# Some Notes on Gene Expression and Proteomic Profile Analysis based on Biological Networks

Limsoon Wong[*]

August 21, 2011

## 1 Introduction

The possibility of using gene expression profiling by microarrays for purposes of diagnostics, prognostics, and understanding the mechanisms of diseases has also generated much excitement in the past decade. Similarly, mass spectrometry (MS)-based proteomics is a widely used and powerful tool for profiling systems-wide protein expression changes (Cox and Mann, 2007). It can be applied to better understand cancer pathogenesis and discover indicative biomarkers for early progression. However, a number of issues persist that hamper the effective use of these technologies. In particular, both microarray gene expression profiling and MS-based proteomics have severe consistency issues; and MS-based proteomics additionally have coverage issue. These issues make it difficult to identify genes and proteins that are meaningful in explaining the difference in disease phenotypes (Soh *et al.*, 2007).

The coverage 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 is 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 (Tang *et al.*, 2006; Mead *et al.*, 2009).

The consistency issue concerns the consistency of gene expression and proteomic profiles at the phenotype level across patient samples. To understand oncobiology and for the discovery of biomarkers, quantitative comparisons of cancerous and non-cancerous samples are performed (Bukhman *et al.*, 2008). Traditional gene expression and post-MS analysis approaches are to select and study only those genes and proteins that are found in most of the samples of the phenotype in question and have a consistently overexpressed or underexpressed ratio. However, genes and proteins with noticeably high or low expression are not necessarily causal or important. At the same time, a mutated gene or protein that drives other genes and proteins to change their levels may not itself report any change in expression or may miss being detected

[*] Department of Computer Science and Department of Pathology, National University of Singapore, Singapore.

by the MS assay. Moreover, many relevant genes and proteins report “swing” ratios, that is, a mixture of both high and low ratios. 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 biomarkers that are consistent and reproducible across different batches of patient 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 on 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 can be 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 5 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% while the chance of $B$–$E$ all being false negatives is $(50\%)^4 = 6\%$. Hence, it is almost 10 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% while the chance of $B$–$E$ all being false positives is $(50\%)^4 = 6\%$. Hence, it is almost 10 times more likely that $A$ is false negative than $B$–$E$ all being false positives.

Each 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 disease (Sohler *et al.*, 2004; Liu *et al.*, 2007; Chuang *et al.*, 2007). Hence the uncertainty in the reliability of the selected proteins from quantitative comparisons of cancer and non-cancer samples can be reduced by considering the molecular functions and the biological processes associated with the genes and proteins (Sivachenko *et al.*, 2007). Such a unifying biological theme is also a basis for inferring the underlying cause of the cancer phenotype. For illustration, let there be 3 cancer samples and 3 controls. Assuming the chance of an arbitrary protein to be found highly expressed in an arbitrary sample is 50%. Then a group of 5 functionally linked proteins that is perfectly correlated to these two groups of samples—e.g., they are all highly expressed in the 3 cancer samples and not in the 3 controls—has $((50\%)^3 * (1 - 50\%)^3)^5 = 8 * 10^{-8}\%$ chance of being a false positive group. On the other hand, if just 1 of these 5 functionally linked proteins was perfectly correlated to the two phenotypes, its chance of being a false positive would be $(50\%)^3 * (1 - 50\%)^3 = 1.5\%$, which is many orders of magnitude higher than when all 5 proteins are simultaneously correlated with the two phenotypes.

Clearly, leveraging on these paradigms can aid in circumventing some of the shortcomings of current proteomics approaches mentioned. In the remainder of this tutorial note, we first describe the different types of biological networks. We then present approaches based on biological networks for improving the coverage of the proteome at the level of individual samples. Following that, we present approaches based on biological networks for improving the consistency of gene expression and proteomic profile analysis at the phenotype level across patient samples.

## 2 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 (Kanehisa *et al.*, 2010), BioCyc (Krummenacker *et al.*, 2005), WikiPathways (Kelder *et al.*, 2009), Reactome (Vastrik *et al.*, 2007), Ingenuity® Knowledge Base (`http://www.ingenuity.com`), NetPro$^{TM}$ (`http://www.molecularconnections.com`), Pathway Commons (Cerami *et al.*, 2011), and PathwayAPI (Soh *et al.*, 2010).

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 genetic interaction 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.* (2009) 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 (Stark *et al.*, 2006), DIP (Xenarios *et al.*, 2002), HPRD (Prasad *et al.*, 2009), IntAct (Aranda *et al.*, 2010), MINT (Chatr-aryamontri *et al.*, 2007), and STRING (Szklarczyk *et al.*, 2011).

The Gene Ontology (GO) was established by the  Gene Ontology Consortium (2001) as an important reference terminology for annotating the function and cellular localization of proteins. 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 this group of proteins are 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 (Ruepp *et al.*, 2010), MIPS (Mewes *et al.*, 2004), and CYC2008 catalogue (Pu *et al.*, 2009).

## 3 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, they may not be reported initially in the first round of data analysis. This occurs frequently in the struggle in the tradeoff between sensitivity and specificity in precursor ion selection for fragmentation. Other potential reasons why these proteins are unreported in the initial round of data analysis 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, as suggested earlier, is to use a database of protein complexes and identify those complexes that have a large overlap with the initial list of detected proteins. The rest of the proteins in these identified complexes are postulated as likely to be present. More sophisticated methods that build on this principle include CEA (Li *et al.*, 2009), MaxLink (Ostlund *et al.*, 2010), shortest-path analysis (Managbanag *et al.*, 2008), and the method of Goh *et al.* (2011) (which we call PEP here). We describe them in subsections below.

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 (Goh *et al.*, 2011). 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 (Ostlund *et al.*, 2010). Another indirect evidence is using databases of gene expression profiles—e.g., Human Protein Atlas (Berglund *et al.*, 2008)—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 (Ostlund *et al.*, 2010).

## 3.1 Clique Enrichment Analysis (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 (Ruepp *et al.*, 2010) is still small. So, one should supplement them with predicted protein complexes and functional modules.

An example that pursues this route is the Clique Enrichment Analysis (CEA) proposed by Li *et al.* (2009). CEA generates cliques—i.e., 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 (Bader and Hogue, 2003).

## 3.2 Proteomics Expansion Pipeline (PEP)

70-80% of proteins share at least one biological process or function with their interaction partners in PPINs and GINs (Titz *et al.*, 2004). A protein is also often observed to participate in a biological process or function that is over-represented in its interaction partners (Hishigaki *et al.*, 2001; Schwikowski *et al.*, 2000). More generally, proteins that are connected or proximal within a biological network often form a functional unit (Chua *et al.*, 2007). 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 (Enright *et al.*, 2002), MCODE (Bader and Hogue, 2003), RNSC (King *et al.*, 2004), CFinder (Palla *et al.*, 2005; Adamcsek *et al.*, 2006), PCP (Chua *et al.*, 2008), and CMC (Liu *et al.*, 2009). These more powerful algorithms can be used in place of clique finding in CEA.

A most recent method that uses a powerful protein complex prediction algorithm is that proposed by Goh *et al.* (2011). We call this method the Proteomics Expansion Pipeline (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 (Palla *et al.*, 2005), 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 MS 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 (Prasad *et al.*, 2009), BioGRID (Stark *et al.*, 2006), IntAct (Aranda *et al.*, 2010), and DIP (Xenarios *et al.*, 2002), as well as data from literature (Rual *et al.*, 2005; Stelzl *et al.*, 2005). While combining PPINs improves coverage of the protein interactome, it may also compound the noise present in them (von Mering *et al.*, 2002). PEP uses the iterated Czekanowski-Dice distance (CD-distance) technique from CMC (Liu *et al.*, 2009) 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. 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.

## 3.3 Maxlink

PPINs have a fairly high level of false positives and false negatives (von Mering *et al.*, 2002). 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, instead of 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.* (2010). 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 (Alexeyenko and Sonnhammer, 2009) (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 (Hishigaki *et al.*, 2001; Schwikowski *et al.*, 2000). Moreover, proteins in the same complex are thought to have more interactions between themselves than with proteins outside the complex (Chen and Yuan, 2006).

### 3.4 Shortest-path network analysis

In a related approach, Managbanag *et al.* (2008) propose 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 (Nikitin *et al.*, 2003).

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 (Witten and Bonchev, 2007). This hypothesis is partially justified by the observation that most proteins share at least one function with their interaction partners (Titz *et al.*, 2004) in a PPIN and thus transitively with the partners of these partners (Nabieva *et al.*, 2005). However, the longer a (shortest) path gets, the more false positives it inevitably contains (Chua *et al.*, 2006).

## 4 Improving consistency using biological networks

Quantitative comparisons of cancerous and non-cancerous samples are central to oncoproteomics (Bukhman *et al.*, 2008). However, biomarkers identified in one batch of patients are quite often not consistent and not reproducible in another batch of patients. 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 paradigm suggests protein interactions constitute a major part of all cellular processes. The extent of interactions between proteins denotes shared functionality (Deng *et al.*, 2003), complex or sub-module participation (Hirsh and Sharan, 2007) and/or co-expression (Stuart *et al.*, 2003). 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 (Zhao and Wang, 2010; Liu *et al.*, 2010; Tusher *et al.*, 2001) 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 (Ein-Dor *et al.*, 2006). For example, for a pair of datasets involving prostate cancer (Lapointe *et al.*, 2004; Singh *et al.*, 2002), Zhang *et al.* (2009) show that the two lists of significant genes identified by running SAM (Tusher *et al.*, 2001) independently on the two datasets have a low overlap of 30% in their top 10 genes and an even lower 15% overlap in their top 100 genes. Methods based on individual-gene analysis—such as SAM Tusher *et al.* (2001)—have a low degree of reproducibility because of their relatively higher level of false positives. For example, suppose genes are selected at $P \leq 0.05$. There are generally $> 10,000$ genes on a microarray. Thus we expect $10,000 * 0.05 = 500$ genes to correlate with the phenotypes in a dataset purely by random chance. Many of these genes may even rank higher than the true relevant genes. But for a different datasets (of the same phenotypes), a different set of false-positive genes is often selected. The two sets of independently selected genes from the two datasets can thus be expected to have a low overlap, resulting in low reproducibility for these methods. 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, gene co-expression is something proven at the protein level, especially when it comes to an induction of a particular function. So, 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.

## 4.1 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 (Khatri and Draghici, 2005) 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. Different test statistics and different thresholds result in a different list of differentially expressed genes. As a result, the outcome of the whole procedure is not stable, leading to potentially low reproducibility. Another problem is that it is not uncommon for a real causal gene underlying a disease phenotype to be not differentially expressed. It thus can never be suggested by these methods. For example, suppose a gene $A$ upregulates both genes $B$ and $C$ in normal people. Suppose also that genes $A$ is observed to be highly expressed in both normal and disease samples; and genes $B$ and $C$ are observed to be highly expressed in normal but not disease samples. Then only genes $B$ and $C$, which are differentially expressed, have a chance to be suggested by these methods. In such as a situation, we have to postulate mutations in $B$ and $C$ to explain their differential expression. However, a more reasonable explanation is that $A$ has a mutation that does not change its expression but changes its ability to upregulate $B$ and $C$.

## 4.2 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 a 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 (Goeman *et al.*, 2004) and GSEA (Subramanian *et al.*, 2005) 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, 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.

## 4.3 Network-based analysis

Network-based analysis methods (Chuang *et al.*, 2007; Sivachenko *et al.*, 2007; Soh, 2010; Sohler *et al.*, 2004; Liu *et al.*, 2007) 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 (Sivachenko *et al.*, 2007). NEA extracts from each biological pathway (usually a 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 (Soh, 2010), 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 (Soh, 2010). The strength of SNet lies in its ability to identify relevant large subnetworks (of known pathways) based on microarray data. As explained earlier, methods such as FCS and GSEA use fixed gene sets and determine whether these gene sets are significant or not. These techniques assume that a gene set is significant only when a substantial proportion of the genes within the gene set is significant. This assumption is not often valid because it is often the case that only a fraction of a gene set is significant; such a gene set will probably go unnoticed if most of the rest of the genes are unaffected. SNet's ability to extract subnetworks based on the microarray data of the phenotypes—and use these as gene sets—ensures that there is sufficient granularity for it to capture portions of pathways or gene sets that are affected. A disadvantage of SNet in the proteomics context—compared to 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.

## 4.4 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 (Wang *et al.*, 2010). So we just briefly describe a few easily accessible methods here. CFinder is based on the clique percolation method described by Palla *et al.* (2005). 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 (Newman and Girvan, 2004). MCL is based on the Markov clustering method (Enright *et al.*, 2002). CMC works by generating maximal cliques from the cleansed network and then merges or removes highly overlapping cliques based on their interconnectivity (Liu *et al.*, 2009).

For the GO term analysis step, it is often done using tools based on the hypergeometric test. Examples include GO East (Zheng and Wang, 2008) and GO Term Finder (Boyle *et al.*, 2004). 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 (Grossmann *et al.*, 2007) and the intuitive "informative GO term" method (Huang *et al.*, 2008).

The parent-child method proposed by Grossmann *et al.* (2007) modifies the hypergeometric test statistics. Instead of the standard hypergeometric distribution, they propose using $P(\sigma_t = k|\sigma_{pa(t)} = n_{pa(t)}) = \binom{m_t}{k}\binom{m_{pa(t)}-m_t}{n_{pa(t)}-k}/\binom{m_{pa(t)}}{n_{pa(t)}}$. Here, $t$ is the GO term that we want 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 $t$; $m_{pa(t)}$ is the number of proteins in the GO database that are annotated to the parent terms of $t$; and $n_{pa(t)}$ 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 (Grossmann *et al.*, 2007), 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 (Chua *et al.*, 2007). This also has the effect of reducing redundancy on GO terms reported as a whole (Huang *et al.*, 2008).

| Database | # nodes,<br># edges | URL | Build<br>Focus | Reference |
|---|---|---|---|---|
| BioGRID | 10k, 40k | `http://thebiogrid.org` | Literature | (Stark *et al.*, 2006) |
| DIP | 2.6k, 3.3k | `http://dip.doe-mbi.ucla.edu` | Literature | (Xenarios *et al.*, 2002) |
| HPRD | 30k, 40k | `http://www.hprd.org` | Literature | (Prasad *et al.*, 2009) |
| IntAct | 56k, 267k | `http://www.ebi.ac.uk/intact` | Literature | (Aranda *et al.*, 2010) |
| MINT | 30k, 90k | `http://mint.bio.uniroma2.it/mint` | Literature | (Chatr-aryamontri *et al.*, 2007) |
| STRING | 5200k, ? | `http://string-db.org` | Literature,<br>Prediction | (Szklarczyk *et al.*, 2011) |

Table 1: Databases of protein-protein interaction networks.

# 5 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.

## 5.1 Reliability of PPINs

The databases of PPINs and GINs have grown rapidly in size over the years, with improved methodologies in testing protein interactions. There 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 Yeast 2 Hybrid (Y2H) and other binding experiments, up to 70% of the reported edges may be false (Deane *et al.*, 2002). Mark Vidal and co-workers tried producing higher quality all-against-all experimental data (Rual *et al.*, 2005; Stelzl *et al.*, 2005), by testing all possible protein pairs in their data set using Y2H. However, these datasets 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.

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 datasets across various sources (Li *et al.*, 2009; Bossi and Lehner, 2009). 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 (2009), 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 (Chua and Wong, 2008). 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) (Nabieva *et al.*, 2005; Sprinzak *et al.*, 2003). 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 (Liu *et al.*, 2009). Other examples are surveyed in (Chua and Wong, 2008). 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 (You *et al.*, 2010), or homologous transfer of edges (Bork and Koonin, 1998).

A harder problem to resolve is the false negative problem—viz., true interactions that are not reported. Chua and Wong (2008) and Shoemaker and Pachenko (2007) provided detailed reviews on approaches for predicting novel protein-protein interactions, including protein primary structures and associated physicochemical properties (Bock and Gough, 2001), interacting domains (Sprinzak and Margalit, 2001), interacting motifs (Li *et al.*, 2006), gene-fusion events (Marcotte *et al.*, 1999), coevolution of proteins or residues (Juan *et al.*, 2008), and the topology of PPINs (Chen *et al.*, 2006).

## 5.2 Completeness of biological pathway databases

The databases of MNs and RNs can be considered as more reliable than PPIN and GIN datasets 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$^{\mathrm{TM}}$), 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.* (2010) 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 (LCS) turned out to be a good compromise. However, Goh *et al.* (2011) 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.

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 want to expand coverage to other species.

| Database | Remarks |
|---|---|
| KEGG | KEGG (`http://www.genome.jp/kegg`) is one of the best known pathway databases (Kanehisa *et al.*, 2010). It consists of 16 main databases, comprising different levels of biological information such as systems, genomic, etc. The data files are downloadable in XML format. At time of writing it has 392 pathways. |
| BioCyc | BioCyc ver 15 (`http://http://biocyc.org`) comprises over 1,129 species-based databases (Krummenacker *et al.*, 2005). An interesting feature of the BioCyc databases is that they are divided into 3 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 (Kelder *et al.*, 2009). It has 1,627 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 (Vastrik *et al.*, 2007). It is one of the largest datasets, with over 4,166 human reactions organized into 1,131 pathways by December 2010. Reactome can be downloaded 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$^{TM}$ | Molecular Connections' NetPro$^{TM}$ (`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 (Cerami *et al.*, 2011). It contains 1,573 pathways across 564 organisms. The data is 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® Knowledge Base (Soh *et al.*, 2010). Data is downloadable as a SQL dump or as a csv file, and is also interfaceable in JSON format. |

Table 2: Databases of biological pathways.

# 6 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 coverage 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 on function based on clusters, rather than average function based solely on differential proteins.

## Acknowledgements

This tutorial note is the result of cutting and pasting from drafts of two papers that I am co-authoring with Donny Soh and Wilson Goh. The works performed in that two papers were supported in part by a Wellcome Trust Scholarship 83701/Z/07/Z (for Wilson Goh), an A*STAR AIP scholarship (for Donny Soh), and a Singapore National Research Foundation grant NRF-G-CRP-2007-04-082(d) and a A*STAR grant SERC 072/101/0016 (both for me),

## References

Adamcsek, B. *et al.* (2006). CFinder: Locating cliques and overlapping modules in biological networks. *Bioinformatics*, **22**(8), 1021–1023.

Alexeyenko, A. and Sonnhammer, E. L. (2009). Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome Research*, **19**, 1107–1116.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S. N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K., and Hermjakob, H. (2010). The IntAct molecular interaction database in 2010. *Nucleic Acids Research*, **38**(Database issue), D525–531.

Bader, G. D. and Hogue, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Berglund, L., Bjorling, E., Oksvold, P., Fagerberg, L., *et al.* (2008). A gene-centric Human Protein Atlas for expression profiles based on antibodies. *Molecular & Cellular Proteomics*, **7**, 2019–2027.

Bock, J. R. and Gough, D. A. (2001). Predicting protein-protein interactions from primary structure. *Bioinformatics*, **17**(5), 455–460.

Bork, P. and Koonin, E. V. (1998). Predicting functions from protein sequences—where are the bottlenecks? *Nature Genetics*, **18**, 313–318.

Bossi, A. and Lehner, B. (2009). Tissue specificity and the human protein interaction network. *Molecular Systems Biology*, **5**, 260.

Boyle, E. I., Weng, S., Gollub, J., Jin, H., *et al.* (2004). GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.

Bukhman, Y. V., Dharsee, M., Ewing, R., Chu, P., *et al.* (2008). Design and analysis of quantitative differential proteomics investigations using LC-MS technology. *Journal of Bioinformatics and Computational Biology*, **6**(1), 107–123.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., *et al.* (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, **39**, D685–D690.

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., and Cesareni, G. (2007). MINT: The Molecular INTeraction database. *Nucleic Acids Research*, **35**, D572–D574.

Chen, J. and Yuan, B. (2006). Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, **22**(18), 2283–2290.

Chen, J., Hsu, W., Lee, M. L., and Ng, S.-K. (2006). Increasing confidence of protein interactomes using network topological metrics. *Bioinformatics*, **22**(16), 1998–2004.

Chua, H. N. and Wong, L. (2008). Increasing the reliability of protein interactomes. *Drug Discovery Today*, **13**(15/16), 652–658.

Chua, H. N., Sung, W.-K., and Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, **22**(13), 1623–1630.

Chua, H. N., Sung, W.-K., and Wong, L. (2007). Using indirect protein interactions for the prediction of gene ontology functions. *BMC Bioinformatics*, **8**(Suppl 4), S8.

Chua, H. N., Ning, K., Sung, W.-K., Leong, H. W., and Wong, L. (2008). Using indirect protein-protein interactions for protein complex prediction. *Journal of Bioinformatics and Computational Biology*, **6**(3), 435–466.

Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, **3**, 140.

Cox, J. and Mann, M. (2007). Is proteomics the new genomics? *Cell*, **130**, 395–398.

Deane, C. M., Salwinski, L., Xenarios, I., and Eisenberg, D. (2002). Protein interactions: Two methods for assessment of the reliability of high-throughput observations. *Molecular and Cellular Proteomics*, **1**, 349–356.

Deng, M., Zhang, K., Mehta, S., Chen, T., and Sun, F. (2003). Prediction of protein function using protein-protein interaction data. *Journal of Computational Biology*, **10**, 947–960.

Dixon, S. J., Costanzo, M., Baryshnikova, A., Andrews, B., and Boone, C. (2009). Systematic mapping of genetic interaction networks. *Annual Review of Genetics*, **43**, 601–625.

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

Enright, A. J., Dongen, S. V., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, **30**(7), 1575–1584.

Gene Ontology Consortium (2001). Creating the gene ontology resource: Design and implementation. *Genome Research*, **11**, 1425–1433.

Goeman, J. J., van de Geer, S. A., de Kort, F., and van Houwelingen, H. C. (2004). A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics*, **20**(1), 93–99.

Goh, W. W. B., Lee, Y. H., Zubaidah, R., Jin, J., Dong, D., Lin, Q., Chung, M., and Wong, L. (2011). A network-based pipeline for analyzing ms data—an application towards liver cancer. *Journal of Proteome Research*, **10**(5), 2261–2272.

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

Hirsh, E. and Sharan, R. (2007). Identification of conserved protein complexes based on a model of protein network evolution. *Bioinformatics*, **23**, e170–e176.

Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., and Takagi, T. (2001). Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, **18**(6), 525–531.

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

Juan, D., Pazos, F., and Valencia, A. (2008). High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc Natl Acad Sci USA*, **105**(3), 934–939.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, **38**(Database Issue), D355–D360.

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

Khatri, P. and Draghici, S. (2005). Ontological analysis of gene expression data: Current tools, limitations, and open problems. *Bioinformatics*, **21**(18), 3587–3595.

King, A. D., Przulj, N., and Jurisica, I. (2004). Protein complex prediction via cost-based clustering. *Bioinformatics*, **20**(17), 3013–3020.

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

Lapointe, J., Li, C., Higgins, J. P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., *et al.* (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. USA*, **101**(3), 811–816.

Li, H., Li, J., and Wong, L. (2006). Discovering motif pairs at interaction sites from sequences on a proteome-wide scale. *Bioinformatics*, **22**(8), 989–996.

Li, J., Zimmerman, L. J., Park, B. H., Tabb, D. L., *et al.* (2009). Network-assisted protein identification and data interpretation in shotgun proteomics. *Molecular Systems Biology*, **5**, 303.

Liu, G., Wong, L., and Chua, H. N. (2009). Complex discovery from weighted PPI networks. *Bioinformatics*, **25**(15), 1891–1897.

Liu, M., Liberzon, A., Kong, S. W., Lai, W. R., Park, P. J., Kohane, I. S., and Kasif, S. (2007). Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genetics*, **3**(6), e96.

Liu, Z. *et al.* (2010). A multi-strategy approach to informative gene identification from gene expression data. *Journal of Bioinformatics and Computational Biology*, **8**(1), 19–38.

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

Marcotte, E. M., Pellegrini, M., Ng, H.-L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**(5428), 751–753.

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

Mewes, H. W. *et al.* (2004). MIPS: Analysis and annotation of proteins from whole genomes. *Nucleic Acids Research*, **32**(Database issue), 41–44.

Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., and Singh, M. (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, **21**(Suppl. 1), i302–i310.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, **69**, 026113.

Nikitin, A., Egorov, S., Daraselia, N., and Mazo, I. (2003). Pathway studio—the analysis and navigation of molecular networks. *Bioinformatics*, **19**(16), 2155–2157.

Ostlund, G., Lindskog, M., and Sonnhammer, E. L. (2010). Network-based identification of novel cancer genes. *Molecular & Cellular Proteomics*, **9**, 648–655.

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

Prasad, T. S. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Kishore, C. J. H., Kanth, S., Ahmed, M., Kashyap, M., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, A. B., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human protein reference database - 2009 update. *Nucleic Acids Research*, **37**, D767–D772.

Pu, S., Wong, J., Turner, B., Cho, E., and Wodak, S. J. (2009). Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, **37**(3), 825–831.

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., *et al.* (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**(7062), 1173–1178.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2010). CORUM: The comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Research*, **38**(Suppl. 1), D497–D501.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. *Nature Biotechnology*, **18**(12), 1257–1261.

Shoemaker, B. A. and Pachenko, A. R. (2007). Deciphering protein-protein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLOS Computational Biology*, **3**(4), e43.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., *et al.* (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Sivachenko, A. Y., Yuryev, A., Daraselia, N., and Mazo, I. (2007). Molecular networks in microarray analysis. *Journal of Bioinformatics and Computational Biology*, **5**(2b), 429–546.

Soh, D. (2010). *Understanding Pathways*. Ph.D. thesis, Department of Computing, Imperial College London, 180 Queens Gate, London SW7 2AZ.

Soh, D., Dong, D., Guo, Y., and Wong, L. (2007). Enabling more sophisticated gene expression analysis for understanding diseases and optimizing treatments. *ACM SIGKDD Explorations*, **9**(1), 3–14.

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

Sohler, F., Hanisch, D., and Zimmer, R. (2004). New methods for joint analysis of biological networks and expression data. *Bioinformatics*, **20**(10), 1517–1521.

Sprinzak, E. and Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interactions. *Journal of Molecular Biology*, **311**(4), 681–692.

Sprinzak, E., Sattath, S., and Margalit, H. (2003). How reliable are experimental protein-protein interaction data? *Journal of Molecular Biology*, **327**(5), 919–923.

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: A general repository for interaction datasets. *Nucleic Acids Research*, **34**(Database issue), D535–539.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., *et al.* (2005). A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, **122**(6), 957–968.

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

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Nat. Acad. Sci. USA*, **102**(43), 15545–15550.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., *et al.* (2011). The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, **39**, D561–D568.

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

Titz, B., Schlesner, M., and Uetz, P. (2004). What do we learn from high-throughput protein interaction data? *Expert Review of Proteomics*, **1**, 111–121.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, **98**(9), 5116–5121.

Vastrik, I., D'Eustachio, P., Schmidt, E., Gopinath, G., *et al.* (2007). Reactome: A knowledge base of biologic pathways and processes. *Genome Biology*, **8**, R39.

von Mering, C., Krause, R., Snel, B., *et al.* (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**(6887), 399–403.

Wang, J., Li, M., Deng, Y., and Pan, Y. (2010). Recent advances in clustering methods for protein interaction networks. *BMC Genomics*, **11**(Suppl. 3), S10.

Witten, T. M. and Bonchev, D. (2007). Predicting aging/longevity-related genes in the nematode *Caenorhabditis elegans. Chemistry & Biodiversity*, **4**, 2639–2655.

Xenarios, I., Salwinski, L., Duan, X., Higney, P., Kim, S., and Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, **30**(1), 303–305.

You, Z.-H., Lei, Y.-K., Gui, J., Huang, D.-S., and Zhou, X. (2010). Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioinformatics*, **26**(21), 2744–2751.

Zhang, M., Zhang, L., Zou, J., Yao, C., Xiao, H., Liu, Q., Wang, J., Wang, D., Wang, C., and Guo, Z. (2009). Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics*, **25**(13), 1662–1668.

Zhao, Y. and Wang, G. (2010). Additive risk analysis of microarray gene expression data via correlation principal component regression. *Journal of Bioinformatics and Computational Biology*, **8**(4), 645–659.

Zheng, Q. and Wang, X. J. (2008). GOEAST: A web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Research*, **36**, W358–W363.