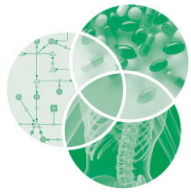
Focus Article



Protein–protein interaction networks and subnetworks in the biology of disease

Rod K. Nibbe,[1*] Salim A. Chowdhury,[2] Mehmet Koyutürk,[1,2] Rob Ewing[1,3] and Mark R. Chance[1,4]

The main goal of systems medicine is to provide predictive models of the pathophysiology of complex diseases as well as define healthy states. The reason is clear—we hope accurate models will ultimately lead to more specific and sensitive markers of disease that will help clinicians better stratify their patient populations and optimize treatment plans. In addition, we expect that these models will define novel targets for combating disease. However, for many complex diseases, particularly at the clinical level, it is becoming increasingly clear that one or a few genomic variations alone (e.g., simple models) cannot adequately explain the multiple phenotypes related to disease states, or the variable risks that attend disease progression. We suggest that models that account for the activities of many interacting proteins will explain a wider range of variability inherent in these phenotypes. These models, which encompass protein interaction networks dysregulated for specific diseases and specific patient sub-populations, will be constructed by integrating protein interaction data with multiple types of other relevant cellular information. Protein interaction databases are thus playing an increasingly important role in systems biology approaches to the study of disease. They present us with a static, but highly *functional* view of the cellular state, and thus give us a better understanding of not only the normal phenotype, but also the overall disease phenotype at the level of the whole organism when certain interactions become dysregulated. *WIREs Syst Biol Med* 2010
DOI: 10.1002/wsbm.121

INTRODUCTION

Protein–protein interaction (PPI) networks are important datasets driving basic and translational research. The datasets may capture direct interactions between proteins (physical), or indirect interactions (functional), and often both. As the regulated activities of proteins (e.g., enzymes, receptors, transcription factors, etc.) are the most important and immediate effectors of molecular phenotype, providing models of their regulated behavior and their interactions with a myriad of environmental factors is critical to understanding the wide spectrum of phenotypes. Of particular importance to researchers, studying the concerted interactions of many proteins as they function together in a network provides specific guidance for validation strategies that are well-suited to establishing the mechanistic cause(s) of disease. Genomic-based studies such as genome-wide profiling for driver mutations, mRNA expression (microarray or RNAseq) profiling, or genome-wide association studies (GWAS), have defined disease susceptibility genes and loci, and in turn have provided important targets for disease classification and mechanistic insight. However, in many cases they have limited application in informing clinical prognosis or driving the discovery of new drug targets. The likely reason

*Correspondence to: rkn6@case.edu

[1]Center for Proteomics and Bioinformatics, Case Western Reserve University, Cleveland, OH, USA

[2]Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH, USA

[3]Department of Genetics, Case Western Reserve University, Cleveland, OH, USA

[4]Department of Physiology and Biophysics, Case Western Reserve University, Cleveland, OH, USA

DOI: 10.1002/wsbm.121

for this is that DNA and RNA are not the proper end-point context in which to fully understand mechanisms of disease or the dysregulated molecular networks that we hypothesize cause and drive its progression. Genomic approaches are not by any means lacking, but function is frequently integrated downstream from genes and thus similar attention must be paid to protein function and integration to provide a complete picture of the phenotype.

Increasingly, we have advanced our technical capability to perform high-dimensional screens of the proteome to detect significant changes in thousands of proteins varying in disease. Similar increases in our understanding of the interactome of model organisms and humans are occurring as are improvements in ascertaining the confidence of these interactions. The integration of multiple data types, along with the wealth of genomic data, in the topological context of a PPI network layers these data within an ideal substrate that is well-suited to powerful computational methods, to create more accurate models of both health and disease. These models will likely drive novel approaches to personalized medicine, where we better understand the genetic background of patients and can predict their proteome response to the environment in the context of their interacting gene products. Still, many significant challenges remain to developing a network-based understanding of biology. These include incompleteness of the human interactome, both in terms of the interactions that occur in a specific context and their regulatory logic, and how this incompleteness impacts our ability to mine these data.

Next, we discuss the extant technologies that have been used to build PPIs, followed by novel examples of the application of PPIs to the research of human disease, and finally a discussion of integrative computational approaches using PPIs. The focus of this review is by no means intended to diminish the importance of studies based on genetic interaction networks—studies that have also furthered our understanding of complex disease phenotypes. For a recent review of genetic interaction networks see Dixon et al.[1]

EXPERIMENTAL STRATEGIES FOR MAPPING PPIs

PPI mapping is a key component of systems-based approaches to understanding cellular function. This section will focus on efforts to map eukaryotic PPIs. Multiple experimental and computational methods have been used to identify and infer PPIs. However, genome-wide protein interaction maps have been largely generated using two complementary technologies: yeast two-hybrid (Y2H) and affinity-purification mass-spectrometry (AP-MS). High-throughput Y2H techniques were developed from initial observations that transcription factor activity can be reconstituted from physically separate activation (AD) and binding (BD) domains brought into close proximity.[2] By fusing the BD and AD to bait and prey proteins, respectively, the interaction of the bait and prey may be tested through reconstitution of transcription factor activity. Large-scale application of this technique is achieved by mating thousands of different yeast strains each expressing a different bait or prey fusion.[3] In contrast to the genetic approach used in Y2H, AP-MS relies on biochemical purification of protein complexes with subsequent identification of complex members using mass-spectrometry.[4] AP-MS combines the specificity of antibody-based protein purification with the sensitivity of mass-spectrometry and enables detection of protein complexes under approximately physiological conditions. In a similar fashion to Y2H, improvements on an initially low-throughput technique have led to the development of work-flows that can identify hundreds or thousands of protein complexes. Current AP-MS strategies typically make use of tagged bait proteins that are expressed in cells and then purified with an antibody against the tag.[5] Aside from their dissimilar experimental approaches, Y2H and AP-MS provide different but complementary views of the protein interactome; Y2H detects binary protein interactions, whereas AP-MS detects co-membership in protein complexes. Both techniques have been applied to multiple model organisms. Initial studies focused on prokaryotes and yeast,[6–10] whereas more recent applications have focused on the human protein interactome.[11–13]

Concomitant with the volumes of data that are generated from protein interactome studies, many studies have focused on the development of computational methods for prediction and analysis of protein interactions. Of particular relevance to this review, are the twin challenges of quality and coverage of current protein interactome maps. Both Y2H and AP-MS generate false positive and false negative results for multiple reasons. For example, in the conventional Y2H assay, bait and prey proteins that are not correctly localized to the nucleus may be recorded as false negatives. In addition, the conventional Y2H assay suffers from auto-activation whereby BD fusions can activate reporter gene expression in the absence of any interaction, generating false positives.[14] On the other hand, sticky proteins that are purified regardless of the identity of the bait protein are a significant source of false

positives in AP-MS data.[12] With both Y2H and AP-MS, false negatives may occur with classes of proteins that are incompatible with the technology (e.g., membrane or extra-cellular proteins) or with weak, transient interactions or interactions only occurring under specific cellular conditions (e.g., phosphorylated states or localization signals). As the technologies mature, systematic quantification of error rates and comparisons of results across studies and technologies are underway.[15] In the latter study, Y2H and AP-MS datasets were concluded to be of similar quality albeit providing different representations of the protein interactome; binary Y2H interactions were found to be enriched for transient signaling interactions and interactions between complexes, whereas AP-MS datasets favored detection of proteins within a complex.

Estimations of the number of interactions occurring, for example, in a typical human cell are useful guides to the scale of the challenge in obtaining a 'complete' interaction map. Different approaches yield different estimates; however, the size of the human interactome (presumably the set of protein–protein interactions occurring under a given set of conditions) has been estimated to be between 154,000 and 369,000[16] or as large as 650,000 interactions.[17] With the magnitude of this challenge and the incompleteness of current interactome maps, several groups have focused on designing more efficient strategies for mapping PPIs. A 'pay-as-you-go' strategy whereby likely network hubs are identified from each successive interaction proteomics experiment and then used as baits themselves was proposed to maximize the efficiency in terms of the number of baits required to cover the interaction network.[18] Alternatively, an approach that combines prioritization of binary interactions according to their probability of occurrence with pooling strategies was proposed to reduce the cost of covering the complete interactome.[19] These authors stressed the need for multiple pass interaction screening to provide sufficient confidence and coverage and emphasized the importance of experimental design in future global studies of the interactome. In the absence of a complete human interactome map, the remainder of this review will focus on data-mining strategies that can and have been used to identify significant subnetworks from incomplete interactome maps.

HUMAN DISEASE—GOING BEYOND GENES

Soon after Francis Crick published the central dogma of biology in 1970, the race was on to discover the genetic basis for a vast number of human diseases, many of which are monogenic.[20] However, many complex diseases such as diabetes and common cancers have polygenic causes. Patients afflicted with these diseases present a major challenge to clinicians, not only because of the genetic complexity but also due to polymorphic variation across patients.[21–24] With respect to genetics, human colorectal cancer (CRC), the second leading cause of cancer death in the USA and UK, is one of the most thoroughly researched of all human cancers. In a recent landmark study,[25] genome-wide profiling was used to identify 69 candidate driver genes significantly mutated in a sizeable cohort of CRC biopsies, yet up to half of the tumor samples used in the screen did not contain mutations in one or more of these genes. However, with the recognition that pathways and not individual genes drive tumorigenesis,[26] a subsequent network analysis of these genes[27] revealed that a number of them co-located on canonical signaling pathways in CRC. *[By pathway we mean a regulated sequence of protein interactions directed toward a particular outcome in the cell, such as the transcription of certain genes, apoptosis, etc. A signaling pathway refers to a special type of pathway that is activated (or deactivated) by the binding of a ligand, typically to a cell surface receptor. For example, the WNT-signaling pathway, which is often constitutively active in CRC]*. This study and others, where genetic variations were identified that associate with disease, reveal the power of PPIs as a context in which to understand how a disparate constellation of mutant gene products, known to cluster in the human PPI network,[28] can shed light on the functional patho-physiology of the disease. Likewise, many expression profiling experiments related to human CRC (>30 found at the Gene Expression Omnibus)[29] have revealed panels of genes whose overall expression profile can classify various CRC disease states, but frequently panels from similar experiments only partially overlap, and even separate analyses of the *same* dataset can lead to strikingly few predictor genes in common.[30] However, when gene expression data are combined with PPI data, the integrative approach may greatly advance our understanding of the functional basis of disease,[31] as can the addition of expression proteomic data[32,33] or, more generally, structural proteomic data.[34] Although common cancers like CRC are thought to be driven by somatic mutations, the pathways and networks that mediate the dysregulation at the level of the proteome are far less well understood. This and similar evidence from large-scale studies in other diseases, plus the rapid growth in the size and annotation of human PPIs, are increasing enthusiasm for network-based approaches for elucidating the mechanistic causes of

many diseases. The shift from single gene targets toward network-based targets is beginning to gain acceptance in the field of drug discovery[35,36] as well as among researchers and clinicians eager for better prognostic markers to improve disease stratification in patients.[37–39]

Beyond their application to the study of individual diseases, PPIs are now being employed to search for modular subnetworks that may play a role in more than one disease, in an effort to identify new disease susceptibility genes. For instance, Goh et al.[40] constructed a human disease network and a disease gene network using data obtained from the Online Mendelian Inheritance in Man (OMIM) database. When overlaid on a high-confidence human PPI, they observed that disorders and genes of the same class are more frequently linked to each other compared to random organizations of both. They also observed that the products of genes essential for life exhibit high degree (number of links to other proteins) in the PPI network compared to non-essential genes, whereas non-essential *disease* genes (1) exhibit lower degrees and (2) their mRNA expression pattern does not correlate with the expression pattern of the rest of genes in the cell. A similar study by Barrenas et al.[41] started with GWAS data obtained from a well-annotated catalog of SNPs associated with human disease. They largely corroborated the finding of Goh et al. and also observed that proteins in the close 'neighborhood' of disease subnetworks provide an important class of new candidate genes. Motivated by these insights, many computational methods that prioritize candidate genes in a linkage interval based on network proximity and connectivity to known disease genes have been developed.[42–44] See Figure 1 for a schematic representation of integrative network-based approaches.

Although the power of network-based approaches paired with equally powerful computational algorithms is promising, the importance of traditional wet-bench validation of the discovered candidate genes, subnetworks, or pathways cannot be

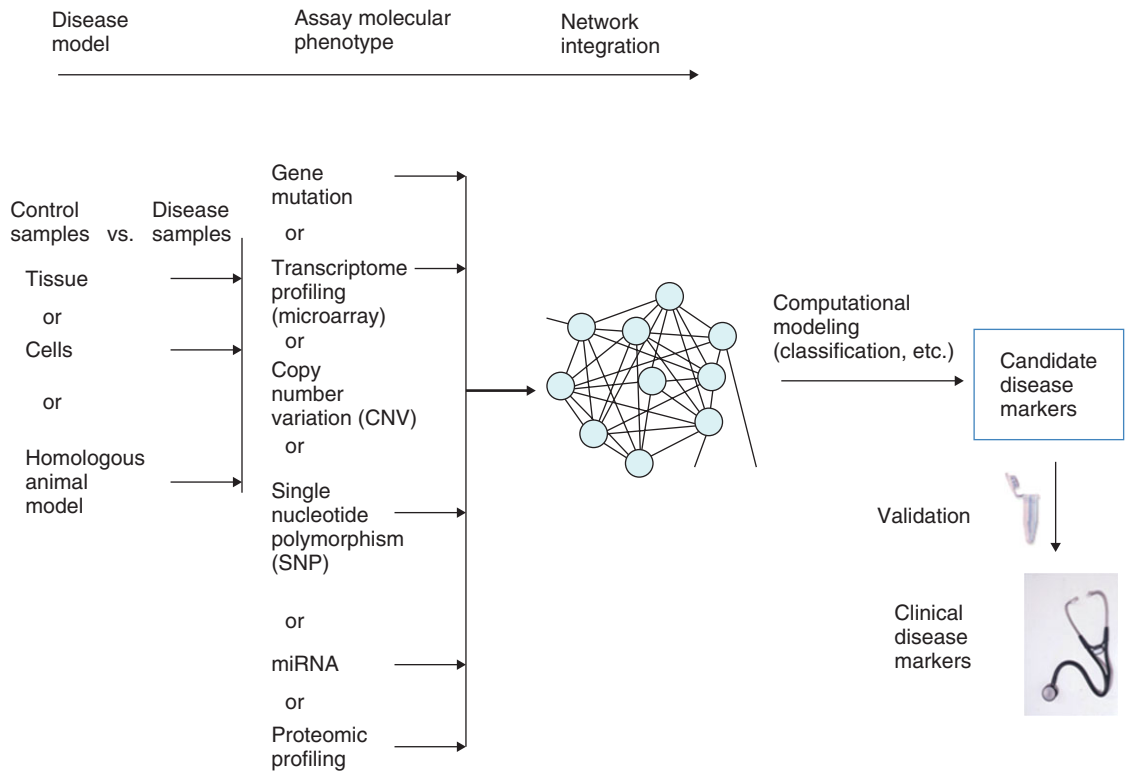


FIGURE 1 | Network-based disease modeling approach. Beginning with a model of disease, e.g., human tissues, cell culture, or relevant animal models, one may assay for significant changes between disease and control (e.g., mutations or differentially expressed genes or proteins or SNPs, etc.). The result of any one of these assays (or, theoretically, more than one) is used to 'seed' a computational search of the PPI for candidate subnetworks discriminative of the disease (see Box 1 for an example). The approach is motivated by the hypothesis that gene products with a role in disease tend to cluster in the interactome. Further computational modeling can be employed to assess the classification power of the candidate subnetworks to discriminate control from disease, and more importantly provide a basis for validation by perturbation analysis (e.g., siRNA screening) to drive validation of disease biomarkers for clinical utility.

understated. The validation of one or more *candidate* gene products and its role in disease is an important and necessary step in the translation of basic research to clinic utility. Next, we will discuss in more detail several integrative approaches that have been used in the context of network biology and disease, approaches that in many cases have provided candidate targets that merit validation.

NETWORK-BASED INTEGRATION OF GENE AND PROTEIN EXPRESSION DATA

PPI networks provide static and qualitative descriptions of the wiring of cellular systems.[45,46] Molecular expression data (e.g., mRNA expression, protein expression), on the other hand, provides quantitative information on the molecular composition of the system in different samples, conditions, tissues, or over time.[47,48] Consequently, it is natural to integrate these two sources of information to gain mechanistic insights on complex diseases. Indeed, integration of molecular expression data with protein–protein interactions is shown to enhance modularization of networks.[49–55] In other words, in the quest for identifying modular groups of proteins, co-expression patterns provide additional information to the connectivity patterns (e.g., high connectivity) indicative of modular function. Taylor et al.[56] also investigate modularity in the context of phenotypic differences and report alterations in network modularity in cancer. Furthermore, interpretation of gene expression in the context of protein–protein interactions is shown to improve the identification of disease genes.[57–59] These results indicate that identification of subnetworks that are dysregulated in a disease of interest may provide new insights in discovery of disease-related genes, improved diagnosis and prognosis, and development of systems-level intervention strategies.

In one of the early algorithmic studies, Ideker et al.[60] propose a method for identifying dysregulated subnetworks with respect to GAL80 deletion in yeast. This method is based on first assessing the dysregulation (differential expression) of each gene individually and then searching for connected subnetworks enriched in dysregulated genes. Variations of this method are shown to be effective in identifying multiple genetic markers in prostate cancer,[61] breast cancer,[56] melanoma,[62] aging,[63–65] Alzheimer's disease[58] and drug response.[66] All of these methods use network information to interpret the dysregulation of individual genes at the systems-level; however, they can be considered univariate analyses from a statistical perspective, as they assess differential expression individually for each gene.[67]

Using a multivariate approach, Chuang et al.[68] attempt to capture sample-specific variation in gene expression while identifying dysregulated subnetworks. They define subnetwork activity as the average mRNA-level expression of the proteins in the subnetwork. They then develop an information-theoretic scheme to assess the dysregulation of a subnetwork in terms of the mutual information between the subnetwork activity and sample class (e.g., normal vs tumor). This captures the *coordination* of multiple genes in discriminating normal and disease samples. The difference between this multivariate approach and earlier univariate approaches is illustrated in Figure 2. When used as features for classification, these *coordinately dysregulated subnetworks* clearly outperform single gene markers in predicting metastasis of breast cancer.[68]

Although quite useful, additive subnetwork activity captures the coordination between the dysregulation of interacting gene products only to a limited extent. Observing that coordinated changes in the mRNA-level expression of interacting proteins can exhibit combinatorial patterns as well, Chowdhury et al.[69] formulate coordinate dysregulation combinatorially and search for *subnetwork state functions* that are indicative of different stages of cancer. When used in conjunction with neural networks to train subnetwork-based classifiers, this method is shown to deliver excellent performance in predicting metastasis of colon cancer. The difference between additive and combinatorial coordinate dysregulation is illustrated in Figure 3. A sample subnetwork indicative of liver metastasis in colorectal cancer, identified using the combinatorial formulation of coordinate dysregulation, is shown in Figure 4.

As a stronger notion, Anastassiou[70] formulates the synergistic dysregulation of a subnetwork as the coordinate dysregulation that is not explained by smaller parts of the subnetwork. Synergy provides a stronger notion than coordinate dysregulation, as it corrects for the coordinate dysregulation of the subsets of the subnetwork, thereby capturing the pattern of dysregulation that emerges only when all genes in the subnetwork are considered (hence the term synergy). Although computation of synergy and identification of synergistic subnetworks are computationally intractable problems for arbitrary subnetwork size,[70] pair-wise assessment of synergy generates synergy networks for complex diseases, which can be interpreted in the context of physical interactions between proteins to gain mechanistic insights.[71] Besides coordinate dysregulation, differential co-expression is also shown

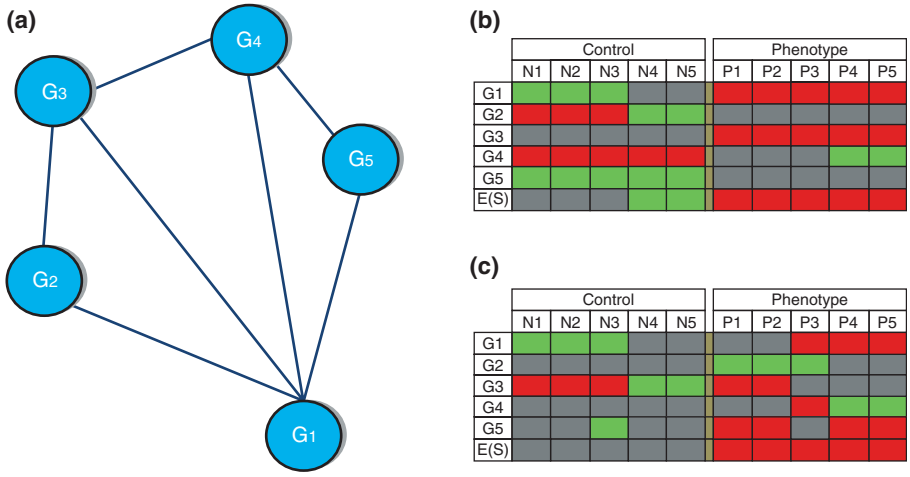


FIGURE 2 | Univariate versus multivariate assessment of subnetwork dysregulation. A hypothetical example illustrating the difference between univariate and multivariate approaches to identifying dysregulated subnetworks. A hypothetical subnetwork *S* of the human PPI network is shown in (a). Genes are shown as nodes; interactions between their products are shown as edges. In (b) and (c), each row shows a coding gene's expression level in control and phenotype samples. (b) and (c) each display different hypothetical scenarios for gene expression data. The last row shows *subnetwork activity*, which is computed as the average of the expression of these five genes in each sample.[68] Dark red shows high expression, light gray shows moderate expression, and light green shows low expression. The dysregulation (differential expression) of a gene (or a subnetwork) is measured in terms of how much its expression profile (or activity) can discriminate phenotype and control. Ideker et al.[60] define subnetwork dysregulation as the aggregate significance of the dysregulation of each gene, normalized by the number of genes in the subnetwork. We consider this a *univariate* approach as the dysregulation of each gene is assessed separately and then the results are combined to assess the dysregulation of the subnetwork. On the contrary, Chuang et al.[68] define the dysregulation of the subnetwork as the mutual information between phenotype and subnetwork activity, i.e., how much the average expression of the genes in the subnetwork can discriminate phenotype and control. We consider this a *multivariate* approach as the dysregulation of all genes in the subnetwork are assessed together to compute the dysregulation of the subnetwork. In (b), all genes exhibit maximum dysregulation, as each of them can perfectly discriminate phenotype and control. Consequently, the univariate approach can correctly identify this subnetwork as a dysregulated subnetwork, as all genes in the subnetwork are dysregulated. On the other hand, in (c), each gene exhibits moderate individual dysregulation, so the subnetwork would not be considered a dysregulated subnetwork by the univariate approach. However, in this case, the subnetwork activity can perfectly discriminate phenotype and control, thereby capturing the *coordinate* dysregulation of the genes in this subnetwork. This example demonstrates the potential of multivariate approaches in discovering dysregulated subnetworks, beyond what can be discovered by a univariate approach.

to be effective in identifying sets of genes that are co-expressed in disease samples although not being co-expressed in control samples and vice versa.[72,73]

Assessment of differential expression in terms of mRNA expression is useful as a proxy to changes in the abundance of functional proteins[74] and enables identification of interacting proteins that are dysregulated at the transcriptional level.[75] However, mRNA expression explains the variation in protein expression only to a limited extent[76–78] and may not capture patterns of posttranslational dysregulation.[79] As network-based analyses primarily focus on interactions among functional proteins, it is valuable to support these analyses with differential analyses of protein expression.[80,81] However common technologies can quantify only a limited fraction of the proteins in the cell at one time.[82] Nibbe et al.[32] address this problem by seeding the search for transcriptionally dysregulated subnetworks with differentially expressed proteins. Namely, they first identify differentially expressed proteins in late stage colorectal cancer using 2D-DIGE and map these proteins on a network of human protein–protein interactions. Subsequently, they identify proteins that exhibit significant crosstalk to these *proteomic seeds*, and assess the coordinate dysregulation of subnetworks composed of these significant crosstalkers. Cross-classification experiments show that this method can identify a compact set of subnetworks that are highly reproducible and very useful as features for classification of tumorigenic phenotype.[33] Proteomic approaches are also useful in characterizing the network mechanisms of cancer, as they can be used to derive causal models for cellular signaling.[83]

CONCLUSION

Systems-based approaches to study human disease serve to remind us that biology is fast becoming an information science. The challenge to construct a more complete and accurate human interactome is large, but so is the opportunity to mine and integrate

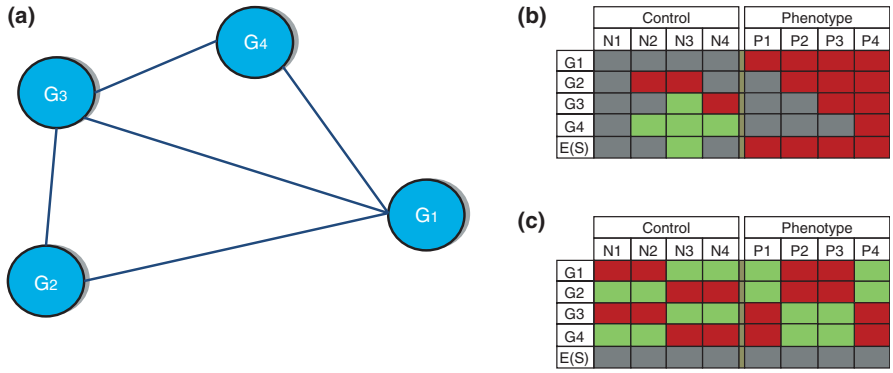


FIGURE 3 | Additive versus combinatorial coordinate dysregulation. A hypothetical example illustrating the difference between additive and combinatorial formulations for subnetwork dysregulation. A hypothetical subnetwork *S* of the human PPI network is shown in (a). Genes are shown as nodes, interactions between their products are shown as edges. In (b) and (c), each row shows a coding gene's expression level in control and phenotype samples. (b) and (c) each display different hypothetical scenarios for gene expression data. The last row shows *subnetwork activity*, which is computed as the average of the expression of these four genes in each sample.[68] Please refer to Figure 2 for description of subnetwork activity and additive coordinate dysregulation.[68] In contrast to additive coordinate dysregulation, combinatorial coordinate dysregulation is defined in terms of how much the expression state of a subnetwork can discriminate control and phenotype samples.[69] Here, the state of a subnetwork refers to the combination of expression levels of all genes in the subnetwork. In (b) and (c), the state of the subnetwork in each sample is given by the first four entries of the column corresponding to that sample (e.g., in (b), the subnetwork has expression state MHML in sample N2, whereas it has state HHMM in sample P2, where H, M, and L, respectively, denote high expression, moderate expression, and low expression). In the case shown in (b), subnetwork activity perfectly discriminates control and phenotype, so the subnetwork is considered dysregulated according to the additive formulation of subnetwork dysregulation. On the other hand, in the case shown in (c), neither the expression of individual genes in *S*,nor the subnetwork activity of *S* can discriminate control and phenotype. However, combination of the expression states of the genes in *S* can perfectly discriminate between control and phenotype (either G1 and G3 or G2 and G4 are expressed in normal samples, whereas in phenotype samples, either G1 and G2 or G3 and G4 are expressed). This example demontsrates the potential power of combinatorial approaches in discovering dysregulated subnetworks, beyond what can be discovered by additive approaches.

the vast data that exist today. In the context of computational models, PPIs present a pivotal point of integration in the overall approach to study the end-point interactions thought to directly cause and sustain the progression of complex human diseases. AP-MS and Y2H will continue to be the workhorse strategies to build a more complete interactome. However, we expect that curated PPIs, where the database of interactions is based on evidence from traditional low-throughput experiments reported in the literature,[84,85] will continue to improve both in terms of coverage and accuracy, notwithstanding their inherent bias toward well-studied proteins.[86] It is critical to systems biology that both high- and low-throughput methods continue to make contributions to high-confidence PPIs, in order to catch up to the development of sophisticated computational models which, having made a head start in other fields, await further development of these databases. Similar to other examples in nature, modeling complex diseases present us with a formidable challenge.[87] We suggest that a more integrative approach to modeling molecular phenotype, where PPIs play an important role, will drive the discovery of better disease markers which, when validated, can be readily translated to the clinic to improve patient outcomes.

BOX 1

A NETWORK INTEGRATION APPROACH

In a given disease state, certain genes may be mutated, or chromosomal aberrations may obtain—genes inserted, deleted, or duplicated. Scores of genes may be differentially expressed, or many proteins differentially expressed or modified. We now know siRNAs have a regulatory role and certain SNPs associate with disease. All these changes ultimately resolve to one or more functional processes in the cell that sustain the phenotype. The PPI network presents an optimal context in which to model the function. Further, the human PPI network is not randomly organized; in cancer, for instance, mutation 'hotspots' exist, certain disease subnetworks have been shown to overlap, and studies of biological networks reveal they have certain mathematical properties that can be exploited by algorithms to identify specific features.

For instance, consider a set of proteomic disease targets, denoted S. To identify proteins that are functionally associated with the proteins S, we use a network of protein interactions.

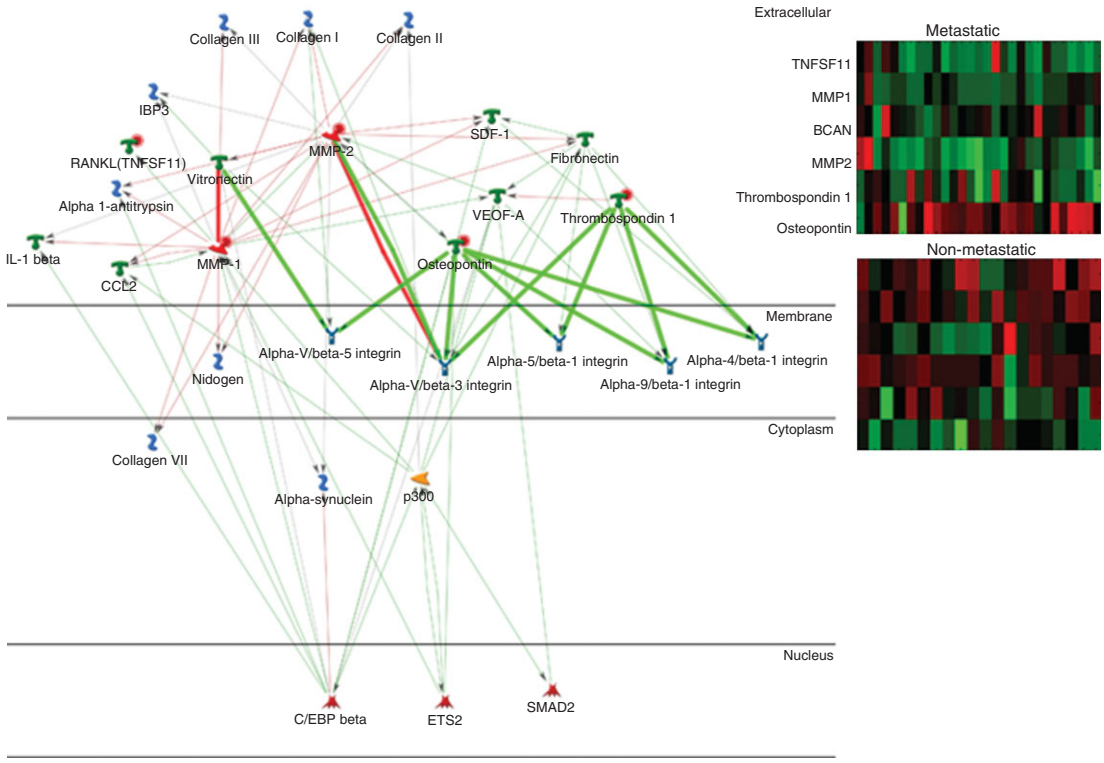


FIGURE 4 | A sample subnetwork of the human PPI network with a state function indicative of liver metastasis in human colorectal cancer. This subnetwork was identified by the CRANE algorithm on the GSE6988 dataset obtained from Gene Expression Omnibus (GEO). The topology of the network that connects the proteins in this subnetwork is shown on the left panel. The mRNA expression profiles of the subnetwork proteins in metastatic and non-metastatic samples are shown on the right panel. For this subnetwork, the state function LLLLLH (in the order of rows of the gene expression matrix, where L and H, respectively, indicate low and high expression) indicates metastasis, i.e., a sample is likely to be metastatic if the first five genes exhibit low expression, but Osteopontin shows high expression. The overall combinatorial coordinate dysregulation of this subnetwork is 0.72. (Reprinted with permission from Ref 69. Copyright 2010 Springer).

Let $G = (V, E)$ denote the network of protein interactions, where V consists of all proteins in the network (S is a subset of V), and an undirected edge $uv \in E$ represents an interaction between proteins $u \in V$ and $v \in V$ Our objective is to compute a score a(v) for each protein $v \in V$, where a(v) quantifies the network crosstalk between v and the proteins in S. Here, network crosstalk is a measure of the network proximity of v to the proteins in S, as well as the multiplicity of network paths in between. This measure of crosstalk is used as an indicator of functional association between proteins. Using this model of crosstalk, and a method analogous to Google's page rank algorithm—an algorithm that scores the importance of a document on the web based on the importance of the documents that link to it, where the importance of other documents are defined similarly in a mutually reinforcing manner—small subnetworks can be discovered that are strongly associated with the disease targets. This model leverages the observation that disease genes often cluster together in the PPI network. The candidate subnetworks may then be rank ordered by an information-theoretic measure (e.g., mutual information) to provide a basis for biological validation.

ACKNOWLEDGEMENTS

This work was supported, in part, by National Institutes of Health Grants UL1-RR024989 from the National Center for Research Resources (Clinical and Translational Science Awards), P30-CA043703 from the Case

Western Reserve University Cancer Center Proteomics Core, and T32-GM008803 from the NIGMS (Institutional National Research Service Award). This work was also supported, in part, by NSF CAREER Award CCF-0953195.

REFERENCES

1. Dixon SJ, Costanzo M, Baryshnikova A, Andrews B, Boone C. Systematic mapping of genetic interaction networks. *Annu Rev Genet* 2009, 43:601–625.
2. Fields S, Song O. A novel genetic system to detect protein–protein interactions. *Nature* 1989, 40:245–246.
3. Chien CT, Bartel PL, Sternglanz R, Fields S. The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc Natl Acad Sci U S A* 1991, 88:9578–9582.
4. Collins MO, Choudhary JS. Mapping multiprotein complexes by affinity purification and mass spectrometry. *Curr Opin Biotechnol* 1998, 19:324–330.
5. Köcher T, Superti-Furga G. Mass spectrometry-based functional proteomics: from molecular machines to protein networks. *Nat Methods* 2007, 4:807–815.
6. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, et al. A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature* 2000, 403:623–627.
7. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 2001, 98:4569–4574.
8. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 2000, 415:180–183.
9. Gavin A, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature* 2006, 440:631–636.
10. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* 2006, 440:637–643.
11. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 2005, 437:1173–1178.
12. Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, McBroom-Cerajewski L, Robinson MD, O'Connor L, Li M, et al. Large-scale mapping of human protein–protein interactions by mass spectrometry. *Mol Syst Biol* 2007, 3:89.
13. Sowa ME, Bennett EJ, Gygi SP, Harper JW. Defining the human deubiquitinating enzyme interaction landscape. *Cell* 2009, 138:389–403.
14. Walhout AJ, Vidal M. A genetic strategy to eliminate self-activator baits prior to high-throughput yeast two-hybrid screens. *Genome Res* 1999, 9:1128–1134.
15. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, et al. High-quality binary protein interaction map of the yeast interactome network. *Science* 2008, 322:104–110.
16. Hart GT, Ramani A, Marcotte E. How complete are current yeast and human protein-interaction networks? *Genome Biol* 2006, 7:120.
17. Stumpf MPH, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C. Estimating the size of the human interactome. *Proc Natl Acad Sci U S A* 2008, 105:6959–6964.
18. Lappe M, Holm L. Unraveling protein interaction networks with near-optimal efficiency. *Nat Biotechnol* 2004, 22:98–103. Epub 2003 Dec 7.
19. Schwartz AS, Yu J, Gardenour KR, Finley RL, Ideker T. Cost-effective strategies for completing the interactome. *Nat Methods* 2009, 6:55–61.
20. According to the World Health Organization in 2010, http://www.who.int/genomics/public/geneticdiseases/en/index2.html/, an estimated 10,000 human diseases are monogenic.
21. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, et al. Variations in DNA elucidate molecular networks that cause disease. *Nature* 2008, 452:429–435.
22. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, et al. Genetics of gene expression and its effect on disease. *Nature* 2008, 452:423–428.
23. Liu ET, Kuznetsov VA, Miller LD. In the pursuit of complexity: systems medicine in cancer biology. *Cancer Cell* 2006, 9:245–247.
24. Hornberg JJ, Bruggeman FJ, Westerhoff HV, Lankelma J. Cancer: a systems biology disease. *Biosystems* 2006, 83:81–90.
25. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, et al. The consensus coding sequences of human breast and colorectal cancers. *Science* 2006, 314:268–274.
26. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. *Nat Med* 2004, 10:789–799.
27. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al.

The genomic landscapes of human breast and colorectal cancers. *Science* 2007, 318:1108–1113.

28. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. *Bioinformatics* 2006, 22:2291–2297.
29. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, et al. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* 2009, 37: D5–15.
30. Tang ZQ, Han LY, Lin HH, Cui J, Jia J, Low BC, Li BW, Chen YZ. Derivation of stable microarray cancer-differentiating signatures using consensus scoring of multiple random sampling and gene-ranking consistency evaluation. *Cancer Res* 2007, 67:9996–10003.
31. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. *Mol Syst Biol* 2007, 3:78.
32. Nibbe RK, Markowitz S, Myeroff L, Ewing R, Chance MR. Discovery and scoring of protein interaction subnetworks discriminative of late stage human colon cancer. *Mol Cell Proteomics* 2009, 8:827–845.
33. Nibbe RK, Koyutürk M, Chance MR. An integrative-omics approach to identify functional sub-networks in human colorectal cancer. *PLoS Comput Biol* 2010, 6:e1000639.
34. Kar G, Gursoy A, Keskin O. Human cancer protein –protein interaction network: a structural perspective. *PLoS Comput Biol* 2009, 5:e1000601.
35. Baggs JE, Hughes ME, Hogenesch JB. *The Network as the Target*. Wiley Interdisciplinary; reviews. 2009.
36. Bader JS. Systems approaches for pharmacogenetics and pharmacogenomics. *Pharmacogenomics* 2008, 9: 257–262.
37. Chen LL, Blumm N, Christakis NA, Barabási AL, Deisboeck TS. Cancer metastasis networks and the prediction of progression patterns. *Br J Cancer* 2009, 101: 749–758.
38. Auffray C, Chen Z, Hood L. Systems medicine: the future of medical genomics and healthcare. *Genome Med* 2009, 1:2.
39. Schadt EE. Molecular networks as sensors and drivers of common human diseases. *Nature* 2009, 461: 218–223.
40. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. *Proc Natl Acad Sci U S A* 2007, 104:8685–8690.
41. Barrenas F, Chavali S, Holme P, Mobini R, Benson M. Network properties of complex human disease genes identified through genome-wide association studies. *PLoS One* 2009, 4:e8090.
42. Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 2008, 82:949–958.
43. Chen J, Aronow B, Jegga A. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics* 2009, 10:73.
44. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* 2010, 6:e1000641+.
45. Barabasi A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004, 5:101–113.
46. Papin JA, Hunter T, Palsson BO, Subramaniam S. Reconstruction of cellular signalling networks and analysis of their properties. *Nat Rev Mol Cell Biol* 2005, 6:99–111.
47. Bossi A, Lehner B. Tissue specificity and the human protein interaction network. *Mol Syst Biol* 2009, 5:260.
48. Shlomi T, Cabili MN, Herrgard MJ, Palsson BO, Ruppin E. Network-basedpredictionofhumantissue-specificmetabolism. *Nat Biotech* 2008, 26:1003–1010.
49. Bebek G, Yang J. PathFinder: mining signal transduction pathway segments from protein–protein interaction networks. *BMC Bioinformatics* 2007, 8:335+.
50. Hanisch D, Zien A, Zimmer R, Lengauer T. Co-clustering of biological networks and gene expression data. *Bioinformatics* 2002, 18(suppl 1):S145–154.
51. Murali TM, Rivera CG. Network legos: building blocks of cellular wiring diagrams, *RECOMB*, San Fransisco, CA, USA, 2007, 47–61.
52. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 2003, 302:249–255.
53. Segal E, Wang H, Koller D. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics. InISMB (Supplement of Bioinformatics)* 2003, 19(suppl 1):264–272.
54. Tornow S, Mewes HW. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res* 2003, 31:6283–6289.
55. Vert J-P, Kanehisa M. Extracting active pathways from gene expression data. *Bioinformatics* 2003, 19:238–244.
56. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature* 2009, 430:88–93.
57. Aragues R, Sander C, Oliva B. Predicting cancer involvement of genes from heterogeneous data. *BMC Bioinformatics* 2008, 9:172.
58. Ma X, Lee H, Wang L, Sun F. CGI: a new approach for prioritizing genes by combining gene expression and protein–protein interaction data. *Bioinformatics* 2007, 23:215–221.
59. Ulitsky I, Karp RM, Shamir R. Detecting disease-specific dysregulated pathways via analysis of clinical

expression profiles. *Res Comput Mol Biol (RECOMB)* 2008, April; 4955:347–359.

60. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 2002, 18(suppl 1):S233–240.
61. Guo Z, Li Y, Gong X, Yao C, Ma W, Wang D, Li Y, Zhu J, Zhang M, Yang D, Rao S, Wang J. Edge based scoring and searching method for identifying condition-responsive protein–protein interaction sub-network. *Bioinformatics* 2007, 23:2121–2128.
62. Serban N, Critchley-Thorne R, Lee P, Holmes S. Gene expression network analysis and applications to immunology. *Bioinformatics* 2007, 23:850–858.
63. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, et al. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* 2004, 430:88–93.
64. Xia K, Dong D, Xue H, Zhu S, Wang J, Zhang Q, Hou L, Chen H, Tao R, Huang Z, et al. Identification of the proliferation/differentiations witch in the cellular network of multicellular organisms. *PLoS Comput Biol* 2006, 2:e145.
65. Xue H, Xian B, Dong D, Xia K, Zhu S, Zhang Z, Hou L, Zhang Q, Zhang Y, Han J-DJ. A modular network model of aging. *Mol Syst Biol* 2007, 3:147.
66. Cabusora L, Sutton E, Fulmer A, Forst CV. Differential network expression during drug and stress response. *Bioinformatics* 2005, 21:2898–2905.
67. Hwang T, Park T. Identification of differentially expressed subnetworks based on multivariate anova. *BMC Bioinformatics* 2009, 10:128+.
68. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 2007, 3:140.
69. Chowdhury SA, Nibbe RK, Chance MR, Koyutürk M. Subnetwork state functions define dysregulated subnetworks in cancer, *Proceedings of 14th International Conference on Research in Computational Biology (RECOMB'10)*, Lisboa, Portugal, 2010,80–95.
70. Anastassiou D. Computational analysis of the synergy among multiple interacting genes. *Mol Syst Biol* 2007, Volume 3, Article number 83.
71. Watkinson J, Wang X, Zheng T, Anastassiou D. Identification of gene interactions associated with disease from gene expression data using synergy networks. *BMC Syst Biol* 2008, 2:10.
72. Fang G, Kuang R, Pandey G, Steinbach M, Myers CL, Kumar V. Subspace differential coexpression analysis: problem definition and a general approach. *Pac Symp Biocomput* 2010, 145–156.
73. Kostka D, Spang R. Finding disease specific alterations in the co-expression of genes. *Bioinformatics (Oxford, England)* 2004, 20(suppl 1):i194–199.
74. Slonim DK. From patterns to pathways: gene expression data analysis comes of age. *Nat Genet* 2002, Dec; 32:502–508
75. Jansen R, Greenbaum D, Gerstein M. Relating whole-genome expression data with protein–protein interactions. *Genome Res* 2002, 12:37–46.
76. Chang J, Chance MR, Nicholas C, Ahmed N, Guilmeau S, Flandez M, Wang DSWD, Nasser S, Albanese JM. Proteomic changes during intestinal cell maturation in vivo. *J Proteomics* 2008, 71:530–546.
77. de Godoy LMF, Olsen JV, Cox J, Nielsen ML, Hubner NC, Fröhlich F, Walther TC, Mann M. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* 2008, 455:1251–1254.
78. Greenbaum D, Colangelo C, Williams K, Gerstein M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* 2003, 4:117.
79. Mata J, Marguerat S, Bähler J. Post-transcriptional control of gene expression: a genome-wide perspective. *Trends Biochem Sci* 2005, 30:506–514.
80. Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L, Aebersold R. Complementary profiling of gene expression at the transcriptome and proteome levels in saccharomyces cerevisiae. *Mol Cell Proteomics* 2002, 1:323–333.
81. Hatzimanikatis V, Lee KH. Dynamical analysis of gene networks requires both mRNA and protein expression information. *Metab Eng* 1999, 1:275–281.
82. Ferguson LP, Smith RD. Proteome analysis by mass spectrometry. *Annu Rev Biophys Biomol Struct* 2003, 32:399–424.
83. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 2005, 308: 523–529.
84. Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis A-R, Simonis N, Rual JF, Borick H, Braun P, Dreze M, et al. Literature-curated protein interaction datasets. *Nat Methods* 2009, 6:39–46.
85. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, et al. Human protein reference database—2006 update. *Nucleic Acids Res* 2006, 34:D411–414.
86. Ramírez F, Schlicker A, Assenov Y, Lengauer T, Albrecht M. Computational analysis of human protein interaction networks. *Proteomics* 2007, 7:2541–2552.
87. Binder PM. Frustration in complexity. *Science* 2008, 320:322–323.