REVIEW

How advancement in biological network analysis methods empowers proteomics

Wilson W. B. Goh[1], Yie H. Lee[2], Maxey Chung[3] and Limsoon Wong[4]

[1] Department of Computing, Imperial College London, UK
[2] Singapore-MIT Alliance for Research and Technology, Singapore
[3] Departments of Biochemistry and Biological Sciences, National University of Singapore, Singapore
[4] Departments of Computer Science and Pathology, National University of Singapore, Singapore

Proteomics provides important information – that may not be inferable from indirect sources such as RNA or DNA – on key players in biological systems or disease states. However, it suffers from coverage and consistency problems. The advent of network-based analysis methods can help in overcoming these problems but requires careful application and interpretation. This review considers briefly current trends in proteomics technologies and understanding the causes of critical issues that need to be addressed – i.e., incomplete data coverage and inter-sample inconsistency. On the coverage issue, we argue that holistic analysis based on biological networks provides a suitable background on which more robust models and interpretations can be built upon; and we introduce some recently developed approaches. On consistency, group-based approaches based on identified clusters, as well as on properly integrated pathway databases, are particularly useful. Despite that protein interactions and pathway networks are still largely incomplete, given proper quality checks, applications and reasonably sized data sets, they yield valuable insights that greatly complement data generated from quantitative proteomics.

1 Introduction

Mass spectrometry (MS)-based proteomics is a widely used and powerful tool for profiling systems-wide protein expression changes [1]. It can be applied for various purposes, e.g. biomarker discovery in diseases and study of drug responses. Although RNA-based high-throughput methods have been useful in providing glimpses into the underlying molecular processes, the evidences they provide

are indirect. Furthermore, RNA and corresponding protein levels have been known to have poor correlations [2]. On the other hand, MS-based proteomics tend to have consistency (poor reproducibility and inter-sample agreement) [3, 4], and coverage [5] (inability to detect the entire proteome) issues that need to be urgently addressed. The former of which implies that multiple analytical runs of the same sample under constant experimental conditions will result in the detection of different but overlapping sets of proteins. Intuitively, this means more LC-MS/MS runs are required to identify a sufficiently large portion of any proteome and is intricately linked to the second issue of inadequate proteome coverage.

Proteomics captures valuable information about the level and existence of individual proteins but the data can be noisy and incomplete. As mentioned, two exigent issues in proteomics are data coverage and consistency. Experimental methods to overcome these issues are technically challenging,

Correspondence: Dr. Limsoon Wong, School of Computing, National University of Singapore, Building COM1, 13 Computing Drive, Singapore 117417, Singapore
E-mail: wongls@comp.nus.edu.sg
Fax: +65-6779-7465

Abbreviations: CEA, Clique Enrichment Analysis; **GIN,** genetic interaction network; **GO,** Gene Ontology; **MN,** metabolic network; **PEP,** proteomics expansion pipeline; **PPIN,** protein–protein interaction network; **RN,** regulatory network; **Y2H,** yeast two hybrid

Colour Online: See the article online to view Fig. 1 in colour.

resource heavy or place an unreasonable heavy dependency on the quality of the initial data set. These include exhaustive fractionation of samples [6, 7], repeated MS runs of the same sample to reach saturation [4, 8] and compilation of MS data specific to a sample type generated and archived from different laboratories [9–11].

The problems are particularly exemplified in a large-scale collaborative study to assess the extent of reproducibility across different laboratories. The results were striking – only 7 out of 27 laboratories correctly reported all 20 proteins, and only 1 laboratory successfully reported all 22 unique peptides [3].

Therefore, alternative approaches are needed to complement the existing experimental approaches to circumvent the stochastic sampling of peptides by MS and increase the comprehensiveness of proteome coverage. Networks provide an informative background or scaffold on which higher confidence assertions can be founded.

A biological network is a set of molecules, e.g. proteins or genes, that are linked together via defined functional relationships. The inter-connections between molecules contain a wealth of information that has yet to be fully exploited in network-based analysis. Deciphering the patterns of wiring in a system allows us to penetrate the apparent complexity, and understand how these wirings could result in coordinated function. Early discoveries suggest that biological networks share common properties with many other natural and man-made systems. For example, it was reported that protein–protein interaction networks (PPINs) are scale-free [12], small-world [13] and disassorted [14]. It was also suggested that highly connected proteins (hubs) were more likely to be essential for cellular survival [15]; and that there were two kinds of hubs – date and party [16].

As our ability to exploit network information improves, some of these early observations are beginning to come under intense scrutiny and revision – especially since they were performed by relatively crude methods that do not capture enough of the complexity underlying biological processes. For example, the existence of date and party hubs [17], or that hubs are also more likely to be essential genes [18], is increasingly disputed. The Barabasi–Albert model, while elegant, does not capture the notion that biological molecules tend to work in complexes or clusters [19]. As of now, there is still no perfect mathematical model for generating a biological network.

Given that network-based analysis methods are still evolving, they must be applied appropriately in order to gain confident biological insight. Network-based analysis in biology is mostly limited to areas where data are more readily accessible or interpretable. Hence, protein–protein interactions, gene-regulation and metabolic systems are more widely studied although, strictly, they are not distinct systems in themselves. A fortunate development is that recent experimental initiatives have increased tremendously the amount of biological network information available about which to perform analysis. For example, groups such

as Marc Vidal's [20] have been generating large-scale yeast two hybrid (Y2H) data in order to build extensive PPINs for model organisms. Also noteworthy are the ascension of large pathway and metabolic databases, as well as integrative platforms.

Currently, not much is known about the true topology of biological networks. And even less is known about how errors such as false positives can adversely affect analysis. Combining networks to include several different types of molecules (e.g. proteins, RNA and metabolites) and interactions (e.g. protein interaction, gene interaction and signaling) to capture various levels of biological complexity is an even taller order.

Despite these difficulties, the theory of networks is an essential next stage in the study of biology. Traditional reductionist methods, while excellent in the study of the individual components of the system, cannot yield its emergent qualities. And it is at the systems level where knowledge on coordination, regulation and control of biological processes can be obtained. Currently, it is increasingly recognized that the understanding of properties that arise from whole-cell function require integrated, theoretical descriptions of the relationships between different cellular components [12].

In this review, we begin in Section 2 with a brief introduction to proteomics methods, protein identification algorithms and experimental planning to suit network-based analysis. This is followed in Section 3 by a general introduction to different types of biological networks. Section 4 covers in greater detail the coverage and consistency issues and why network-based approaches are suitable. Sections 5 and 6 introduce some methods for dealing with the former and latter issues, respectively, with a focus on protein-interaction networks and biological pathway networks. In Section 7, we take issue on the importance of data quality of the reference network, and what are the caveats to note in order to maximize analytical outcome.

2 Quantitative advances in proteomics and algorithms for protein identification

2.1 Proteomics methods

Proteomics can be pursued in many different flavors (e.g. 2D gels and shotgun proteomics) and forms (e.g. protein structures, activities, expressions and interactions). 2D gels were traditionally favored but lack reproducibility and are resource heavy [21].

Recent MS-based methods have higher sensitivity, increased throughput and greater automation. These include shotgun proteomics, differential isotope labeling, label-free quantitation and targeted proteomics. Further details and other recent advances in MS technologies can be found in, e.g. the review by Mann and Kelleher [22].

2.2 Bioinformatics for peptide and protein identification

The detection of a peptide and the determination of its amino acid sequence can be done using two types of algorithms. The first type is database search algorithms that work by matching the mass spectrum of the peptide to a database of known peptide sequences. Examples of these algorithms include MASCOT [23], Protein Prospector [24], SEQUEST [25] and Paragon [26].

A second type performs de novo sequencing of peptides from mass spectra. Examples of these algorithms include PEAKS [27], ADEPTS [28], Lutefisk [29], PepNovo [30] and GST-SPC* [31].

3 Types of biological networks

A biological network is a simplified model that describes the inter-relationships between a set of functional entities such as genes, proteins or metabolites. For the purpose of this review, we broadly regard the followings as biological networks: metabolic pathways (MNs), regulatory pathways (RNs), protein–protein interactions (PPINs), genetic interactions (GINs), protein complexes and proteins annotated to the same Gene Ontology (GO) terms.

MNs link two proteins in a directed relationship if the product of one is the substrate of the other. RNs refer to transcriptional relationships or other indirect relationships where one protein controls the expression or repression of the other. MNs and RNs are thus natural biological pathways. Popular databases of MNs and RNs include KEGG [32], BioCyc [33], WikiPathways [34], Reactome [35], Ingenuity® Knowledge Base (http://www.ingenuity.com), NetPro™ (http://www.molecularconnections.com), Pathway Commons [36] and PathwayAPI [37].

In PPINs, a relationship between two proteins exists if they are experimentally verified to interact physically. In GINs, a gene interacts with another if a combined mutation between them results in a more severe phenotype as opposed to a single mutation in either of them. A GIN may imply a physical interaction (as part of a complex) or a complete ablation of functions across two compensatory pathways. GINs are only beginning to be better understood but remain difficult to study empirically; see Dixon et al. [38] for an excellent review on GINs. Unlike MNs and RNs, PPINs and GINs are purely pairwise interaction information and cannot yet be put into the context of a natural biological pathway. Important databases of PPINs and GINs include BioGRID [39], DIP [40], HPRD [41], IntAct [42], MINT [43] and STRING [44].

The GO was established by the Gene Ontology Consortium as an important reference terminology for annotating the function and cellular localization of proteins [45]. GO terms are organized into three separate hierarchical ontologies — viz., cellular component terms (CC), molecular function terms (MF) and biological process terms (BP). A protein that is annotated by a particular GO term is considered to be annotated by all ancestor terms (in the corresponding hierarchical ontology) of that GO term; that is, the so-called "through-path" rule is applied. Associated with the GO is a large and well-organized database of proteins annotated to GO terms. In particular, when a group of proteins are annotated to a CC, BP or MF term, it means that this group of proteins is localized to that cellular compartment (corresponding to the CC term), participate in that biological process (corresponding to the BP term), or participate in that molecular function (corresponding to the MF term), respectively.

Protein complexes and proteins annotated to the same GO terms are not actually networks. Nevertheless, proteins that are in the same complex or annotated to the same GO terms are functionally linked and can be considered to form functional linkage networks. The larger databases of protein complexes include CORUM [46], MIPS [47] and CYC2008 catalogue [48].

4 Two issues in proteomic profile analysis that call for a more holistic analysis based on biological networks

In this section, we highlight two important issues in proteomic profile analysis that need to be addressed and suggest a more holistic proteomic profile analysis utilizing biological networks and pathways.

The first issue concerns the coverage of the proteome at the level of an individual sample. In particular, even as the advancement of MS technologies continues, certain limitations to current proteomics approaches remain that hamper the complete mapping of the proteome in a sample. Like many high-throughput methods, proteomics data are noisy. Furthermore, due to demanding technological and manpower requirements, as well as limited sample availability, often there are few repeats to guarantee that the results are not false positives due to chance. Consequently, stringent score thresholding is generally used in various steps of peptide detection and identification to reduce noise. However, more stringent thresholds also reduce coverage of the proteome. For example, a relevant protein may escape reporting because it does not meet a required threshold on its dynamic range. A relevant protein may also escape detection because it does not meet a require threshold on its signal intensity, perhaps due to imperfect prediction of MS-amendable transitions [49, 50].

The second issue concerns the consistency of proteomic profiles at the phenotype level across samples. To understand proteome biology and/or for the discovery of biomarkers, quantitative comparisons – e.g. of cancerous and non-cancerous samples – are an important aspect of proteomics [51]. Analogous to DNA/RNA microarrays and common to all the labeling methods mentioned earlier,

protein quantification is usually expressed as fold change ratio. The traditional post-MS analysis approach is therefore to select and study only those proteins that are found in most of the samples of the phenotype in question and have a consistently over-expressed or under-expressed ratio. However, proteins with noticeably high or low expression are not necessarily causal or important. At the same time, a mutated protein that drives other proteins to change their levels may not itself report any change in expression or may miss being detected. Moreover, many relevant proteins report "swing" ratios, that is, a mixture of both high and low ratios across samples. These factors are further compounded by the noise and coverage of the proteome at the level of individual samples. Hence, one often fails to find key proteins, much less biomarkers that are consistent and reproducible across different batches of samples.

Proteins usually function as combinatorial units. At a fine granularity, these units are protein complexes; at a coarser granularity, these units are biological pathways. We shall generically refer to these combinatorial units of proteins as "biological networks".

Biological networks are critical to understanding the function of genes and proteins in a more holistic way. Thus, the appearance in recent years of many databases containing information about biological networks may offer innovative solution to the two issues above.

As proteins in the same functional unit – e.g. a protein complex – interact with each other in some manner, these proteins are expected to be expressed in a correlated or coordinated manner. Therefore, it is reasonable to postulate that detected proteins in a proteomic screen that form a known functional unit are likely to be involved in biological function, while isolated proteins are noise. This postulate can be applied to improve coverage of a proteomic screen and remove noise.

For illustration, let A, B, C, D and E be the five proteins that function as a group and thus are normally correlated in their expression. Suppose only A is detected in a proteomics screen and B–E are not detected. Suppose also that the screen has 50% reliability. Then, A's chance of being false positive is 50%, whereas the chance of B–E being all false negatives is $(50\%)^4 = 6\%$. Hence, it is almost ten times more likely that A is noise than B–E all being missed. Conversely, suppose only A is not detected and all of B–E are detected. Then, A's chance of being false negative is 50%, whereas the chance of B–E all being false positives is $(50\%)^4 = 6\%$. Hence, it is almost ten times more likely that A is false negative than B–E all being false positives.

Each biological state – e.g., in disease – generally has some underlying causes. Thus, it is reasonable to postulate that there should be some unifying biological themes – certain biological networks or subnetworks – for genes and proteins that are truly associated with the state [52–54]. Hence, the uncertainty in the reliability of the selected proteins from quantitative comparisons of disease and non-disease samples can be reduced by considering the mole-

cular functions and the biological processes associated with the genes and proteins [55]. Such a unifying biological theme is also a basis for inferring the underlying cause of the disease phenotype.

For illustration, let there be three disease samples and three controls. Assume that the chance of an arbitrary protein found to be highly expressed in an arbitrary sample is 50%. Then, a group of five functionally linked proteins that is perfectly correlated to these two groups of samples – e.g. they are all highly expressed in the three disease samples and not in the three controls – has $((50\%)^3 \times (1–50\%)^3)^5 = 9.3 \times 10^{-8}\%$ chance of being a false-positive group. On the other hand, if just one of these five functionally linked proteins was perfectly correlated to the two phenotypes, its chance of being a false positive would be $(50\%)^3 \times (1–50\%)^3 = 1.6\%$, which is many orders of magnitude higher than when all five proteins are simultaneously correlated with the two phenotypes.

Furthermore, network-based approaches to proteomic profiles analysis are able to significantly reduce the number of samples needed in a proteomic study. To appreciate this, let us illustrate with the following a simplified scenario. Assume again that an arbitrary protein has equal chance to be up or down-regulated in a sample. Suppose that there are $2n$ samples, with n samples in each of the two phenotypes. Suppose also that there are 1000 proteins being tested in each sample. Then, for a simple method that tests each protein individually, the random chance of a protein that is perfectly correlated with the two phenotypes is $(1/2)^n \times (1/2)^n$. Thus, the expected number of false-positive genes that are perfectly correlated with the phenotypes is $1000 \times (1/2)^{2n}$. In contrast, for a method that tests a group of proteins at a time, the random chance of a group of k genes that are perfectly correlated with the phenotypes is $((1/2)^n \times (1/2)^n)^k$. In theory, there are $^{1000}C_k$ possible groups of k genes; and so the expected number of false-positive groups of k genes is $(1/2)^{2nk} \times 1000!/(k! \times (1000-k)!)$. In practice, the group-based methods that we will describe (e.g. FCS, GSEA) do not test all possible groups. Instead, they define each pathway in a database to be a group; and they only test these groups. As a typical pathway database has <1000 pathways, the expected number of false-positive groups of k genes is reduced significantly to $1000 \times (1/2)^{2nk}$. Since $1000 \times (1/2)^{2nk}/1000 \times (1/2)^{2n} = (1/2)^k$, we can estimate that, given the same number of samples, the group-based methods achieve $(1/2)^k$ times less false positives than individual-gene methods. Conversely, to achieve the same number of false positives, the number of samples needed by group-based methods is $(1/2)^k$ times less than that needed by individual-protein methods. For example, at $k = 5$, the number of samples needed by group-based methods is $(1/2)^5 = \sim3\%$ that of individual-protein methods, while delivering a comparable level of sensitivity and specificity.

Clearly, leveraging on these network-based paradigms can aid in circumventing some of the shortcomings of current proteomics approaches mentioned.

5 Improving coverage using biological networks

There are cases where the mass spectra may identify some particular proteins, but, because their scores are below the defined cutoff threshold, may not be reported initially in the first round of data analysis. This occurs frequently in the tradeoff between sensitivity and specificity in precursor ion selection for fragmentation. Other potential reasons why these proteins are unreported include: (i) not satisfying the minimum two unique peptides requirement for confident protein identification – that is, the protein is identified by a single peptide; (ii) the proteins are short in amino acid composition and subsequently are identified only by short peptides; and/or (iii) they are not consistently found in patient samples.

Network-based analysis can allow expansion of the detected proteome to uncover and/or discover novel proteins. This is critical in recovering missing proteins in known pathways or complexes. It is even more important in uncovering less-abundant proteins commonly shrouded in shotgun proteomics.

A simple network-based method is to use a database of protein complexes and identify those complexes that have a large overlap with the initial list of detected proteins. A significance value can then be calculated via generating randomized clusters of equal size to the cluster. If significant, then the rest of the proteins in the complex are postulated as likely to be present. This method is also referred to as functional class scoring [56].

More sophisticated methods that build on this principle include Clique Enrichment Analysis (CEA) [57], Maxlink [58], shortest-path analysis [59] and the method of Goh et al. [60] (which we call Proteomics Expansion Pipeline (PEP) here). Regardless of the methods used, they are all a form of "guilt by association". Hence, the list of recovered proteins should be validated using some additional evidence. The most direct evidence is by returning to the original mass spectra to verify the quality of the corresponding *y*- and *b*-ion assignments [60]. Proteins with low copy numbers and high cellular turnover such as transcription factors and some protein kinases may still not be located through retrospective assessment of the original MS/MS data. Therefore, other validation methods such as immunological assays may be used on interesting targets. A less direct evidence is to check whether these recovered proteins are annotated to a list of GO terms that are enriched in the initial list of high-confidence proteins [58]. Another indirect evidence is using databases of gene expression profiles – e.g. Human Protein Atlas [61] – to check whether these recovered proteins show a pattern of differential expression between relevant disease samples and normal samples that is similar to that shown by the initial list of high-confidence proteins [58].

5.1 CEA

The simple network-based method suggested earlier is to shortlist non-confidence proteins in protein complexes that contain many high-confidence proteins. However, the number of known protein complexes available in protein complex databases such as CORUM [46] is still small. So, one should supplement them with predicted protein complexes and functional modules.

An example that pursues this route is the CEA proposed by Li et al. [57]. CEA generates cliques – that is, fully connected subnetworks – from a PPIN. Those cliques that are enriched with high-confidence proteins are considered detected. Non-confident proteins in these cliques are thus rescued. The use of cliques from PPINs is reasonable because cliques in a PPIN often correspond to proteins at the core of complexes [62].

5.2 PEP

About 70–80% of proteins share at least one biological process or function with their interaction partners in PPINs and GINs [63]. A protein is also often observed to participate in a biological process or function that is over-represented in its interaction partners [64, 65]. More generally, proteins that are connected or proximal within a biological network often form a functional unit [66]. On the basis of these observations, many algorithms have been developed for predicting protein complexes and functional modules from PPINs and GINs – e.g. MCL [67], MCODE [62], RNSC [68], CFinder [69, 70], PCP [71] and CMC [72]. These more powerful algorithms can be used in place of clique finding in CEA.

A more recent method that uses a powerful protein complex prediction algorithm is that proposed by Goh et al. [60]. We call this method the PEP. PEP first identifies the group of high-confidence proteins from the proteomic screen. It then maps these proteins to nodes in a large integrated PPIN. Next, it generates an expanded subnetwork by taking the immediate neighbors of these seeds in the PPIN. The subnetwork is then clustered using CFinder [69], which overlaps closely related cliques. Each cluster is then ranked based on the average expression value of the proteins it contains. Proteins (in high-ranking clusters) not found in the proteomics screen are then screened against the original mass spectra for evidence of existence.

A notable aspect of PEP is the PPIN that it uses. The PPIN is one of the most comprehensive to date. It comprises data from HPRD [41], BioGRID [39], IntAct [42] and DIP [40], as well as data from the literature [73, 74]. While combining PPINs improves coverage of the protein interactome, it may also compound the noise present in them [75]. PEP uses the iterated Czekanowski-Dice distance (CD-distance) technique from CMC [72] to eliminate potential noise edges from the integrated PPIN. Although the CD-distance technique assesses the reliability of an edge in a

PPIN purely based on the local topology of the edge, it is very effective. In particular, while this method eliminates about 50% of the edges from the integrated PPIN, it doubles the level of functional and localization coherence in the remaining edges in the PPIN.

5.3 Maxlink

PPINs have a fairly high level of false positives and false negatives [75]. This has an impact on the sensitivity of clique finding and other protein complex prediction algorithms mentioned earlier. For example, a single missing edge in the PPIN is sufficient to exclude a protein from a clique in clique finding.

To achieve greater sensitivity, rather than requiring an entire protein complex to be predicted before testing for enrichment in high-confidence proteins, one can test for a more relaxed condition. In particular, one can instead test whether a protein is likely to be part of the same complex with a group of already known high-confidence proteins, without requiring knowing what the other proteins in the complex are.

Maxlink is a method for identifying cancer genes introduced by Ostlund et al. [58]. Although not explicitly tested on proteomics data, it can be considered as an example that follows this more relaxed route. Maxlink first requires the identification of a set of high-confidence seeds. It then generates, scores and ranks a list of new candidates based on the number of links in FunCoup [76] (which is a PPIN database) to the seed set. The more the number of connections to seeds, and the less the number of connections to non-seeds, the higher the score. This approach is justified because a protein is often observed to participate in the same biological process, biological function or protein complex that is over-represented in its interaction partners [64, 65]. Moreover, proteins in the same complex are thought to have more interactions between themselves than with proteins outside the complex [77].

5.4 Shortest-path network analysis

In a related approach, Managbanag et al. [59] proposed using shortest paths to recover genes that lie between two high-confidence seeds. In their study, they first define a set of seeds previously reported to be associated with the disease in question. They then extract a shortest-path composite network from PATHWAY STUDIO 5.0, a commercial PPIN database and software suite [78].

This approach is based on the hypothesis that proteins connecting pairs of other proteins with a well-defined biological function have a higher probability to share that function than randomly selected proteins [79]. This hypothesis is partially justified by the observation that most proteins share at least one function with their interaction partners [63] in a PPIN and thus transitively with the partners of these partners [80]. However, the longer a (shortest) path gets, the more false positives it inevitably contains [81].

6 Improving consistency using biological networks

Quantitative comparison of samples is central to proteomics. However, biomarkers identified in one batch are quite often not consistent and not reproducible in another batch of samples. This is likely due to (i) the noise and coverage of the proteome at the level of individual samples and (ii) limitation of current statistical techniques as a result of insufficient sample size.

In order to qualitatively improve the statistical power of proteomic analysis methods and the reliability of the results, additional dimensions present in the problem have to be brought into consideration. In particular, current paradigms suggest that protein interactions constitute a major part of all cellular processes. The extent of interactions between proteins denotes shared functionality [82], complex or submodule participation [83] and/or co-expression [84]. In the case of metabolic and biochemical relationships, extensive validation studies have established with higher confidence relationships between proteins in a pathway; and it is reasonable to postulate shared functionalities between such proteins even though, in pathways, an edge can mean different things such as regulation or signaling. Thus, a comparative proteomic profile analysis that incorporates such information from biological networks, as suggested earlier, is useful in identifying results that are more consistent, more reproducible and more biologically coherent.

An analogous situation exists in gene expression profile analysis. Many approaches [85–87] have been proposed for identifying differentially expressed genes useful for diagnosis of diseases and prognosis of treatment response. However, these methods often produce gene lists that are inconsistent when they are applied to different data sets of the same disease phenotypes [88]. For example, for a pair of data sets involving prostate cancer [89, 90], Zhang et al. [91] show that the two lists of significant genes identified by running SAM [87] independently on the two data sets have a low overlap of 30% in their top 10 genes and an even lower 15% overlap in their top 100 genes. In order to overcome the uncertainty in the reliability of the selected genes, over the years, the gene expression analysis community has developed powerful methods that analyze gene expression profiles with respect to biological networks. Although gene expression (DNA and RNA) does not always directly correlate with protein expression [92], gene co-expression is something proven at the protein level, especially when it comes to an induction of a particular function [93]. Hence, some of these methods from the gene expression community can be adapted for proteomic profile analysis.

In the following subsections, we briefly introduce three types of approaches – viz., overlap analysis, direct group analysis and network-based analysis – for identifying significant pathways from the gene expression analysis community. We also briefly describe approaches for identifying and characterizing significant novel protein clusters.

6.1 Overlap analysis

Overlap analysis methods are well known. A list of differentially expressed genes or proteins is first determined. This list is then intersected with each biological pathway (usually a protein complex, MN or RN) in a database. The statistical significance of the overlap is computed using, e.g., the hypergeometric test. The subsets of differentially expressed genes that have a statistically significant intersection with a pathway are declared candidate biomarkers. ORA [94] is a representative of overlap analysis methods.

These methods have a shortcoming in that they are sensitive to the thresholds used in determining the differentially expressed genes or proteins.

6.2 Direct group analysis

Direct group analysis methods work on a different principle to avoid the shortcoming above. In direct group analysis, each reference biological pathway (usually an MN, RN or protein complex) is checked to establish whether the pathway is differentially expressed as a whole. This is achieved by comparing the distributions of expression values of genes and proteins on the pathway with the distributions of expression values of all the other genes and proteins, e.g. by a weighted Kolmogorov–Smirnov test. FCS [95] and GSEA [96] are examples of the direct group analysis methods.

These methods are able to detect more subtle changes in gene and protein expression profiles. For example, if the majority of genes and proteins on the biological pathway have small but correlated expression level changes, then they can still result in a high statistical significance of the biological pathway under a direct group analysis method. Nevertheless, direct group analysis methods have a key shortcoming in that they work on a whole-pathway basis. Thus, they are unable to declare a large pathway to be significant when only a small subnetwork within that pathway is truly responsible for the disease phenotype.

6.3 Network-based analysis

Network-based analysis methods [52–55, 97] are newer developments in gene expression analysis. The advantage of these methods is that, rather than using pathways as a whole, they identify subnetworks that are significantly differentially expressed. Although gene expression (DNA and RNA) is known not to correlate directly with protein expression, the concepts behind these network-based techniques are applicable to proteomics profile analysis.

An early example of these network-based methods is NEA [55]. NEA extracts from each biological pathway (usually an MN, RN or PPIN) a set of subnetworks, by treating each regulator in a pathway and all its direct targets in the pathway as a separate group. Each such subnetwork is then tested – using a direct group analysis method like FCS or GSEA – to see whether the genes and proteins in the subnetwork are differentially expressed as a whole. A significant subnetwork potentially provides a more precise hypothesis that explains the disease phenotype than an entire pathway. A shortcoming of NEA is that it tends to produce small subnetworks as each subnetwork comprises only a regulator and its immediate regulatees.

The latest addition to this family of methods is SNet [97], which is able to find larger subnetworks than NEA. SNet first maps the genes or proteins that are highly expressed in most samples of the disease phenotype in question to biological pathways (usually MNs, RNs or PPINs). It then discards other genes and proteins in these pathways and networks, causing these pathways to fragment into separate subnetworks. The subnetworks are scored against the disease cases and the controls. Those subnetworks showing a significant difference in scores between cases and controls are declared significant. Experiments have shown that SNet produces subnetworks that are both much more substantial in size and much more consistent cross independent data sets of the same disease phenotypes than other methods [97]. A disadvantage of SNet in the proteomics context – compared with NEA, FCS, GSEA, etc. – is that it requires the subnetworks to be scored against individual samples; thus it may not be straightforward to adapt SNet for situations where samples are pooled.

6.4 Identifying and characterizing novel protein clusters

The methods mentioned earlier – viz., ORA, FCS, GSEA, NEA, SNet, etc. – are dependent on both the quality and comprehensiveness of the reference pathway databases. Hence, they cannot yield good result if the underlying cause of the disease phenotype is a novel functional module or pathway. So, they need to be complemented by methods for identifying and characterizing novel functional modules.

A simple approach for identifying novel functional modules is to first map the differentially expressed proteins to a PPIN. Then, a protein complex prediction method is run on the mapped portion of the PPIN to produce a list of predicted protein clusters, each comprising some subsets of the differentially expressed proteins. These protein clusters are potentially novel protein complexes and functional modules. After that, these predicted protein clusters are characterized using some form of GO term analysis.

For the protein complex prediction step, there is no dearth of methods. A detailed review covering newer methods can be found in [98]. So, we just briefly describe a few easily accessible methods here. CFinder is based on the clique percolation method described by Palla et al. [69]. It relaxes the constraint on cluster definition by first identifying cliques and then scoring those that overlap using a standard component analysis procedure. MoNet is an implementation of the Girvan–Newman method based on betweenness centrality [99]. MCL is based on the Markov clustering method [67]. CMC works by generating maximal cliques from the cleansed network and then merges or removes highly overlapping cliques based on their interconnectivity [72].

For the GO term analysis step, it is often performed using tools based on the hypergeometric test. Examples include GO East [100] and GO Term Finder [101]. These tools essentially test predicted protein clusters against the reference protein sets defined by GO terms. If a predicted protein cluster is enriched in some GO terms, the proteins in the cluster can be considered to consistently show a function described by these GO terms. However, many times, given the incompleteness of GO annotations and the complexity of the GO tree structure, the returned GO term lists can be perplexing and difficult to analyze. Many significant GO terms may also be returned; this creates a misleading picture that the cluster is heterogeneous when, in fact, many of the returned GO terms could be closely related.

There are other methods that can improve the resolution of GO analysis. The two simplest are the parent–child method [102] and the intuitive "informative GO term" method [103].

The parent–child method proposed by Grossmann et al. [102] modifies the hypergeometric test statistics. Instead of the standard hypergeometric distribution, they propose using

$$P(\sigma_t = k | \sigma_{\mathrm{pa}(t)} = n_{\mathrm{pa}(t)}) = \binom{m_t}{k} \binom{m_{\mathrm{pa}(t)} - m_t}{n_{\mathrm{pa}(t)} - k} / \binom{m_{\mathrm{pa}(t)}}{n_{\mathrm{pa}(t)}}$$

(1)

where t is the GO term that we need to establish whether it is enriched in the predicted protein cluster; m_t is the number of proteins in the GO database that are annotated to t; $m_{\mathrm{pa}(t)}$ is the number of proteins in the GO database that are annotated to the parent terms of t; and $n_{\mathrm{pa}(t)}$ is the number of proteins in the predicted protein cluster that are annotated to the parent terms of t. This approach reduces the dependencies between individual term's measurements and avoids producing false positives due to inheritance problems [102], thereby increasing the stringency for significance reporting.

The "informative GO terms" method decreases the number of terms reported by introducing a threshold on the GO tree itself. Only terms that are annotated to at least 30 genes, and each of whose direct child has no more than 30 genes, are considered informative. This way, each GO term considered is at the finest resolution possible while being annotated to a sufficiently large number of proteins for a valid analysis [66]. This also has the effect of reducing redundancy on GO terms reported as a whole [103].

7 Use of biological networks: What to watch out for

The use of biological network databases for improving proteomics analysis is very promising. Nevertheless, we should be aware of a number of caveats, especially with respect to the reliability and completeness of these databases.

7.1 Reliability of PPINs

The databases of PPINs and GINs have grown rapidly in size over the years, with improved methodologies in testing protein interactions; see Fig. 1. The prominent PPIN and GIN databases include HPRD, BioGRID, MINT, IntAct, STRING and DIP; see Table 1 for details. It should be noted that STRING corresponds more to protein functional associations than to physical protein interactions.

In spite of the growth of PPIN databases, it is difficult to ascertain quality. In fact, given high false-positive rates in Y2H and other binding experiments, up to 70% of the reported edges may be false [104]. Mark Vidal and co-workers tried producing higher quality all-against-all experimental data [73, 74], by testing all possible protein pairs in their data set using Y2H. However, these data sets are a select subset of the entire proteome, and are not reflective of the whole PPIN. It also does not eliminate false positives reported by Y2H.
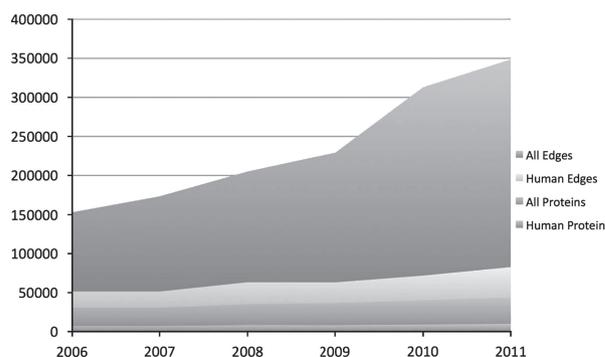
Figure 1. Growth of BioGrid from 2006 to current. All data points are taken in July except in 2011 (taken in April). While the growth in human protein interaction has been steady, it does not significantly contribute to the large growth in recent years. This is in part due to the incorporation of new species and data from other model organisms.

Table 1. Databases of protein–protein interaction networks

Database	# Nodes, # edges	URL	Build focus	References
BioGRID	10k, 40k	http://thebiogrid.org	Literature	[39]
DIP	2.6k, 3.3k	http://dip.doe-mbi.ucla.edu	Literature	[40]
HPRD	30k, 40k	http://www.hprd.org	Literature	[41]
IntAct	56k, 267k	http://www.ebi.ac.uk/intact	Literature	[42]
MINT	30k, 90k	http://mint.bio.uniroma2.it/mint	Literature	[43]
STRING	5200k, ?	http://string-db.org	Literature, prediction	[44]

Using a poor-quality PPIN is likely to skew analytical outcome. Network coverage needs to be sufficiently extensive in order to enhance resolution. In recent works, it is common to merge data sets across various sources [57, 105]. However, simple integration may lead to compounded errors for which confidence is not certain due to different or poorly defined study parameters.

A walk around this problem, as demonstrated by Bossi and Lehner [105], is to repeat the analysis on two networks and check for consistency. The first is a lower confidence construct using edges supported by at least one publication source. The second is a higher confidence construct using edges supported by at least two publications. However, experiment-based filtering is biased, and two papers utilizing the same flawed technique may also give rise to the same erroneous result. Hence, more robust methods for evaluating the network quality are needed.

A good way to assess the reliability of an edge in a PPIN is based on GO term coherence. That is, we check whether the two proteins connected by that edge are annotated to an informative GO term in common [106]. The overall reliability of a PPIN can in turn be assessed based on the fraction of its edges that have coherent GO term annotations. This approach is reasonable because two interacting proteins should be in the same cellular compartment (i.e. share an informative CC term) and participate in the same biological function or process (i.e. share an informative MF or BP term) [80, 107]. Limitations of this method include incomplete GO term annotation, unresolved bona fide localization of proteins, and the dynamic distribution of proteins in different physiological states.

Another way to assess the reliability of an edge in a PPIN is based on the hypothesis that if two proteins interact, it is also likely that they share common neighbors in the PPIN. This hypothesis follows naturally from the more fundamental postulate that proteins usually function as a group. One early example of this "topological" approach is given by the CD-distance, which is calculated as the number of interaction partners shared between two proteins divided by the set of interaction partners of both proteins [72]. Other examples are surveyed in [106]. Since topological cleaning approaches rely on network intra-connectivity, they do not perform well on sparse networks. It is possible that improvements could be achieved via manifold embedding [108], or homologous transfer of edges [109].

A harder problem to resolve is the false negative problem – viz., true interactions that are not reported. Chua and Wong [106] and Shoemaker and Pachenko [110] provided detailed reviews on approaches for predicting novel protein–protein interactions, including protein primary structures and associated physicochemical properties [111], interacting domains [112], interacting motifs [113], gene-fusion events [114], coevolution of proteins or residues [115] and the topology of PPINs [116].

7.2 Completeness of biological pathway databases

The databases of MNs and RNs can be considered as more reliable than PPIN and GIN data sets due to higher levels of curation and experimental evidence. In today's research landscape, the major ones include single-lab curation efforts (KEGG, BioCyc), collaborating labs (WikiPathways, Reactome) and commercially compiled databases (Ingenuity®, NetPro™), as well as integrative databases that merge information from other databases. The details of these databases are given in Table 2.

It was a surprising revelation that none of the pathway databases proved comprehensive in terms of coverage. For example, comparison of human apoptosis pathway in humans between Ingenuity® Knowledge Base, KEGG and WikiPathways showed only a small 32–46% gene overlap and an even more alarming 11–16% edge overlap.

Soh et al. [37] demonstrated the difficulties associated with integrating pathway databases. Merging pathways via gene or reaction overlap proved inefficacious: A low threshold resulted in many false positives while too high produced many false negatives. Combining pathways via longest common substring match in pathway names (longest common substring, LCS) turned out to be a good compromise. However, Goh et al. [60] found that some redundancies still persist within and between databases during functional analyses. This suggests limitations in LCS that could be further improved and built upon in future works. Since pathway edges have been verified by expert knowledge and experimental verification, they likely have low false-positive rates. Hence, in combining same pathways across different databases, it is acceptable to simply take the union of their genes, proteins and reactions.

Table 2. Databases of biological pathways

Database	Remarks
KEGG	KEGG (http://www.genome.jp/kegg) is one of the best known pathway databases [32]. It consists of 16 main databases, comprising different levels of biological information such as systems and genomic. The data files are down-loadable in XML format. At time of writing it has 392 pathways
BioCyc	BioCyc ver 15 (http://http://biocyc.org) comprises over 1129 species-based databases [33]. An interesting feature of the BioCyc databases is that they are divided into three tiers, where tier 1 is high-confidence manually curated, tier 2 is computer generated with moderate curation, and tier 3 has minimal curation. BioCyc can be downloaded via BIOPax, SBML among other formats
WikiPathways	WikiPathways (http://www.wikipathways.org) is a Wikipedia-based collaborative effort among various labs [34]. It has 1627 pathways of which 369 are human. The content is downloadable in GPML format
Reactome	Reactome (http::/www.reactome.org) is also a collaborative effort like WikiPathways [35]. It is one of the largest datasets, with over 4166 human reactions organized into 1131 pathways by December 2010. Reactome can be down-loaded in BioPax and SBML among other formats
Ingenuity®	Ingenuity® Knowledge Base (http://www.ingenuity.com) is a repository of biological interactions accessible via its proprietary interface. Information is returned as an image file
NetPro™	Molecular Connections' NetPro™ (http://www.molecularconnections.com) is a commercial manually curated database. It contains more than 320 000 protein–protein interactions and small molecule–protein interactions across 20 organisms. Data can be downloaded in XML-format files or via SQL queries
Pathway Commons	Pathway Commons (http://www.pathwaycommons.com) collects information from various databases but does not unify the data [36]. It contains 1573 pathways across 564 organisms. The data are returned in BioPax format
PathwayAPI	PathwayAPI (http://www.pathwayapi.com) contains over 450 unified human pathways obtained from a merge of KEGG, WikiPathways and Ingenuity R Knowledge Base [37]. Data are downloadable as a SQL dump or as a csv file, and is also interfaceable in JSON format

Integration problems aside, there are specific problems associated with different pathway databases that still prove a challenge to resolve fully. For example, WikiPathways lack a stable and useful API. Extracting data from the coordinate-based XML file is also rather challenging. In Ingenuity® Knowledge Base, only image-based maps can be retrieved. In previous efforts, we used manual curation to extract the data. But this is inefficient and non-scalable if we need to expand coverage to other species.

8   Final remarks

The use of biological networks is an extremely powerful tool for enhancing proteomics analysis. Although protein clusters and metabolic pathways are topologically different, they should yield complementary results that can augment the functional characterization of the proteome.

Data quality is paramount in determining the resolution and power of analysis. Due to different coverages of various databases, it is advisable to use all available information for network construction. A caveat is that quality of information should also be checked. This can be performed by using measures such as GO term coherence, or topology-based edge scoring methods such as CD-distance.

Pathway databases are fragmented, and merging such information is harder than in PPINs. Although we addressed some of the inherent problems, more work remains to be done in ensuring higher quality data extraction and merging.

Another point to address is on expansion of the proteome. Given the fragmented nature of the recovered proteins, they usually give rise to a relatively sparse network. Shortest distance approaches, or identification of whether the differential protein belongs to a clique, followed by recovery of lower confidence proteins can help to alleviate the problem of data wastage. It can also better capture information about function based on clusters, rather than average function based solely on differential proteins.

The authors have declared no conflict of interest.

9   References

[1] Cox, J., Mann, M., Is proteomics the new genomics? *Cell* 2007, *130*, 395–398.

[2] Pascal, L. E., True, L. D., Campbell, D. S., Deutsch, E. W. et al., Correlation of mRNA and protein levels: cell type-specific gene expression of cluster designation antigens in the prostate. *BMC Genomics* 2008, *9*, 246.

[3] Bell, A. W., Deutsch, E. W., Au, C. E., Kearney, R. E. et al., A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat Methods* 2009, *6*, 423–430.

[4] Liu, H., Sadygov, R. G., Yates, J. R., III, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 2004, *76*, 4193–4201.

[5] White, M. Y., Brown, D. A., Sheng, S., Cole, R. N. et al., Parallel proteomics to improve coverage and confidence in the partially annotated *Oryctolagus cuniculus* mitochondrial proteome. *Mol. Cell. Proteomics* 2011, *10*, M110 004291.

[6] Wang, H., Clouthier, S. G., Galchev, V., Misek, D. E. et al., Intact-protein-based high-resolution three-dimensional quantitative analysis system for proteome profiling of biological fluids. *Mol. Cell. Proteomics* 2005, *4*, 618–625.

[7] de Godoy, L. M., Olsen, J. V., Cox, J., Nielsen, M. L. et al., Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* 2008, *455*, 1251–1254.

[8] Schrimpf, S. P., Weiss, M., Reiter, L., Ahrens, C. H. et al., Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol.* 2009, *7*, e48.

[9] Desiere, F., Deutsch, E. W., Nesvizhskii, A. I., Mallick, P. et al., Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* 2005, *6*, R9.

[10] Craig, R., Cortens, J. P., Beavis, R. C., Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* 2004, *3*, 1234–1242.

[11] Martens, L., Hermjakob, H., Jones, P., Adamski, M. et al., PRIDE: the proteomics identifications database. *Proteomics* 2005, *5*, 3537–3545.

[12] Albert, R., Scale-free networks in cell biology. *J. Cell Sci.* 2005, *118*, 4947–4957.

[13] Aloy, P., Russell, R. B., Taking the mystery out of biological networks. *EMBO Rep.* 2004, *5*, 349–350.

[14] Khor, S., Concurrency and network disassortativity. *Artif. Life* 2010, *16*, 225–232.

[15] Albert, R., Jeong, H., Barabasi, A. L., Error and attack tolerance of complex networks. *Nature* 2000, *406*, 378–382.

[16] Han, J. D., Bertin, N., Hao, T., Goldberg, D. S. et al., Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* 2004, *430*, 88–93.

[17] Agarwal, S., Deane, C. M., Porter, M. A., Jones, N. S., Revisiting date and party hubs: novel approaches to role assignment in protein interaction networks. *PLoS Comput. Biol.* 2010, *6*, e1000817.

[18] Zotenko, E., Mestre, J., O'Leary, D. P., Przytycka, T. M., Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality. *PLoS Comput. Biol.* 2008, *4*, e1000140.

[19] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., Barabasi, A. L., Hierarchical organization of modularity in metabolic networks. *Science* 2002, *297*, 1551–1555.

[20] Simonis, N., Rual, J. F., Carvunis, A. R., Tasan, M. et al., Empirically controlled mapping of the *Caenorhabditis elegans* protein–protein interactome network. *Nat. Methods* 2009, *6*, 47–54.

[21] Santoni, V., Molloy, M., Rabilloud, T., Membrane proteins and proteomics: unamour impossible? *Electrophoresis* 2000, *21*, 1054–1070.

[22] Mann, M., Kelleher, N. L., Precision proteomics: The case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. USA* 2008, *105*, 18132–18138.

[23] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, *20*, 3551–3567.

[24] Clauser, K. R., Baker, P., Burlingame, A. L., Role of accurate mass measurement ($\pm$10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* 1999, *71*, 2871–2882.

[25] Eng, J. K., McCormack, A. L., Yates, J. R., III, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[26] Shilov, I. V., Seymour, S. L., Patel, A. A., Loboda, A. et al., The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol. Cell. Proteomics* 2007, *6*, 1638–1655.

[27] Ma, B., Zhang, K., Hendrie, C., Liang, C. et al., PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 2003, *17*, 2337–2342.

[28] He, L., Ma, B., ADEPTS: Advanced peptide de novo sequencing with a pair of tandem mass spectra. *J. Bioinform. Comput. Biol.* 2010, *8*, 981–994.

[29] Taylor, J. A., Johnson, R. S., Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 1997, *11*, 1067–1075.

[30] Frank, A., Pevzner, P., PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* 2005, *77*, 964–973.

[31] Ning, K., Ye, N., Leong, H. W., On preprocessing and antisymmetry in de novo peptide sequencing: improving efficiency and accuracy. *J. Bioinform. Comput. Biol.* 2008, *6*, 467–492.

[32] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., Hirakawa, M., KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 2010, *38*, D355– D360.

[33] Krummenacker, M., Paley, S., Mueller, L., Yan, T., Karp, P. D., Querying and computing with BioCyc databases. *Bioinformatics* 2005, *21*, 3454–3455.

[34] Kelder, T., Pico, A. R., Hanspers, K., van Iersel, M. P. et al., Mining biological pathways using WikiPathways web services. *PLoS One* 2009, *4*, e6447.

[35] Vastrik, I., D'Eustachio, P., Schmidt, E., Gopinath, G. et al., Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.* 2007, *8*, R39.

[36] Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I. et al., Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.* 2006, *39*, D685–D690.

[37] Soh, D., Dong, D., Guo, Y., Wong, L., Consistency, comprehensiveness, and compatibility of pathway databases. *BMC Bioinformatics* 2010, *11*, 449.

[38] Dixon, S. J., Costanzo, M., Baryshnikova, A., Andrews, B., Boone, C., Systematic mapping of genetic interaction networks. *Annu. Rev. Genet.* 2009, *43*, 601–625.

[39] Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L. et al., BioGRID: A general repository for interaction datasets. *Nucleic Acids Res.* 2006, *34*, D535–D539.

[40] Xenarios, I., Salwinski, L., Duan, X. J., Higney, P. et al., DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 2002, *30*, 303–305.

[41] Prasad, T. S., Kandasamy, K., Pandey, A., Human protein reference database and human proteinpedia as discovery tools for systems biology. *Methods Mol. Biol.* 2009, *577*, 67–79.

[42] Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I. et al., The IntAct molecular interaction database in 2010. *Nucleic Acids Res.* 2010, *38*, D525–D531.

[43] Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G. et al., MINT: the molecular interaction database. *Nucleic Acids Res.* 2007, *35*, D572–D574.

[44] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M. et al., The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* 2011, *39*, D561–D568.

[45] GOConsortium, Creating the gene ontology resource: Design and implementation. *Genome Res.* 2001, *11*, 1425–1433.

[46] Ruepp, A., Waegele, B., Lechner, M., Brauner, B. et al., CORUM: the comprehensive resource of mammalian protein complexes – 2009. *Nucleic Acids Res.* 2009, *38*, D497–D501.

[47] Mewes, H. W., Amid, C., Arnold, R., Frishman, D. et al., MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.* 2004, *32*, D41–D44.

[48] Pu, S., Wong, J., Turner, B., Cho, E., Wodak, S. J., Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.* 2009, *37*, 825–831.

[49] Mead, J. A., Bianco, L., Ottone, V., Barton, C. et al., MRMaid, the web-based tool for designing multiple reaction monitoring (MRM) transitions. *Mol. Cell. Proteomics* 2009, *8*, 696–705.

[50] Tang, H., Arnold, R. J., Alves, P., Xun, Z. et al., A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics* 2006, *22*, e481–e488.

[51] Bukhman, Y. V., Dharsee, M., Ewing, R., Chu, P. et al., Design and analysis of quantitative differential proteomics investigations using LC-MS technology. *J. Bioinform. Comput. Biol.* 2008, *6*, 107–123.

[52] Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., Ideker, T., Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 2007, *3*, 140.

[53] Liu, M., Liberzon, A., Kong, S. W., Lai, W. R. et al., Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genet.* 2007, *3*, e96.

[54] Sohler, F., Hanisch, D., Zimmer, R., New methods for joint analysis of biological networks and expression data. *Bioinformatics* 2004, *20*, 1517–1521.

[55] Sivachenko, A. Y., Yuryev, A., Daraselia, N., Mazo, I., Molecular networks in microarray analysis. *J. Bioinform. Comput. Biol.* 2007, *5*, 429–456.

[56] Raghavan, N., Amaratunga, D., Cabrera, J., Nie, A. et al., On methods for gene function scoring as a means of facilitating the interpretation of microarray results. *J. Comput. Biol.* 2006, *13*, 798–809.

[57] Li, J., Zimmerman, L. J., Park, B. H., Tabb, D. L. et al., Network-assisted protein identification and data interpretation in shotgun proteomics. *Mol. Syst. Biol.* 2009, *5*, 303.

[58] Ostlund, G., Lindskog, M., Sonnhammer, E. L., Network-based identification of novel cancer genes. *Mol. Cell. Proteomics* 2010, *9*, 648–655.

[59] Managbanag, J. R., Witten, T. M., Bonchev, D., Fox, L. A. et al., Shortest-path network analysis is a useful approach toward identifying genetic determinants of longevity. *PLoS One* 2008, *3*, e3802.

[60] Goh, W. W., Lee, Y. H., Zubaidah, R. M., Jin, J. et al., Network-based pipeline for analyzing MS data: an application toward liver cancer. *J. Proteome Res.* 2011, *10*, 2261–2272.

[61] Berglund, L., Bjorling, E., Oksvold, P., Fagerberg, L. et al., A genecentric Human Protein Atlas for expression profiles based on antibodies. *Mol. Cell. Proteomics* 2008, *7*, 2019–2027.

[62] Bader, G. D., Hogue, C. W., An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003, *4*, 2.

[63] Titz, B., Schlesner, M., Uetz, P., What do we learn from high-throughput protein interaction data? *Expert Rev. Proteomics* 2004, *1*, 111–121.

[64] Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., Takagi, T., Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast* 2001, *18*, 525–531.

[65] Schwikowski, B., Uetz, P., Fields, S., A network of protein–protein interactions in yeast. *Nat. Biotechnol.* 2000, *18*, 1257–1261.

[66] Chua, H. N., Sung, W. K., Wong, L., Using indirect protein interactions for the prediction of gene ontology functions. *BMC Bioinformatics* 2007, *8*, S8.

[67] Enright, A. J., Van Dongen, S., Ouzounis, C. A., An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 2002, *30*, 1575–1584.

[68] King, A. D., Przulj, N., Jurisica, I., Protein complex prediction via cost-based clustering. *Bioinformatics* 2004, *20*, 3013–3020.

[69] Palla, G., Derenyi, I., Farkas, I., Vicsek, T., Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 2005, *435*, 814–818.

[70] Adamcsek, B., Palla, G., Farkas, I. J., Derenyi, I., Vicsek, T., CFinder: Locating cliques and overlapping modules in biological networks. *Bioinformatics* 2006, *22*, 1021–1023.

[71] Chua, H. N., Ning, K., Sung, W. K., Leong, H. W., Wong, L., Using indirect protein–protein interactions for protein complex prediction. *J. Bioinform. Comput. Biol.* 2008, *6*, 435–466.

[72] Liu, G., Wong, L., Chua, H. N., Complex discovery from weighted PPI networks. *Bioinformatics* 2009, *25*, 1891–1897.

[73] Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T. et al., Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 2005, *437*, 1173–1178.

[74] Stelzl, U., Worm, U., Lalowski, M., Haenig, C. et al., A human protein–protein interaction network: a resource for annotating the proteome. *Cell* 2005, *122*, 957–968.

[75] von Mering, C., Krause, R., Snel, B., Cornell, M. et al., Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 2002, *417*, 399–403.

[76] Alexeyenko, A., Sonnhammer, E. L., Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome Res.* 2009, *19*, 1107–1116.

[77] Chen, J., Yuan, B., Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics* 2006, *22*, 2283–2290.

[78] Nikitin, A., Egorov, S., Daraselia, N., Mazo, I., Pathway studio – the analysis and navigation of molecular networks. *Bioinformatics* 2003, *19*, 2155–2157.

[79] Witten, T. M., Bonchev, D., Predicting aging/longevity-related genes in the nematode *Caenorhabditis elegans*. *Chem. Biodivers.* 2007, *4*, 2639–2655.

[80] Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M., Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* 2005, *21*, i302–i310.

[81] Chua, H. N., Sung, W. K., Wong, L., Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. *Bioinformatics* 2006, *22*, 1623–1630.

[82] Deng, M., Zhang, K., Mehta, S., Chen, T., Sun, F., Prediction of protein function using protein–protein interaction data. *J. Comput. Biol.* 2003, *10*, 947–960.

[83] Hirsh, E., Sharan, R., Identification of conserved protein complexes based on a model of protein anetwork evolution. *Bioinformatics* 2007, *23*, e170–e176.

[84] Stuart, J. M., Segal, E., Koller, D., Kim, S. K., A gene-coexpression network for global discovery of conserved genetic modules. *Science* 2003, *302*, 249–255.

[85] Zhao, Y., Wang, G., Additive risk analysis of microarray gene expression data via correlation principal component regression. *J. Bioinformat. Comput. Biol.* 2010, *8*, 645–659.

[86] Liu, Z., Phan, S., Famili, F., Pan, Y. et al., A multi-strategy approach to informative gene identification from gene expression data. *J. Bioinform. Comput. Biol.* 2010, *8*, 19–38.

[87] Tusher, V. G., Tibshirani, R., Chu, G., Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* 2001, *98*, 5116–5121.

[88] Ein-Dor, L., Zuk, O., Domany, E., Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. USA* 2006, *103*, 5923–5928.

[89] Lapointe, J., Li, C., Higgins, J. P., van de Rijn, M. et al., Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. USA* 2004, *101*, 811–816.

[90] Singh, D., Febbo, P. G., Ross, K., Jackson, D. G. et al., Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 2002, *1*, 203–209.

[91] Zhang, M., Zhang, L., Zou, J., Yao, C. et al., Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics* 2009, *25*, 1662–1668.

[92] Vogel, C., Abreu Rde, S., Ko, D., Le, S. Y. et al., Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* 2010, *6*, 400.

[93] Giallourakis, C., Cao, Z., Green, T., Wachtel, H. et al., A molecular-properties-based approach to understanding PDZ domain proteins and PDZ ligands. *Genome Res.* 2006, *16*, 1056–1072.

[94] Khatri, P., Draghici, S., Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 2005, *21*, 3587–3595.

[95] Goeman, J. J., van de Geer, S. A., de Kort, F., van Houwelingen, H. C., A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004, *20*, 93–99.

[96] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S. et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* 2005, *102*, 15545–15550.

[97] Soh, D., Dong, D., Guo, Y., Wong, L., Finding consistent disease subnetworks across microarray datasets. *BMC Bioinformatics* 2011, *12*, S15.

[98] Wang, J., Li, M., Deng, Y., Pan, Y., Recent advances in clustering methods for protein interaction networks. *BMC Genomics* 2010, *11*, S10.

[99] Newman, M. E., Girvan, M., Finding and evaluating community structure in networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 2004, *69*, 026113.

[100] Zheng, Q., Wang, X. J., GOEAST: A web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.* 2008, *36*, W358–W363.

[101] Boyle, E. I., Weng, S., Gollub, J., Jin, H. et al., GO::Term-Finder – open source software for accessing Gene Ontol-

ogy information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* 2004, *20*, 3710–3715.

[102] Grossmann, S., Bauer, S., Robinson, P. N., Vingron, M., Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics* 2007, *23*, 3024–3031.

[103] Huang, W. L., Tung, C. W., Ho, S. W., Hwang, S. F., Ho, S. Y., ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. *BMC Bioinformatics* 2008, *9*, 80.

[104] Deane, C. M., Salwinski, L., Xenarios, I., Eisenberg, D., Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics* 2002, *1*, 349–356.

[105] Bossi, A., Lehner, B., Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.* 2009, *5*, 260.

[106] Chua, H. N., Wong, L., Increasing the reliability of protein interactomes. *Drug Discov. Today* 2008, *13*, 652–658.

[107] Sprinzak, E., Sattath, S., Margalit, H., How reliable are experimental protein–protein interaction data? *J. Mol. Biol.* 2003, *327*, 919–923.

[108] You, Z. H., Lei, Y. K., Gui, J., Huang, D. S., Zhou, X., Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioinformatics* 2010, *26*, 2744–2751.

[109] Bork, P., Koonin, E. V., Predicting functions from protein sequences – where are the bottlenecks? *Nat. Genet.* 1998, *18*, 313–318.

[110] Shoemaker, B. A., Panchenko, A. R., Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.* 2007, *3*, e43.

[111] Bock, J. R., Gough, D. A., Predicting protein–protein interactions from primary structure. *Bioinformatics* 2001, *17*, 455–460.

[112] Sprinzak, E., Margalit, H., Correlated sequence-signatures as markers of protein–protein interaction. *J. Mol. Biol.* 2001, *311*, 681–692.

[113] Li, H., Li, J., Wong, L., Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics* 2006, *22*, 989–996.

[114] Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W. et al., Detecting protein function and protein–protein interactions from genome sequences. *Science* 1999, *285*, 751–753.

[115] Juan, D., Pazos, F., Valencia, A., High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl. Acad. Sci. USA* 2008, *105*, 934–939.

[116] Chen, J., Hsu, W., Lee, M. L., Ng, S. K., Increasing confidence of protein interactomes using network topological metrics. *Bioinformatics* 2006, *22*, 1998–2004.