Disease Markers 28 (2010) 253–266
DOI 10.3233/DMA-2010-0695
IOS Press

Separating the drivers from the driven: Integrative network and pathway approaches aid identification of disease biomarkers from high-throughput data

Jason E. McDermott[a,*], Michelle Costa[a], Derek Janszen[a], Mudita Singhal[b] and Susan C. Tilton[a]
[a]*Computational Biology and Bioinformatics Group, Pacific Northwest National Laboratory, Richland, WA, USA*
[b]*Data Intensive Scientific Computing, Pacific Northwest National Laboratory, Richland, WA, USA*

Abstract. The recent advances in high-throughput data acquisition have driven a revolution in the study of human disease and determination of molecular biomarkers of disease states. It has become increasingly clear that many of the most important human diseases arise as the result of a complex interplay between several factors including environmental factors, such as exposure to toxins or pathogens, diet, lifestyle, and the genetics of the individual patient. Recent research has begun to describe these factors in the context of networks which describe relationships between biological components, such as genes, proteins and metabolites, and have made progress towards the understanding of disease as a dysfunction of the entire system, rather than, for example, mutations in single genes. We provide a summary of some of the recent work in this area, focusing on how the integration of different kinds of complementary data, and analysis of biological networks and pathways can lead to discovery of robust, specific and useful biomarkers of disease and how these methods can help shed light on the mechanisms and etiology of the diseases being studied.

Keywords: Networks, systems biology, high-throughput data, topological analysis

1. Introduction

Traditional analysis of high-throughput (HT) data, which relies on identification of differentially expressed genes or proteins, has been very successful at identifying biomarker candidates that reflect the downstream effects of the process and/or pathology being studied. Recently, the availability of different types of data for the same system, e.g. transcriptomic, proteomic and metabolomic measurements, provides the opportunity to more fully characterize biological systems. Bioinformatics approaches are addressing these challenges by representing the system as biological networks and pathways which allows multi-source HT experimental data to be more easily integrated. These advances have been used in identification of biomarkers that represent the mediators of the processes or pathologies of interest, rather than their effects, which can provide more robust and biologically relevant information about the systems and diseases being studied. A number of recent publications highlight the potential for this emerging field as well as the current limitations of its application. In this review we highlight several ways in which network representations of HT data are being used to identify biomarkers of disease.

New ways of integrating disparate HT data types and representing the relationships between the components to which they correspond will prove to be an essential aspect of biomarker discovery. These approaches hold the promise of gaining insights into the causes

*Corresponding author: Jason McDermott, PhD, Computational Biology and Bioinformatics Group, Pacific Northwest National Laboratory, MSIN: J4-33, 902 Battelle Boulevard, PO Box 999, Richland, WA 99352, USA. Tel.: +1 509 372 4360; Fax: +1 509 372 4720; E-mail: Jason.McDermott@pnl.gov.

ISSN 0278-0240/10/$27.50

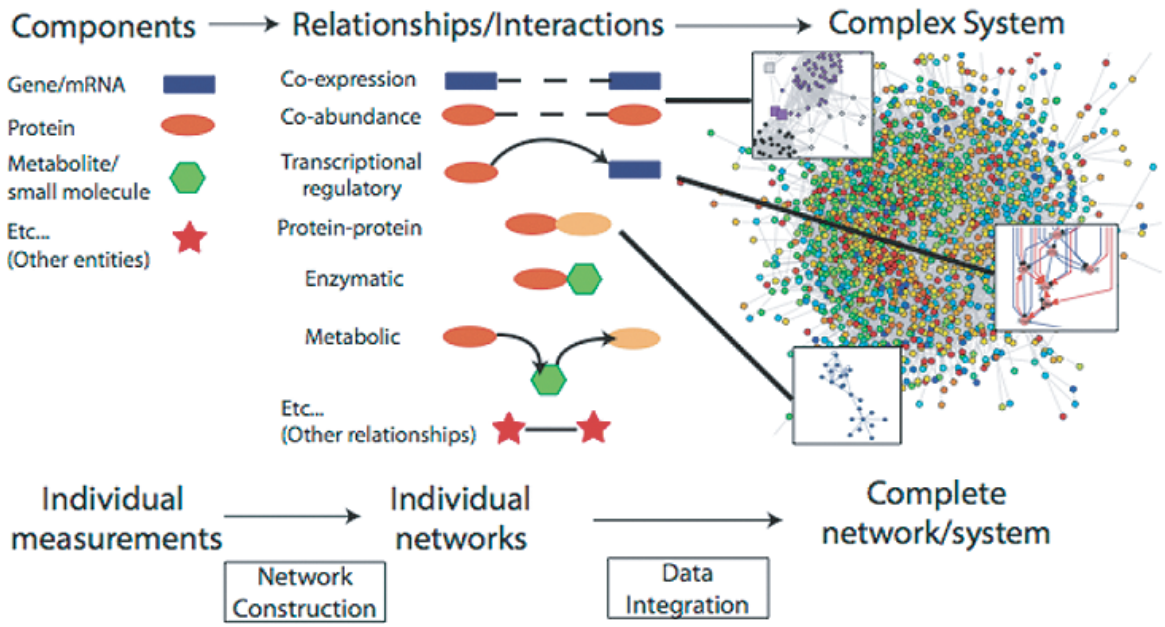

Fig. 1. Levels of complexity in analysis of high-throughput biological data. Measurements of individual biological components can be assembled into networks. Each of these kinds of networks is really a different 'view' of the underlying complex system. Better representations of this complete network/system can be obtained by integrating different kinds of network views.

of complex diseases from a systems-level standpoint. A primary tool of this work is the biological network, composed of various kinds of relationships among biological entities (Fig. 1). The disease state arises from disruptions in the interplay of the underlying biological network, which is shaped by the genetic background of the individual, and the environment. It is now understood that most diseases involve the interaction of many biological components and can generally not be traced back to one, or even a small number of, "responsible" genes, proteins, or metabolites [69]. This realization has profound implications for the application of HT methodologies to human disease. Identification of individual biological entities that cause a particular disease is not possible in many cases, rather requiring the consideration of groups of entities acting in concert. Additionally, focusing efforts solely on the most differentially regulated entities will result in limited insights about the etiology of the disease at the molecular level. Instead, more sophisticated methods are required to elucidate important biomarkers of disease and subnetworks associated with the development, progression and pathology of disease. We describe the current state of the field for identification of disease-related entities and summarize some recent work aimed at distinguishing the drivers of disease from the downstream components that are driven by disease.

2. What is a biological network?

Different types of biological networks have been characterized and each is generally based on the underlying technology for determination of relationships. These networks are conceptually distinct, but actually represent different incomplete views of one biological network (Fig. 1). The networks represent their biological components as 'nodes' and the relationships between them as 'edges'. Components of a biological system are genes, coding and non-coding RNAs, proteins, and metabolites; anything that has a function in the system. Possible relationships between these components are numerous and interrelated. Networks represent the overall connectivity structure of the interactions among the components at the global or population level [69]. In practice, the use of biological networks is limited to a particular view of the real network that can be provided by the experimental and/or computational method(s) used to interrogate it. Common views of a biological network include protein-protein interaction (PPI) networks, transcriptional regulatory networks, signaling networks, and metabolic networks. Each of these views of the biological network is predicated on a physical interaction between the components represented.

One important type of biological network is the physical PPI network, determined by either two-hybrid

methods or physical association in co-immunoprecipitation or pull-down experiments. PPI networks have been very informative about the nature of biological networks for a limited number of model organisms [8, 33,72]. Though there are a large number of PPIs that have been experimentally determined in the human interactome, the utility of this network is limited for several reasons. The first is that the interactions are a view of the network from a large number of different cell types under different conditions. Though these can be used in combination with disease- and tissue- specific information from transcriptomics or proteomics (discussed further below), the specificity of interactions is limited. The second is that the network is false-negative rich; it is likely that the vast majority of interactions have not yet been described [67,73]. Finally these networks have potentially high false-positive rates associated with the underlying experimental methods, each of which has its own characteristic artifacts [6,7].

An alternate view of the network is provided by statistical inference methods [58,90] that determine relationships between components in the system based on shared expression or abundance patterns over a range of different conditions (here called associative networks). These kinds of network views can be very specific to the disease states being studied, are relatively inexpensive to generate, and potentially include many more relevant biological components than can be currently considered using existing physical interaction data. In terms of providing information about disease states and etiology, these associative networks are turning out to be every bit as valid and useful as the network views which are based more on physical interactions. The top-down approach of constructing networks from significant correlations between expression or abundance patterns of components over a range of experimental conditions has been enabled by advances in HT methodology and the resulting explosion of data. In these approaches, relationships are predicted between pairs of components based on statistical association [27,49,83]. Although association does not imply causation, statistically significant associations are considered to be indicative of a relationship between two components that is mediated by a range of possible interactions and/or intermediaries. For example, co-expressed genes and gene products may physically interact with each other (e.g. in a complex), may be involved in a regulatory network, or may share a metabolic compound that links the two components. Other significant relationships are possible as well, for example the activation of an ion channel could be linked to an ion-responsive regulator through changes in ion concentration.

Finally, an alternate approach to the consideration of biological networks is as a tool for identification of important nodes, disease mechanisms, or other factors of interest, rather than as a strictly faithful representation of the true relationships between biological components. In this sense, the individual relationships between biological components are not as important as the ability of the network to reveal the underlying functioning of the system. Associative networks have been used in this way [14,25,53,70,91] and we discuss several applications using this approach below.

To clarify our discussion we define several terms, stressing that these areas are not clearly delineated and overlap in a number of places. We discuss "network analysis" as dealing with relatively unstudied biological networks derived from large-scale analyses (e.g. two-hybrid protein-protein interactions) or inferred from HT data (e.g. compendia of microarray experiments). We contrast this kind of analysis with "pathway analysis", in which HT data is overlaid onto well-characterized pathways that have been previously defined in databases or literature, for example signal transduction pathways. The two approaches, sometimes referred to as top-down (network analysis) and bottom-up (pathway analysis), have complementary strengths and weaknesses. Network analysis allows utilization of entire datasets and produces hypotheses about uncharacterized novel targets. Although it can suggest mechanisms and further avenues of investigation, it may not provide a readily interpretable biological story, as many of the components involved may lack confident functional annotations and/or solid experimental evidence. Pathway analysis provides biological insight into individual interactions and mechanisms of action that, in some cases, have been extensively studied. However, this information is available for a very limited number of components, preventing identification of novel targets. Additionally, these pathway maps are generally canonical in nature and understood to be limited in terms of pathway cross-talk and inputs and outputs. Network and pathway analyses are complementary approaches, presenting a more complete picture of the system under study.

In this review we first discuss the integration of disparate kinds of HT data, then review some recent advances in the analysis of pathways and networks that can help reveal the biological basis of disease processes. Important biomarkers can be elucidated through the integration of HT data and application of network and pathway analysis techniques (Fig. 2). Network analysis provides insight into information flow in the system,

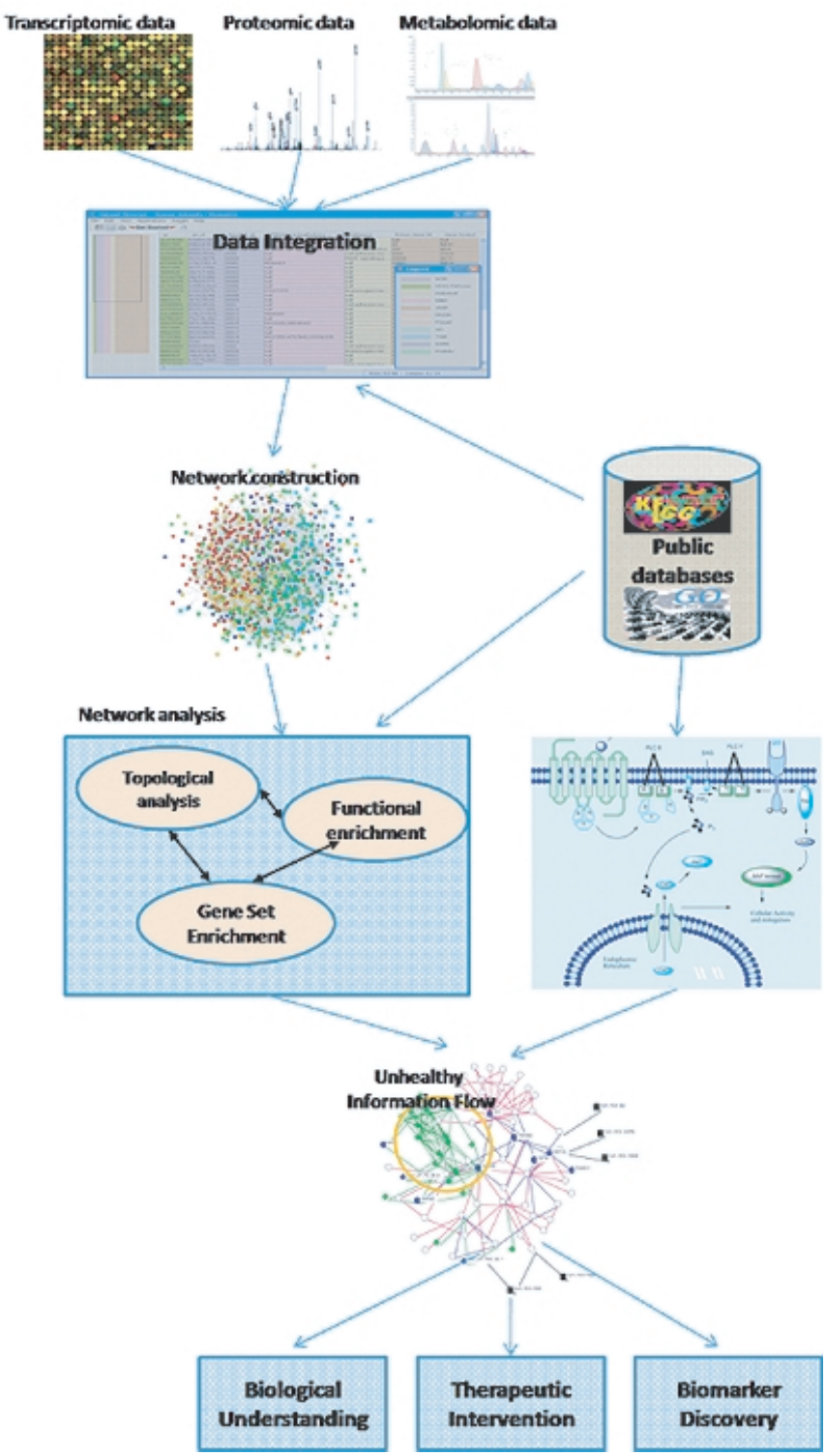

Fig. 2. Bioinformatics workflow leading to biomarker discovery. Different types of measurements can be integrated using a variety of different kinds of tools. Biological networks are then inferred and/or constructed from the data and information in public databases. Networks and pathways are analyzed using a variety of techniques to elucidate biological understanding of the disease, discover biomarkers, and provide possible avenues for therapeutic intervention.

and how it is affected by the diseased state. Restoring the system to its normal, "healthy" state is the goal of targeted drug design. Thus a more detailed understanding of the dynamics of the system in both normal and diseased states is critical to the success of drug targeting efforts.

3. Integrating disparate data

Since different experimental approaches and their resulting networks represent conceptualizations of a single underlying system, it is important to integrate different types of data and network representations to provide a more complete picture of this system. Recent advances in methods for HT data acquisition have resulted in the ability to concomitantly measure many different aspects of the same system. Transcriptomics or next-generation sequencing can be used to assay the levels of coding and non-coding mRNAs. Mass spectrometry-assisted proteomics can provide global measurements of the abundance of proteins, small molecule metabolites and lipids. Genetic assays, such as variants on the yeast two-hybrid method, and mass-spectrometry have been used to provide global information about physical PPIs and complexes in several systems. A number of other HT methods provide large amounts of different kinds of data for specific applications: Chromatin immunoprecipitation (CHIP)-chip and the related CHIP-seq methods, both provide global measurements of binding sites targeted by a specific regulator; siRNA libraries coupled with other methods can provide information about the functioning of individual genes in the system of interest. Integration of these disparate types of data is a prerequisite to elucidating relationships and associations, and more complete and nuanced understanding of biological systems.

Integrating data from different sources is difficult for a number of reasons. One principal reason for this is the plethora of databases and the lack of standardization, for example, common shared identities and names, shared semantics, or shared and stable access mechanisms [31,78], for mining and extracting the data contained therein. A secondary reason is the lack of tools (http://www.ngpharma.com/article/So-Much-Data-so-Little-Time). Inroads are being made to develop tools to combine data preprocessing with downstream analysis (e.g., Bioconductor [29]), to meaningfully integrate genomic and proteomic data (e.g. [78]), and to promote interdisciplinary research.

Many approaches for linking datasets rely on linking individual components. If undertaken manually, this can be a prolonged process of searching through multiple databases to find common linking variables. The challenges facing computational tools are similar; there are a multitude of sequence identifiers from both nucleotide and protein sequences that must be cross-referenced with each other, and there are many different variants and redundancies in the data including splice variants and post-translational modifications. There are several tools and databases [16,40,42], which allow annotating the data and adding cross-reference identifiers to facilitate the integration, but very few allow doing the batch integration in an automated manner. The Bioinformatics Resource Manager (BRM) [71] is a tool we have developed in-house which allows for merging heterogeneous data across platforms, such as from different microarray platforms, and also merge datasets between data types (e.g. microarray and proteomics) using cross-referencing information from gene and protein sequencing databases.

After these technical hurdles have been crossed different sources of data can be compared for related entities (a gene and its cognate protein, e.g.). However, care must be taken to consider the biological meaning of the questions being posed. It must be remembered that correlation does not mean causation and there may be no straightforward connection between the observed correlations and the underlying network. Strong correlations may be seen between components that are, in truth, distant, whereas no correlation may be observed between components that might be expected to be highly correlated. One notable example of this is the well-known disparity between mRNA levels and associated protein levels [78].

In some cases a metric, correlation coefficient or other similarity measure, is used to compare the two components based on the two different sources of data. Such a measure can then be used for other purposes, to create a visualization of the network, as in a correlation network [3], or in an analytical context to test whether associations between disparate sources of functional genomic data are significant [5].

Novel and useful insights have been gained by integrating multiple types of data together [3,10,48,82]. Wienkoop, et al. [82] detected a significant association between the protein ATGRP7 and metabolite cold-acclimation markers, proline and glutamine. Additional analyses allowed them to detect protein-metabolite co-regulation and the assignment of a circadian output regulated RNA-binding protein to these processes.

Bushel et al. [10] simultaneously clustered gene expression, clinical chemistry, and histopathology data to demonstrate that the biological process of cell growth and maintenance, amine metabolism, and stress response could discern levels of acetaminophen-induced centrilobular necrosis. Using correlation network analysis for data integration, Adourian et al. [3] identified various plasma molecules as suitable biomarkers of drug-induced hepatic alterations of lipid metabolism and urea cycle processes in rats administered a toxic compound. It is interesting to note that some of these biologically relevant biomarkers did not have very high correlations with other components. That is, important biomarkers of disease states may not be members of large co-expressed clusters. We have utilized a Bayesian integration model of high-throughput proteomics and metabolomics data to improve early detection of microbial infections [79].

A data integration method that is not dependent on probe or sequence annotation is co-inertia analysis [17]. It combines the information from different layers of the data (e.g., samples, genes, proteins, metabolites, GO information), and links multiple datasets on the sample level, rather than the individual component level (genes/proteins/metabolites). It therefore includes all variables in the analysis and thus there is no loss of information from any one component. This approach has been applied to gene expression and protein abundance measurements to discriminate melanomas from epithelial carcinomas [26].

Integration of data of different types can be thought of as a first step to revealing the underlying system in greater detail. Though we have summarized some research in this area, it is important to note that data integration on the HT scale is a relatively new area but will continue to increase in importance as a tool to understand disease processes and also to identify more robust biomarkers of the disease state.

4. Characterization of disease networks

Several studies have attempted to gain insight into networks of interactions associated with disease states using combinations of computational and experimental approaches. Lim et al. [46] developed an interaction network for cerebellar Purkinje cell (PC) degeneration encompassing a small number of proteins associated with a variety of ataxias. Using a stringent yeast two-hybrid screen they were able to identify PPIs to establish a network for ataxia. A similar study [1] was able to connect known angiogenesis related genes with unknown signaling components generating a global network pattern for vascular homeostasis. These studies established experimentally-determined interactomes around a central disease phenotype but this approach requires a considerable investment of time and effort.

A more common approach is to combine high-throughput measurements with PPI networks to highlight subnetworks that are significantly differentially regulated in the disease state. In this approach subnetworks are identified from PPIs and assigned a score based on the expression levels of the subnetwork member genes. High scoring subnetworks can then be associated with the disease state, and may include both genes/proteins known to be involved in the disease and novel predictions. Variations on this approach have been used to describe modules associated with diabetes [47], neurodegenerative disorders [32] and cancer [77], as well as modules involved in aging that are relevant to disease [86]. In a related approach, known associations between diseases and genes have been combined with PPI networks to provide insight into the commonalities and relationships between complex diseases and to identify novel proteins involved in specific diseases [44,84].

Coexpression networks have also been used to identify disease-associated subnetworks, in a similar fashion. The approach involves inferring gene association networks from microarray data from a number of different samples, identifying coexpressed modules, then examining these modules for associations with specific disease processes. An example of this approach is described in Keller, et al. where coexpression network analysis was used to characterize modules involved in diabetes [39].

The origins of disease are dependent on a large number of factors such as genetic predisposition, diet, lifestyle, infection, accidental trauma, and stress. Variations in disease states and effective treatments for particular patients have long been associated in various ways with these factors through longitudinal and other types of associative studies. Through HT genetic association studies, biomarkers have been identified for a wide range of diseases [4], mainly using traditional statistical methods that aim to associate a very small number of genetic elements (e.g. genes or loci) with a disease state. These approaches have been very successful for a small number of diseases, such as diabetes, myocardial infarction, and some cancers. It is clear that the interaction of complex biological networks with en-

vironmental and genetic variations form the basis of most common human diseases [25]. Therefore, combining genetic approaches with network and pathway analysis has emerged as a powerful new method for investigating complex disease.

The combination of quantitative trait loci (QTL) mapping with gene expression profiling, termed expression QTL (eQTL), is a powerful tool to elucidate the connection between genotype, expression, and phenotype for complex diseases [9]. Compared to traditional QTL mapping techniques, which map genetic loci, possibly containing many functional regions (genes, regulatory elements, or small RNAs), to particular disease phenotypes, eQTL can make it easier to identify the causative components of disease and to identify robust biomarkers. Several approaches have been used to identify networks and pathways that mediate disease. For example, we have used a combination of eQTL data and PPIs to identify the likely mediators of differential expression and determine a pathway of molecular interactions that explains this phenotype [65] in yeast. Chen, et al. (2008) described an approach that identifies modules from associative networks that are perturbed by QTLs and thus lead to obesity, diabetes, and atherosclerosis [13]. This study suggests that there are large networks that can modulate susceptibility to metabolic disease and that the disease may be an emergent property of the underlying network.

5. Network topology and disease mediators

One promising method to identify causative mediators (drivers) of systems-level phenotypes is topological analysis of biological networks. Instead of focusing on the most differentially regulated components, which are likely to be reactive effectors (the driven), this approach utilizes the structure of the network to identify the mediators of system phenotypes. We believe that appropriate analyses of high-throughput data can allow discrimination of the drivers of disease from those biomarkers and processes that are driven by disease.

This kind of analysis of biological networks was first reported in the yeast PPI network where topological hubs and bottlenecks were found to be significantly enriched in proteins that were essential for growth in rich media [88]. Hubs are defined as proteins that have many interactions. Bottlenecks are highly central proteins that link different parts of the network, similar to a bridge that links different parts of a city. Bottlenecks are thought to be potential points of information flow through the system (Fig. 3). This study showed that the architecture of networks was biologically relevant, since network properties were correlated with organismal phenotypes. Similar approaches were used to show that hubs and bottlenecks from the human interactome are more likely than other proteins to be targeted for modulation by pathogens [21] and are more likely to be successful drug targets [87]. A study of cancer related genes found that their protein products were more likely to be hubs in the human PPI network [64]. Finally, an analysis of topological properties of the human PPI network was able to confidently identify genes known to be involved in a variety of hereditary diseases and to predict novel disease genes [85]. Collectively, these studies show that insight into systems-level phenotypes relevant to disease processes can be obtained from the topology of PPI networks.

Due to the limitations of PPI networks discussed above, we have extended this approach to networks inferred from HT data. We found that the topology of these inferred networks revealed genes important in virulence for *Salmonella enterica* serovar Typhimurium [51]. In this study we constructed networks from transcriptional data using a statistical inference method [27]. We then showed that bottlenecks in the network were significantly enriched in genes known to be essential for virulence. Other bottleneck genes were also implicated in virulence, showing that the inferred network could be used as a tool to identify important genes. This study also indicated that bottlenecks might be mediators of information flow in the system, a point that is discussed further below.

We have used a network-topology-based approach for data integration to investigate important proteins and lipid components in Hepatitis C virus (HCV) infection [20]. We first constructed a co-abundance network based on a global proteomics profile of HCV-infected cells at various time points post-infection and found that bottlenecks in this network were significantly enriched in known targets of pathogens in general, and specifically in known HCV targets. We then used a simple method to integrate relationships from lipids identified by mass-spectrometry, and known PPIs. We evaluated the integrated network and found that it had significantly increased the enrichment of bottlenecks in known pathogen targets. In this application we used the topological qualities of the network to predict important proteins and lipid moieties for the functioning of the system. This application does not rely on the 'truth' of individual interactions or associations in the networks generated, but rather evaluates network quality based

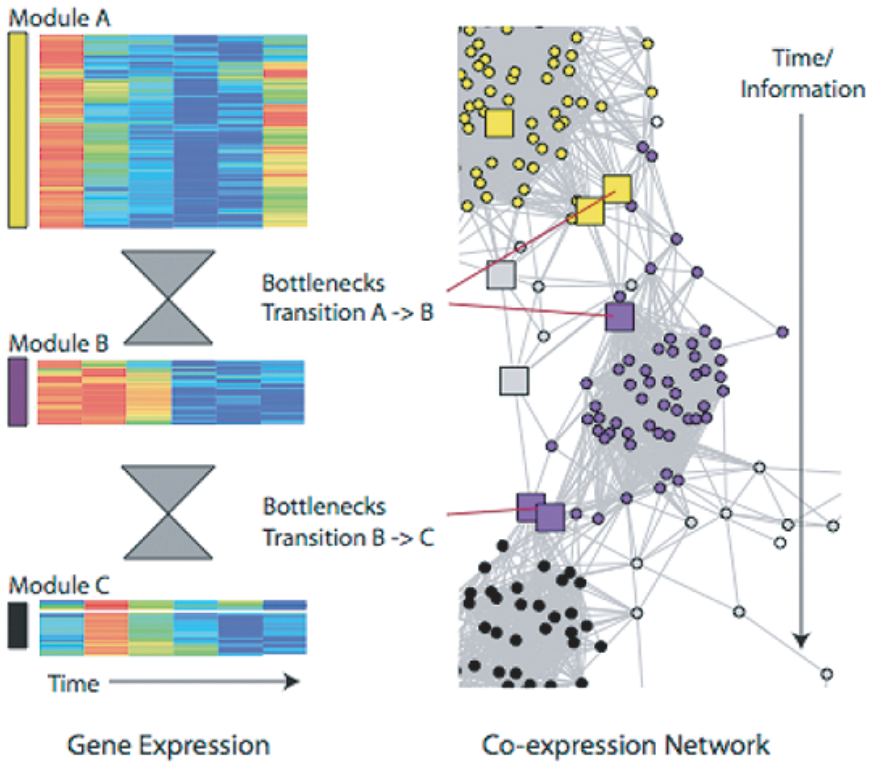

Fig. 3. Topology of coexpression networks and its impact on information flow in the system. A heat map of gene expression during a time course is shown on the left broken into co-expressed modules. On the right is the corresponding co-expression network with topological bottlenecks shown as squares. The bottlenecks link modules and thus represent drivers of the transitions between states of the system, in this case temporal states.

its ability to make accurate predictions about the importance of individual components based on their topological role in the network. An intriguing implication is that integrating different types of data, even in a simple way, can improve the quality of the network. This finding highlights the important point that each method for experimental determination is measuring a different aspect of the same underlying biological network.

Using topological analysis in association networks is a powerful tool in the study of disease progression and can provide valuable hypotheses specific to the system being studied. We have applied this approach to a mouse model of neuroprotection during stroke. Briefly, we inferred coexpression networks from expression data gathered from mouse brain over a time course under several different neuroprotection treatments [50]. We then identified bottlenecks from these networks and characterized modules of coexpressed genes. We found that bottlenecks were significantly more likely to be conserved in other organisms than other genes in the network, indicating that they are important for the functioning of the system (McDermott et al., manuscript in preparation). We also found that many of the bottlenecks with the highest centrality have been reported to be involved in stroke including the important hypoxia inducing factor (HIF-1α), a transcription factor that has been shown to be a primary mediator of response to stroke [66].

Our results from these studies have suggested that analysis of the topological properties of associative networks can provide valuable insight into biomarkers that represent drivers of disease processes. Because of their greater level of evolutionary conservation and important roles in system functioning, these biomarkers may be more robust indicators of underlying disease and may be associated with early, generative processes of disease. Our results also show that topological analyses can be used to integrate data effectively and shed light on the mechanisms of disease processes.

6. Pathway analysis

Pathways represent molecular processes at an individual level as opposed to networks which are more

of a global (population-level) representation of the biological entities [69]. Knowledge of canonical signaling and metabolic pathways has been gathered from years of published studies and is curated in a number of databases and tools [24,37,41]. Pathway analysis is commonly performed on a manual basis to analyze experimental data (HT or otherwise) in the context of known pathways. However, some approaches have used pathway knowledge to characterize HT data in a less biased manner.

Pathway analysis has been combined with HT data in order to study signal transduction pathways [80]. Starting with a known signal transduction pathway, HT techniques are applied to measure signal transduction products in terms of gene expression levels (transcriptomics) or to observe relative protein abundance and/or phosphorylation states (proteomics) [38]. In a study investigating the ErbB receptor signaling pathway, antibody arrays were used to monitor activation, uptake, and signaling of pathway components. By overlaying relative abundance with phosphorylation, they were able to determine the fraction of phosphorylated receptors as a function of time [55]. In a similar study Chan et al., developed a multiplexed reverse phase protein (RPP) microarray platform to simultaneously measure site-specific phosphorylation for numerous signaling proteins [12]. The implications of performing HT experiments to develop and refine signal transduction pathways could be catalytic for obtaining predictive models. It also represents a medium for merging "bottom-up" and "top-down" approaches and further refining biological networks.

Methods have been developed to utilize the wealth of functional annotation data to map genes or proteins to pathways. There are two complementary approaches for conducting pathway analysis from expression data using data mining methods. One approach is to cluster the expression data to identify sets of genes following similar expression patterns and conducting functional enrichment to assign meaning to the gene clusters [19, 34]. This approach is good for getting an overview of the data by highlighting the significant pathways in the data. A complementary approach is to use gene set enrichment analysis (GSEA) techniques which group genes/proteins based on pathways of interest and detect differentially expressed pathways across datasets. This approach provides a way to detect modest but consistent changes in expression of a group of related genes, for example a known pathway that would be missed if only a few individual genes appeared significant [75,76].

In a related application, the cross-ontological analysis (XOA) tool [62] provides a sophisticated method for associating biological components based on their gene ontology assigned functions. This association is based on a sophisticated metric that can provide measures of similarity across ontologies (e.g., between cellular component and biological process). We have used XOA to provide explanatory information for the inferred functional relationships in the stroke study described previously. We identified a subnetwork from this analysis focused on the important regulator Hif1α. In the resulting network XOA provides an extra layer of biological explanation [68]. When combined with the other methods described, this approach can provide useful biological insight based on both top-down and bottom-up approaches.

In prokaryotic systems the analysis of metabolic networks and pathways has been very successful [60]. This success has been due to the availability of a large amount of data from closely aligned model systems, such as *E. coli*, as well as the ability to easily perturb the organism in multiple ways and assay the resulting effects on the system. A popular computational approach, flux balance analysis (FBA), seeks to solve an underdetermined biological system using physicochemical constraints [22,23]. Assuming that the system has reached steady state the network is optimized, and convex analysis approaches (elementary modes, and the canonical pathway descriptions called extreme pathways) are used to sample the state space [59,61]. Such approaches have been used to identify important enzymes in metabolic networks, targets for drug development and modes of metabolism [30,35,63]. Due to the lack of availability of comprehensive metabolic network information in higher eukaryotes, and the difficulties associated with generating the appropriate conditional data from these organisms, the use of metabolic FBA in the study of disease has been very limited in the number of components considered [11,15,18, 28]. However, as metabolic models improve and new methods of measuring appropriate components, such as metabolites, are developed, this kind of analysis will be very important in dissecting the causes of metabolic human diseases.

7. Information flow in disease

Bottlenecks in association networks link functional modules composed of coexpressed components, which often can be associated with phenotypes, including disease states. Thus, topological bottlenecks represent mediators of the transitions between these system

states/coexpression modules. Disease states are thus brought about when information flow in the system becomes dysregulated, and the system is stuck in a limited number of system states [89]. Identifying the points of control in the diseased system and, importantly, the mediators of transition from a normal to diseased phenotype, is important for effective treatment of the disease.

Emergent behavior arising from transitions between system states has been described in the literature. Mitchell et al. (2009) observed "anticipatory regulation" when subjecting *E. coli* to temperature increases followed by a drop in oxygen availability [52]. The concept of anticipatory regulation or stochastic switching suggests that the natural temporal order of stimuli is embedded in the wiring of regulatory networks [2, 43,57] and favors functional modularity. Though the same observation has not yet been described in disease it is likely that cells from eukaryotes have similar patterned behavior and that disease may result from the dysregulation of such programs. Dynamic transitions between system states represent both adaptability and stability within the normal phenotype. In an interesting study Nykter et al. [56] showed that systems transitioning through different states, in this case macrophages responding to stimulation with Toll-like receptor agonists, display "critical" behavior. Criticality is system behavior that is balanced between two states, stochastic and highly structured, and thus is thought to allow biological networks to be both stable to perturbation and adaptable to new situations. Importantly, this finding suggests that disease states may represent a deviation from criticality, and that this deviation may be apparent by monitoring a limited number of components of the system, the biomarkers of the disease state.

8. Conclusions and future directions

We have presented a range of applications of network analysis to the elucidation of biomarkers of disease etiology and state. These analyses provide useful information about the processes resulting from disease states, the functions and endpoints driven by underlying disease processes. They are also useful to begin to identify the drivers of disease, that is, the early mediators of disease. In single-gene hereditary diseases the driver could be considered to be the causative gene. However, the majority of human diseases result from the interaction of environmental factors (diet, pathogen or toxin exposure, *etc.*) and a complex network of biological components that is determined by individual genetics. For these diseases, new approaches must be developed such as the network topology approaches described here and the genetic network methods described in [69], to allow determination of the causes of the disease. This problem has a direct impact on biomarker discovery. Useful biomarkers are highly specific for the disease or state for which they are indicators. Downstream components and processes driven by disease may not provide the robust and specific prediction of disease state because they represent the endpoints of disease processes and may be activated by different disease states. Identification of the disease drivers provides a specific snapshot of the disease, potentially at an early stage. The distinction between drivers and driven is, of course, a simplification, and well-controlled traditional biomarker studies may be able to identify important disease drivers. However, as understanding unfolds of complex diseases as dysfunctions of the system, which are greater than the sum of the contributions of individual components, network-based models of disease will be essential.

Our brief review of the emerging field of systems biomarkers of disease has highlighted a number of hurdles that must be addressed to allow full realization of the potential. The first is a set of technical problems raised by the plethora of different databases, identifiers, and other descriptors for different biological components. The barriers imposed by these issues on the study of systems are often not reported [81] but become more significant when data from different sources need to be directly compared. A second hurdle is the integration of different types of data, HT and otherwise, to allow identification of useful biomarkers and to facilitate understanding of mechanisms of disease. The rapid progress in developing new HT techniques has outstripped methods to interpret these types of data in thoughtful and non-trivial ways. The growing availability of matched datasets for the same system from different methods provides an opportunity for development and testing of novel methodologies. The tendency of researchers from all backgrounds to focus on determining simple causes for diseases (i.e. a mutation in a single gene that causes or is associated with a disease) is driven by the lack of understanding of the behavior of systems from both theoretical and practical standpoints. An improved understanding of systems in terms of emergent properties, dynamics, and methods to analyze systems properties both computationally and experimentally, is needed to provide a cultural background for considering complex disease. Finally, many

of the studies discussed above investigate intracellular networks and pathways, using cells from various tissues or sources. Many tissues may not be sufficiently accessible to provide clinically useful biomarker candidates. Thus a better understanding of networks and pathways in accessible tissues or compartments, such as blood, is needed.

In the emerging field of systems identification of biomarkers of disease the selection of the appropriate input data and analysis methods is very important to provide relevant, robust, and specific biomarkers. The flavors of network analysis we have outlined each have their own technical and biological implications. Networks of PPIs have been used in a variety of applications related to disease. A significant issue with these networks is that they are likely both false positive and false negative rich. Therefore, results and predictions must be compared with other existing sources of data and carefully validated experimentally. Associative networks inferred from HT data have emerged, in part, as a way to circumvent the technical issues associated with HT PPI networks. However, associative networks represent a very different, though overlapping, view of the underlying system. Association does not imply causation and further experimental validation is always necessary to confirm associations observed in HT data. A second consideration of associative networks is that the meaning of the network and associations between its components is determined by the HT data sets that were used to generate it. For experiments containing a limited number of related samples (e.g., microarrays from different patients with the same underlying disease) associative networks can be relevant to the processes and dynamics of the disease being studied. Conversely, associative networks based on a large number of samples from different phenotypes, genetic backgrounds, and/or environmental conditions, will provide insight into more general regulatory pathways. Pathway analysis approaches are more appropriate for studies in which there is a large body of existing knowledge about the underlying mechanism of the disease. Pathway analysis methods are also useful when dealing with a limited amount of HT data. For example, functional enrichment of differentially regulated gene sets is widely used to provide biological information about HT data. In some cases the amount of novel insight into the disease that can be gained from such analysis may be limited and is not likely to identify new components of importance that have not been well studied. Investigators naturally focus on what is familiar and understood when presented with an excess of information.

A real concern with the development of computational methods for determination of biomarkers of human disease is that they are very dependent on the algorithms used for the analysis. Conversely, method development on existing datasets suffers from the issue that the developers actually know the answers they are looking for (in some cases), and this can mean that the novel methods developed for a particular dataset(s) will not generalize well to new datasets. Other scientific communities have attempted to address this problem through the use of community-wide competitions, in which novel datasets are presented to participants, and the results of experimental validation are available, but held from the participants until after the competition is completed. In this way the community as a whole can evaluate the performance of its best methods on real-world problems in a largely unbiased fashion. This process also drives the development of better and more robust algorithms for data analysis, and the discussions following these competitions further refine the collective thinking and goals of the community. The protein structure prediction community hosts the critical assessment of structural prediction (CASP) every other year to assess the state of the field of prediction of protein structure from primary protein sequence [54]. The closely-related critical assessment of prediction of interactions (CAPRI) evaluates the ability of the community to determine structures of bound complexes and protein interactions [45]. The dialogue for reverse engineering assessments and methods (DREAM) provides a forum for evaluation of methods for inference of networks from HT data [74]. Finally, the critical assessment of microarray data analysis (CAMDA) provides a forum for competitive evaluation of microarray analysis techniques for specific applications [36]. The field of computational identification of biomarkers of complex diseases would greatly benefit from such a forum. Careful consideration would have to be given to selection and/or generation of appropriate datasets for use in such a competition, since these datasets and their validation are likely to be more complex than the examples listed above for a variety of reasons.

In summary, it is becoming clear that the majority of human diseases are driven not by a single gene or protein, or even a small collection of components, but rather by a complex interplay between environmental considerations and complex networks that are shaped by the genetics of individuals. In light of this emerging understanding it is clear that more sophisticated methods, such as those discussed here, are necessary for understanding human disease and for identification of robust, specific and useful biomarkers of disease state.

References

[1] A. Abdollahi, C. Schwager, J. Kleeff, I. Esposito, S. Domhan, P. Peschke, K. Hauser, P. Hahnfeldt, L. Hlatky, J. Debus, J.M. Peters, H. Friess, J. Folkman and P.E. Huber, Transcriptional network governing the angiogenic switch in human pancreatic cancer, *Proc Natl Acad Sci U S A* **104** (2007), 12890–12895.

[2] M. Acar, J.T. Mettetal and A. van Oudenaarden, Stochastic switching as a survival strategy in fluctuating environments, *Nat Genet* **40** (2008), 471–475.

[3] A. Adourian, E. Jennings, R. Balasubramanian, W.M. Hines, D. Damian, T.N. Plasterer, C.B. Clish, P. Stroobant, R. McBurney, E.R. Verheij, I. Bobeldijk, J. van der Greef, J. Lindberg, K. Kenne, U. Andersson, H. Hellmold, K. Nilsson, H. Salter and I. Schuppe-Koistinen, Correlation network analysis for data integration and biomarker selection, *Mol Biosyst* **4** (2008), 249–259.

[4] D. Altshuler, M.J. Daly and E.S. Lander, Genetic mapping in human disease, *Science* **322** (2008), 881–888.

[5] R. Balasubramanian, T. LaFramboise, D. Scholtens and R. Gentleman, A graph-theoretic approach to testing associations between disparate sources of functional genomics data, *Bioinformatics* **20** (2004), 3353–3362.

[6] A. Beyer, S. Bandyopadhyay and T. Ideker, Integrating physical and genetic maps: from genomes to interaction networks, *Nat Rev Genet* **8** (2007), 699–710.

[7] R. Bonneau, Learning biological networks: from modules to dynamics, *Nat Chem Biol* **4** (2008), 658–664.

[8] P. Bork, L.J. Jensen, C. von Mering, A.K. Ramani, I. Lee and E.M. Marcotte, Protein interaction networks from yeast to human, *Curr Opin Struct Biol* **14** (2004), 292–299.

[9] R.B. Brem, J.D. Storey, J. Whittle and L. Kruglyak, Genetic interactions between polymorphisms that affect gene expression in yeast, *Nature* **436** (2005), 701–703.

[10] P.R. Bushel, R.D. Wolfinger and G. Gibson, Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes, *BMC Syst Biol* **1** (2007), 15.

[11] T. Cakir, S. Alsan, H. Saybasili, A. Akin and K.O. Ulgen, Reconstruction and flux analysis of coupling between metabolic pathways of astrocytes and neurons: application to cerebral hypoxia, *Theor Biol Med Model* **4** (2007), 48.

[12] S.M. Chan, J. Ermann, L. Su, C.G. Fathman and P.J. Utz, Protein microarrays for multiplex analysis of signal transduction pathways, *Nat Med* **10** (2004), 1390–1396.

[13] L. Chen, T. Tong and H. Zhao, Considering dependence among genes and markers for false discovery control in eQTL mapping, *Bioinformatics* **24** (2008), 2015–2022.

[14] Y. Chen, J. Zhu, P.Y. Lum, X. Yang, S. Pinto, D.J. MacNeil, C. Zhang, J. Lamb, S. Edwards, S.K. Sieberts, A. Leonardson, L.W. Castellini, S. Wang, M.F. Champy, B. Zhang, V. Emilsson, S. Doss, A. Ghazalpour, S. Horvath, T.A. Drake, A.J. Lusis and E.E. Schadt, Variations in DNA elucidate molecular networks that cause disease, *Nature* **452** (2008), 429–435.

[15] L. Coquin, J.D. Feala, A.D. McCulloch and G. Paternostro, Metabolomic and flux-balance analysis of age-related decline of hypoxia tolerance in Drosophila muscle tissue, *Mol Syst Biol* **4** (2008), 233.

[16] R.G. Cote, P. Jones, L. Martens, S. Kerrien, F. Reisinger, Q. Lin, R. Leinonen, R. Apweiler and H. Hermjakob, The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases, *BMC Bioinformatics* **8** (2007), 401.

[17] A.C. Culhane, G. Perriere and D.G. Higgins, Cross-platform comparison and visualisation of gene expression data using co-inertia analysis, *BMC Bioinformatics* **4** (2003), 59.

[18] R.K. Dash, Y. Li, J. Kim, D.A. Beard, G.M. Saidel and M.E. Cabrera, Metabolic dynamics in skeletal muscle during acute reduction in blood flow and oxygen supply to mitochondria: in-silico studies using a multi-scale, top-down integrated model, *PLoS One* **3** (2008), e3168.

[19] G. Dennis, Jr., B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane and R.A. Lempicki, DAVID: Database for Annotation, Visualization, and Integrated Discovery, *Genome Biol* **4** (2003), P3.

[20] D.L. Diamond, A.J. Syder, J.M. Jacobs, C.M. Sorensen, K. Walters, S.C. Proll, J.E. McDermott, M.A. Gritsenko, Q. Zhang, R. Zhao, T.O. Metz, D.G. Camp, 2nd, K.M. Waters, R.D. Smith, C.M. Rice and M.G. Katze, Temporal proteome and lipidome profiles reveal hepatitis C virus-associated reprogramming of hepatocellular metabolism and bioenergetics, *PLoS Pathogens* **6**(1) (2010), e1000719.

[21] M.D. Dyer, T.M. Murali and B.W. Sobral, The landscape of human proteins interacting with viruses and other pathogens, *PLoS Pathog* **4** (2008), e32.

[22] J.S. Edwards, M. Covert and B. Palsson, Metabolic modelling of microbes: the flux-balance approach, *Environ Microbiol* **4** (2002), 133–140.

[23] J.S. Edwards and B.O. Palsson, Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions, *BMC Bioinformatics* **1** (2000), 1.

[24] S. Ekins, Y. Nikolsky, A. Bugrim, E. Kirillov and T. Nikolskaya, Pathway mapping tools for analysis of high content data, *Methods Mol Biol* **356** (2007), 319–350.

[25] V. Emilsson, G. Thorleifsson, B. Zhang, A.S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G.B. Walters, S. Gunnarsdottir, M. Mouy, V. Steinthorsdottir, G.H. Eiriksdottir, G. Bjornsdottir, I. Reynisdottir, D. Gudbjartsson, A. Helgadottir, A. Jonasdottir, U. Styrkarsdottir, S. Gretarsdottir, K.P. Magnusson, H. Stefansson, R. Fossdal, K. Kristjansson, H.G. Gislason, T. Stefansson, B.G. Leifsson, U. Thorsteinsdottir, J.R. Lamb, J.R. Gulcher, M.L. Reitman, A. Kong, E.E. Schadt and K. Stefansson, Genetics of gene expression and its effect on disease, *Nature* **452** (2008), 423–428.

[26] A. Fagan, A.C. Culhane and D.G. Higgins, A multivariate analysis approach to the integration of proteomic and gene expression data, *Proteomics* **7** (2007), 2162–2171.

[27] J.J. Faith, B. Hayete, J.T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J.J. Collins and T.S. Gardner, Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles, *PLoS Biol* **5** (2007), e8.

[28] J.D. Feala, L. Coquin, D. Zhou, G.G. Haddad, G. Paternostro and A.D. McCulloch, Metabolism as means for hypoxia adaptation: metabolic profiling and flux balance analysis, *BMC Syst Biol* **3** (2009), 91.

[29] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y. Yang and J. Zhang, Bioconductor: open software development for computational biology and bioinformatics, *Genome Biol* **5** (2004), R80.

[30] E.P. Gianchandani, A.R. Joyce, B.O. Palsson and J.A. Papin, Functional states of the genome-scale Escherichia coli transcriptional regulatory system, *PLoS Comput Biol* **5** (2009), e1000403.

[31] C. Goble, R. Stevens, D. Hull, K. Wolstencroft and R. Lopez, Data curation + process curation = data integration + science, *Brief Bioinform* **9** (2008), 506–517.
[32] J. Goni, F.J. Esteban, N.V. de Mendizabal, J. Sepulcre, S. Ardanza-Trevijano, I. Agirrezabal and P. Villoslada, A computational analysis of protein-protein interaction networks in neurodegenerative diseases, *BMC Syst Biol* **2** (2008), 52.
[33] J. Hollunder, A. Beyer and T. Wilhelm, Identification and characterization of protein subcomplexes in yeast, *Proteomics* **5** (2005), 2082–2089.
[34] W. Huang da, B.T. Sherman and R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat Protoc* **4** (2009), 44–57.
[35] R.U. Ibarra, P. Fu, B.O. Palsson, J.R. DiTonno and J.S. Edwards, Quantitative analysis of Escherichia coli metabolic phenotypes within the context of phenotypic phase planes, *J Mol Microbiol Biotechnol* **6** (2003), 101–108.
[36] K. Johnson and S. Lin, Call to work together on microarray data analysis, *Nature* **411** (2001), 885.
[37] M. Kanehisa and S. Goto, KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res* **28** (2000), 27–30.
[38] A. Kel, N. Voss, R. Jauregui, O. Kel-Margoulis and E. Wingender, Beyond microarrays: Finding key transcription factors controlling signal transduction pathways, *BMC Bioinformatics* **7**(Suppl 2) (2006), S13.
[39] M.P. Keller, Y. Choi, P. Wang, D.B. Davis, M.E. Rabaglia, A.T. Oler, D.S. Stapleton, C. Argmann, K.L. Schueler, S. Edwards, H.A. Steinberg, E. Chaibub Neto, R. Kleinhanz, S. Turner, M.K. Hellerstein, E.E. Schadt, B.S. Yandell, C. Kendziorski and A.D. Attie, A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility, *Genome Res* **18** (2008), 706–716.
[40] P.J. Kersey, J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney and R. Apweiler, The International Protein Index: an integrated database for proteomics experiments, *Proteomics* **4** (2004), 1985–1988.
[41] T.S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D.S. Somanathan, A. Sebastian, S. Rani, S. Ray, C.J. Harrys Kishore, S. Kanth, M. Ahmed, M.K. Kashyap, R. Mohmood, Y.L. Ramachandra, V. Krishna, B.A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady and A. Pandey, Human Protein Reference Database–2009 update, *Nucleic Acids Res* **37** (2009), D767–D772.
[42] P. Khatri, C. Voichita, K. Kattan, N. Ansari, A. Khatri, C. Georgescu, A. Tarca and S. Draghici, Onto-Tool: New Additions and Improvements in 2006, *Nucleic Acids Res* **25** (2007), W206–W211.
[43] E. Kussell and S. Leibler, Phenotypic diversity, population growth, and information in fluctuating environments, *Science* **309** (2005), 2075–2078.
[44] K. Lage, E.O. Karlberg, Z.M. Storling, P.I. Olason, A.G. Pedersen, O. Rigina, A.M. Hinsby, Z. Tumer, F. Pociot, N. Tommerup, Y. Moreau and S. Brunak, A human phenome-interactome network of protein complexes implicated in genetic disorders, *Nat Biotechnol* **25** (2007), 309–316.
[45] M.F. Lensink, R. Mendez and S.J. Wodak, Docking and scoring protein complexes: CAPRI 3rd Edition, *Proteins* **69** (2007), 704–718.
[46] J. Lim, T. Hao, C. Shaw, A.J. Patel, G. Szabo, J.F. Rual, C.J. Fisk, N. Li, A. Smolyar, D.E. Hill, A.L. Barabasi, M. Vidal and H.Y. Zoghbi, A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration, *Cell* **125**(4) (2006), 801–814.
[47] M. Liu, A. Liberzon, S.W. Kong, W.R. Lai, P.J. Park, I.S. Kohane and S. Kasif, Network-based analysis of affected biological processes in type 2 diabetes models, *PLoS Genet* **3** (2007), e96.
[48] L.J. Lu, Y. Xia, A. Paccanaro, H. Yu and M. Gerstein, Assessing the limits of genomic data integration for predicting protein networks, *Genome Res* **15** (2005), 945–953.
[49] A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera and A. Califano, ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *BMC Bioinformatics* **7**(Suppl 1) (2006), S7.
[50] B. Marsh, S.L. Stevens, A.E. Packard, B. Gopalan, B. Hunter, P.Y. Leung, C.A. Harrington and M.P. Stenzel-Poore, Systemic lipopolysaccharide protects the brain from ischemic injury by reprogramming the response of the brain to stroke: a critical role for IRF3, *J Neurosci* **29** (2009), 9839–9849.
[51] J.E. McDermott, R.C. Taylor, H. Yoon and F. Heffron, Bottlenecks and hubs in inferred networks are important for virulence in Salmonella typhimurium, *J Comput Biol* **16** (2009), 169–180.
[52] A. Mitchell, G.H. Romano, B. Groisman, A. Yona, E. Dekel, M. Kupiec, O. Dahan and Y. Pilpel, Adaptive prediction of environmental changes by microorganisms, *Nature* **460** (2009), 220–224.
[53] M.F. Moffatt, M. Kabesch, L. Liang, A.L. Dixon, D. Strachan, S. Heath, M. Depner, A. von Berg, A. Bufe, E. Rietschel, A. Heinzmann, B. Simma, T. Frischer, S.A. Willis-Owen, K.C. Wong, T. Illig, C. Vogelberg, S.K. Weiland, E. von Mutius, G.R. Abecasis, M. Farrall, I.G. Gut, G.M. Lathrop and W.O. Cookson, Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma, *Nature* **448** (2007), 470–473.
[54] J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost and A. Tramontano, Critical assessment of methods of protein structure prediction – Round VIII, *Proteins* **77 Suppl 9** (2009), 1-4.
[55] U.B. Nielsen, M.H. Cardone, A.J. Sinskey, G. MacBeath and P.K. Sorger, Profiling receptor tyrosine kinase activation by using Ab microarrays, *Proc Natl Acad Sci U S A* **100** (2003), 9330–9335.
[56] M. Nykter, N.D. Price, M. Aldana, S.A. Ramsey, S.A. Kauffman, L.E. Hood, O. Yli-Harja and I. Shmulevich, Gene expression dynamics in the macrophage exhibit criticality, *Proc Natl Acad Sci U S A* **105** (2008), 1897–1900.
[57] E. Oxman, U. Alon and E. Dekel, Defined order of evolutionary adaptations: experimental evidence, *Evolution* **62** (2008), 1547–1554.
[58] W. Pan, A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments, *Bioinformatics* **18** (2002), 546–554.
[59] J.A. Papin, N.D. Price and B.O. Palsson, Extreme pathway lengths and reaction participation in genome-scale metabolic networks, *Genome Res* **12** (2002), 1889–1900.
[60] J.A. Papin, N.D. Price, S.J. Wiback, D.A. Fell and B.O. Palsson, Metabolic pathways in the post-genome era, *Trends Biochem Sci* **28** (2003), 250–258.
[61] J.A. Papin, J. Stelling, N.D. Price, S. Klamt, S. Schuster and B.O. Palsson, Comparison of network-based pathway analysis methods, *Trends Biotechnol* **22** (2004), 400–405.
[62] C. Posse, B. Sanfilippo, R. Gopalan, R. Riensche, N. Beagley and B. Baddeley, Cross-Ontological Analytics: Combining Associative and Hierarchical Relations in the Gene Ontologies to Assess Gene Product Similarity, *Lecture notes in computer*

science **3992** (2006), 871–878.
[63] N.D. Price, J.A. Papin, C.H. Schilling and B.O. Palsson, Genome-scale microbial in silico models: the constraints-based approach, *Trends Biotechnol* **21** (2003), 162–169.
[64] D. Rambaldi, F.M. Giorgi, F. Capuani, A. Ciliberto and F.D. Ciccarelli, Low duplicability and network fragility of cancer genes, *Trends Genet* **24** (2008), 427–430.
[65] I. Rashid, J. McDermott and R. Samudrala, Inferring molecular interactions pathways from eQTL data, *Methods Mol Biol* **541** (2009), 211–223.
[66] R.R. Ratan, A. Siddiq, N. Smirnova, K. Karpisheva, R. Haskew-Layton, S. McConoughey, B. Langley, A. Estevez, P.T. Huerta, B. Volpe, S. Roy, C.K. Sen, I. Gazaryan, S. Cho, M. Fink and J. LaManna, Harnessing hypoxic adaptation to prevent, treat, and repair stroke, *J Mol Med* **85** (2007), 1331–1338.
[67] J.F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G.F. Berriz, F.D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D.S. Goldberg, L.V. Zhang, S.L. Wong, G. Franklin, S. Li, J.S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R.S. Sikorski, J. Vandenhaute, H.Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M.E. Cusick, D.E. Hill, F.P. Roth and M. Vidal, Towards a proteome-scale map of the human protein-protein interaction network, *Nature* **437** (2005), 1173–1178.
[68] A. Sanfilippo, B. Baddeley, N. Beagley, J.E. McDermott, R. Riensche, R.C. Taylor and B. Gopalan, Using the gene ontology to enrich biological pathways, *International Journal of Computational Biology and Drug Design* **2**(3) (2009), 221–235.
[69] E.E. Schadt, Molecular networks as sensors and drivers of common human diseases, *Nature* **461** (2009), 218–223.
[70] E.E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. Guhathakurta, S.K. Sieberts, S. Monks, M. Reitman, C. Zhang, P.Y. Lum, A. Leonardson, R. Thieringer, J.M. Metzger, L. Yang, J. Castle, H. Zhu, S.F. Kash, T.A. Drake, A. Sachs and A.J. Lusis, An integrative genomics approach to infer causal associations between gene expression and disease, *Nat Genet* **37** (2005), 710–717.
[71] A.R. Shah, M. Singhal, K.R. Klicker, E.G. Stephan, H.S. Wiley and K.M. Waters, Enabling high-throughput data management for systems biology: the Bioinformatics Resource Manager, *Bioinformatics* **23** (2007), 906–909.
[72] E. Sprinzak, Y. Altuvia and H. Margalit, Characterization and prediction of protein-protein interactions within and between complexes, *Proc Natl Acad Sci U S A* **103** (2006), 14718–14723.
[73] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F.H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach and E.E. Wanker, A human protein-protein interaction network: a resource for annotating the proteome, *Cell* **122** (2005), 957–968.
[74] G. Stolovitzky, R.J. Prill and A. Califano, Lessons from the DREAM2 Challenges, *Ann N Y Acad Sci* **1158** (2009), 159–195.
[75] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander and J.P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proc Natl Acad Sci U S A* **102** (2005), 15545–15550.
[76] J. Tomfohr, J. Lu and T.B. Kepler, Pathway level analysis of gene expression using singular value decomposition, *BMC Bioinformatics* **6** (2005), 225.
[77] E. Wang, A. Lenferink and M. O'Connor-McCourt, Cancer systems biology: exploring cancer-associated genes on cellular networks, *Cell Mol Life Sci* **64** (2007), 1752–1762.
[78] K.M. Waters, J.G. Pounds and B.D. Thrall, Data merging for integrated microarray and proteomic analysis, *Brief Funct Genomic Proteomic* **5** (2006), 261–272.
[79] B.J. Webb-Robertson, L.A. McCue, N. Beagley, J.E. McDermott, D.S. Wunschel, S.M. Varnum, J.Z. Hu, N.G. Isern, G.W. Buchko, K. McAteer, J.G. Pounds, S.J. Skerrett, D. Liggitt and C.W. Frevert, A Bayesian integration model of high-throughput proteomics and metabolomics data for improved early detection of microbial infections, *Pac Symp Biocomput* (2009), 451–463.
[80] A.D. Weston and L. Hood, Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine, *J Proteome Res* **3** (2004), 179–196.
[81] D. Wichadakul, J. McDermott and R. Samudrala, *Prediction and integration of regulatory and protein-protein interactions*, in: *Computational Systems Biology*, J. Mcdermott et al., eds, Humana Press: New York, 2009, pp. 101–144.
[82] S. Wienkoop, K. Morgenthal, F. Wolschin, M. Scholz, J. Selbig and W. Weckwerth, Integration of metabolomic and proteomic phenotypes: analysis of data covariance dissects starch and RFO metabolism from low and high temperature compensation response in Arabidopsis thaliana, *Mol Cell Proteomics* **7** (2008), 1725–1736.
[83] C.J. Wolfe, I.S. Kohane and A.J. Butte, Systematic survey reveals general applicability of guilt-by-association within gene coexpression networks, *BMC Bioinformatics* **6** (2005), 227.
[84] X. Wu, R. Jiang, M.Q. Zhang and S. Li, Network-based global inference of human disease genes, *Mol Syst Biol* **4** (2008), 189.
[85] J. Xu and Y. Li, Discovering disease-genes by topological features in human protein-protein interaction network, *Bioinformatics* **22** (2006), 2800–2805.
[86] H. Xue, B. Xian, D. Dong, K. Xia, S. Zhu, Z. Zhang, L. Hou, Q. Zhang, Y. Zhang and J.D. Han, A modular network model of aging, *Mol Syst Biol* **3** (2007), 147.
[87] L. Yao and A. Rzhetsky, Quantitative systems-level determinants of human genes targeted by successful drugs, *Genome Res* **18** (2008), 206–213.
[88] H. Yu, P.M. Kim, E. Sprecher, V. Trifonov and M. Gerstein, The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics, *PLoS Comput Biol* **3** (2007), e59.
[89] A. Yuryev, (ed.), Pathway analysis for drug discovery, in: *Introduction to Pathway Analysis*, John Wiley & Sons, eds, Inc: Hoboken, 2008.
[90] S. Zang, R. Guo, L. Zhang and Y. Lu, Integration of statistical inference methods and a novel control measure to improve sensitivity and specificity of data analysis in expression profiling studies, *J Biomed Inform* **40** (2007), 552–560.
[91] J. Zhu, B. Zhang, E.N. Smith, B. Drees, R.B. Brem, L. Kruglyak, R.E. Bumgarner and E.E. Schadt, Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks, *Nat Genet* **40** (2008), 854–861.