CHAPTER 13

DISCOVERING PROTEIN-PROTEIN INTERACTIONS

Soon-Heng Tan
Institute for Infocomm Research
soonheng@i2r.a-star.edu.sg

See-Kiong Ng
Institute for Infocomm Research
skng@i2r.a-star.edu.sg

The genomics and proteomics efforts have helped identify many new genes and proteins in living organisms. However, simply knowing the existence of genes and proteins does not tell us much about the biological processes in which they participate. Many major biological processes are controlled by protein interaction networks. A comprehensive description of protein-protein interactions is therefore necessary to understand the genetic program of life. In this chapter, we provide an overview of the various current methods for discovering protein-protein interactions experimentally and computationally.

ORGANIZATION.

Section 1. We introduce the term "protein interactome", which is the complete set of protein-protein interactions in the cell.

Section 2. Then we describe some common experimental approaches to detect protein interactions. The approaches described include traditional experimental methods such as co-immunoprecipitation and synthetic lethals. The approaches described also include high-throughput experimental methods such as yeast two-hybrid, phage display, affinity purification and mass spectrometry, and protein microarrays.

Section 3. Next we present various computational approaches to predict protein interactions. The approaches presented include structure-based predictions such as structural homology. The approaches presented also include sequence-based predictions such as interacting orthologs and interactiing domain pairs. Another class of approaches presented are the genome-based predictions such as gene neighborhood, gene fusion, phylogenetic profiles, phylogenetic tree similarity, and correlated mRNA expression.

Section 4. Lastly, we conclude with a caution on the need to counter check detected and/or predicted protein interactions using multiple approaches.

1. Introduction

Identifying and sequencing the genes is the monumental task that has been undertaken and completed by the Human Genome Project. However it does not provide sufficient information to develop new therapies. In the cells, genes are merely blueprints for the construction of proteins who are the actual workhorses for the different biological processes occurring in the cells. Protein-protein interactions—as elementary constituents of cellular protein complexes and pathways—are the key determinants of protein functions. For the discovery of new and better drugs for many diseases, a comprehensive protein-protein interaction map of the cell is needed to fully understand the biology of the diseases.

The term proteome, coined in 1994 as a linguistic equivalent to the concept of genome, is used to describe the complete set of proteins that is expressed by the entire genome in a cell. The term proteomics refers to the study of the proteome using technologies for large-scale protein separation and identification. The nomenclature has been catching on. The generation of messenger RNA expression profiles, which revolves around the process of transcription, has been referred to as transcriptomics, while the set of mRNAs transcribed from a cell's genome is called the transcriptome. In a similar vein, we can use the term "interactome" to describe the set of biomolecular interactions occurring in a cell. Since many of the key biological processes are controlled by protein interaction networks, we use the phrase "protein interactome" to refer to the complete set of protein-protein interactions in the cell. Other biomolecular interactions in a cell include protein-DNA and protein-small molecule interactions. This chapter is devoted to providing an overview of "protein interactomics"—the dissection of the protein interactome using technologies of large-scale protein interaction detection. Both experimental and computational methods are discussed, as computational approaches are rapidly becoming important tools of the trade in the molecular biology laboratories in the post-genome era.

2. Experimental Detection of Protein Interactions

In this section, we describe the various common experimental approaches to detect protein interactions. We classify the experimental methods into two classes: traditional experimental methods and high-throughput detection methods. The former contains methods for assaying protein interactions with limited throughput (sometimes only individually), while the latter describe technologies addressing the

genome era's push for large-scale data generation. We cover the various computational means for predicting protein interactions in another section.

2.1. *Traditional Experimental Methods*

Traditionally, protein-protein interactions can be assayed biochemically by a variety of co-purification, gradient centrifugation, native gel, gel overlay, and column chromatography methods. Alternatively, protein-protein interactions can also be assessed indirectly by investigating their corresponding genetic interaction at the genome level. In this section, we highlight two representative experimental methods using biochemical and genetic approaches.

2.1.1. *Co-Immunoprecipitation*

A protein-protein interaction can be detected biochemically by selectively picking up one of the proteins from a mixture and then showing that the other protein, the interacting partner, is also picked up from the mixture.

In "co-immunoprecipitation",[296] the biochemical agent used to pick up selected proteins are specific antibodies. First, protein mixtures containing potential interacting protein partners are prepared in a cell lysate. An antibody designed to pick up—that is, immunoprecipitate—a specific protein is then applied. If the protein had been involved in a protein-protein interaction, its interacting protein partners would have also been picked up, or co-immunoprecipitated, along with it. The presence of this interacting protein partner can then be separated and identified using gel electrophoresis and mass spectrometry techniques.

The co-immunoprecipitation method is a laborious process. It is also restricted by the need of having specific antibodies against the proteins of interest. As a result, co-immunoprecipitation is not very amenable for the systematic large-scale detection of protein interactions which has become a necessary consideration in the post-genome era. As such, scientists have been working on ways to adapt it for large scale analysis of interactomes. One recent adaptation attempts to eliminate the limitation of having to have specific antibodies by using short protein sequences as "tags"[391,615] to attach to the proteins of interest. In this way, tag-specific antibodies instead of protein-specific antibodies can be used for co-immunoprecipitating any proteins of interest. With this and other creative technological refinements, co-immunoprecipitation has the potential to develop into a high-throughput protein interaction detection method suitable for post-genome discoveries.

2.1.2. *Synthetic Lethal Screening*

Unlike co-immunoprecipitation which directly assays for the protein interactions, synthetic lethal screening is a genetic method that detects functional linkages between two proteins from which possibility of interactions can be suggested.[28,74] Using mutations in the genes that encode the proteins of interest (*e.g.*, gene deletions), we can observe the phenotypic effects of mutations in a pair of proteins from which functional linkages can be inferred. In synthetic lethal screening, the strategy is to screen for cases in which a mutation in a single protein is non-lethal but cell survival is destroyed when it is coupled with a mutation in another protein. Such a synthetic lethality occurrence can be explained by two scenarios:

(1) The two proteins are playing back-up or redundant roles in an essential pathway—therefore, loss of function only occurs when both are simultaneously disabled.
(2) The two proteins are performing discrete steps in an essential pathway. A mutation in either of the proteins only weakens the functioning of the pathway, but the combined detrimental effect of concurrent mutations in both proteins is sufficient to eliminate the essential pathway from the cell.

The second scenario can suggest the potential existence of physical interaction between the two proteins. Note that this method is only applicable for proteins that are involved in essential pathways. It is therefore not amenable for proteome-wide investigation of the interactome. While the method also does not provide direct information regarding the "biochemical distance" between the proteins—two proteins involved in a synthetic lethal interaction could be as close as interacting subunits of a protein complex or be dozens of steps away in a complex branching pathway—its results can still be useful for providing further "hints" or evidences for the exploration of the vastly uncharted protein interactome.

2.2. *High Throughput Experimental Methods*

The Human Genome Project—with its ambitious goal of assembling the entire sequence of the human genome—has ignited the now-prevalent emphasis on high-throughput data generation in molecular biology. It has catalyzed a major paradigm shift in modern biology: the scale of experimental investigations in biology has taken a great leap from studying single genes, proteins, and interactions to screening whole genomes, proteomes, and interactomes. Biologists can now study the living systems in both comprehensive systemic scope and exquisite molecular details. In this chapter, we describe several high-throughput experimental methods suitable for large-scale detection of the always formidable interactome. The

practical bioinformatician is often tasked to analyze data generated from such high-throughput detection methods—it is useful to understand how these data are generated to understand better their various strengths as well as weaknesses.

2.2.1. *Yeast Two-Hybrid*

Yeast geneticists have developed a clever way of seeing whether two proteins can physically associate using the yeast as an *in vivo* platform. To do so, they enlist the service of a third protein—called a transcriptional activator—that has the ability to cause specific detectable "reporter genes" to be switched on. The scientists can experimentally separate an activator protein into two functional fragments, and then attach them separately to each of the candidate interacting proteins. If the two proteins—or rather, the two "hybrid" proteins, since they each has a part of an activator protein attached—interact, then the two fragments of the activator are reunited and switch on the associated "reporter gene" which produces a color change in the yeast cells. This is called the yeast two-hybrid (or Y2H) method, [250] the "two-hybrid" referring to the usage of the two hybrid candidate interacting proteins in the detection process.

The yeast scientists separate the transcriptional activator protein used in Y2H systems based on its two key functional parts: a DNA-binding domain and a trans-activation domain. The DNA-binding domain of a transcriptional activator is fused to a candidate protein known as the "bait", while its trans-activation domain is fused to the candidate protein's potential interacting protein partners known as the "prey". Since the yeast has two sexes, the "baits" and "prey" can easily be introduced into the same yeast cell by mating. If they physically interact, the DNA-binding and trans-activation domains are closely juxtaposed and the reconstituted transcriptional activator can mediate the switching-on of the associated reporter gene; see Figure 1.

Yeast two-hybrid was first described in 1989 by Fields and Song from State University of New York. [250] It has since become a routine method in biological labs to detect interaction between two proteins, albeit in a rather low-throughput manner. In recent years, Y2H has been successfully adapted for systematic high-throughput screening of protein-protein interaction. The first major high throughput genome-wide analysis of protein-protein interaction using yeast two-hybrid was applied to the yeast (or *Saccharomyces cerevisiae*) proteome itself. [386, 849] Of course, yeast is only an *in vivo* platform, detection of protein-protein interaction is not restricted to only the yeast proteome. In fact, large-scale identification of protein interaction using yeast two-hybrid has been carried out successfully on non-yeast proteomes, such as the proteomes of *Caenorhabditis elegans* [868] and

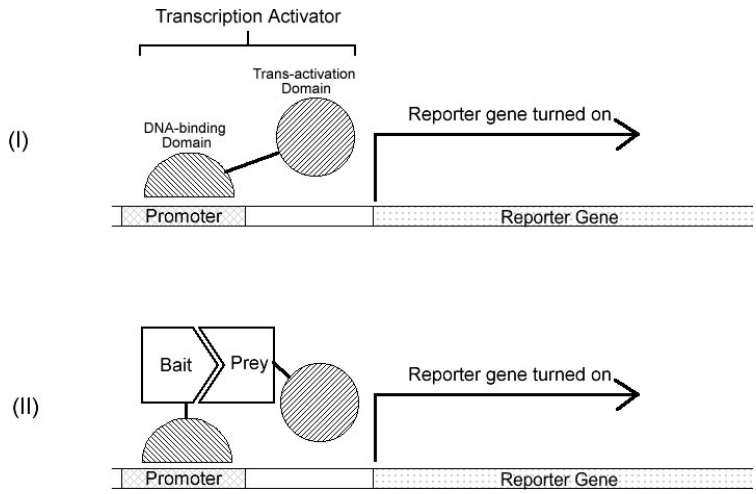

Fig. 1. Interaction Detection by Yeast Two-Hybrid Assay. (I) Activation of reporter gene by transcriptional activator. (II) Activation of reporter gene by reconstituted transcriptional activator.

Helicobacter pylori.[695]

Large-scale studies of protein-protein interaction detection using the yeast two-hybrid method have revealed many interactions not detected previously by any genetic and biochemical studies. However, a prudent bioinformatician must not take all of the detected interactions at their face values. Several recent studies on the reliability of high-throughput detection of protein interaction using yeast-two hybrids have revealed high error rates,[483,864] some reporting as high as 50% false positive rates.[793] There are inherent limitations in the Y2H method that can lead to false positives or biologically meaningless interactions:

- Some proteins exhibit transcriptional properties and can cause the reporter gene to be switched on by themselves, leading to artifactual interactions in which a positive signal is detected even though the two proteins do not interact with each other.
- Hybrid proteins can also adopt non-native interacting folds as a result of the fusion or in a foreign environment, giving rise to artificial interactions that may not occur naturally in the cellular environment.

As a choice method for large-scale genome-wide screening for protein-protein interactions, the yeast two-hybrid also suffers in terms of coverage. Neither of the two key comprehensive yeast interactome studies by Ito *et al.*[386] and Uetz*et al.*[849] using yeast-two-hybrid assays have recapitulated more than $\sim$13% of the pub-

lished interactions detected by the yeast biologist community using conventional single protein analyses.[337] The high false negative rate could be due to inherent experimental limitations of the Y2H method such as:

- In yeast two-hybrid systems, interactions are detected in the nucleus where transcription occurs. The method is therefore weak in detecting interactions for cytoplasmic proteins.
- Just as the non-native foldings of the hybrid proteins can give rise to artificial interactions, they can also prevent the interaction of two interacting proteins.
- Many proteins require post-translation modification for interaction, but this is not accommodated by the yeast two-hybrid approach.

In general, a prudent bioinformatician must be aware of the potential errors in experimental data, especially those generated by high-throughput methods. The detection of a protein-protein interaction—experimentally or computationally—must always be confirmed by at least two or more independent means. It is therefore important to develop other alternative methods for protein interaction detection and discovery. We describe a few other high-throughput experimental methods below, and leave the alternative computational methods for the next section.

2.2.2. *Phage Display*

Phages (or rather, bacteriophages) are viruses that infect bacterial cells and take over the hosts' cellular machinery for its reproduction. A phage is a very simple and efficient parasitic machine, made up of a genetic material (in form of either RNA or DNA) encapsulated by a protein coat assembled from viral proteins; see Part I of Figure 2. A key feature of phages is that they can accommodate segments of foreign DNA—a gene segment from another species, or stretches of chemically synthesized DNA—as "inserts" in their DNA. As the virus' DNA is replicated in the bacteria host, the foreign insert is also replicated along with it as a sort of passenger. This makes phages a choice vehicle in the laboratory for replicating other types of DNA.

In a phage display vector, we use the phages' DNA insertion templates to program the host bacteria cells to synthesize specific foreign peptides. By choosing one of the genes that make coat proteins for the phages to insert a foreign DNA into, hybrid fusion coat proteins that contain the foreign peptides are synthesized and used to construct the protein coats of the replicated phages. In this way, the expressed foreign peptides are "displayed" on the outer surface of the replicated phages for easy detection; see Part II of Figure 2.

Unlike yeast two-hybrid that detects interaction between two full length

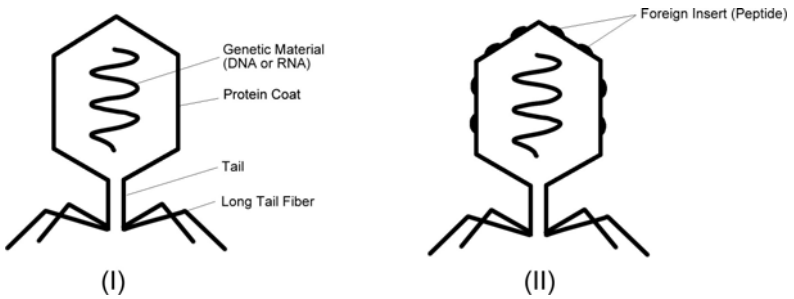

Fig. 2. Schematic diagrams of (I) a phage; and (II) interaction detection by phage display.

proteins, phage display can be used to determine the binding of short protein sequences—for example, only around 10 amino acids in length—to proteins. To detect what peptide sequences bind to a protein, we immobilize the protein on a solid surface as a "bait", and then expose it to a large library of phages with different display peptide sequences. Phages with display peptides that bind to the "bait" protein can then be selected for infection on a bacterial host, to amplify the binding sequence to an appropriate amount to allow for identification by sequencing.

Phage Display was first reported by Smith in 1985.[778] To-date, phage display has been used effectively to determine short binding sequences for the SH3,[243] WW,[419] SH2,[169] and PDZ[477] protein domains. While phage display is best used for determining short binding sequences, it can also be used to detect—or rather, predict—protein-protein interaction based on the fact that most protein-protein interactions involve direct contact of very small numbers of amino acids. One example is the combined use of phage display and yeast two-hybrid technologies to respectively predict and validate a network of interactions between most of the SH3 domain containing proteins in yeast by Tong *et al.*[833] In this study, the SH3 domain binding motifs derived from phage display are used to predict a network of hypothetical interactions—between proteins with SH3 domain and those with sequences that matches the detected binding motifs—that are then experimentally validated using yeast two-hybrid screens.

Phage display is a powerful tool to identify partners of protein-protein interactions—it is one of the most established techniques to generate lead molecules in drug discovery. The method is easily amenable for rapid high-throughput, combinatorial detection. Phage display libraries containing 10^6 to 10^{10} independent clones can be readily constructed, with each clone carrying a different foreign DNA insert and therefore displaying a different peptide on its surface. However, detection of protein-protein interaction by phage display is an indi-

rect method that predicts interaction between two proteins containing the detected short binding peptide sequences. Furthermore, phage display is most suitable for detecting interactions with short sequences instead of full-length proteins—as such, it may not work for all proteins. As in the other experimental approaches, phage display needs to work in combination with other complementary interaction detection methods—experimental and computational—in order to map out the vast and complicated interactome fully and accurately.

2.2.3. *Affinity Purification and Mass Spectrometry*

Interactions between proteins are not limited to pair-wise interactions such as those detected by the above methods—several proteins (sometimes as many as 20 or more) can come together to form a multimeric protein complex. Many functional pathways in the cell involve multi-protein complexes. The detection of protein interactions in the form of multi-protein complexes is therefore important for understanding the biochemical mechanisms of the living cell.

The so-called "affinity purification" process can be used to identify groups of proteins that interact together to form a complex.[67] To do so, a "bait" protein is first immobilized on a matrix or a solid surface such as the internal of a column. This can be done by attaching an affinity tag to the bait protein which helps stick it to the solid surface. Then, a mixture of candidate proteins passes through the column: proteins binding to the immobilized protein are thus retained and captured while non-interacting proteins are eluted away. The captured proteins in turn serve as additional baits to capture other proteins, leading to formation of protein complexes; see Figure 3. The bound proteins are subsequently collected from the column by washing it with a solution that decreases the binding affinity of the bound proteins, or using an enzyme to cleave the affinity tag to remove the bound proteins from the column. As with the genome-wide two-hybrid system, robotics made the assays high-throughput.

Traditional protein identification methods such as Edman sequencing and Western blots are tedious, time-consuming and not easily scalable for large-scale identification of proteins. For throughput, mass spectrometry provides a fast and accurate means for dissecting the protein complexes. To detect the mass fingerprint of a protein, the protein is first cleaved into many short-sequence peptides using proteases that cut proteins at specific sites. The masses of these cleaved peptide fragments are then determined in a mass spectrometry process known as Matrix-Assisted Laser Desorption/Ionization (MALDI) to generate a series of peaks, each describing the molecular mass of a single peptide in the mixture. Because the proteases cut the protein at specific sites, it is possible to know exactly

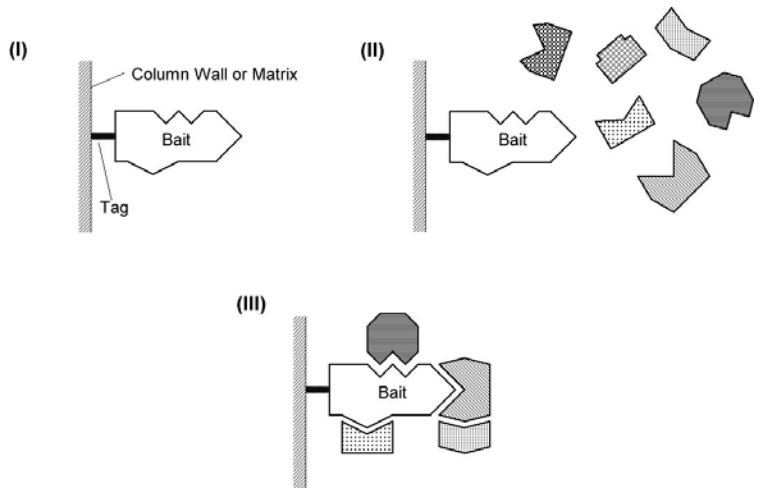

Fig. 3. Interaction Detection by Affinity Purification.

which cleaved peptide fragments any given protein can generate. Each protein in the proteome can therefore be characterized with a "fingerprint" consisting of the set of peptide masses resulting from the mass spectrometry experiment. With the recent completion of many large-scale sequencing projects, huge sequence databases are now available to enable researchers to compare an observed peptide fingerprint with, for example, every possible human protein to identify an unknown protein. For identifying a mixture of proteins such as a protein complex from affinity purification studies, the proteins are first separated using gel electrophoresis and then individually identified using mass spectrometry.

Researchers have applied this approach on a proteome-wide scale. Gavin *et al.*[278] have found 232 yeast protein complexes using an affinity purification process. Subsequent protein identification using MADLI has revealed 1,440 distinct captured proteins, covering about 25% of the yeast's proteins. Ho *et al.* [358] have applied the same general approach to identify 3,617 interactions involving 1,578 yeast proteins. As in the other experimental methods, protein coverage is still a limitation. Furthermore, it has the caveat that a single bait protein may occur in more than one complex in a cell: it may therefore link with two or more proteins that never actually occur in the same complex, giving the illusion that the detected protein complex is bigger and more elaborate than it actually is. Both groups also report a significant number of false-positive interactions, while failing to identify many known associations.[462] So, as the results from most large-scale

studies have illustrated, there are no perfect detection methods for mapping the interactome. It is essential to integrate data from many different sources in order to obtain an accurate and comprehensive understanding of protein networks.

2.2.4. *Protein Microarrays*

In the past, the study of molecular biology focused on studying a single gene or protein at a time. Today, with the many advancements in high-throughput technologies, it is now possible for biologists to perform global informational analyses in their discovery pursuits. For example, the DNA microarray technology has made possible the analysis of the expression levels of hundreds and thousands of genes simultaneously, allowing the biologists to analyze gene expression behaviors at the whole-genome level. As we will see in Section 3.3.5, scientists have even been able to use gene expression data to decipher the encoded protein networks that dictate cellular function. However, most cellular functions are manifested by the direct activities of the translated proteins and not by the genes themselves. In fact, protein expression levels often do not correlate with mRNA expression levels.[310] Expression analysis at the proteomic level is a more superior approach as proteins are one step closer to biochemical activities than genes are.

Researchers have recently begun to focus on developing protein microarray methods for the high-throughput analysis of proteins. Just like the gene microarrays, a protein microarray consists of tens to thousands of proteins, individually spotted at unique addresses in a micro- or even nano-scale matrix, so that interactions between the bait proteins and the test samples can easily be identified. The detection process in protein chip is very similar to the affinity purification technique described in the previous section. The bait proteins are purified and spotted separately onto a small solid surface such as a glass slide for capturing testing proteins in solution. The solid surface is then overlaid with a testing protein for interaction with the baits, washed and then assayed for protein binding at each microarray spot. Usually, the testing protein is attached to a suitable dye or enzyme that makes it easy for the bound proteins to be detected.

MacBeath and Schreiber[534] describe a proof-of-principle work in 2000 of spotting purified proteins onto glass slides using the existing DNA microarrayer and scanning tool, and showing that the purified proteins retained their activities when spotted onto chemically-treated glass slides. Since then, many researchers have worked on using protein microarray to detect protein-protein interaction on a massive scale. For example, Zhu *et al.*[939] construct a genome-wide protein chip and use it to assay interactions of proteins and phospholipids in yeast. A total of 5,800 predicted yeast's ORFs are cloned and 80% of these are purified to a de-

tectable amount and then spotted on glass slides to construct a yeast proteome microarray to screen for their ability to interact with proteins and phospholipids. Their results illustrate that microarrays of an entire eukaryotic proteome can be prepared and screened for diverse biochemical activities.

Protein microarrays hold great promise for revolutionizing the analysis of entire proteomes, just as what DNA microarrays have done for functional genomics. However, developing protein microarrays is a much harder problem than making DNA microarrays. Proteins are heterogeneous, making it difficult to develop methods to attach them to biochips and have them remain functional. Proteins are also more difficult to synthesize than DNA and are more likely to lose structural or functional properties in different environments or when modified. Unlike DNA, where the sequence is all that matters, a protein's three-dimensional structure must be preserved. However, one can be confident that novel technologies will continue to expand the power of protein arrays so that it will soon play a major role—together with the other protein-protein interaction techniques described in this chapter–in deciphering the protein networks that dictate cellular functions.

3. Computational Prediction Protein Interaction

As we have seen in the previous sections, even the best experimental methods for detecting protein-protein interactions are not without their limitations. As such, the detection—or rather, prediction—of protein-protein interactions using computational approaches in a rapid, automatic, and reasonably accurate manner would complement the experimental approaches. Toward this end, bioinformaticians have developed many different computational approaches to screen entire genomes and predict protein-protein interactions from a variety of sources of information:

(1) *Structure-based predictions.* Interactions between proteins can be deemed as biophysical processes whereby the shapes of the molecules play a major role. Structural biologists have long been exploiting the structural information of the protein molecules to determine whether they interact. However, the determination of the three-dimensional structure of proteins is still a major bottleneck today, greatly limiting the use of structure-based prediction for unraveling protein-protein interactions at the proteome level as the structural data of most proteins are still unavailable.
(2) *Sequence-based predictions.* On the other hand, the genetic and amino acid sequences of most proteins are now available. This has prompted to resourceful bioinformaticians to find ways to predict protein-protein interactions based on sequence information of the proteins alone.

(3) *Genome-based predictions.* The complete sequences of many genomes are also available. To-date, according to the Entrez website at the US National Center for Biotechnology Information, more than 170 species already have their complete genetic sequences mapped. These entire genomes of multiple species can be used to screen for genome-level contextual information such as gene co-localizations, phylogenetic profiles, and even gene expression to infer interactions between proteins.

In this section, we describe a variety of computational methods that bioinformaticians have developed under each of the above categories. While it is clear that computational approaches will never be able to replace experimental methods, by combining the results from multiple approaches—*in silico* or otherwise—we can improve both the quantity and quality of protein interaction detected by leveraging on the complementary strengths of the different detection methods. *E.g.*, experimental methods typically suffer from limited coverage, whereas computational methods usually have broad coverage as they are less sensitive to the *in vivo* and *in vitro* biochemical intricacies. It is thus important for the bioinformaticians to continue to develop computational methods for the detection of protein-protein interactions. The combination of experimental and computational data will eventually lead to the complete set of information for us to understand the underlying interaction networks that govern the functioning of the living cell.

3.1. *Structure-Based Predictions*

Much of the focus in structure-based predictions related to protein-protein interactions is in the prediction of interaction sites, also known as the docking problem if the structures of the two interacting proteins are known. Docking is the process whereby two molecules fit together in three-dimensional space. However, knowing the induced fit based on the unbound, isolated structures of two protein molecules do not immediately imply that the two proteins will interact, because proteins undergo conformational changes upon binding. As such, most docking algorithms are used mainly to predict whether and how small molecules, such as drug candidates, interact with known protein targets.

However, even if we can solve the docking problem for protein-protein interactions, it is still hindered by the very small number of protein structures available. In order to handle genome-wide protein interaction prediction, structure-based methods must be able to infer from proteins whose structures are not yet known based on knowledge derived from limited number of known structures of protein-protein interactions—usually complexes—in an approach similar to sequence homology.

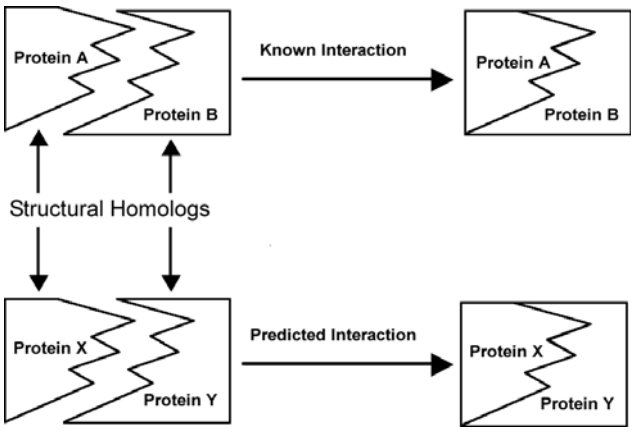

Fig. 4. Structure-based Prediction of Protein-Protein Interactions.

3.1.1. *Structural Homology*

The premise behind protein-protein interaction prediction by structural homology is fairly straightforward: if protein A interacts with protein B, and two new proteins X and Y each looks structurally like proteins A and B respectively, then protein X might also interact with protein Y; see Figure 4. Given that most proteins do not have known structures, the first step is to predict the structure of a protein from its primary sequence. This can be done by a computational process known as "threading", in which we align the sequence of the protein of interest to a library of known folds and find the closest matching structure. If the structures are known to interact—from existing 3D structures of protein complexes, say—we can then compute the interfacial energy and electrostatic charge to further confirm whether the partners form a stable complex. Lu *et al.*[527] use such a threading-based algorithm to assign putative structures for predicting interaction between yeast proteins. They have predicted 2,865 interactions, with 1,138 interactions verified in the Database of Interacting Proteins (DIP).[910]

However, as we are only interested in the parts of the proteins' structures that are involved in the interactions, we can focus on the structures on key areas of proteins such as the protein domains that are most likely to be involved in the protein-protein interactions. Aloy and Russell[22] use pairs of interacting Pfam domains[64] from known three-dimensional complex structures for prediction. Pairs of proteins with sequences homologous to a known interacting domain pair can then be scored for how well they preserve the atomic contacts at the predicted interaction interface by using empirical potentials to confirm the predicted interactions.

The current lack of protein 3-D structures clearly limits the global application of structure-based approaches for genome-wide protein-protein interaction predictions, even with an approach that uses structural homology. With the increasing pace of structure determination and structural genomics efforts, we hope that the structures for many more protein complexes will be available in the future. In the meantime, the vast amount of information in protein and gene sequences can be used as an alternative source for inferring protein-protein interactions.

3.2. *Sequence-Based Predictions*

While protein structures may be the most informative source for protein-protein interactions, protein sequences can also be used, together with existing protein interaction data, for predicting new protein interactions. In this section, we describe two representative sequence-based approaches: one approach is based on conventional sequence homology across various species, while a second approach uses protein domains as an abstraction of proteins for their interactions, and then reduces protein-protein interactions into domain-domain interactions which can in turn be used for predicting new interactions.

3.2.1. *Interacting Orthologs*

A widely used approach of assigning function to newly sequenced genes is by comparing their sequences with that of annotated proteins in other species. If the new gene or protein's sequence bears significant similarity to the sequence of a gene or protein—namely, its ortholog—in an annotated database of another species, it can be assumed that the two proteins are either the same genetic instantiation, or at the very least, share very similar properties and functions. As such, if protein A interacts with protein B, then the orthologs of A and B in another species are also likely to interact.

A study by Matthews *et al.*[552] has investigated the extent to which a protein interaction map generated in one species can be used to predict interactions in another species under the interacting orthologs or "interologs" principle. In their study, Matthews *et al.* compare protein-protein interactions detected in *S. cerevisiae* to interactions detected in *C. elegans* using the same experimental method yeast two-hybrid. Although only 31% of the high-confidence interactions detected in *S. cerevisiae* are also detected in *C. elegans*, it confirmed that some interactions are conserved between organisms, and we should expect more interologs between more closely related species than *S. cerevisiae* and *C. elegans*.

3.2.2. *Interacting Domain Pairs*

The interolog method described above scans proteins full-length to look for co-evolved interactions. Since protein interactions usually involve only small regions of the interacting molecules, conservation of interactions theoretically only requires that these key subregions on the interacting proteins be conserved. One approach is to treat proteins as collections of conserved domains, where each domain is responsible for a specific interaction with another domain. Protein domains are modules of amino acid sequence on proteins with specific evolutionarily conserved motifs—these protein domains are therefore quite likely the structural or functional units that participate in intermolecular interactions. As such, the existence of certain domains in proteins can be used to suggest the possibility of two proteins to interact or form a stable complex. In fact, Wojcik and Schächter [892] have shown that the use of domain profile pairs can provide better prediction of protein interactions than the use of full-length protein sequences.

Researchers have begun to use domain-domain interactions to predict protein-protein interactions with promising results. [199, 298, 609] For example, Deng *et al.* [199] predict yeast protein-protein interactions using inferred domain-domain interactions, and they achieve 42.5% specificity and 77.6% sensitivity using the combined data of Uetz *et al.* [849] and Ito *et al.* [386] showing that interacting domain pairs can be useful for computational prediction of protein-protein interactions. Note that the relatively low specificity may be caused by the fact that the observed protein-protein interactions in the Uetz-Ito combined data represent only a small fraction of all of the real interactions. However, one major drawback of this approach is that there are currently no efficient experimental methods for detecting domain-domain interactions—the number of experimentally derived interacting domain pairs is highly limited. As such, researchers can only used inferred domain-domain interactions in the prediction of protein-protein interactions, the accuracy of which may be further thwarted by the inference errors associated with the inferred domain-domain interactions.

3.3. *Genome-Based Predictions*

Given that a rapidly increasing number of genomes have already been sequenced, we can transcend conventional homology-based methods such as those described in the previous sections, and take into account the genomic context of proteins and genes within complete genomes for the prediction of interactions. By mining entire genomes of different species, we can discover cross-genome contextual information that are useful for predicting protein-protein interactions—usually indirectly through functional linkages—such as:

(1) *Gene locality context.* We can track the localities of genes in different species and use such information to infer functional linkages and possible interactions. We can explore the idea of co-localization or gene neighborhood, which is based on notion that genes which interact or are at least functionally associated will be kept in physical proximity to each other on the genome.[184] Alternatively, we can search for gene fusion events, whereby the fusion of two genes in one species can indicate possible interactions.
(2) *Phylogenetic context.* Instead of tracking the spatial arrangements of genes on the genomes, we can also track the evolutionary patterns of the genes, using the notion that genes that are functionally related tend to be inherited together through evolution.[660]

A third source of genome-wide information for protein-protein interaction prediction can also be gleaned, albeit indirectly, from gene expression experimental data:

(3) *Gene expression context.* Microarray technologies has enabled quantitative measurement of genome-wide gene expression levels simultaneously. To reveal the various functions of the genes and proteins, the gene expression profiles of a series of experimental conditions can be analyzed so that the genes can be grouped into clusters based on the similarity in their patterns of expression. The co-expression clusters can then be interpreted as potential functional linkages from which we may infer protein interactions.

Below, we describe the use of these three categories of genome-based information for the *in silico* detection of protein-protein interaction in details.

3.3.1. *Gene Locality Context: Gene Neighborhood*

One of the earlier attempts at genome-based prediction of protein-protein interactions is based on the notion of conservation of gene neighborhood. We can predict functional linkage between a pair of genes if their orthologs tend to be in close physical proximity in many genomes, as shown in Figure 5. In fact, studies have revealed that genes that participate in the same biological pathway tend to be neighbors or be clustered into discrete region along the genomic DNA. The most well-known example occurs in the bacterial and archael genomes, which are organized into regions such as operons that code for functionally-related proteins.[85]

As functionally-related proteins are clearly more likely to interact than unrelated ones, genes conserved as neighbors across genomes indicate possible interactions between their protein products. In a study by Dendekar *et al.*,[184] $\sim$300 genes were identified to be conserved in neighboring clusters across different bac-

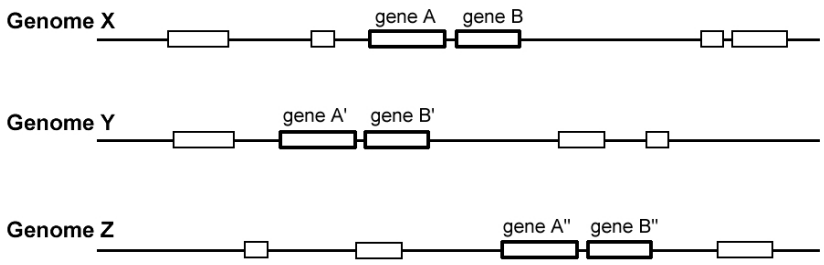

Fig. 5. Interaction Prediction by Gene Neighborhood.

terial genomes. Out of these ∼300 genes, 75% have been previously described to be physically interacting, while another 80% of the remaining conserved neighbors have functions that are highly indicative of interactions between them. Their results show that gene neighborhood can be a powerful method for inferring protein-protein interactions in bacteria. In fact, Overbeek *et al.* [641] has successfully used this method to detect missing members of metabolic pathways in a number of prokaryotic species. While the gene neighborhood method has worked well with bacteria, this method may not be directly applicable to the higher eukaryotic species, in which the correlation between genome order and biological functions is less pronounced since the co-regulation of genes is not imposed at the genome structure level. For these other species, alternative genome-based methods must be used instead.

3.3.2. *Gene Locality Context: Gene Fusion*

One alternative method that is quite similar to the gene neighborhood approach is the so-called Rosetta Stone[541] or gene fusion[234] method. In fact, the complete fusion of two genes into one single unit can be deemed the ultimate form of gene proximity. It has been observed that many genes become fused through the course of evolution due to selective pressure—for example, fusion of two genes may allow the metabolic channeling of substrates or decrease the regulatory load in the cell. Gene fusion events have been observed frequently in evolution; some well-known examples include the fusion of tryptophan synthetase α and β subunits from bacteria to fungi,[126] and that of TrpC and TrpF genes in *E. coli* and

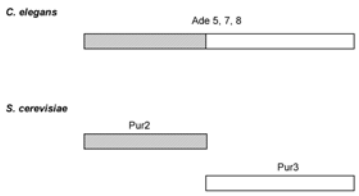

Fig. 6. Interaction Prediction by Gene Fusion.

H. influenzae.[720] Figure 6 depicts another example gene fusion event: the proteins *Pur2* and *Pur3* are two separate interacting proteins in yeast (*S. cerevisiae*), but their orthologs are found fused into one protein in *C. elegans*.

The numerous observed examples suggest that the protein products of the fused genes either physically interact or are at least closely functionally associated. As such, computational detection of gene-fusion events in complete genomes can be used to infer functional linkage or even physical interaction between proteins. In a study by Marcotte *et al.*,[541] they detected ∼7,000 putative protein-protein interactions in *E. coli*, and ∼45,500 putative protein-protein interactions in yeast by gene fusion analysis, demonstrating that the gene fusion phenomenon is quite widespread.

Recently, a similar approach based on protein domains has been proposed. As we have explained in Section 3.2.2, protein domains are evolutionarily conserved modules of amino acid sequence on proteins that can be deemed the structural or functional units that participate in intermolecular interactions. To exploit the gene fusion concept in protein interaction prediction, we can treat a protein as a set of conserved domains, where each domain is responsible for a specific interaction with another one.[417,541] In this domain fusion method, we computationally detect fused composite proteins in a reference genome with protein domains that correspond to individual full-length component proteins in other genomes.

Marcotte *et al.*[541] use predefined protein domains as a basis for searching fused—*i.e.*, multi-domain—proteins to detect gene fusion from a database of protein sequences. Using the SWISS-PROT database[86] annotated with domain information from ProDom,[765] they have detected ∼7,842 so-called Rosetta Stone domain fusion links in yeast and ∼750 high-confidence ones in *E. coli*, indicating that the domain fusion phenomenon—even using only ProDom domains—is widely observed and suitable as a basis for predicting protein interactions. How-

ever, the use of pre-defined domains such as those in ProDom may limit the coverage of the approach, since a portion of proteins may not have pre-assigned domains. Enright and Ouzounis[233] used an alternative approach that employed sequence alignment techniques to detect regions of local similarities between proteins from different species instead of using pre-defined domains. They have successfully detected 39,730 domain fusion links between 7,224 proteins from the genomes of 24 species.[233]

Unlike the gene neighborhood method described in the previous section, the gene or domain fusion method does not require the individual genes to be proximal along the chromosomes. As such, the method can be applied to eukaryotic genomes.[234] The occurrence of shared domains in distinct proteins is a phenomenon whose true extent in prokaryotic organisms is still unclear,[853] limiting the use of the domain fusion method for protein-protein interaction predictions in the prokaryotes. This shows that just as it is in the case for experimental approaches, the coverage of various computational detection methods can also differ—it is therefore necessary to explore multiple complementary approaches such that complete information about interaction networks can be obtained.

3.3.3. *Phylogenetic Context: Phylogenetic Profiles*

During evolution, functionally-linked proteins tend to be either preserved or eliminated in a new species.[660] This means that if two proteins are functionally associated, their corresponding orthologs will tend to occur together in another genome. We can exploit such evolutionary patterns to predict if proteins interact.

One approach, called phylogenetic profiling, is to detect the presence or absence of genes in related species for suggesting possible interaction. The phylogenetic profiling method is based on the notion that interacting or functionally linked proteins must be jointly present or jointly absent in different organisms. A phylogenetic profile describes an occurrence of a certain protein in a set of organisms. Proteins whose genes have highly correlated phylogenetic profiles can then be inferred as physically interacting or at least functionally linked.

The phylogenetic profile of a protein is typically represented as a string that encodes the presence or absence—in form of 1 or 0—of a protein in a given number of genomes; see Figure 7. Using this binary vector representation, the phylogenetic profiles of proteins are computationally constructed across different genomes. Then, proteins that share similar profiles are clustered together, and functional linkage or even physical interaction can be predicted for proteins that are clustered together, as shown in the figure.

In a study by Pellegrini *et al.*,[660] they apply this method to detect possible

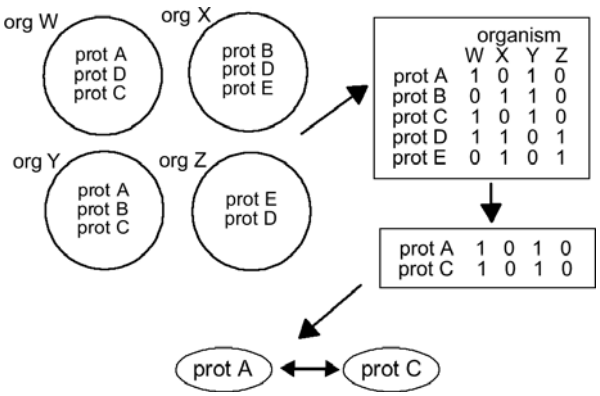

Fig. 7. Interaction Prediction by Phylogenetic Profiling.

functional linkages between 4,290 *E. coli* proteins using 16 different genomes. They demonstrate that comparing with random groups of proteins, the clusters of proteins formed by similar phylogenetic profiles tend to share the same functional annotation under SWISS-PROT. Since this method requires the detection of the absence of a protein in a genome, it can only be applied to complete genomes. However, this limitation should not be a key concern as an increasing number of complete genome sequences are becoming available. In fact, the method is expected to become more powerful since more completely-sequenced genomes will allow for larger and more accurate profiles to be constructed for each protein.

One limitation with phylogenetic profiling is its inability to detect linkages for proteins that are essential and common to most species. This group of proteins constitute a major portion of entire gene set in a genome. Also, as with most of the other computational methods, phylogenetic profiling can only be used to suggest possible functional linkages—a direct physical interaction between the proteins is not necessarily implied. It is prudent for the practical bioinformatician to be mindful when using these information for further discoveries.

3.3.4. *Phylogenetic Context: Phylogenetic Tree Similarity*

The co-evolution of interacting protein pairs has long been observed in such well-known interacting protein pairs as dockerins and cohexins,[644] as well as insulin and its receptors[266]—the corresponding phylogenetic trees of these proteins show a significantly greater degree of similarity than non-interacting proteins are expected to show. As such, phylogenetic tree similarity is another suitable form of

evolutionary information for inferring possible interaction between two proteins.

The phylogenetic tree similarity method is based on the notion of co-related residues changes between two proteins across different genomes. The reasoning is as follows: if an incurred residue change in one protein disrupts its interaction with its partner, some compensatory residue changes must also occur in its interacting partner in order to sustain the interaction or they will be selected against and eliminated. As a result, a pair of interacting proteins in the course of evolution would go through similar series of changes, whereas the residue changes for non-interacting proteins would be totally uncorrelated. This means that the phylogenetic tree of interacting proteins would be very similar, reflecting their similarity in their evolutionary histories.

While phylogenetic profiling looks for proteins that co-exist (or otherwise) in different genomes, the phylogenetic tree similarity method looks for co-related residues changes between two proteins across different genomes. Although the name of the method may imply a direct comparison of the phylogenetic tree structures, we can measure tree similarity by comparing the correlation between the distance matrices of protein orthologs from different species. Distance matrices are typically used in the construction of phylogenetics trees—a distance matrix is an $n \times n$ matrix that contains the pairwise distances between n sequences in a set. The distance between two sequences are measured by their sequence alignment scores, which could simply be the number of mismatchs in the alignment. In this way, we can account for and compare the structure of the underlying phylogenetic trees between orthologous proteins. Note that the phylogenetic profiling method described in the previous section can be considered as a simplification of the phylogenetic tree similarity method, where the "distance matrix" for each protein is merely a binary vector indicating the presence or absence of the protein ortholog in a particular species.

In a study by Goh *et al.*,[294] they apply this procedure—also known as mirrortree—to the two interacting domains of phosphoglycerate kinase. They found a high correlation coefficient of 0.8 between two corresponding distance matrices between the two interacting protein domains. This value was later confirmed by Pazos *et al.*[650] in a larger scale experiment for predicting protein interactions in *E. coli*. In a control set of 13 known interactions, they found that the interacting protein pairs have high correlation values in their distance matrices—in fact, 9 out of 13 have correlation coefficient values higher than 0.77. A total of 67,000 unknown pairs of proteins are then compared across 14 genomes; 2,742 pairs have correlation coefficient values greater than 0.8—they can therefore be inferred as interacting protein pairs.

The basic steps in this method are shown in Figure 8. To determine if a pair of

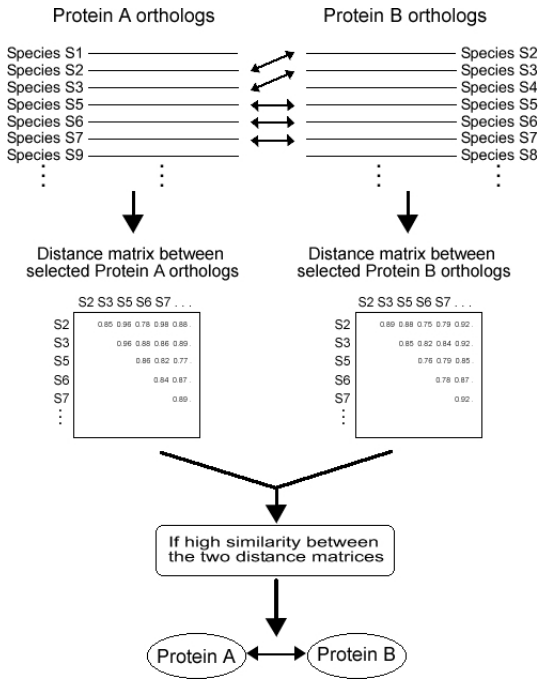

Fig. 8. Interaction Prediction by Phylogenetic Tree Similarity.

candidate proteins interact, we search for their orthologs across different genomes to form as many ortholog pairs from coincident species as possible. For each protein, we construct its distance matrix by pairwise sequence comparison between its various orthologs found. If the two proteins have a high correlation coefficient value between their two distance matrices, they can be predicted to be a possibly interacting pair.

A clear advantage of this method over the phylogenetic profiling method is that it does not require the presence of fully-sequenced genomes. Furthermore, while many of the other genome-based methods detect functional linkages and infer interactions indirectly, the phylogenetic tree similarity method predicts interactions that are more likely to be physically interacting. However, a main limitation of this method is that we must find a sufficient number orthologs pairs in order to make a reasonable postulation about interaction. In the study by Pazos *et al.*, [650] 11 was the minimum number of species required. As the sequence of many other species become available, this requirement will soon not be a constraint of the phylogenetic tree similarity method.

3.3.5. *Gene Expression: Correlated mRNA Expression*

Perhaps more than any other method, the emergence of DNA microarrays has transformed genomics from a discipline traditionally restricted to large sequencing labs to a cottage industry practiced by labs of all sizes. DNA microarrays allow investigators to measure simultaneously the level of transcription for every gene in a genome. For the first time, we can collect data from a whole genome as it responds to its environment. Such global views of gene expression can be used for elucidating the functional linkages of the various genes by applying clustering techniques to group together genes with expression levels that correlate with one another under different experimental conditions. As in many other genome-based methods we have described, the detected functional linkages can then be used to infer—albeit indirectly—possible interactions between the proteins that they encode. Several researchers have shown global evidence that genes with similar expression profiles are more likely to encode interacting proteins. For example, Grigoriev[301] demonstrates that there is indeed a significant relationship between gene expression and protein interactions on the proteome scale—the mean correlation coefficients of gene expression profiles between interacting proteins are higher than those between random protein pairs, in both the genomically simplistic bacteriophage T7 and the more complex *Saccharomyces cerevisiae* (yeast) genomes. In a separate study by Ge *et al.*[280] on yeast, they compare the interactions between proteins encoded by genes that belong to common expression-profiling clusters with those between proteins encoded by genes that belong to different clusters, and found that proteins from the intra-group genes are more than 5 times likely to interact with each other than proteins from the inter-group genes.

In another work to relate whole-genome expression data with protein-protein interactions, Jansen *et al.*,[390] find that while the subunits of the permanent protein complexes do indeed share significant correlation in their RNA expression, the correlation expression method is understandably relatively weak in detecting transient interactions. However, they have also observed weak correlated RNA expression patterns between interacting proteins determined by genome-wide yeast two-hybrid studies, indicating potential limitations in using this approach for protein-protein interaction prediction. On the other hand, while this method by itself is relatively weak for accurate interaction detection, it can serve as an excellent complementary method to validate interaction generated from other experimental methods. In a comprehensive study conducted by Kemmeren *et al.*, [426] up to 71% of biologically-verified interactions can be validated with the gene co-expression approach. Integration of expression and interaction data is thus a way to improve

the confidence of protein-protein interaction data generated by high-throughput technologies.

4. Conclusion

Before the advent of high-throughput experimental and computational methods, protein-protein interactions have always been studied in the molecular biology laboratories in a relatively small scale using conventional experimental techniques. However, in order to understand, model, and predict the many unfathomable rules that govern protein-protein interactions inside the cell on the genomic level, large scale protein interaction maps must be generated. In this chapter, we have provided an overview of the various current high-throughput protein-protein interaction detection methods. In particular, we have shown that both the conventional experimental approaches and the new computational approaches can be useful for mapping the vast interactomes. We have also shown that there is no single best method for large-scale protein-protein interaction detection—each method, experimental or otherwise, has its own advantages and disadvantages.

The advent of the various high-throughput detection and prediction technologies has brought about a major paradigm shift in modern molecular biology research from single-molecule experiments to genome and proteome-level investigations. With the current high throughput approaches powerful enough to generate more data than those accumulated over many decades from small scale experiments, predictive research has become a mainstay of knowledge discovery in modern molecular biology. This has led to the tendency for experiments to be technology-driven rather than hypothesis-driven, with datasets routinely generated without much specific knowledge about the functions of genes being investigated. This can be problematic because the high-throughput data have been shown to exhibit high error rates. For example, a recent rigorous study by Sprinzak *et al.*[793] has revealed that the reliability of the popular high-throughput yeast-two-hybrid assay is only about 50%. Another comprehensive survey on current protein-protein interaction detection technologies done by von Mering *et al.* [864] showed that different experimental methods cover rather different classes of protein interactions. This indicates the possibility of high false negative rates in the interaction data in addition to the many false positive detections. As practicing bioinformaticians, we should always be mindful about how the data that we are analyzing are generated in order to have a good grasp of the data quality. In this way, we can then be sufficiently vigilant in detecting the inherent data artifacts to avoid making spurious conclusions.

Interaction data from traditional small-scale experiments are generally more

reliable because their biological relevance is often very thoroughly investigated by the researchers. In fact, the published results are oftentimes based on repeated observations by multiple research groups. The current explosive rate of data generation fueled by the powerful high-throughput interaction detection technologies has made it impractical for their verification by traditional methods in small scale experiments. Nevertheless, we can still generate high-quality interaction data by using an integrative approach. In the von Mering study, [864] interactions confirmed by two or more detection methods are found to have a higher percentage of true positives than those that are detected by only individual methods. Interactions confirmed by three or more detection methods have an even higher degree of accuracy. This means that in order to generate an accurate map of the interactomes, each experiment indicating a particular protein-protein interaction must be confirmed by at least two or more independent means, computationally and/or experimentally. Fortunately, as we have shown in this chapter, the concerted efforts by the industrious biologists and bioinformaticians have already resulted in a wide array of methods for discovering protein-protein interactions in high throughput, each with its own strengths and specialties. These methods, together with the continuing efforts by investigators in developing and refining further innovative interaction discovery techniques—for example, automatic text mining for discovering annotated protein interactions from the literature, [536, 543, 608] and the formulation of mathematical measures for assessing the reliability of protein-protein interactions in terms of the underlying interaction network topology, [734, 735] to name just a couple—will, in the near future, lead us to a complete and accurate map of the interactome.