Molecular Systems Biology 3; Article number 124; doi:10.1038/msb4100163
Citation: ***Molecular Systems Biology*** **3:124**

www.molecularsystemsbiology.com

molecular systems biology

PERSPECTIVE

Human disease classification in the postgenomic era: A complex systems approach to human pathobiology

Joseph Loscalzo[1,2,*], Isaac Kohane[2,3] and Albert-Laszlo Barabasi[4]

[1] Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA,
[2] Harvard Medical School, Boston, MA, USA,
[3] Department of Pediatrics, Children's Hospital, Boston, MA, USA and
[4] Department of Physics and Computer Science, University of Notre Dame, Notre Dame, IN, USA
* Corresponding author. Department of Medicine, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115, USA. Tel.: +1 617 525 4833; Fax: +1 617 525 4830; E-mail: jloscalzo@partners.org

Received 12.3.07; accepted 4.5.07

Contemporary classification of human disease derives from observational correlation between pathological analysis and clinical syndromes. Characterizing disease in this way established a nosology that has served clinicians well to the current time, and depends on observational skills and simple laboratory tools to define the syndromic phenotype. Yet, this time-honored diagnostic strategy has significant shortcomings that reflect both a lack of sensitivity in identifying preclinical disease, and a lack of specificity in defining disease unequivocally. In this paper, we focus on the latter limitation, viewing it as a reflection both of the different clinical presentations of many diseases (variable phenotypic expression), and of the excessive reliance on Cartesian reductionism in establishing diagnoses. The purpose of this perspective is to provide a logical basis for a new approach to classifying human disease that uses conventional reductionism and incorporates the non-reductionist approach of systems biomedicine.
Molecular Systems Biology 10 July 2007; doi:10.1038/msb4100163
Subject Categories: metabolic and regulatory networks; molecular biology of disease
Keywords: genotype; network; pathology; phenotype

Introduction

Contemporary classification of human disease dates to the late 19th century, and derives from observational correlation between pathological analysis and clinical syndromes. Characterizing disease in this way established a classification schema that has served clinicians well to the current time, relying on observational skills to define the syndromic phenotype. Throughout the last century, this approach became more objective, as the molecular underpinnings of many disorders were identified and definitive laboratory tests became an essential part of the overall diagnostic paradigm. Yet, this classic diagnostic strategy has widely recognized shortcomings that reflect both a lack of sensitivity in identifying preclinical disease and a lack of specificity in defining disease unequivocally. Some have argued that the lack of specificity is a consequence of the false positive rate of any objective diagnostic test. For this reason, probabilistic frameworks have been used to improve diagnostic accuracy (Schwartz *et al*, 1981); however, these frameworks also rely on choices grounded in reductionism and, thus, suffer from the limitations of a reductionist approach. The purpose of this perspective, then, is to provide a logical argument for a new approach to classifying human disease that both appreciates the uses and limits of reductionism and incorporates the tenets of the non-reductionist approach of complex systems analysis.

Disease classification: history and shortcomings

Current disease classification and medical diagnosis are the direct consequence of inductive generalization predicated on Occam's razor. This scientific approach has served clinicians well in their effort to establish syndromic patterns that streamline the number of phenotypes to consider. In addition, owing to the dearth of quantitative information available to refine and parse these phenotypes further, the diagnostic exercise was intrinsically limited but tractable for the individual practitioner.

These diagnostic limitations, however, will soon become a historical footnote. With the complete sequence of the human genome a reality, and with a growing body of transcriptomic, proteomic, and metabolomic data sets in health and disease, we are now in a unique position in the history of medicine to define human disease precisely, uniquely, and unequivocally, with optimal sensitivity and specificity. Theoretically, this precise molecular characterization of human disease will allow us to understand the basis for disease susceptibility and environmental influence; to offer an explanation for the different phenotypic manifestations of the same disease; to define disease prognosis with greater accuracy; and to refine and, ideally, individualize disease treatment for optimal therapeutic efficacy.

The importance of redefining human disease in this postgenomic era cannot be overemphasized. Several examples serve to prove this point well. Subcategorizing histologically similar cancers by differences in surface biomarkers, transcription profiling, or proteomic analysis is currently being applied to several malignancies, including lymphomas (Dave *et al*, 2006) and adenocarcinoma of the breast (Hedenfalk *et al*, 2001; Hall *et al*, 2006), in an effort to provide better information about prognosis and response to therapy. This approach defines the expanding field of molecular pathology,

in which molecular signatures replace histopathology to diagnose disease and predict outcome.

As an example of a disease whose genetic basis not only is felt to be much simpler than that of malignancies, but also is affected by host genomic and environmental complexities, consider sickle cell disease. This classic Mendelian disease is, by definition, viewed as a monogenic disorder, in which all affected individuals have a single point mutation at position 6 of the beta-chain of hemoglobin, leading to a substitution of valine for glutamic acid. This single mutation changes the oxygen affinity of hemoglobin, and leads to its ability to form polymers under hypoxic conditions, which, in turn, deform the erythrocyte into the characteristic sickle shape. Yet, despite this well-defined mutation, and its biochemical and physiological consequences, the genotype simply cannot invariably predict the phenotype of patients with the disease. Patients with this mutation are not at all homogeneous in their clinical presentations: some develop principally painful crises with or without bony infarcts; others are prone to hemolytic crises; some develop vasoocclusive crises, including stroke; still others develop acute chest syndrome; while many are phenotypically normal, except for mild anemia. There are many reasons for these different clinical phenotypes, including the presence of disease-modifying genes (Sebastiani *et al*, 2005) and environmental influences (ambient oxygen concentration, infection, dehydration), which can interact to yield different phenotypes (Kato *et al*, 2007). This example points out that our true understanding of even the most straightforward of genetic disorders is quite limited.

As a second example, consider familial pulmonary arterial hypertension (PAH) (Farber and Loscalzo, 2004). This disorder is associated with mutations in members of the TGF-β receptor superfamily, including bone morphogenetic protein receptor-2 (BMPR-2), Alk-1, and endoglin. In this disorder, there is a common phenotype, but many different genotypes yielding it: for example, over 50 different mutations in BMPR-2 have been identified. While all of these mutations are in the same gene, the mechanisms by which they confer the phenotype are not entirely clear and range from dominant negative effects to haplo-insufficiency. Importantly, only approximately one-quarter of individuals with the mutations manifest the disease; this incomplete penetrance is also likely a consequence of the effects of disease-modifying genes, environmental influences, or both in a given individual.

As a third example, consider yet another disorder with a phenotype common to many different mutations, familial hypertrophic cardiomyopathy. In contrast to familial PAH, hypertrophic cardiomyopathy can be caused by mutations in several different genes that code for different sarcomeric proteins, including myosin heavy chain, myosin light chain, tropomyosin, and troponin C (Seidman and Seidman, 2001), as well as non-sarcomeric proteins. Clearly, in this particular disorder, the common phenotype is misleading, suggesting a single disease as its cause when, in fact, the pathophenotype comprises multiple genetically distinct diseases. Another lesson learned from hypertrophic cardiomyopathy is knowing the sarcomeric protein involved and the specific mutation does not invariably provide prognostic information about the course of the disease, including the risk of sudden cardiac death. The logical reductionist approach to this disorder suggesting that different mutations have different effects on myocardial function, clinical disease course, and outcome risk is an erroneous oversimplification. The reasons for this failing range from insufficient information from which to predict system behavior to the inability of a complex biologic system to be reductively explicable by the basal elements (genes, proteins) from which it emerges (Kim, 2006) (*vide infra*). One example of this latter principle is the property of robustness, or the ability of a complex biological system to maintain stable function in the face of perturbation (Kitano, 2004), a property that cannot be predicted without understanding the component parts and their (non-linear) interactions.

Definition and determinants of disease phenotype

These illustrative examples compel us to consider what precisely defines disease phenotype. Let us begin by stating the obvious fact that all disease phenotypes reflect the consequences of defects in a complex genetic network operating within a dynamic environmental framework. Even classical Mendelian disorders have different clinical phenotypes that are a consequence of polymorphic or mutant disease-modifying genes and their interactions with environmental factors. The disease-modifying genes can be subclassified into two groups, those whose actions are uniquely affected by the primary genetic mutation (e.g., they are in the biochemical/molecular module) and those whose actions reflect generic responses to organism stress evoked by the principal mutation and/or environmental exposures. These generic responses define intermediate phenotypes that comprise to varying extents all basic pathobiologies and include inflammation, thrombosis and hemorrhage, fibrosis, the immune response, proliferation, and apoptosis/necrosis (Figure 1).

Together with this network of genetic mutations and polymorphisms, we must consider the environmental factors to which the individual is exposed, either acutely or chronically. From an evolutionary perspective, organisms evolved to accommodate a range of environmental stresses, and have done so with varying degrees of success depending upon the duration and magnitude of the stress. Viewed in this way, environmental exposures include temperature, radiation, hydration and tonicity, oxygen tension, micro- and macronutrients, infective agents and parasitism, and toxins. The interaction between the human host and microbes reflects a unique exposure that not only can lead directly to disease expression, but also can lead to changes in human host phenotype that is not directly pathogenic. Recent work by Gordon's group, for example, demonstrates that the commensal gut microbiome of an individual influences the efficiency of energy harvesting from ingested foodstuffs and, as a result, can directly influence the extent of weight gain associated with food consumption (Turnbaugh *et al*, 2006).

With this brief background, let us consider how one might use this information to classify disease. We begin with four different modular networks within and between which nodes (in this case, genes, mRNA, proteins, biochemical or physiological properties, or environmental factors) interact to yield the ultimate phenotype. The first of these networks is that

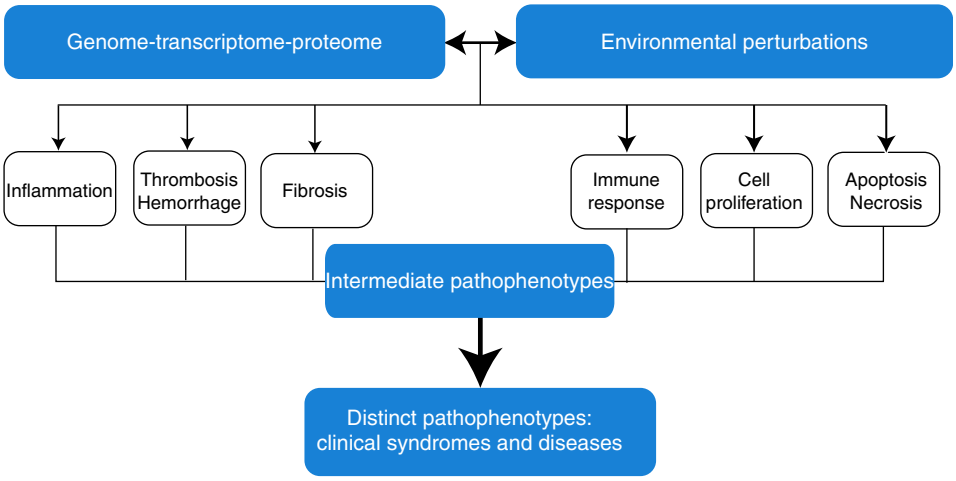

Figure 1 Diagram indicating associations among genetic and environmental factors reduce and their interactions with intermediate phenotypes to yield distinct pathophenotypes. The intermediate phenotypes determine, in part, variation in disease expression and clinical presentation among individuals with equivalent underlying genetic or environmental exposures that predispose to a disease state.

comprising the principal molecular abnormalities (genetic or acquired), when known, that have been associated with the general phenotype (*primary disease genome or proteome, G*). In the classic Mendelian case, this is a network of two nodes, one for each allele; in the case of a complex trait, such as PAH, there will be many nodes (~100 currently) comprising known mutations in *BMPR-2* and *Alk-1* alleles. The second modular network comprises known disease-modifying genes (*secondary disease genome or proteome, D*) and their polymorphisms or haplotypes. In the case of sickle cell disease, for example, this network will include the hemoglobin F gene, the hemoglobin C gene, thalassemic beta-chain deletion, glucose-6-phosphate dehydrogenase gene mutations, and single nucleotide polymorphisms in three genes in the TGF-β pathway found to be associated with stroke in sickle cell disease (Sebastiani *et al*, 2005). The third modular network incorporates known polymorphisms or haplotypes that influence each of the generic responses to organism stress (*intermediate phenotype, I*, or *response genome or proteome*), and will define, for example, the extent to which an individual can mount an inflammatory response, develop thrombosis, or accommodate oxidant stress. The fourth modular network comprises *environmental determinants, E*. The interactions among nodes of these modular networks define all disease phenotypes. Environmental factors interact with the different subgenomes to modify the transcription of their component genes and to modulate the translation of protein products and their posttranslational modification, yielding changes in protein and cellular function and metabolism, and defining an intermediate phenotype. The patterns of these polymorphic genes and their expression profiles comprise the molecular signatures of unique pathophenotypes, offering the promise of diagnostic, prognostic, and therapeutic specificity; the definition of disease becomes 'personalized', as does its specific therapeutic targets. These intermediate phenotypes combine to define the *pathophysiological states (PS)*, which, in turn, underlie all disease phenotypes (*pathophenotype, P*; Figure 2A). Examples of this approach are included in Table I and Figure 2B for sickle cell disease, and in Table II for PAH. In the latter case, note that both heritable (BMRP-2 mutations) and acquired (anorexigen induced) forms of the disease are incorporated in the table; also note that this approach allows one to define genetic susceptibility to environmentally induced phenotypes.

Importantly, cataloging disease in this way is only the beginning of a rigorous analytical process that can lead to defining prognostic determinants and better-individualized therapeutic responses. Nicholson (2006) has long been a proponent of the application of systems analysis to diagnostics and therapeutics, focusing, in particular, on the metabolome and its environmental (including microbial) modulation. To achieve this level of insight will require a network-based analysis of the associations among the individual genes, proteins, metabolites, intermediate phenotypes, and environmental factors that conspire to yield the pathophenotype. Defining the network interactions among these modular elements (and their probabilistic relationships, where appropriate) not only will account for the ultimate pathophenotype but also can lead to the identification of potential regulatory nodes within the network that can modify phenotype (i.e., potential therapeutic target). Representative examples of modular network representations of disease are shown in Figure 3.

Network principles in biological systems

Experimental feasibility and analytical tractability have been the driving forces behind reductionism in medical science. Over the past 50 years, the biomedical community has operated fully within this scientific framework, applying it successfully to basic molecular medicine and clinical trials. With the advent of the genomic era, however, biological investigators are confronted daily with ever-growing data sets that contain potentially useful functional information which cannot readily be analyzed optimally with the conventional approach. Most analytical approaches to this problem are

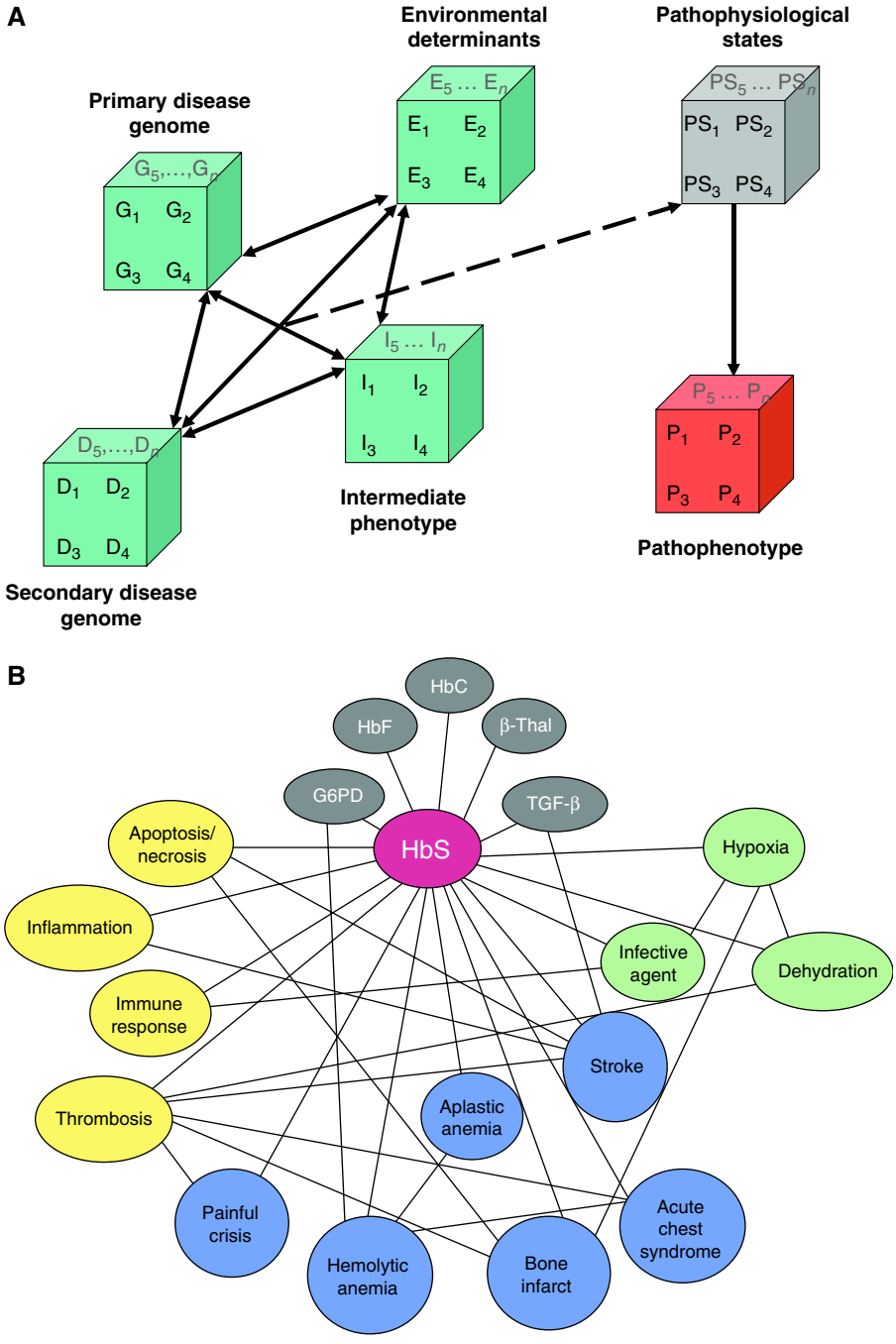

Figure 2 (**A**) Theoretical human disease network illustrating the relationships among genetic and environmental determinants of the pathophenotypes. Key: G, primary disease genome or proteome; D, secondary disease genome or proteome; I, intermediate phenotype; E, environmental determinants; PS, pathophysiological states leading to P, pathophenotype. (**B**) Example of this theoretical construct applied to sickle cell disease. Key: red, primary molecular abnormality; gray, disease-modifying genes; yellow, intermediate phenotypes; green, environmental determinants; blue, pathophenotypes.

rather rudimentary and involve calculating simple correlations among the genes whose expression changes in response to a perturbation, clustering the genes by molecular or known functional class, and drawing crude inferences about mechanism on that basis. This is not a very conceptually satisfying situation and does little to advance our understanding of the mechanism(s) underlying the changes in gene expression in response to the perturbation or its implications for clinical phenotype.

Until recently, system-based analysis of complex biological networks was limited by the quantitative tools available, and by the grossly incomplete knowledge of the network nodes and their interactions. With the development of novel quantitative approaches to network analysis, and with the

Table I Sickle cell disease

Primary molecular abnormality (disease genome or proteome)	Disease-modifying genes or proteins (secondary disease genome or proteome)	Intermediate phenotype (response genome or proteome and pathological manifestations)	Environmental determinant	Pathophenotype
Hb A Val6Glu	Hb F Hb C β-Thalassemia G6PD TGF-β	Thrombosis Inflammation Immune response Fibrosis Apoptosis/necrosis	Hypoxia Dehydration Infective agent	Hemolytic anemia Aplastic anemia Stroke Bone infarction Painful crisis Acute chest syndrome

Abbreviation: Hb, hemoglobin.

Table II Pulmonary arterial hypertension

Primary molecular abnormality (disease genome or proteome)	Disease-modifying genes or proteins (secondary disease genome or proteome)	Intermediate phenotype (response genome or proteome and pathological manifestations)	Environmental determinant	Pathophenotype
BMPR-2 mutations Alk-1 mutations Endoglin mutations	5-HT2B 5-HTT Thromboxane synthetase	Thrombosis Vasospasm Inflammation	Hypoxia Infective agent (HIV, HHV-8) *Crotolaria sp.*/toxin	Pulmonary hypertension Cor pulmonale Pulmonary thromboembolism
	Prostacyclin synthetase 5-Lipoxygenase NADPH oxidase Endothelin Hemoglobinopathies Hereditary spherocytosis HHT Thrombocytosis	Fibrosis Proliferation Immune response Apoptosis/necrosis	Cocaine Anorexigens Alcoholic cirrhosis	

Abbreviations: 5-HTT, serotonin transporter; 5-HT2b, serotonin 2b receptor; HHT hereditary hemorrhagic telangiectasias.

explosion of currently available genomic, transcriptomic, proteomic, and metabolomic data sets, the scientific landscape has changed considerably and biological network analysis is now a tractable exercise (Barabási and Oltvai, 2004). We will next consider the methods available to analyze biological networks, demonstrate the application of these principles to selected problems in biology and medicine, and discuss the implications of biological network theory as a construct for characterizing human disease and defining novel therapies.

Networks can be briefly classified as random or scale free (Albert and Barabási, 2002). Random networks are those in which the connections among nodes are driven by chance, each node having the same likelihood of being connected to any other node, with the resulting probability, P, of the numbers of links per node following a Poisson distribution. In scale-free networks, the probability of the number of links per node follows a power law (or scale free) distribution ($P(k)=k^{-\gamma}$, where k is the number of links per node and γ is the slope of the log $P(k)$ versus log(k) plot) (Barabási and Albert, 1999; Albert and Barabási, 2002). A power law distribution decreases more slowly than the exponential distribution of random networks; in scale-free networks some nodes are highly connected (hubs), while the majority of nodes have few connections.

Real networks are scale free because they evolve with new nodes added one at a time to nodes that are already highly linked (Barabási and Oltvai, 2004). In biological systems in particular, this scale-free addition of new nodes is likely a consequence of gene duplication (Qian *et al*, 2001), and is also affected by alternate splicing and posttranslational modification in protein networks (Qian *et al*, 2001; Bhan *et al*, 2002; Pastor-Satorras *et al*, 2003; Vazquez *et al*, 2003), as well as the variable chemical versatility of the metabolic intermediates in metabolic networks.

There are many beneficial consequences of scale-free networks in biological systems. They facilitate chemical diversity at minimal energy cost and minimize the transition time between metabolic states (Wagner and Fell, 2001). They recapitulate natural selection and evolution: in complex gene networks, mutations or deletions of highly linked (hub) genes lead to embryonic lethality, while mutations of weakly linked genes account for biological variability and natural selection (Oikonomou and Cluzel, 2006). Driven by random mutation and selection, scale-free networks are capable of evolving rapidly toward an optimal functional state, without any tuning (Albert *et al*, 2000). Scale-free networks also minimize the consequences of most biochemical or genetic errors (Wagner and Fell, 2001), and accommodate perturbations to the network with minimal effects on critical functions (Pastor-Satorras and Vespignani, 2002), unless hubs are the targets of the perturbations (Albert, 2005) (*vide infra*).

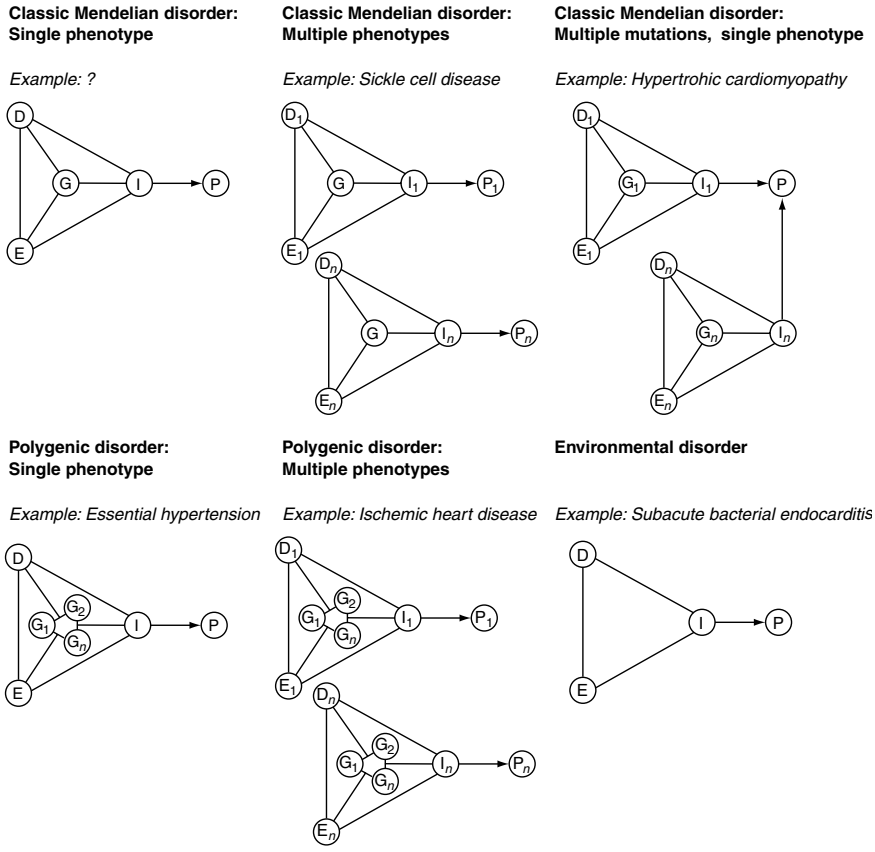

Figure 3 Examples of modular network representations of disease. Key: G, primary disease genome or proteome; D, secondary disease genome or proteome; E, environmental determinants; I, intermediate phenotype; P, pathophenotype.

The application of network analysis to problems in biological systems is evolving rapidly. This approach provides a method for analyzing the connections and interactions among elements in the ever-expanding genetic, proteomic, and metabolic data sets of organisms. In addition, it has been used to identify key regulatory elements in complex gene and metabolic networks. The dynamic response of a biological network can be quantitated using a variety of approaches, including reverse engineering (Basso *et al*, 2005), non-linear differential equations coupled with multiple linear regression (Gardner *et al*, 2003), Bayesian analysis (Yu *et al*, 2004), and cellular automata (Wurthner *et al*, 2000).

Network approaches to human disease

These methods have recently been applied successfully to human disease. Network analysis has been used to characterize the spread of epidemics (Pastor-Satorras and Vespignani, 2001; Eubank *et al*, 2004; Madar *et al*, 2004) and to determine ways to control them (Pastor-Satorras and Vespignani, 2001; Dezso and Barabasi, 2002; Madar *et al*, 2004). Systems approaches have also been used to identify novel targets that influence the metastatic propensity and lethality of adenocarcinoma of the prostate (Ergun *et al*, 2007). Using a reverse-engineered analysis of gene networks involved in malignant transformation of prostate cancer cells coupled with expression (transcription) profiling, these investigators computed the likelihood that genes within the network and associated pathways mediate disease pathogenesis. The results of this analysis identified a novel pathway and genetic mediator of metastasis for adenocarcinoma of the prostate, the androgen receptor gene. This conclusion is biologically plausible because androgen suppression therapy is a standard approach to the treatment of primary prostate cancer, and recurrence invariably is associated with a loss of growth-dependence on androgens.

Lim *et al* (2006) have also recently used systems analysis of protein–protein interaction networks to identify potential disease-modifying proteins that are common to a wide range of neurodegenerative disorders causing ataxia. Inherited ataxias are associated with gain-of-function or loss-of-function mutations in many (over 23) seemingly unrelated genes. These investigators used network analysis to demonstrate that many ataxia-causing proteins share proteins with which they interact, some of which can modify neurodegenerative responses in animal models.

Lu *et al* (2007) have used a network approach to analyze the allergic response in experimental asthma. They devised a biological interaction network using the Biomolecular Object Network Databank database of molecular interactions curated from the biomedical literature, then mapped differentially expressed genes from expression array data onto the network. They next analyzed the topological characteristics of the

differentially expressed genes, and then determined the correlation between topology and biological function using the Gene Ontology classifications. Using this approach, these investigators found that nodes (genes) with high connectivity tend to have lower levels of change in expression than peripheral nodes, consistent with the notion that disease-causing genes are typically not central hubs in a molecular module.

Recently, Goh *et al* (in press) have used network analysis methods to characterize the set of disease-gene associations documented in the Online Mendelian Inheritance in Man database. They observed that genes associated with similar disorders (e.g., cataracts and cardiomyopathies) show a greater likelihood of association between their products and greater similarity among their transcription profiles than those not associated with similar disorders (Figure 4). Similarly, proteins that are associated with the same disease show a 10-fold increased tendency to interact with each other than those not associated with the same disease. These observations support the concept of disease-specific functional modules, which comprise a comprehensive network of known genetic diseases. These investigators also demonstrated that the vast majority of disease genes are not essential and show no tendency to encode highly connected protein hubs, rather being localized to the functional periphery of the network. In contrast, essential genes whose defects often lead to lethality *in utero* or in early extrauterine life tend to encode hubs and to occupy a central position in the network.

Network approaches to therapeutics

The development of network strategies for the analysis of biological systems raises the question of whether one can use these approaches to characterize and treat human disease. Identifying the molecular causes of disease represented a major breakthrough in the history of medicine, moving the discipline from pattern recognition and therapeutic strategies based on syndromic pathophysiology to molecular mechanism and evidence-based therapies derived from clinical trials designed on the basis of molecular mechanism. Clearly, this transition reflects the success of the conventional scientific method, upon which medical research has been based, and cast the discipline of medicine in an entirely different light as scientifically rigorous, rational, and deterministic.

Notwithstanding this record of success, the medical literature is rife with counterexamples that fail to support a straightforward approach to pharmaco-therapeutics derived from reductionist principles. An example will serve to illustrate this point for a therapeutic trial. Hyperhomocysteinemia, a know risk factor for atherothrombosis, can be treated by facilitating the methylation of homocysteine to methionine. For this reason, several large-scale clinical trials were initiated to test the hypothesis that lowering homocysteine with folic acid and vitamin B_{12} can reduce the risk of atherothrombotic events in individuals with established vascular disease and hyperhomocysteinemia. Unfortunately, three of these trials recently completed yielded negative results: while homocysteine levels were reduced with the therapy, event rates were unchanged compared with those in the population treated with placebo. A possible reason for this unexpected outcome, in retrospect, is that the assumed exclusivity of the homocysteine-lowering effect of supplemental folic acid and vitamin B_{12} dramatically oversimplifies their potential effects in this complex system (Loscalzo, 2006): these vitamin cofactors not only lower homocysteine but also promote DNA synthesis, thereby supporting cell proliferation; they can also enhance methylation potential in the setting of hyperhomocysteinemia, by increasing the ratio of *S*-adenosylmethionine to *S*-adenosylhomocysteine, which can alter gene expression by modulating the methylation status of CpG-rich promoter regions.

As with diagnostics, this example suggests that reductionist approaches to therapeutics have their limitations and can, in the worst case, be misleading. Optimizing therapeutic approaches to human disease will clearly require the application of network analysis (Morel *et al*, 2004; Kitano, 2007): network analysis can be used to identify new drug targets (e.g., the androgen receptor in prostate cancer; Ergun *et al*, 2007), to determine the appropriate dosing of a drug, based on metabolomic profiling (Nicholson, 2006), and to ascertain the causes of resistance to therapies or enhanced toxicities of drugs based on the robustness-fragility trade-off inherent in the system (Kitano, 2007).

Organizational principles of biological networks and their application to human disease networks

Are there organizational principles at the molecular level that govern biological networks and their transition to disease from which we can develop rational therapies? A key principle is that cellular functions are conducted in a highly modular manner (Hartwell *et al*, 1999; Ravasz *et al*, 2002). In general, modularity refers to a group of physically or functionally linked nodes (in this case molecules) that work together to achieve a distinct functional phenotype. Biology is rife with examples of modularity: the overwhelming majority of molecules in a cell is either part of an intracellular complex with modular activity, such as the ribosome, or participates in an extended (functional) module as a temporally regulated element of a relatively distinct process (e.g., signal amplification in a phosphorylation-mediated signaling pathway). The identification of the specific functional modules in a network is complicated by the fact that at face value, the scale-free organization and modularity seem to be internally inconsistent network properties. Modules by definition imply the existence of groups of nodes that are relatively isolated from the rest of the system. Yet, in a scale-free network, hubs are in contact with a high fraction of nodes, making the existence of relatively isolated modules unlikely. Clustering and hubs can naturally coexist; however, if topological modules are not independent but combine to form an hierarchical network in which small, highly integrated modules assemble into larger modules each of which combines in an hierarchical fashion into even larger modules (Hartwell *et al*, 1999). Signatures of such hierarchical modularity are present in all cellular networks that have been investigated to date, ranging from metabolic (Ravasz *et al*, 2002) to protein–protein (Yook *et al*, 2004) interaction and regulatory networks.

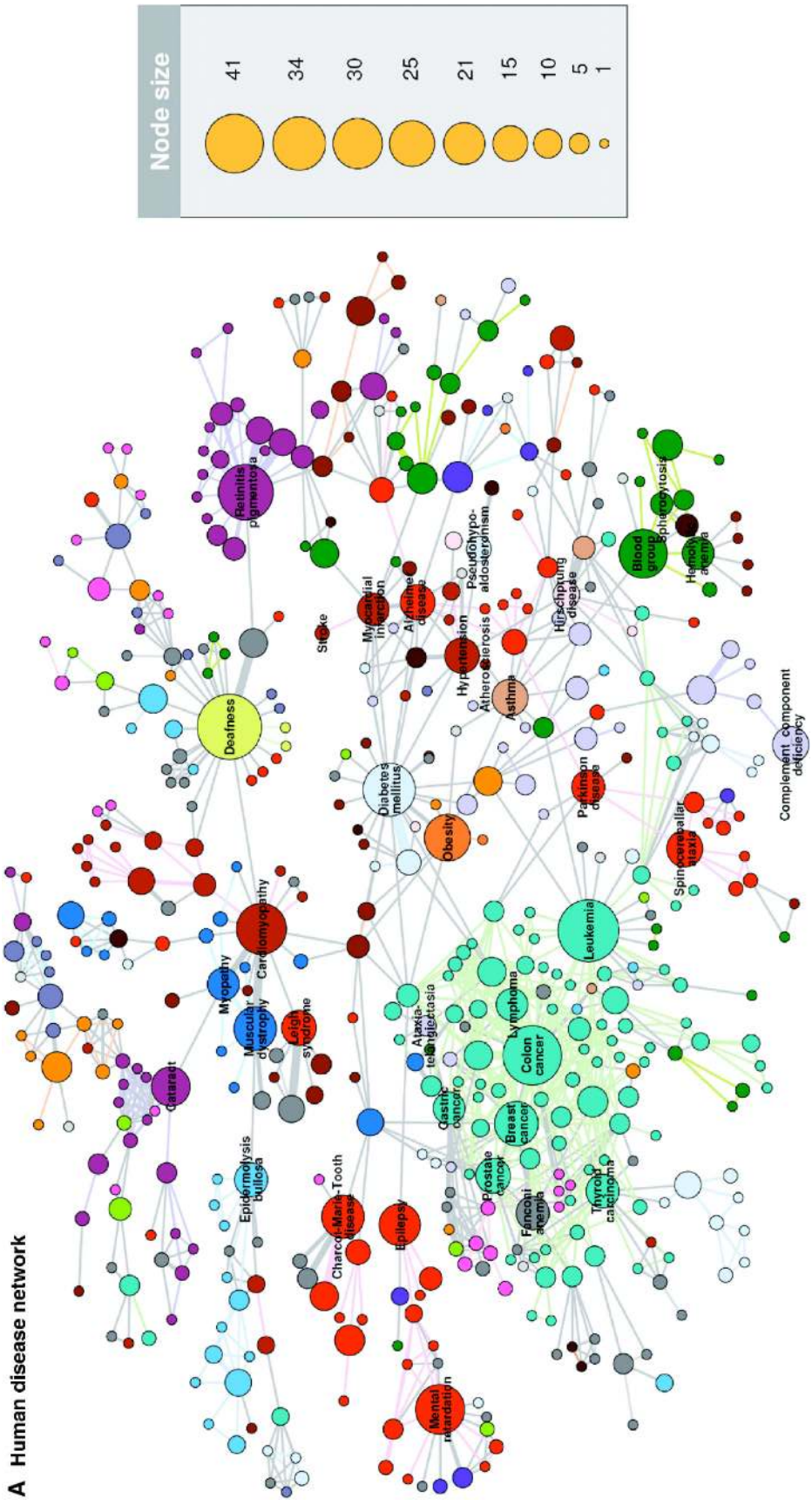

Figure 4 (**A**) Human disease network. Each node corresponds to a specific disorder colored by class (22 classes, shown in the key to (**B**)). The size of each node is proportional to the number of genes contributing to the disorder. Edges between disorders in the same disorder class are colored with the same (lighter) color, and edges connecting different disorder classes are colored gray, with the thickness of the edge proportional to the number of genes shared by the disorders connected by it. (B) Disease gene network. Each node is a single gene, and any two genes are connected if implicated in the same disorder. In this network map, the size of each node is proportional to the number of specific disorders in which the gene is implicated. (Reproduced with permission from the National Academies Press; Goh *et al*, in press.)

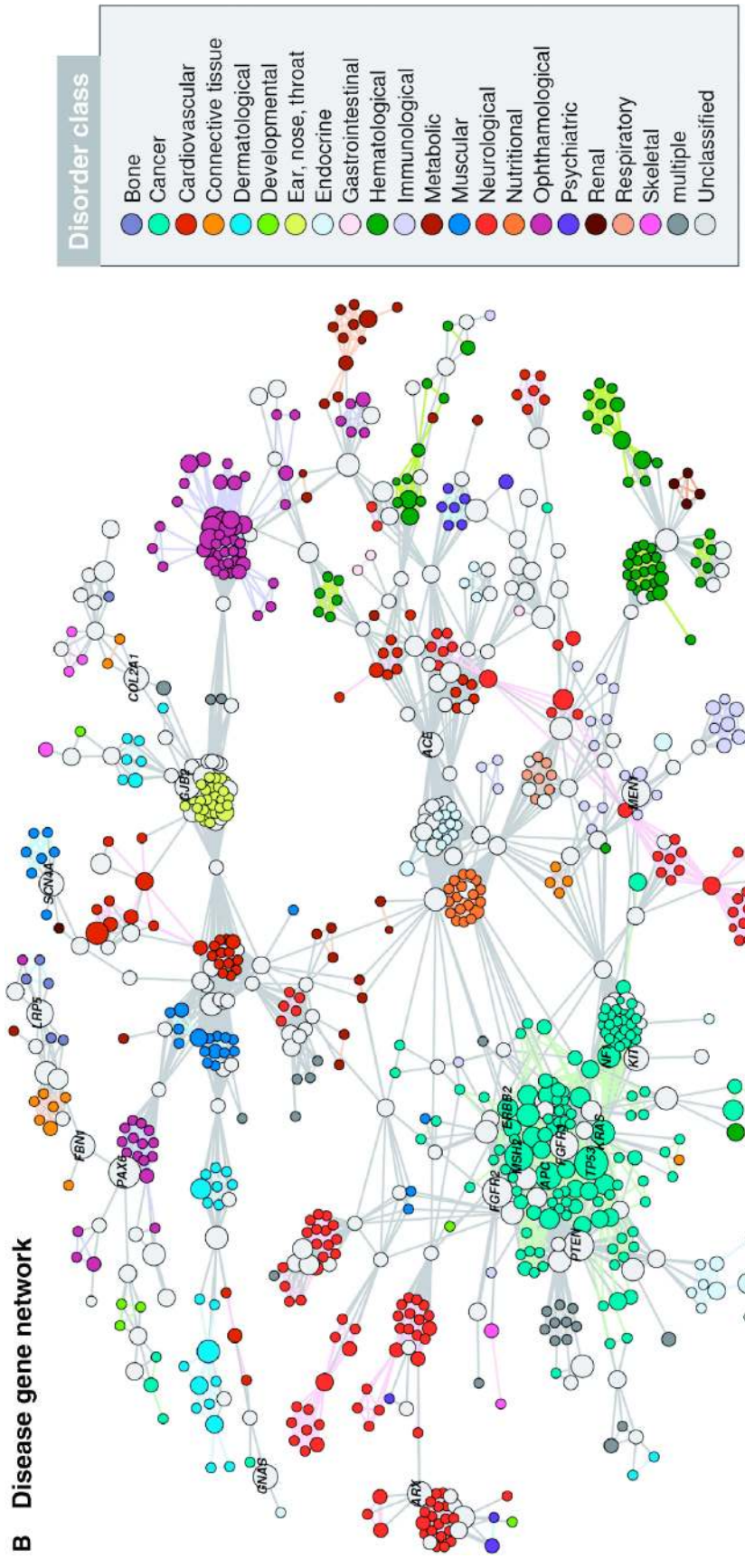

Figure 4 continued.

What can this rapidly evolving knowledge of cellular networks tell us about diseases? For several disorders known to arise from mutations in any one of a few distinct genes, the corresponding protein products have been shown to participate in the same cellular pathway, molecular complex, or functional module. For example, Fanconi anemia arises from mutations in a set of genes encoding proteins involved in DNA repair, many of them forming a single heteromeric complex. Recent findings indicate that such association between disorders and distinct functional modules is more than anecdotal. Indeed, the protein products of the genes that belong to common disorder classes tend to interact with each other via protein–protein interactions, to display high coexpression levels, and to exhibit synchronized expression as a group (Yook *et al*, 2004). Taken together, these findings support the idea of a global functional relatedness for disease genes and their products, and offer a network-based model for the disease propensity in an individual (Goh *et al*, in press). Cellular networks are modular, consisting of groups of highly interconnected proteins responsible for specific cellular functions. In this construct, a disease represents the perturbation or breakdown of a specific functional module caused by variation in one or more of the components producing recognizable developmental and/or physiological abnormalities. This model offers a simple explanation for the emergence of complex or polygenic disorders: a phenotype often correlates with the inability of a particular functional module to carry out its basic functions. For extended modules, many different combinations of perturbed genes could incapacitate the module, as a result of which mutations in different genes will appear to lead to the same phenotype (e.g., hypertrophic cardiomyopathy). This correlation between disease and functional modules can also inform our understanding of cellular networks by helping us to identify which genes are involved in the same cellular function or network module. Importantly, this association of disease with functional modules can also inform our choice of rational therapeutic targets.

Conclusion

What, then, is the benefit of a network analysis of disease and its treatment? First, systems-based network analysis can identify those determinants (nodes) or combinations of determinants that strongly influence network behavior and disease expression or phenotype. Second, these regulatory determinants may not always be obvious from reductionist principles, and, thus, the analysis provides unique insight into disease mechanism and potential therapeutic targets. Third, network analysis of disease gives one the opportunity to consider with quantitative rigor the relationships within the network genome, environmental exposures, and environmental effects on the proteome (posttranslational proteome) that define the specific pathophenotype. In this construct, disease can be considered the result of a modular collection of genomic, proteomic, metabolomic, and environmental networks that interact to yield the pathophenotype. Fourth, disease network analysis ultimately provides a mechanistic basis for defining phenotypic differences among individuals with the same disease through consideration of unique genetic and environmental factors that govern intermediate phenotypes contributing to disease expression. Lastly, disease network analysis offers a unique method for identifying therapeutic targets or combinations of targets that can alter disease expression. In short, this approach offers a novel method for classifying human disease. The novelty in this approach rests not simply in nosology, but in defining disease expression on the basis of its molecular and environmental elements in a holistic and fully deterministic way. As we have reviewed here, the application of these principles to specific diseases is in its infancy, but the early concepts are internally consistent and the early results encouraging.

Acknowledgements

We thank Susan Vignolo-Collazzo, Katherine Seropian, and Stephanie Tribuna for expert secretarial assistance.

References

Albert R (2005) Scale-free networks in cell biology. *Journal Cell Sci* **118:** 4947–4957

Albert R, Barabási AL (2002) Statistical mechanics of complex networks. *Rev Mod Phys* **74:** 47–97

Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. *Nature* **406:** 378–382

Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* **286:** 509–512

Barabási AL, Oltvai Z (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5:** 101–113

Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A (2005) Reverse engineering of regulatory networks in human B cells. *Nat Genet* **37:** 382–390

Bhan A, Galas DJ, Dewey TG (2002) A duplication growth model of gene expression networks. *Bioinformatics* **18:** 1486–1493

Dave SS, Fu K, Wright GW, Lam LT, Kluin P, Boerma EJ, Greiner TC, Weisenburger DD, Rosenwald A, Ott G, Muller-Hermelink HK, Gascoyne RD, Delabie J, Rimsza LM, Braziel RM, Grogan TM, Campo E, Jaffe ES, Dave BJ, Sanger W, Bast M, Vose JM, Armitage JO, Connors JM, Smeland EB, Kvaloy S, Holte H, Fisher RI, Miller TP, Montserrat E, Wilson WH, Bahl M, Zhao H, Yang L, Powell J, Simon R, Chan WC, Staudt LM (2006) Molecular diagnosis of Burkitt's lymphoma. *N Eng J Med* **354:** 2431–2442

Dezso Z, Barabasi AL (2002) Halting viruses in scale-free networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **65**, [E-pub]

Ergun A, Lawrence CA, Kohanski MA, Brennen TA, Collins JJ (2007) A network biology approach to prostate cancer. *Mol Syst Biol* **3**, [E-pub]

Eubank S, Guclu H, Kumar VS, Marathe MV, Srinivasan A, Toroczkai Z, Wang N (2004) Modeling disease outbreaks in realistic urban social networks. *Nature* **429:** 180–184

Farber H, Loscalzo J (2004) Pulmonary hypertension. *N Engl J Med* **351:** 1655–1665

Gardner TS, di Bernardo D, Lorenz D, Collins JJ (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301:** 102–105

Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabasi A-L##The human disease network. *Proc Natl Acad Sci USA* (in press)

Hall P, Ploner A, Bjohle J, Huang F, Lin CY, Liu ET, Miller LD, Nordgren H, Pawitan Y, Shaw P, Skoog L, Smeds J, Wedren S, Ohd J, Bergh J (2006) Hormone-replacement therapy influences gene expression profiles and is associated with breast-cancer prognosis: a cohort study. *BMC Med* **4:** 16

Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. *Nature* **402:** C47–C52

Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J, Raffeld M, Yakhini Z, Ben-Dor A, Dougherty E, Kononen J, Bubendorf L, Fehrle W, Pittaluga S, Gruvberger S, Loman N, Johannsson O, Olsson H, Sauter G (2001) Gene-expression profiles in hereditary breast cancer. *N Eng J Med* **344:** 539–548

Kato GJ, Gladwin MT, Steinberg MH (2007) Deconstructing sickle cell disease: reappraisal of the role of hemolysis in the development of clinical subphenotypes. *Blood Rev* **21:** 37–47. e-published (2006).

Kim J (2006) Emergence: core ideas and issues. *Synthese* **151:** 547–559

Kitano H (2004) Biological robustness. *Nat Rev Genet* **5:** 826–837

Kitano H (2007) Innovation: a robustness-based approach to systems-oriented drug design. *Nat Rev Drug Discov* **4:** 202–210

Lim J, Hao T, Shaw C, Patel AJ, Szabo G, Rual J-F, Fisk CJ, Li N, Smolyar A, Hill DE, Barabasi A-L, Vidal M, Zoghbi HY (2006) A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* **125:** 801–814

Loscalzo J (2006) Homocysteine trials—clear outcomes for complex reasons. *N Engl J Med* **354:** 1629–1632

Lu X, Jain VV, Finn PW, Perkins DL (2007) Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. *Mol Syst Biol* **3:**98

Madar N, Kalisky T, Cohen R, Ben-Avraham D, Havlin S (2004) Immunization and epidemic dynamics in complex networks. *Eur Phys J B* **38:**269–276

Morel NM, Holland JM, van der Greef J, Marple EW, Clish C, Loscalzo J, Naylor S (2004) Primer on medical genomics. Part XIV: introduction to systems biology—a new approach to understanding disease and treatment. *Mayo Clin Proc* **79:** 651–658

Nicholson JK (2006) Global systems biology, personalized medicine and molecular epidemiology. *Mol Syst Biol* **2:**52

Oikonomou P, Cluzel P (2006) Effect of topology on network evolution. *Nat Phys* **2:** 532–536

Pastor-Satorras R, Smith E, Sole R (2003) Evolving protein interaction networks through gene duplication. *J Theor Biol* **222:** 199–210

Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. *Phys Rev Letts* **86:** 3200–3203

Pastor-Satorras R, Vespignani A (2002) Immunization of complex networks. *Physical Review E* **65:** 036104

Qian J, Luscombe NM, Gerstein M (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J Mol Biol* **313:** 673–681

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL (2002) Hierarchical organization of modularity in metabolic networks. *Science* **297:** 1551–1555

Schwartz WB, Wolfe HJ, Pauker SG (1981) Pathology and probabilities: a new approach to interpreting and reporting biopsies. *N Eng J Med* **305:** 917–923

Sebastiani P, Ramoni MF, Nolan V, Baldwin CT, Steinberg MH (2005) Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nat Genet* **37:** 435–440

Seidman JC, Seidman C (2001) The genetic basis for cardiomyopathy from mutation identification to mechanistic paradigms. *Cell* **104:** 557–567

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444:** 1027–1031

Vazquez A, Flammini A, Maritan A, Vespignani A (2003) Modeling of protein interaction networks. *ComPlexUs* **1:** 38–44

Wagner A, Fell DA (2001) The small world inside large metabolic networks. *Proc R Soc Lond B* **268:** 1803–1810

Wurthner JU, Mukhopadhyay AK, Peimann CJ (2000) A cellular automaton model of cellular signal transduction. *Comput Biol Med* **30:** 1–21

Yook SH, Oltvai ZN, Barabási AL (2004) Functional and topological characterization of protein interaction networks. *Proteomics* **4:** 928–942

Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED (2004) Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* **20:** 3594–3603