comprehensive analysis of protein complexes. In contrast, MS-based approaches allow for the isolation of large protein complexes and for the detection of networks of protein interactions. However, MS-based approaches are biased towards highly abundant, stable complexes, whereas the two-hybrid system is particularly useful for the detection of weak or transient interactions.

It is clear that yeast two-hybrid and MS-based techniques have both independently made significant impacts on our understanding of the interactome. However, both techniques have many associated problems that, if used alone, limit the information provided.

Known macromolecular complexes provide a defined and objective set of protein interactions with which to compare biochemical and genetic data validation. Edwards *et al.* (2002) show that a significant fraction of the individual interactions reported in the literature are inconsistent with the known 3D structures of three complexes (RNA polymerase II, Arp2/3 and the proteasome).

The proteasome and Arp2/3 subunits were also analyzed as part of several genome-wide yeast two-hybrid screens. In the first, carried out by Uetz *et al.* 2002 and colleagues, five interactions involving proteasome subunits and other proteins were uncovered, but not even one interaction between two known proteasomal subunits was found. This dataset also did not contain any interaction between Arp2/3 subunits. Ito *et al.* 1997 identified 30 interactions between proteasome subunits and non-proteasome proteins, but only one intra-complex interaction. Their screen correctly identified one interaction between Arp2/3 subunits.

For the three complexes they studied, the *in vivo* pull down method was quite successful in identifying subunits that interact within the complex. In the case of RNA polymerase, two of ten polymerase subunits were tagged and the co-purifying proteins identified. Half of the subunits known to interact directly with these subunits were detected, for a false-negative rate of 50%. When a subset of the Arp2/3 and the proteasome subunits were tagged, there were no false negatives; all interactions present in the crystal structure were found in the datasets.

von Meering *et al.* (2002) demonstrated that each large-scale interaction study produces a unique distribution of interaction with respect to functional categories of interacting proteins. Our view of the interactome can only be enhanced when data from yeast two-hybrid and MS approaches, in combination with other data, are integrated.

III: Protein microarrays

A more direct approach for the global identification of biochemical activities uses functional protein microarrays. In this technique, sets of proteins or an entire proteome are overexpressed, purified, distributed in an addressable array format, and then assayed.

The major limitation in the production of functional protein microarrays has been the preparation of proteins to analyze. This requires high-quality and comprehensive expression libraries and array production methods that yield a large number of functionally active proteins. The basic technologies necessary for producing this protein content (i.e. cDNA cloning, PCR, recombinant protein expression and purification) have been in place for a number of years. However, various efforts to integrate these technologies into automated, high-throughput systems for the industrial-scale production of proteins have only recently got underway. These efforts range from the *C. elegans* ORFeome project to the various public efforts to generate human gene collections (Vaglio *et al.* 2003). One such publicly funded collection is the Unigene set, which has been generated from a human fetal brain cDNA library (Bussow *et al.* 2000). Other collections include the Full-length Expression (FLEX) Gene Repository (Briazuela *et al.* 2001, 2002), the Integrated Molecular Analysis of Genomes and their Expression (IMAGE) cDNA collection and the associated sequence verified Mammalian Gene Collection (MGC) Full-length Clone Collection (Lennon *et al.* 1996).

The most common substrates for functional protein microarrays are glass microscope slides, as these are compatible with many commercial scanners. Proteins are attached to the surface via either direct covalent methods, linkers, or affinity tags. The bound proteins are then assayed for binding or enzymatic activities. MacBeath and Schreiber (2000) used this format to demonstrate that they could detect antibody-antigen

interactions, protein kinase activities, and protein interactions with small molecules.

Protein microarrays are very useful since the experimental conditions can be well controlled. For example, different cofactors or inhibitors can be introduced in the binding assays to adjust the stringency of the binding activities. Another advantage of this approach is that these highly parallel assays are not biased towards abundant proteins. In addition, with proper detection methods, proteome chips can be used to identify the downstream targets of various enzymes such as kinases, phosphatases, methyl transferases, and proteases.

Zhu *et al.* 2001 have provided a conclusive demonstration that proteome microarrays can be used to screen for protein–protein interactions (Zhu *et al.* 2001). For example, 39 calmodulin-interacting proteins were identified by adding biotinylated calmodulin to a chip printed with the majority of the yeast proteome. Within this set, several expected calmodulin-binding proteins as well as 33 novel calmodulin-interacting proteins were revealed. In a single experiment, a new consensus binding site for calmodulin was defined.

Several of the interactions with calmodulin were missed in the large-scale yeast two-hybrid and/or the affinity purification-MS studies, including the well-known interaction of calmodulin with calmodulin kinase.

IV: Bioinformatics methods

The development of experimental techniques for detecting protein–protein interactions has generated an extensive amount of data from which biological and biomedical research will greatly benefit. This has created the dilemma of how to effectively utilize the vast amount of information gathered through these large-scale studies. Thus, there is clearly a need to develop systematic bioinformatics approaches that can determine protein function and define how these macromolecules interact within complex networks.

These bioinformatics methods include well known and novel approaches: data mining, annotation by sequence similarity, phylogenetic profiling, metabolic pathway mapping, gene neighbor and domain name fusion analyses.

Development of protein–protein interaction databases

Building an accurate and complete cellular map, tantamount to a dynamic high dimensional information matrix, will require the integration of many layers of systematic cellular and molecular biology experimental efforts. This will require powerful information storage, query capacity and analysis engines to effectively manipulate data.

Until recently, most of the biomolecular interaction and pathway data were stored in printed journal

articles where the information is difficult to manage and compute upon. Therefore, various groups have attempted to compile databases of protein–protein interaction data. Data repositories such as BIND (Bader *et al.* 2003), DIP (Xenarios *et al.* 2002), GRID (Breitkreutz *et al.* 2003a), SGD (Christie *et al.* 2004) and HPRD (Peri *et al.* 2004) aim to integrate the diverse body of experimental knowledge about interacting proteins into a single, easily accessible database. The primary goal of these databases is to extract and integrate the wealth of information about protein–protein interactions accessible in many different scientific journals and in archives such as MEDLINE (National Library of Medicine, MD, USA). These databases also offer tools to visualize networks of interactions, to map pathways across taxonomic branches, and to generate information for kinetic simulations.

However, practically all medium- to large-scale projects develop their own systems for the storage, representation and analysis of protein interaction data. In addition to the duplication of work, this results in a high degree of incompatibility between different protein interaction datasets. Consequently, the need for common formats to allow data exchange between both public and commercial database systems has been recognized, as has a growing need for the establishment of public data repositories in which published data can be deposited and retrieved by scientists working in the field and wishing to further analyze this information. The Proteomics Standards Initiative (PSI) of the Human Proteome Organization (HUPO) are developing a standard data format for the representation of protein interaction data (Orchard *et al.* 2004).

Building and maintaining databases require a substantial amount of effort. Thus, creating a database large enough to capture cell map information will require massive community investment and commitment and innovation, ranging from the individual researcher (biologists to computer scientists and database developers) to the funding agencies and journals. To date, the majority of protein–protein interactions reported in many databases are for *S. cerevisiae*, for which the most comprehensive protein–protein interaction datasets are available. The reliability of the data appears to be limited by the high number of false-positive interactions (Legrain *et al.* 2001), which complicate the identification of biologically relevant interactions. There are, however, compilations of data from other species. Suzuki *et al.* (2003) described the development of a mammalian protein–protein interaction (PPI) database. In this database are stored the mammalian PPIs identified through their own PPI assays (represented as internal PPIs), as well as those extracted and processed from publicly available data sources, such as the DIP and BIND databases and MEDLINE abstracts (represented as external PPIs). The internal and external PPIs

have been integrated into the PPI database, which is linked to the main FANTOM2 (Kasukawa *et al.* 2003) viewer. The PPI viewer can be very useful for identifying interactions of significant biological interest in the network of interactions. Using the PPI viewer, the author found a novel interaction partner of TRAF2, a key adaptor molecule involved in tumor necrosis factor (TNF)-induced signaling pathway. The author has verified the interaction experimentally and designated the partner as T2 BP (TRAF2 binding protein (Kanamori *et al.* 2002)). Another example of the utility of the PPI database and viewer is single stranded DNA-binding protein 2 (SSBP2). Kasukawa *et al.* (2003) identified in their PPI assay 14 interaction partners of these proteins. But, when they modified parameters (luciferase reporter activity), they found that this protein interacts with four proteins containing LIM domains, three of which are LIM heterodomain transcription factors; the remaining one is LIM domain only (LMO) protein LMO4. The SSBP2 is suggested to possess tumor-suppressor activity through gene dosage or other epigenetic mechanisms (Castro *et al.* 2002). In addition to the evidence that deregulated expression of the LMO family proteins LMO1, LMO2 and LMO4 are associated with oncogenesis, several reports suggest that other LIM domain proteins are also involved in leukemogenesis (Wu *et al.* 1996, Rabbitts 1998, Visvader *et al.* 2001, Kawamata *et al.* 2002). These results suggest that interaction of the LIM proteins with SSBP2 plays an important role in the molecular mechanism of tumor-suppressor activity of SSBP2.

Another resource freely available to academic institutions is the Human Protein Reference Database (HPRD), which integrates information relevant to the functions of human proteins in health and disease (Peri *et al.* 2003). Protein information such as interactions, post-translational modifications, enzyme/substrate relationships, disease associations, tissue expression, and subcellular localization has been incorporated into this database. Information has been extracted from the literature by biologists who have read and interpreted >300 000 published articles to date. The HPRD continues to grow at a rapid pace and aims to become a comprehensive source of information for human proteins. At the time of writing, HPRD comprises 8346 entries representing pairs of proteins known to mutually bind.

In the field of molecular endocrinology, there is no specific database for protein–protein interaction. However, Duarte *et al.* (2002) have developed Nurebase (<http://www.ens-lyon.fr/LBMC/laudet/nurebase/nurebase.html>), a database containing protein and DNA sequences, reviewed protein alignments and phylogenies, taxonomy and annotations for all nuclear receptors. The importance of nuclear receptors has prompted the accumulation of rapidly increasing data from a great diversity of fields of research: sequences, expression

patterns, three dimensional structures, protein–protein interactions. The aim of this database is not to develop a specific protein–protein interaction database but to present an integrated database with a unique, interactive interface, centralizing up-to-date information about nuclear receptors for the specialist and non-specialist.

In conclusion, it is essential to develop and maintain protein–protein interaction databases. These databases constitute a platform for data mining and visualization of protein interaction networks. Also, it is important to develop databases that focus on other organisms, such as mammals, because large-scale analysis will soon produce a huge number of protein–protein interactions.

Literature mining for protein–protein interactions

Mining the literature for information is essential for transforming discoveries reported in the literature into a database. Several automated text mining tools have been developed. These methods identify journal articles containing information about biomolecular interactions and confirm sentences that mention specific protein–protein interactions. The most popular methods for document categorization are based on Natural Language Processing (NLP), Naïve Bayes (Marcotte *et al.* 2001), decision trees (Wilcox & Hripcsak 1999), neural networks, nearest neighbor and support vector machine (SVM). Due to the complexity and variety of the English language, the Natural Language Processing approaches are inherently difficult and therefore less applicable (Eisenberg *et al.* 2000). Moreover, several methods for the extraction of interaction information and other biological relationships from the literature have recently been described (Humphreys *et al.* 2000, Proux *et al.* 2000, Rindfleisch *et al.* 2000, Stapley & Benoit 2000, Thomas *et al.* 2000, Friedman *et al.* 2001, Nagashima *et al.* 2003).

PreBind (Donaldson *et al.* 2003) is an approach complementary to BIND for finding interaction data in the over 14 million PubMed abstracts. PreBind functions as follows: (1) SVM technology is used to identify articles about biomolecular interactions and confirm sentences that mention specific protein–protein interactions; (2) protein names and their gene symbols are derived from a non-redundant sequence database (RefSeq) and from the *S. cerevisiae* Genome Database (SGD); (3) this information extraction system is coupled to a human-reviewed data-entry queue for a publicly available BIND.

The PreBind parser collects synonyms for proteins present in the NCBI RefSeq sequence database and the *S. cerevisiae* genome database and stores them in the PreBind database. The PubMed literature database is then queried for each of these protein names, and the PubMed identifiers for abstracts returned by these searches are also stored in the PreBind database. These abstract identifiers are retrieved and assigned a score that describes the relative likelihood that the article

contains molecular interaction information. This step is accomplished using textomy, or text anatomy (<http://www.litminer.ca/>), a text processing software that uses SVM to capture the statistical pattern of words used in papers that have previously been presented to the machine as ‘papers of interest’. Textomy is employed in a second round to score the likelihood that an interaction is described for any given pair-wise combination of proteins mentioned in an abstract. These potential interactions are stored in the PreBind database. Once the user has reviewed the potential interactions, they could subsequently submit them to the BIND database.

Computational methods

For predicting functional associations (including direct binding), the current growth in completed genomes offers unique opportunities through the use of so-called genomic context or non-homology based inference methods. These methods are based on the fact that functionally associated proteins are encoded by genes that share similar selection pressures. The genes need to be maintained together, and regulated together, such that the encoded proteins can interact at the same time and place in the cell.

Marcotte *et al.* (1999) have investigated computationally whether protein–protein interactions can be recognized from genome sequences. Genes whose protein products need to interact closely in the cell have a noticeable tendency to be fused into a single gene, encoding a combined polypeptide in which the proteins have a higher chance of interacting productively. For example, their analyses showed that two proteins (Gyr A and Gyr B in *E. coli*) are fused into a single chain in another organism (topoisomerase II in yeast: Rosetta Stone protein). In this way, the domain fusion analysis makes two distinct predictions. It predicts potential protein–protein interactions, and it predicts protein pairs that have related biological functions (e.g. proteins that participate in a common structural complex, metabolic pathway or biological process). Furthermore, predicted potential protein–protein interactions may be constructed into putative pathways, as shown in Fig. 4. Pathways B and D are constructed from the proteins in pathways A and C with connections predicted by the domain fusion method.

The genetic requirement for maintaining functionally associated genes together becomes apparent when they are conserved phylogenetically. The genes tend to be either present together or absent together. That is, they have the same phylogenetic profiles. This is the computational approach described by Pellegrini *et al.* (1999) and Huynen and Bork (1998). For example, phylogenetic profiles show that if two proteins have homologs in the same subset of fully sequenced organisms, they are likely to be functionally linked. The authors describe a new tool for identifying the complex

Figure 4 Reconstruction of two metabolic pathways in *E. coli*. Pathways A and C are the known pathways for biosynthesis of shikimate and purine. Pathways B and D are constructed from the proteins in pathways A and C with connections predicted by the domain fusion method to interact with other proteins of the pathway. Taken from Marcotte *et al.* (1999) and reproduced with permission from *Science*.

or pathway in which a protein participates. This type of analysis is useful because functionally linked proteins may have no amino acid sequence similarity with each other, and therefore they cannot be linked by conventional sequence alignment techniques.

Dandekar *et al.* (1998) described an approach based on conserved gene pairs and correlated prokaryotic interacting gene products. The need for similar regulation is often reflected in a tendency for functionally associated genes to be close neighbors in prokaryotic genomes, where they generally have the same transcriptional orientations and little or no sequence similarity. This study analyzed nine completely sequenced genomes and demonstrated that a conserved gene order could be correlated with physical interaction between the encoded proteins. The comparison provides additional strong support for this concept.

These computational methods have some limitations. Because investigators concentrate on different organisms, or reporting is confined to partially hypothesized interaction results, it is difficult to compare the predictive power of these various computational methods in an objective manner.

Bock and Gough (2001) reported a method to expand the range of prediction to whole proteomes using computational statistical learning theory. In contrast to the above-cited investigations, the methodology reported by Bock and Gough takes an entirely different approach to computational prediction of protein interactions. Given a database of known protein–protein interaction pairs, a machine learning system was trained to

recognize interactions based solely on primary structure and associated physiological properties. The prediction methodology generates a binary decision about potential protein-protein interactions. This suggests the possibility of proceeding directly from the automated identification of a cell's gene products to inference of the protein interaction pairs, facilitating protein function and cellular signaling pathway identification.

Although much effort has been put into methods for identifying interacting partners, there has been a limited focus on comparing these interactions with those known from three-dimensional (3D) structures. Decades of x-ray crystallography have produced hundreds of structures for protein complexes, and these structures provide a rich source of data for learning principles of how proteins interact and for validating interactions determined by other methods. Although the number of complexes of known 3D structures is relatively small, it is possible to expand this set by considering homologous proteins. Aloy & Russell (2002) have used 3D complex structures to model putative interactions in order to assess the structural compatibility of proposed protein-protein interactions. The method can correctly predict interactions within several systems. Given a known 3D complex structure and homologous sequences for each interacting protein, the method can rank the likelihood of all the possible interactions between the homologs of the same species.

A recent publication by Tien *et al.* (2004) described a systematic approach that can predict potential targets *in silico* to aid in understanding how complex biological systems work. The authors combined experimental large-scale interaction data, data mining and bioinformatics methods. The goal was to create a virtual protein-protein interaction map for Aurora family kinases. This involved a combination of extensive literature searches, database analyses of various yeast protein-protein interactions, homology searches and analysis of expression databases. The potential interactors were subjected to laboratory approaches to validate the predicted biochemical interactions.

Finally, databases that integrate different computational methods to predict protein interactions have been developed. For example, the STRING (von Mering *et al.* 2003) and POINT (Tien *et al.* 2004) databases can predict interactions between proteins. Although the STRING database can predict interactions, this database focuses on identifying the neighboring genes in the genomic context and does not include any experimental protein-protein interactions. The POINT database provides novel insights into protein-protein interaction networks in combination with publicly accessible microarray databases.

V: Interactome

The use of network visualization and modeling tools such as Cytoscape (Fig. 5) (<http://www.cytoscape.org>), Os-

prey (Breitkreutz *et al.* 2003b) and Biolayout (Enright & Ouzounis 2001) is critical for understanding data relationships. Network visualization can be very useful for identifying interactions of significant biological interest. Typically, interaction networks are viewed within a graphic application that represents genes as nodes and interactions as edges between nodes. Such a viewer has been developed by Suzuki *et al.* (2003). Their visualization tool is useful because it can incorporate interactions from various resources into a single window. Information about protein-protein interactions beyond the protein of interest may be viewed in the interaction network. Proteins of interest can be searched by either accession numbers or identifiers or keywords. Using their PPI viewer, they found a novel interaction partner of TRAF2, a key adaptor molecule involved in the TNF-induced signaling pathway. Several additional interactions whose biological significance was successfully suggested through the use of the PPI viewer are also identified. Indeed, these visualization tools must be developed as the interactive entry point to the integrated cell map, where a gene of interest connects directly to the latest information about that gene and its relationships.

VI: Bioinformatics applications

The determination of protein/peptide sequences is a basic requirement for biomedical research, including protein-protein interactions. It is absolutely essential for the characterization and identification of proteins and peptides. The investigator will want to find out as much as possible about specific proteins or peptides. The first step in this process is to look for similarities with already discovered peptide sequences/proteins. This is accomplished by comparing the novel sequence with those contained in biological databases.

Biological databases

Biological databases are archives of consistent data that are stored in a uniform format and an efficient manner. These databases contain data from a broad spectrum of molecular biology areas. Primary or archived databases contain information and annotation of DNA and protein sequences, DNA and protein structures, and DNA and protein expression profiles.

Specialized databases for specific subjects have been set-up: for example the non redundant (nr) database from NCBI (National Center for Biotechnology Information: <http://www.ncbi.nlm.nih.gov>), EMBL database from EMBL (European Molecular Biology Laboratory: <http://www.embl.org>), UniProt/Swiss-Prot protein database (<http://www.uniprot.org>) and PDB (<http://www.rcsb.org/pdb/>) a 3D protein structure database. There are also specialized databases for protein-protein interactions such as: BIND (<http://www.bind.ca/>),

Figure 5: Screenshot from Cytoscape (website: <http://www.cytoscape.org>) showing a portion of an interaction network.

DIP (Database Interacting Proteins: <http://dip.docmbl.ucla.edu/>), HPRD (<http://www.hprd.org/>) and Nurebase for molecular endocrinology (<http://www.ens-lyon.fr/LBMC/laudet/nurebase/nurebase.html>)

Scientists also need to be able to integrate the information obtained from the underlying heterogeneous databases in a sensible manner in order to be able to get a clear overview of their biological subject. Entrez from NCBI (<http://www.ncbi.nlm.nih.gov/Entrez/index.html>) or SRS (Sequence Retrieval System: <http://srs.ebi.ac.uk>) from EBI are powerful, querying tools providing links to information from more than 150 heterogeneous resources.

For the construction of protein–protein interaction networks, identification of the protein–protein interactions may use the keyword search functions provided at BIND, HPRD or DIP. Proteins without assigned names or without characterized functions as listed at genome (<http://www.stanford.edu>) are excluded. Protein annotation is also provided at the above website and in Gene Ontology (www.geneontology.org), permit-

ting selection of proteins that fit the inclusion criteria (mitochondrial, nucleus, spindle pole ..).

Another tool which is very useful to find protein–protein interactions in the literature is PreBIND. This data mining tool locates biomolecular interaction information in the scientific literature (<http://www.blueprint.org/products/prebind/>).

Once all the biological data is stored and is easily available to the scientific community, the requirement is then to provide methods for extracting meaningful information from the data. Bioinformatics tools are software that are designed to carry out this analysis step.

There exists different databases that integrate different computational methods to predict protein interactions. For example, STRING (<http://string.embl.de/>) is a database of predicted functional associations among genes/proteins.

Finding protein interaction homologs for different model organisms will be accomplished by similarity searching tools such as BLAST (<http://www.ncbi.nlm.nih.gov/BLAST/>). This tool can be used

to identify similarities between novel query sequences of unknown structure and function and database sequences whose structure and function have been elucidated. For refining models about protein–protein interaction, the BLAST search provided by BIND (BINDBlast. <http://www.bind.ca/BINDBlast>) could be used.

After putative protein–protein interaction homologs have been found using the above tools, the protein sequences may be translated from their sequences in the GenBank database and aligned using the CLUSTALW program (<http://www.ebi.ac.uk/clustalw/>).

To determine the protein function, many groups of programs enable protein sequences to be compared with a secondary protein database that contains information on motifs, signatures and protein domains. For example, Pfam (<http://www.sanger.ac.uk/Software/Pfam/>) is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families, Prodom (<http://protein.toulouse.inra.fr/prodom/current/html/home.php>) and PROSITE (http://www.expasy.org/prosite/prosite_details.html) are databases of protein families and domains. Highly significant hits against these different pattern databases allow the approximation of the biochemical function of your query protein.

Conclusions

The methods presented here provide a toolbox for the elucidation of protein interactions. In an organism with tens of thousands of genes, it is difficult to establish comprehensive protein–protein interaction data sets, because the total number of experiments is estimated to be far larger than for budding yeast. It will be necessary to analyze the experimental protein–protein interaction data together with publicly available information on mammalian protein–protein interaction. The diverse nature of the interactions discovered thus far illustrates that the methods presented here are highly complementary, and all are needed to cover the diverse protein synergies in the cell.

The mapping of protein interactions will be the key to better understanding of human protein function and diseases. Novel experimental and bioinformatics tools have accelerated the deciphering of protein interactions. In addition, emerging technologies such as protein microarrays, live cell microarrays, and RNAi hold the promise of systematically studying the entire human proteome. As high-throughput functional genomics and proteomics technology and bioinformatics develop concurrently, these technologies will become more accessible to the individual laboratory. Researchers will thus be empowered to ask increasingly more interesting and complex biological questions. Mapping protein complexes combined with other biological and genomic information will provide the

framework of a physical map of the cell that can be filled in with ever-increasing details to encompass metabolic and signaling pathways. Given the advantages provided by an *in silico* approach, it seems reasonable to propose that it will become an essential tool for initially evaluating novel hypotheses and will offer an improved rationale for target prioritization. The goal is that only the most promising targets will be subjected to empirical testing.

A collection of protein interaction database web sites

BIND, Biomolecular Interaction Network Database: <http://www.bind.ca>

DIP, Database of Interacting Proteins: <http://dip.doe-mbi.ucla.edu/>

PIM – Hybrigenics: <http://pim.hybrigenics.com/pimridext/common/index.jsp>

PathCalling Yeast Interaction Database: <http://portal.curagen.com/extpc/com.curagen.portal.servlet.Yeast>

MINT, a Molecular Interaction Database: <http://mint.bio.uniroma2.it/mint/>

GRID, The General Repository for Interaction Datasets: <http://biodata.mshri.on.ca/grid/servlet/Index>

InterPreTS, Protein interaction prediction through tertiary structure: <http://www.russell.embl.de/interprets/>

STRING, predicted functional associations among gene/proteins: <http://string.embl.de/>

Mammalian protein–protein interaction database (PPI): <http://fantom21.gsc.riken.go.jp/PPI/>

InterDom, Database of putative interacting protein domains: <http://interdom.lit.org.sg/>

FusionDB, database of bacterial and archaeal gene fusion events: <http://igs-server.cnrs-mrs.fr/FusionDB/>

IntAct Project: <http://www.ebi.ac.uk/intact/index.html>

The Human Protein Interaction Database: <http://wilab.inha.ac.kr/hpid/>

ADVICE, Automated Detection and Validation of Interaction by Co-Evolution: <http://advice.i2r.a-star.edu.sg/>

InterWeaver, protein interaction reports with online evidence: <http://interweaver.i2r.a-star.edu.sg/>

PathBLAST, alignment of protein interaction networks: <http://www.pathblast.org/bioc/pathblast/blastpathway.jsp>

ClusPro, a fully automated algorithm for protein–protein docking: <http://nrc.bu.edu/cluster/>

HPRD, Human Protein Reference Database: <http://www.hprd.org/>

Funding

This research was supported by the Canadian Institutes of Health Research. The authors declare that there is no

conflict of interest that would prejudice the impartiality of this scientific work.

References

- Aloy P & Russell RB 2002 Interrogating protein interaction networks through structural biology. *PNAS* **99** 5896–5901.
- Anderson NL, Matheson AD & Steiner S 2000 Proteomics: applications in basic and applied biology. *Current Opinion in Biotechnology* **11** 408–412.
- Angers S, Salahpour A, Joly E, Hilairt S, Chelsky D, Dennis M & Bouvier M 2000 Detection of beta 2-adrenergic receptor dimerization in living cells using bioluminescence resonance energy transfer (BRET). *PNAS* **97** 3684–3689.
- Bader GD, Betel D & Hogue CW 2003 BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research* **31** 248–250.
- Baniahmad C, Nawaz Z, Baniahmad A, Gleeson MA, Tsai MJ & O'Malley BW 1995 Enhancement of human estrogen receptor activity by SPT6: a potential coactivator. *Molecular Endocrinology* **9** 34–43.
- Bartel PL, Roecklein JA, SenGupta D & Fields S 1996 A protein linkage map of *Escherichia coli* bacteriophage T7. *Nature Genetics* **12** 72–77.
- Bauer A & Kuster B 2003 Affinity purification-mass spectrometry. Powerful tools for the characterization of protein complexes. *European Journal of Biochemistry* **270** 570–578.
- Bock JR & Gough DA 2001 Predicting protein–protein interactions from primary structure. *Bioinformatics* **17** 455–460.
- de Boer E, Rodriguez P, Bonte E, Krijgsvelde J, Katsantoni E, Heck A, Grosveld F & Strouboulis J 2003 Efficient biotinylation and single-step purification of tagged transcription factors in mammalian cells and transgenic mice. *PNAS* **100** 7480–7485.
- Breitkreutz BJ, Stark C & Tyers M 2003a The GRID: the General Repository for Interaction Datasets. *Genome Biology* **4** R23.
- Breitkreutz BJ, Stark C & Tyers M 2003b Osprey: a network visualization system. *Genome Biology* **4** R22.
- Brizuela L, Braun P & LaBaer J 2001 FLEXGene repository: from sequenced genomes to gene repositories for high-throughput functional biology and proteomics. *Molecular and Biochemical Parasitology* **118** 155–165.
- Brizuela L, Richardson A, Marsischky G & Labaer J 2002 The FLEXGene repository: exploiting the fruits of the genome projects by creating a needed resource to face the challenges of the post-genomic era. *Archives of Medical Research* **33** 318–324.
- Bucher MH, Evdokimov AG & Waugh DS 2002 Differential effects of short affinity tags on the crystallization of *Pyrococcus furiosus* maltodextrin-binding protein. *Acta Crystallographica Section D Biological Crystallography* **58** 392–397.
- Bussow K, Nordhoff E, Lubbert C, Lehrach H & Walter G 2000 A human cDNA library for high-throughput protein expression screening. *Genomics* **65** 1–8.
- Cagney G & Emili A 2002 *De novo* peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. *Nature Biotechnology* **20** 163–170.
- Castro P, Liang H, Liang JC & Nagarajan L 2002 A novel, evolutionarily conserved gene family with putative sequence-specific single-stranded DNA-binding activity. *Genomics* **80** 78–85.
- Causier B & Davies B 2002 Analysing protein–protein interactions with the yeast two-hybrid system. *Plant Molecular Biology* **50** 855–870.
- Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE *et al.* 2004 *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Research* **32** (Database issue) D311–D314.
- Cronan JE Jr 1990 Biotinylation of proteins *in vivo*. A post-translational modification to label, purify, and study proteins. *Journal of Biological Chemistry* **265** 10327–10333.
- Cronan JE Jr & Reed KE 2000 Biotinylation of proteins *in vivo*: a useful posttranslational modification for protein analysis. *Methods in Enzymology* **326** 440–458.
- Dandekar T, Snel B, Huynen M & Bork P 1998 Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences* **23** 324–328.
- Deng M, Sun F & Chen T 2003 Assessment of the reliability of protein–protein interactions and protein function prediction. *Pacific Symposium Biocomputing* 140–151.
- Donaldson I, Martin J, De Bruijn B, Wolting C, Lay V, Tuckam B, Zhang S, Baskin B, Bader GD, Michalickova K *et al.* 2003 PreBIND and Textomy-mining the biomedical literature for protein–protein interactions using a support vector machine. *BMC Bioinformatics* **4** 11.
- Duarte J, Perriere G, Laudet V & Robinson-Rechavi M 2002 NUREBASE: database of nuclear hormone receptors. *Nucleic Acids Research* **30** 364–368.
- Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J & Gerstein M 2002 Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends in Genetics* **18** 529–536.
- Einhauer A & Jungbauer A 2001 The FLAG peptide, a versatile fusion tag for the purification of recombinant proteins. *Journal of Biochemical and Biophysical Methods* **49** 455–465.
- Eisenberg D, Marcotte EM, Xenarios I & Yeates TO 2000 Protein function in the post-genomic era. *Nature* **405** 823–826.
- Enright AJ & Ouzounis CA 2001 BioLayout - an automatic graph layout algorithm for similarity visualization. *Bioinformatics* **17** 853–854.
- Fenyó D 2000 Identifying the proteome: software tools. *Current Opinion in Biotechnology* **11** 391–395.
- Fields S & Song O 1989 A novel genetic system to detect protein–protein interactions. *Nature* **340** 245–246.
- Figues D 2002 Functional proteomics: mapping protein–protein interactions and pathways. *Current Opinion in Molecular Therapeutics* **4** 210–215.
- Figues D 2003 Novel approaches to map protein interactions. *Current Opinion in Biotechnology* **14** 119–125.
- Flajolet M, Rotondo G, Daviet L, Bergametti F, Inchauspe G, Tiollais P, Transy C & Legrain P 2000 A genomic approach of the hepatitis C virus generates a protein interaction map. *Gene* **242** 369–379.
- Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al.* 1995 Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269** 496–512.
- Förster T 1948 Intramolecular energy migration and fluorescence. *Annales de Physique* **2** 55–75.
- Friedman C, Kra P, Yu H, Krauthammer M & Rzhetsky A 2001 GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* **17** (Suppl 1) S74–S82.
- Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM *et al.* 2002 Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415** 141–147.
- Gill RK, Atkins LM, Hollis BW & Bell NH 1998 Mapping the domains of the interaction of the vitamin D receptor and steroid receptor coactivator-1. *Molecular Endocrinology* **12** 57–65.
- Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E *et al.* 2003 A protein interaction map of *Drosophila melanogaster*. *Science* **302** 1727–1736.
- Giovannone B, Lee E, Laviola L, Giorgino F, Cleveland KA & Smith RJ 2003 Two novel proteins that are linked to insulin-like growth factor (IGF-I) receptors by the Grb10 adaptor and

- modulate IGF-I signaling. *Journal of Biological Chemistry* **278** 31564–31573.
- Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH & Aebersold R 1999 Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology* **17** 994–999.
- Ho Y, Grubler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K *et al.* 2002 Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415** 180–183.
- Humphreys K, Demetriou G & Gaizauskas R 2000 Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. *Pacific Symposium on Biocomputing* 505–516.
- Huynen MA & Bork P 1998 Measuring genome evolution. *PNAS* **95** 5849–5856.
- Ing NH, Beekman JM, Tsai SY, Tsai MJ & O'Malley BW 1992 Members of the steroid hormone receptor superfamily interact with TFIIB (S300-II). *Journal of Biological Chemistry* **267** 17617–17623.
- Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S & Sakaki Y 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. *PNAS* **97** 1143–1147.
- Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M & Sakaki Y 2001 A comprehensive two-hybrid analysis to explore the yeast protein interactome. *PNAS* **98** 4569–4574.
- Ito T, Ota K, Kubota H, Yamaguchi Y, Chiba T, Sakuraba K & Yoshida M 2002 Roles for the two-hybrid system in exploration of the yeast protein interactome. *Molecular Cell Proteomics* **1** 561–566.
- Johnsson N & Varshavsky A 1994 Split ubiquitin as a sensor of protein interactions *in vivo*. *PNAS* **91** 10340–10344.
- Kaelin WG Jr, Pallas DC, DeCaprio JA, Kaye FJ & Livingston DM 1991 Identification of cellular proteins that can interact specifically with the T/E1A-binding region of the retinoblastoma gene product. *Cell* **64** 521–532.
- Kanamori M, Suzuki H, Saito R, Muramatsu M & Hayashizaki Y 2002 T2 BP, a novel TRAF2 binding protein, can activate NF-kappaB and AP-1 without TNF stimulation. *Biochemical and Biophysical Research Communications* **290** 1108–1113.
- Kasukawa T, Furuno M, Nikaido I, Bono H, Hume DA, Bult C, Hill DP, Baldarelli R, Gough J, Kanapin A *et al.* 2003 Development and evaluation of an automated annotation pipeline and cDNA annotation system. *Genome Research* **13** 1542–1551.
- Kawamata N, Sakajiri S, Sugimoto KJ, Isobe Y, Kobayashi H & Oshimi K 2002 A novel chromosomal translocation t(1;14)(q25;q32) in pre-B acute lymphoblastic leukemia involves the LIM homeodomain protein gene, Lhx4. *Oncogene* **21** 4983–4991.
- Keegan L, Gill G & Ptashne M 1986 Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein. *Science* **231** 699–704.
- Knuesel MP, Wan Y, Xiao Z, Holinger EP, Lowe N, Wang W & Liu X 2003 Identification of novel protein–protein interactions using a versatile mammalian TAP expression system. *Molecular Cell Proteomics* **2** 1225–1233.
- Legrain P, Wojcik J & Gauthier JM 2001 Protein–protein interaction maps: a lead towards cellular functions. *Trends in Genetics* **17** 346–352.
- Lennon G, Auffray C, Polymeropoulos M & Soares MB 1996 The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* **33** 151–152.
- MacBeath G & Schreiber SL 2000 Printing proteins as microarrays for high-throughput function determination. *Science* **289** 1760–1763.
- McCraith S, Holtzman T, Moss B & Fields S 2000 Genome-wide analysis of *Vaccinia* virus protein–protein interactions. *PNAS* **97** 4879–4884.
- Mak HY, Hoare S, Henttu PM & Parker MG 1999 Molecular determinants of the estrogen receptor-coactivator interface. *Molecular and Cellular Biology* **19** 3895–3903.
- Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO & Eisenberg D 1999 Detecting protein function and protein–protein interactions from genome sequences. *Science* **285** 751–753.
- Marcotte EM, Xenarios I & Eisenberg D 2001 Mining literature for protein–protein interactions. *Bioinformatics* **17** 359–363.
- von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S & Bork P 2002 Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417** 399–403.
- von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P & Snel B 2003 STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research* **31** 258–261.
- Nagashima T, Silva DG, Petrovsky N, Socha LA, Suzuki H, Saito R, Kasukawa T, Kurochkin IV, Konagaya A & Schonbach C 2003 Inferring higher functional information for RIKEN mouse full-length cDNA clones with FACTS. *Genome Research* **13** 1520–1533.
- Nishikawa J, Saito K, Goto J, Dakeyama F, Matsuo M & Nishihara T 1999 New screening methods for chemicals with hormonal activities using interaction of nuclear hormone receptor with coactivator. *Toxicology and Applied Pharmacology* **154** 76–83.
- Orchard S, Hermjakob H, Julian RK Jr, Runte K, Sherman D, Wojcik J, Zhu W & Apweiler R 2004 Common interchange standards for proteomics data: public availability of tools and schema. *Proteomics* **4** 490–491.
- Pandey A & Mann M 2000 Proteomics to study genes and genomes. *Nature* **405** 837–846.
- Parrott MB & Barry MA 2000 Metabolic biotinylation of recombinant proteins in mammalian cells and in mice. *Molecular Therapeutics* **1** 96–104.
- Parrott MB & Barry MA 2001 Metabolic biotinylation of secreted and cell surface proteins from mammalian cells. *Biochemical and Biophysical Research Communications* **281** 993–1000.
- Patton WF, Schulenberg B & Steinberg TH 2002 Two-dimensional gel electrophoresis; better than a poke in the ICAT? *Current Opinion in Biotechnology* **13** 321–328.
- Pazos F, Helmer-Citterich M, Ausiello G & Valencia A 1997 Correlated mutations contain information about protein–protein interaction. *Journal of Molecular Biology* **271** 511–523.
- Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D & Yeates TO 1999 Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *PNAS* **96** 4285–4288.
- Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjana V, Muthusamy B, Gandhi TK, Gronborg M *et al.* 2003 Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research* **13** 2363–2371.
- Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TK, Chandrika KN, Deshpande N, Suresh S *et al.* 2004 Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Research* **32** (Database issue) D497–D501.
- Perkins DN, Pappin DJ, Creasy DM & Cottrell JS 1999 Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20** 3551–3567.
- Phizicky EM & Fields S 1995 Protein–protein interactions: methods for detection and analysis. *Microbiology Reviews* **59** 94–123.
- Pirson I, Jacobs C, Vandenbroere I, El Housni H, Dumont JE & Perez-Morga D 1999 Use of two-hybrid methodology for

- identifying proteins of interest in endocrinology. *Molecular and Cellular Endocrinology* **151** 137–141.
- Porter W, Saville B, Hoivik D & Safe S 1997 Functional synergy between the transcription factor Sp1 and the estrogen receptor. *Molecular Endocrinology* **11** 1569–1580.
- Proux D, Rechenmann F & Julliard L 2000 A pragmatic information extraction strategy for gathering data on genetic interactions. *Proceedings International Conference Intelligent Systems for Molecular Biology* **8** 279–285.
- Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M & Seraphin B 2001 The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **24** 218–229.
- Rabbitts TH 1998 LMO T-cell translocation oncogenes typify genes activated by chromosomal translocations that alter transcription and developmental processes. *Genes and Development* **12** 2651–2657.
- Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V *et al.* 2001 The protein–protein interaction map of *Helicobacter pylori*. *Nature* **409** 211–215.
- Ranish JA, Yi EC, Leslie DM, Purvine SO, Goodlett DR, Eng J & Aebersold R 2003 The study of macromolecular complexes by quantitative proteomics. *Nature Genetics* **33** 349–355.
- Rechsteiner M 1988 Regulation of enzyme levels by proteolysis: the role of pest regions. *Advances in Enzyme Regulation* **27** 135–151.
- Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M & Seraphin B 1999 A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology* **17** 1030–1032.
- Rindfleisch TC, Tanabe L, Weinstein JN & Hunter L 2000 EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pacific Symposium on Biocomputing* 517–528.
- Schwikowski B, Uetz P & Fields S 2000 A network of protein–protein interactions in yeast. *Nature Biotechnology* **18** 1257–1261.
- Shibata H, Spencer TE, Onate SA, Jenster G, Tsai SY, Tsai MJ & O'Malley BW 1997 Role of co-activators and co-repressors in the mechanism of steroid/thyroid receptor action. *Recent Progress in Hormone Research* **52** 141–164; discussion 164–145.
- Smith DB & Johnson KS 1988 Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. *Gene* **67** 31–40.
- Stapley BJ & Benoit G 2000 Bibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pacific Symposium on Biocomputing* 529–540.
- Suzuki H, Saito R, Kanamori M, Kai C, Schonbach C, Nagashima T, Hosaka J & Hayashizaki Y 2003 The mammalian protein–protein interaction database and its viewing system that is linked to the main FANTOM2 viewer. *Genome Research* **13** 1534–1541.
- Tatusov RL, Koonin EV & Lipman DJ 1997 A genomic perspective on protein families. *Science* **278** 631–637.
- Thomas J, Milward D, Ouzounis C, Pulman S & Carroll M 2000 Automatic extraction of protein interactions from scientific abstracts. *Pacific Symposium on Biocomputing* 541–552.
- Tien AC, Lin MH, Su LJ, Hong YR, Cheng TS, Lee YC, Lin WJ, Still IH & Huang CY 2004 Identification of the substrates and interaction proteins of aurora kinases from a protein–protein interaction model. *Molecular Cell Proteomics* **3** 93–104.
- Tremblay A, Tremblay GB, Labrie F & Giguere V 1999 Ligand-independent recruitment of SRC-1 to estrogen receptor beta through phosphorylation of activation function AF-1. *Molecular Cell* **3** 513–519.
- Truong K & Ikura M 2001 The use of FRET imaging microscopy to detect protein–protein interactions and protein conformational changes *in vivo*. *Current Opinion in Structural Biology* **11** 573–578.
- Uetz P 2002 Two-hybrid arrays. *Current Opinion in Chemical Biology* **6** 57–62.
- Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P *et al.* 2000 A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403** 623–627.
- Vaglio P, Lamesch P, Reboul J, Rual JF, Martinez M, Hill D & Vidal M 2003 WorFDB: the *Caenorhabditis elegans* ORFeome Database. *Nucleic Acids Research* **31** 237–240.
- Visvader JE, Venter D, Hahm K, Santamaria M, Sum EY, O'Reilly L, White D, Williams R, Armes J & Lindeman GJ 2001 The LIM domain gene LMO4 inhibits differentiation of mammary epithelial cells *in vitro* and is overexpressed in breast cancer. *PNAS* **98** 14452–14457.
- Wajant H & Scheurich P 2001 Tumor necrosis factor receptor-associated factor (TRAF) 2 and its role in TNF signaling. *International Journal of Biochemistry and Cell Biology* **33** 19–32.
- Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N & Vidal M 2000 Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287** 116–122.
- Wasinger VC, Cordwell SJ, Cerpa-Poljak A, Yan JX, Gooley AA, Wilkins MR, Duncan MW, Harris R, Williams KL & Humphrey-Smith I 1995 Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis* **16** 1090–1094.
- Wehrman T, Kleaveland B, Her JH, Balint RF & Blau HM 2002 Protein–protein interactions monitored in mammalian cells via complementation of beta-lactamase enzyme fragments. *PNAS* **99** 3469–3474.
- Wilcox A & Hripscak G 1999 Classification algorithms applied to narrative reports. *Proceedings AMLA Symposium* 455–459.
- Wojcik J & Schachter V 2001 Protein–protein interaction map inference using interacting domain profile pairs. *Bioinformatics* **17** (Suppl 1) S296–S305.
- Wu HK, Heng HH, Siderovski DP, Dong WF, Okuno Y, Shi XM, Tsui LC & Minden MD 1996 Identification of a human LIM-Hox gene, hLH-2, aberrantly expressed in chronic myelogenous leukaemia and located on 9q33–34.1. *Oncogene* **12** 1205–1212.
- Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM & Eisenberg D 2002 DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research* **30** 303–305.
- Xu Y, Piston DW & Johnson CH 1999 A bioluminescence resonance energy transfer (BRET) system: application to interacting circadian clock proteins. *PNAS* **96** 151–156.
- Yates JR 3rd, Eng JK & McCormack AL 1995a Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Analytical Chemistry* **67** 3202–3210.
- Yates JR 3rd, Eng JK, McCormack AL & Schieltz D 1995b Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical Chemistry* **67** 1426–1436.
- Zhou H, Ranish JA, Watts JD & Aebersold R 2002 Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. *Nature Biotechnology* **20** 512–515.
- Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T *et al.* 2001 Global analysis of protein activities using proteome chips. *Science* **293** 2101–2105.

Received 22 November 2004

Accepted 20 December 2004

Made available as an Accepted Preprint 7 January 2005