8038 J. Med. Chem. 2009, 52, 8038–8046
DOI: 10.1021/jm901001p

Journal of Medicinal Chemistry Article

Drug Effects Viewed from a Signal Transduction Network Perspective

Anton F. Fliri,[†,‡] William T. Loging,[§] and Robert A. Volkmann*[,†,‡]

[†]*Pfizer Global Research and Development, Eastern Point Road and* [‡]*SystaMedic Inc., 1084 Shennecossett Road, Groton, Connecticut 06340, and* [§]*Boehringer Ingelheim Pharmaceuticals, Inc., 900 Ridgebury Road, Ridgefield, Connecticut 06877*

Received July 7, 2009

Understanding how drugs affect cellular network structures and how resulting signals are translated into drug effects holds the key to the discovery of medicines. Herein we examine this cause−effect relationship by determining protein network structures associated with the generation of specific *in vivo* drug-effect patterns. Medicines having similar *in vivo* pharmacology have been identified by a comparison of drug-effect profiles of 1320 medicines. Protein network positions reached by these medicines were ascertained by examining the coinvestigation frequency of these medicines and 1179 protein network constituents in millions of scientific investigations. Interestingly, medicine associations obtained by comparing by drug-effect profiles mirror those obtained by comparing drug−protein coinvestigation frequency profiles, demonstrating that these drug−protein reachability profiles are relevant to *in vivo* pharmacology. By using protein associations obtained in these investigations and independent, curated protein interaction information, drug-mediated protein network topology models can be constructed. These protein network topology models reveal that drugs having similar pharmacology profiles reach similar discrete positions in cellular protein network systems and provide a network view of medicine cause−effect relationships.

Introduction

Understanding how medicines work has always been a challenge in drug discovery.[1] In the past, scientists typically relied on observations of organism's response to drug treatment for predicting drug effects. Contemporary approaches, on the other hand, rely largely on preclinical data and assessments of cellular behaviors emerging from the interaction between medicines and protein network components. Unfortunately, despite the wealth of information gained in recent years, statistics show that the limited information gathered from preclinical studies is insufficient for predicting the full spectrum of drug-effect observations in organisms. One likely reason for this predicament is that current drug discovery paradigms generally do not consider the plasticity of biological systems, which adapt to or compensate for loss or decline in specific protein functions by rerouting the information flow in organism network systems.[2] While the plasticity of living organisms increases chance of survival in case of injury, it also creates problems for drug-effect predictions that are based on the examination of mechanism of action-centered cause−effect relationships.[3] Thus, increasing current success rates of drug discovery likely requires consideration of information provided by the examination of system-wide, drug-effect relationships.[4−8] Working toward this goal, we and others have recently described the development of methods providing quantitative comparisons of broad preclinical and clinical drug-effect information for medicines.[9−13] These methods provide interesting and important associations between medicines, between proteins, and between effects and rely on enormous amounts of information from disparate data sources. In our investigations, all of the data are converted into information spectra and normalized, and these spectra are sorted by hierarchical clustering to provide drug, protein, and effect associations.

Identifying the proteins which play a role in generating drug-mediated organism effects is central to deciphering system-wide drug-effect relationships. For identifying proteins that are in some way contributing to a medicine's biological profile, we have relied on the coinvestigation frequency of proteins with medicines in millions of reported studies. Accumulating this information for a large number of medicines provides a useful comparative tool for analyzing medicines from a protein reachability perspective and, in addition, provides a mechanism for understanding functional relationships between proteins. By sorting medicine−protein coinvestigation frequency profiles via hierarchical clustering, both medicine and protein associations are obtained. Medicine associations group pharmacological agents by their ability to reach similar proteins and affect their function. Protein associations obtained in this fashion are an indicator of protein functional coupling. This assumption is based on the premise that proteins have a higher probability of being coinvestigated and associated in structure−function studies or are likely co-occurring with similar frequency in scientific publications, if their pertinent functions, effects, or properties are coupled or correlated.

In this investigation, we use dendrogram relationships of proteins provided by the hierarchical clustering of the coinvestigation frequency spectra of 1320 drugs (Supporting Information) and 1179 proteins (Supporting Information) to identify the functional coupling of proteins. For the analysis of this network reachability information, coinvestigation counts greater than 100 were set to 100, since higher values

*Address correspondence to this author at SystaMedic Inc. Tel: 860-912-6101. Fax: 860-405-9031. E-mail: ravolkmann@gmail.com.

 Published on Web 11/05/2009

would not increase the certainty of medicines reaching a particular protein network position. This normalization strategy provides directly comparable network reachability information spectra for each of the 1320 medicines and mitigates existing variations in information density obtained, in particular, with medicines more frequently investigated. By utilizing over a thousand medicines and over a thousand proteins, the overall results and interpretations from these investigations are not affected by the accuracy of individual data points but rather determined by the overall shape of information spectra ("fingerprint or discriminative properties of information"). The methodology is particularly well suited for investigating interactions between complex protein network systems which require analysis of heterogeneous, incomplete, and noisy information sources. Herein, we describe a platform for generating protein network topology map models which describe the linkage between protein network components associated with the generation of specific *in vivo* drug-effect patterns.

Results

Aligning Preclinical and Clinical Information Spectra of Medicines. Identification of medicines with similar effect similarities is a prerequisite for evaluating the protein network components associated with the characteristic *in vivo* drug-effect patterns of these medicines. COSTART and MedDRA descriptors, which are clinically validated international medical terminology descriptors, have been used for this purpose, and the pharmacological relevance of these descriptors is well-known.[9−13] By determining the cocitation frequency of 1082 COSTART-effect descriptor terms (Supporting Information) and 1320 medicines in millions of published investigations and creating COSTART-effect spectra for each medicine, system-wide *in vivo* drug-effect profiles of these medicines can be compared. Close inspection of these COSTART drug-effect profiles reveals that, while many medicines with similar molecular targets and chemical architectures often produce very similar clinical drug-effect profiles, a significant number of drugs with different molecular targets and having different chemical architectures also produce similar clinical effect profiles. For example, the COSTART-based effect information profile of rosiglitazone (**1**), a potent agonist of peroxisome proliferator-activated receptor (PPAR)-gamma, exhibits 95% drug-effect information profile similarity with glimepiride (**2**), which is an insulin secretagogue and, like rosiglitazone, used for treatment of diabetes. In contrast, the drug-effect spectra profile associated with ciglitazone (**5**), another PPAR-gamma agonist, exhibits only a 75% drug-effect profile similarity with its mechanism equivalent, rosiglitazone (Figure 1C). Moreover, comparison of the COSTART-effect profile of rosiglitazone with the COSTART profiles of all 1320 medicines reveals that medicines with different chemical structures and mechanisms of action, such as, for example, the PPAR-alpha agonists, fenofibrate and benzafibrate, HMGCoA reductase inhibitors, simvastatin and atorvastatin, and the biguanide, metformin, share more than 70% effect profile similarity with rosiglitazone. By extending this scrutiny to other COSTART profiles in the entire 1320 medicine database, many additional instances of *in vivo* drug-effect similarities between structurally and mechanistically distinct medicines were revealed, indicating that mechanism of action and/or molecular structure similarity are (is) not sufficient for explaining drug-effect similarities between medicines.

The drug−protein coinvestigation frequency spectra for medicines with similar effect spectra similarity were identified. These protein-based spectra are useful for capturing and comparing cellular (preclinical) drug−protein information of this medicine cohort.[12] For example, examination of drug−protein coinvestigation frequency spectra profiles and the drug-effect cocitation frequency spectra profiles of the 12 medicines having the greatest effect spectra similarity to rosiglitazone (Figure 1C) revealed that the two drug-effect information profiles are highly correlated ($R^2 = 0.917$; $p = 0.083$). Interestingly, comparable correlations between *in vivo*/*in vitro* drug-effect information profile similarities were also observed with other groups of mechanistically and structurally distinct medicines [i.e., clozapine and fluoxetine (Figure 1D)]. For shedding light on the origin of these *in vitro* and *in vivo* drug-effect similarities, hierarchical clustering of the *in vitro* (drug−protein) and *in vivo* (drug-effect) spectra profiles of 1320 medicines was carried out.

Identifying Protein Network Components Responsible for Drug-Effect Patterns. Hierarchical clustering of the COSTART-effect profiles associated with the entire cohort of 1320 medicines using the UPGMA algorithm and cosinus correlation as similarity measurement simultaneously sorts 1320 medicines by similar effect spectra and 1082 effect terms using these medicines. Accordingly, the "*Y*-axis" dendrogram (cluster) identifies associations of "COSTART effects" found with these medicines and the "*X*-axis" dendrogram identifies medicines sharing characteristic drug-effect patterns. Similarly, hierarchical clustering of the protein co-occurrence frequency information of the 1320 medicines identifies on the "*X*-axis" groups of medicines that are most frequently coinvestigated with similar proteins using the 1179 different protein network monitoring positions. The vertically displayed dendrogram, in turn, identifies groups (associations) of proteins that are most frequently coinvestigated in structure−function studies with 1320 medicines (Figure 2D). Anticipating that high coinvestigation frequency between specific drugs and specific proteins generally reflects the capacity of these medicines to modulate pertinent protein functions, close inspection of the distribution of the clustered drug−protein coinvestigation frequency information reveals that associated medicines affect the functions of discrete groups of proteins. In this way, the clustered coinvestigation frequency information between 1320 drugs and 1179 proteins not only provides evidence that associated medicines in this data set reach discrete positions in the cellular protein network system but they also affect the functions of discrete sets of proteins.[14−16]

Comparing the distribution of the drug−protein coinvestigation frequency information associated with medicines producing similar *in vivo* effects indicates that the generation of characteristic drug-effect patterns *in vivo* is paralleled by the generation of characteristic drug-effect patterns on discrete sets of proteins (Figure 1C,D). Close inspection of respective drug-effect patterns suggests that medicines sharing similar *in vitro* (protein) and *in vivo* (COSTART) drug-effect information profiles reach similar positions in respective network systems. Furthermore, since the generation of specific drug-effect patterns in organisms is preceded by the generation of discrete functional relationships between specific protein network components, the observed correlation between these two different sets of network reachability information suggests cause−effect relationships and hence provides information on the functions of underlying

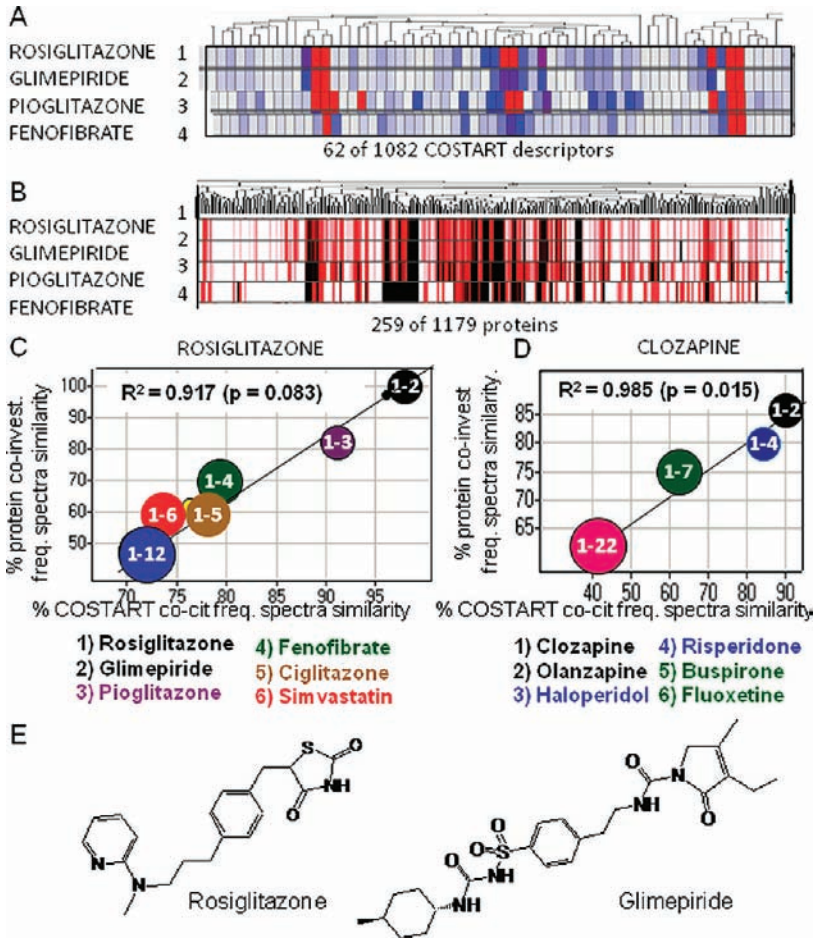

Figure 1. Investigating COSTART cocitation and protein coinvestigation frequency profiles of medicines **1**−**4**. (A) A portion of the COSTART-effect spectra profiles of rosiglitazone (**1**), glimepiride (**2**), pioglitazone (**3**), and fenofibrate (**4**). The cocitation frequency between drugs and effects is used to assess drug-effect similarity (gray = 0, blue = 50, and red = > 100 COSTART citations). (B) A portion of the protein coinvestigation frequency spectra profiles for medicines **1**−**4**. The coinvestigation frequency between drugs and proteins is used to assess medicines' effects on protein function (white = 0, red = 5, and black = > 20 protein citations). (C) A correlation exists between preclinical (protein: *Y*-axis) and clinical (COSTART: *X*-axis) drug-effect information profiles for 12 medicines sharing greater than 70% COSTART-effect spectra profile similarity correlations with rosiglitazone (**1**). Medicines with the greatest spectra correlations to each standard are shown in order by color below each graph. (D) Correlation between the two profiles for 22 medicines sharing more than 40% COSTART-effect spectra profile similarity with clozapine. (E) Molecular structures of rosiglitazone and glimepiride.

large-scale network structures. Anticipating that the dynamic modulation of the topology of the signal transduction network may hold the key for understanding these cause−effect relationships, a new approach was examined for constructing the router-level connectivity of these network structures.

On Establishing the Router-Level Connectivity of Network Structures. For identifying network structures involved in information processing in cellular systems, several new approaches have recently been developed.[17,18] Most of these approaches rely on information derived from yeast two-hybrid experiments and the assumption that knowledge pertaining to specific protein−protein interactions in yeast can be extended to higher organisms.[5,6,19−21] However, in higher organisms, signal transduction network topologies have been shown to vary with cell type, sex, age, time, expression level of receptors/effectors/targets, nature of posttranscriptional modifications, disease background, variations in environmental backgrounds, and many other factors modifying transcellular and intracellular communication pathways.[5,6,19−23] Thus, investigating functions of signal transduction networks encounters the formidable challenge of assessing the pharmacological significance of vast amounts of protein network connectivity information residing in protein−protein interaction databases. Hence, in the absence of correlated functional-effect information, connectivity information derived from yeast two-hybrid experiments describes, in principle, architectures of nearly infinite sets of signal-inducible protein network topologies.[22,23]

Methods developed for identifying network topologies of large-scale communication networks[24−26] seemed appropriate for shedding light on the router-level connectivity of signal transduction networks. One of these methods, frequently referred to as measurement-induced network topology (MINT), relies on the identification of network positions that can be reached by tracer probes (Figure 2A) sent from a specific source to a specific destination.[25,27] This approach uses the clustering of network reachability information associated with individual tracer probes for identifying network nodes that are most frequently encountered by tracer probes routed through the network system. Anticipating that the most direct routes for information transfer involve neighboring network positions, the clustering of network reachability information is used in our case for identifying

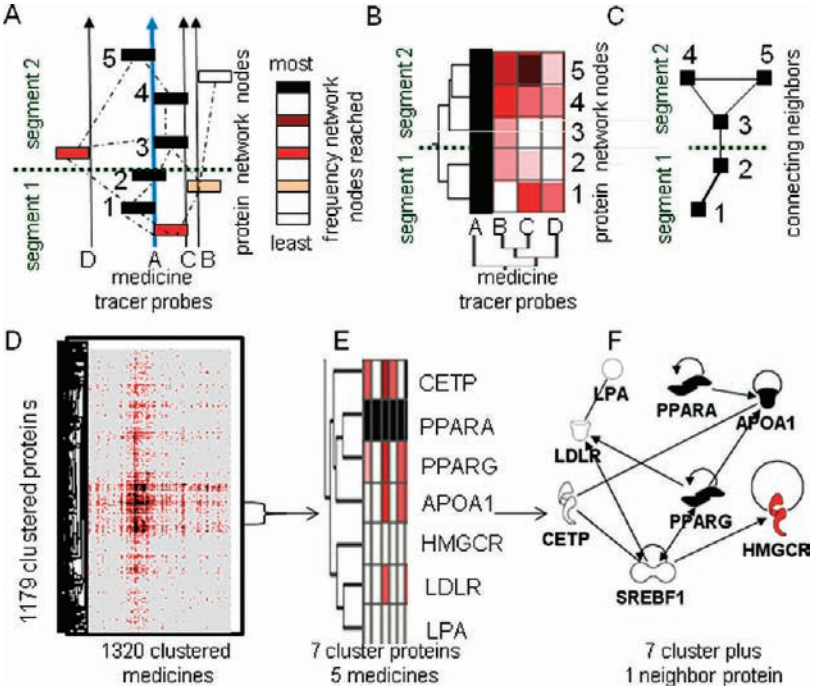

Figure 2. Overall strategy for generating drug-induced protein network topology maps. (A) Text mining-derived drug−protein coinvestigation frequency information associated with medicines **A**−**D** is used for identifying protein network positions that can be reached during drug treatment. This network reachability information is used for identifying the average shortest path distance (blue arrow) for transferring drug-induced signals through the protein network. (B) Clustering network reachability information identifies a group of proteins, **1**−**5**, most frequently coinvestigated with medicines **A**−**D**. Inspecting dendrogram relationships identifies (a) cellular protein network positions that can be reached by medicines **A**−**D** (tracer probes with similar pharmacology) and (b) associated protein clusters (e.g., proteins **1**−**5**) most frequently involved in conducting drug-inducible signals through the network system. (C) Translation of cluster proximity relationships between proteins into network topology information. (D) A heat map obtained by clustering protein coinvestigation frequency spectra of 1320 medicines using 1179 proteins identifies protein and medicine associations. (E) An example of protein dendrogram relationships (*Y*-axis) obtained in clustering showing seven functionally coupled proteins. (F) A protein network topology map is generated by adding the minimal number of neighbor proteins (in this case one: SREBF1) identified using curated protein interaction databases to directly connect all the dendrogram proteins shown in panel E.

associations of protein network nodes sharing high cotransmission frequencies and for identifying associations of medicine tracer probes encountering similar network positions (Figure 2B). Connection of protein network nodes that are most frequently encountered by discrete associations of medicine tracer probes (tracer probes using similar routes for the "end-to-end" transfer of information through the network system) identifies the shortest routes (paths) for information transfer through the communication network.[28] By analysis of associations provided by the clustering of network reachability information linked with all tracer probes, the entire router-level network connectivity can be identified (Figure 2C).

Accordingly, in lieu of tracer probes, herein we used 1320 medicines residing in published structure−function information (drug−protein coinvestigation frequencies) for identifying which of the 1179 protein network positions can be reached by each of them. For identifying associations of medicines likely using similar routes for transferring information through the effect network systems, medicines having high *in vivo*/*in vitro* correlations as illustrated in Figure 1C,D were used. In this respect, the hierarchical clustering of drug-COSTART cocitation frequency and drug−protein coinvestigation frequency information were used for identifying "end" (having similar effects on organ systems) and "start" (reaching similar positions in protein networks) positions of the information transfer associated with medicines with similar *in vivo* and *in vitro* pharmacology.[29−32] Concomitantly, we examined if protein associations provided by clustering of coinvestigation frequency information identify router-level connectivities of the signal transduction network.[19,29,30]

Protein Associations Identified by the Hierarchical Clustering of Drug−Protein Spectra. The horizontal (*X*-axis) dendrogram obtained by clustering 1320 medicine−protein cooccurrence frequency spectra (Figure 3A) identifies groups of medicines with similar *in vitro* structure−function information, and their spectra provide information on how often specific signal transduction network positions are reached by a particular group of medicines. The vertically displayed dendrogram, in turn, identifies the frequency of information exchanges between discrete groups of proteins and provides a mechanism for assessing which groups of proteins are reachable by medicine groups (Figure 3).

Anticipating that the network reachability information provided by 1179 proteins (a subset of the proteome) is not sufficient for identifying the large-scale framework of the signal transduction network, the strategy adopted was to use the 1179 proteins and curated protein interaction information (yeast two-hybrid data) to identify additional proteins not in our original data set that are capable of directly interacting with the monitoring proteins. These directly interacting, nearest neighbor proteins were anticipated to be capable of directly transmitting the drug-induced information flow between the 1179 monitoring positions. Therefore, by determining protein associations via clustering of

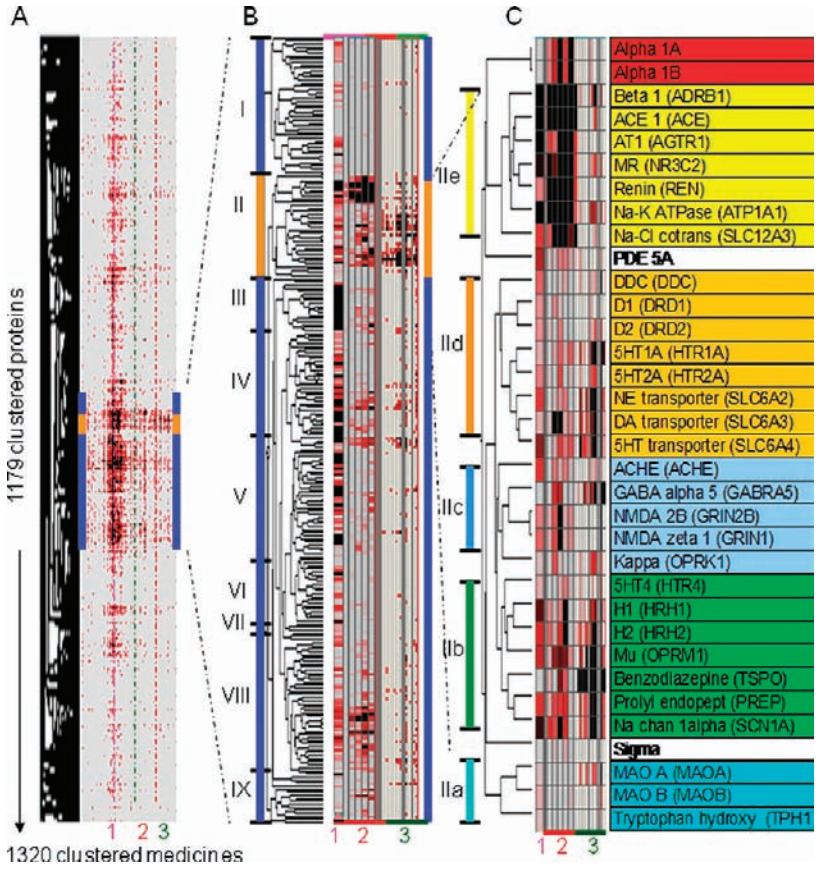

Figure 3. Protein associations resulting from hierarchical clustering of the coinvestigation frequency information of 11 million structure−function studies involving 1179 proteins and 1320 medicines. (A) The horizontal dendrogram (not shown) identifies protein coinvestigation frequency spectra similarity of 1320 medicines, and the vertical dendrogram (shown) identifies medicine coinvestigation frequency similarity of 1179 proteins. Color shadings are used for identifying coinvestigation frequencies: red denotes frequency measures in the range 1−20 and black exceeding >20 literature citations. Network reachability information (protein profiles) is shown for (1) a statin, (2) 5 dihydropyridines, and (3) 15 benzodiazepines. (B) An enlargement of a vertical dendrogram section denoted as I−IX is shown for identifying a cohort of 259 proteins that are most often coinvestigated in structure−function studies with the 1320 medicines. (C) Further enlargement of a subsection of this dendrogram, denoted as II, for illustrating that the coinvestigation frequency information for medicines in the three highlighted pharmacology classes exhibits characteristic coinvestigation frequency patterns (evidence that similar medicines reach discrete protein network positions).

protein network reachability information and adding selected nearest neighbor proteins, a mechanism was in place for identifying shortest path distances conducting the information flow induced by 1320 medicines through the signal transduction network system.[6,20−23]

Starting with the premise that the coinvestigation frequencies between proteins engaged in coupled functions are higher than the coinvestigation frequency of proteins executing unrelated functions, the first goal in the analysis was to investigate if protein associations identified in the various vertical dendrogram sections occupy proximate protein network positions.[19] Accordingly, inspection of the protein associations defined in the vertical displayed dendrogram section of 1179 proteins identifies a prominent group of 259 proteins (Figure 3B), sharing a confidence in cluster similarity value (CCS) of >0.426 [algorithm cluster scoring wherein 0 = lowest (no similarity) to 1 = highest (identical)].[12] Close inspection of relationships between these 259 proteins identifies nine large protein associations (dendrogram clusters), denoted as I, II, III, IV, V, VI, VII, VIII, and IX (Supporting Information). Moreover, inspections of groups I−IX indicate that each of these main dendrogram sections is partitioned into several smaller protein associations, wherein each of these smaller associations contains proteins that are most frequently coinvestigated in structure−function studies. For example, dendrogram section II (Figure 3B) contains protein associations IIa−IIe (Figure 3C). Moreover, inspecting the coinvestigation frequency relationship of proteins in associations IIa−IIe indicates that proteins residing in individual associations are frequently coinvestigated not only with members residing in the same subcluster but also with proteins positioned in the adjacent subcluster sections. This observation indicates that proteins in IIa−IIe (some of which are drug targets for marketed cardiovascular, analgesic, and anxiolytic medicines) are engaging in drug-induced information exchanges involving multiple combinations of different groups of proteins and that each of these proteins is, in principle, capable of affecting either directly or indirectly functions of any one of the comembers of protein associations IIa−IIe. Moreover, inspecting the coinvestigation frequency relationships of proteins in II and proteins located in associations I, III, IV, V, VI, VII, VIII, and IX indicates proteins in II are also frequently coinvestigated with proteins residing in associations I, III, IV, V, VI, VII, VIII, and IX. This observation indicates that proteins in II are not only capable of affecting the functions of other comembers but also have the capacity of affecting the functions of proteins residing in associations I and III−IX (Figure 3B). These observations suggest that these 259 proteins form an integrated interaction network.

Using Protein Associations Established by Hierarchical Clustering To Construct Protein Network Topology Maps. Clustering-derived protein associations (Figure 3B,C) identify proteins which are functionally coupled. In many cases, as anticipated, the direct connectivity between these proteins can be corroborated by curated protein information. The pertinent data used to identify direct protein−protein interactions can be extracted from the Ingenuity Pathways Knowledge Base (Ingenuity Systems, www.ingenuity.com), a reliable source of independent protein-interaction information. Thus, the "direct" protein-interaction information used for corroborating the functional coupling of dendrogram proteins includes activation of function, changes in protein expression levels, inhibition of protein functions, changes in a protein's localization, changes in the degree of phosphorylation, effects on protein−DNA interactions, effects on protein−protein interactions, effects on protein−RNA interactions, the proteolysis of binding partners, the regulation of a protein's ligand binding capacities, and changes in transcription rates and effects on the translocation of interacting proteins as criteria for direct protein coupling. If dendrogram proteins are not directly coupled, additional proteins (nearest neighbor proteins) are identified in the Ingenuity platform and used to link all of the dendrogram-associated proteins. This particular protein network connectivity strategy was selected based on the anticipation that direct interactions between nearest neighbor proteins would provide the least ambiguous means for ascertaining shortest path, distance-network topology, enabling efficient information transfer.[19] Accordingly, for determining network proximity relationships[19] between the 259 member proteins in I−IX, nine network fragments, I^n−IX^n, were constructed. The construction of respective network fragments involved connecting all of the proteins in individual dendrogram sections I−IX either directly or by adding nearest neighbor proteins to produce network fragments I^n−IX^n via shortest path distance (see Experimental Section for details).

Close inspection of the topology of protein network fragments IIa^n−IIe^n (Figure 4) reveals that the dendrogram proteins of each of the five subsets of II (Figure 3C) can be placed in integrated protein networks having direct protein−protein interactions once neighbor proteins are added. Interestingly, proteins in IIe^n are also found in IIb^n−IId^n (Figure 4) providing a mechanism for protein network overlap. The fact that proteins residing in cluster dendrograms IIa−IIe are reachable by many medicines and are capable of directly interacting with the same nearest neighbor proteins suggests that the protein connectivities identified in dendrogram sections IIa−IIe are capable of directly transmitting information to respective cluster comembers. Since network topologies IIa^n−IIe^n are based on the clustered structure−function information provided by all of the 1320 medicines, network topologies IIa^n−IIe^n (forming the integrated protein network topology II^n) identify the likely average, shortest path distances (fastest route) of the information transfer, induced by the 1320 medicines through this particular signal transduction network section.[15]

Likewise, for determining routes for the drug-inducible information transfer between proteins identified in dendrogram section II and proteins identified in dendrogram sections I and III−IX (Figure 5A), the network topology overlap between these different network fragments was determined. This network topology overlap was determined

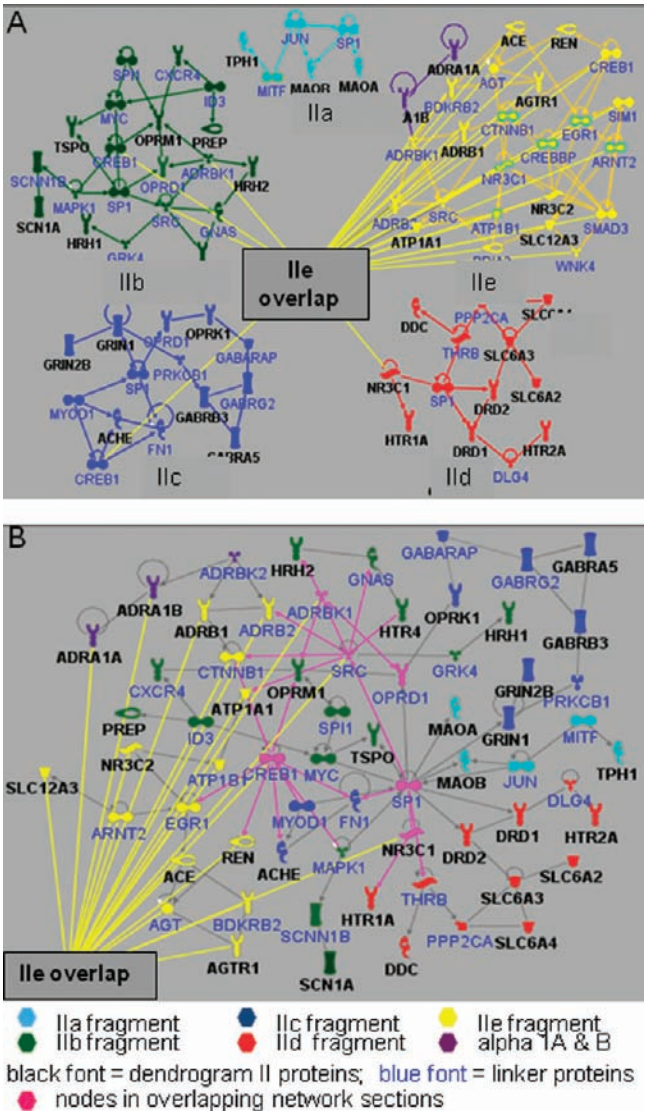

Figure 4. The construction of protein network fragments using proteins in dendrogram subsections IIa−IIe (Figure 3C). (A) Network fragments IIa^n−IIe^n are constructed using proteins (black font) in dendrogram sections IIa−IIe and by identifying directly interacting nearest neighbor proteins (shown in blue font) using the Ingenuity platform. Only 2 of the 34 proteins in dendrogram section II (sigma and PDE 5A) could not be connected using Ingenuity's direct protein interaction information. Each of the five network fragments of II contains a collection of functionally linkable protein network nodes and is depicted by respective colors: light blue (IIa), green (IIb), dark blue (IIc), red (IId), and yellow (IIe). (B) The topology overlap between IIa^n−IIe^n creates a larger network fragment II^n, seamlessly integrating fragments IIa^n−IIe^n. Proteins (shown in magenta) residing in overlapping network sections (e.g., SP1, CREB1, SRC) are not unique to a particular fragment and create overlapping network topologies.

by identifying protein network nodes shared by network fragments I^n−IX^n (Supporting Information). Inspection of connectivity between network nodes shared by different network fragments indicates that fragments I^n−IX^n have overlapping network topologies (Figure 5B). Moreover, examining proximity relationships between network nodes in fragments I^n−IX^n indicates that the relationships between proteins identified through the clustering of medicine−protein coinvestigation frequency information involve either direct interactions between cluster proteins or direct

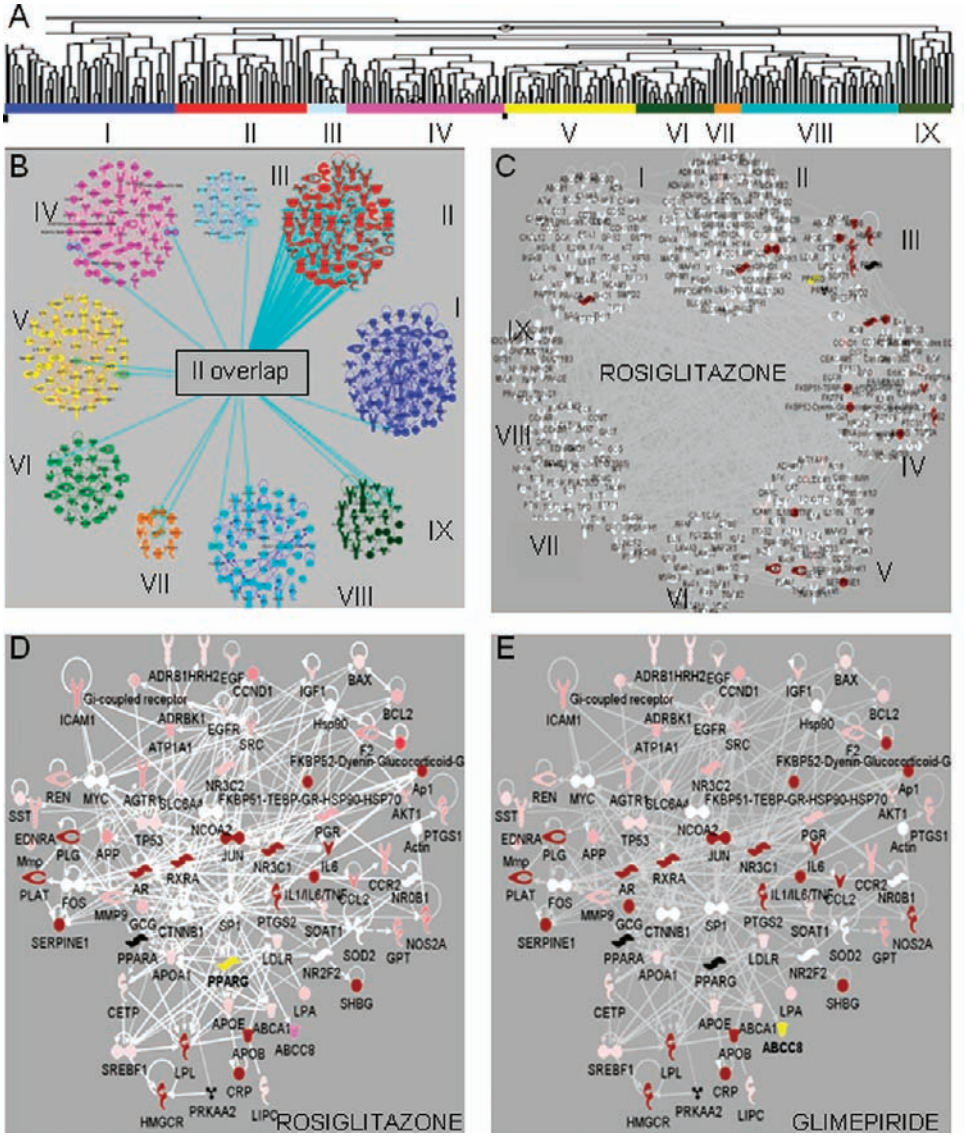

Figure 5. Investigating drug-effect similarities using drug-induced protein network topology maps. (A) Dendrogram sections I−IX containing 259 proteins (Figure 3B) are used for generating a protein network topology map I^n−IX^n containing these proteins. (B) Comember proteins in II^n and I^n and III^n−IX^n create overlapping network topologies. (C) By superimposing individual drug−protein coinvestigation frequency information (white = 0, red = 20, and black = > 50) on the drug-induced protein network topology map I^n−IX^n, shortest path distances for information transfer can be described for each of the 1320 medicines. A protein network topology map for rosiglitazone is shown. (D) An integrated protein network topology map is shown for rosiglitazone containing the minimal number of proteins in I^n−IX^n required to connect all of its network reachable proteins. The drug target (PPARG) of rosiglitazone is shown in yellow. (E) An integrated protein network topology map for glimepiride, whose putative drug target (ABCC8), is highlighted.

interactions of the cluster proteins with the same nearest neighbor proteins. This observation indicates that all of the 259 proteins residing in dendrogram sections I−IX are capable of directly exchanging information via shortest path distances. From a drug-effect prediction perspective, most noteworthy is the observation that the topology describing interactions between 259 proteins in the I^n−IX^n protein network fragment enables information transfer between most known biochemical pathways. Hence, this particular network topology is anticipated to be capable of modulating a plethora of different cellular functions.

Drug-Induced Protein Network Topology Models. The large-scale, router-level protein network topology (I^n−IX^n) map (shown in Figure 5B,C) is a working model which describes the average shortest path distance for routing drug-inducible signals for each of the 1320 medicines through the protein network. This "average shortest path distance" topology serves as standard foundation for constructing network topologies induced by individual medicines. This model is used to investigate how perturbations on this network topology through pharmacological means might affect organisms' response to drug treatment. Use of this model for examining cause−effect relationships of medicines assumes that drugs are freely distributed throughout the body's organ systems and that the router-level, signal transduction network topology between different cell types and differentiation stages is similar.

Construction of the protein network topologies that can be induced by an investigated medicine starts with the constructed, standard protein network topology map (I^n−IX^n) and identification of network nodes that can be reached by the investigated medicine. Nodes reachable for a

medicine are identified by a medicine's protein co-occurrence frequency reports and are color coded to reflect the frequency of protein−medicine literature reports. Topology maps unique to each of the 1320 medicines in our investigation can be constructed in this way.

For investigating the origin of the drug-effect similarity between rosiglitazone (**1**) and medicines **2**−**12**, which have similar preclinical and clinical drug-effect information profiles (Figure 1C), protein network topology maps for **1**−**12** were constructed, each identifying the shortest path for routing signals through the network. This construction involves identification of network nodes reachable by medicines **1**−**12** and connection of network reachable nodes in a manner that creates shortest path distances between network-reachable nodes. Structure-effect comparison between medicines **1**−**12** requires comparison of these 12 topologies. Facilitating this comparison is the fact that the construction of individual drug-inducible topologies uses, as standard framework, the router-level connectivity of the signal transduction network, connecting signals induced by 1320 medicines (I^n−IX^n). As illustrated in Figure 5D,E, rosiglitazone (**1**) and glimepiride (**2**), which have different protein targets but very similar drug−protein coinvestigation frequency spectra, reach identical protein network positions and hence use very similar "shortest path distance" protein network topologies for conducting signals through the protein network system. The induction of similar shortest path distance topologies for conducting signals induced by medicines through the network is anticipated to lead to the cross-linking of similar biochemical pathways and concomitant generation of similar cellular output signals. Distribution of these output signals throughout the body would lead to similar drug effects on organ systems. Supporting this premise is the statistically significant correlation between protein network reachability information and *in vivo* drug-effect patterns characterizing medicines **1**−**4** (Figure 1C). This premise also provides a rationale for why rosiglitazone (**1**) and glimepiride (**2**), which have different chemical architectures and different putative mechanisms of actions, have 95% *in vivo* drug-effect profile similarity, since these medicines induce nearly identical shortest path distance topologies for routing respective signals through the protein network (Figures 1A−C and 5D,E).

Discussion

Analysis of broad *in vivo* and *in vitro* structure−function information associated with 1320 medicines suggests that the shortest path distance routing of the drug-induced, information flow through the signal transduction network determines the pharmacology of medicines. Accordingly, the clustering of coinvestigation frequency information between medicines and proteins has been used for creating a working model describing the router-level connectivity of the cellular signal transduction network. By utilizing both protein network reachability information as well as clinical-effect information of medicines, cause−effect relationships of medicines can be evaluated from a protein network perspective. For validating this approach, cause−effect relationships of drugs with different mechanisms of action and molecular architectures were examined. These examinations reveal that medicines produce similar *in vivo* drug effects if they induce similar shortest path distances for transferring information in the signal transduction network, leading to the cross-linking of similar biochemical pathways and the generation of similar cellular output signals (similar effects on the body's organ systems). Moreover, this analysis suggests that similar shortest path distance topologies can be involved in conducting signals originating at different protein network positions. This observation also suggests that the compensatory properties of biological systems, resulting from rerouting of cellular information flows in case of injury, may have their roots in the router-level connectivity of the signal transduction network, enabling the steering of external information along similar, shortest path distances. Hence, the perspective gained from examining medicines **1**−**12** also suggests that "off-target activities" of medicines may be an attribute of the signal transduction network connectivity and not necessarily a medicine's target promiscuity.

In the end, comparison of protein network reachability information in combination with targeted *in vitro* and *in vivo* experimentation may well provide a cost-effective avenue for assessing if experimental medicines will indeed produce clinical effects that differ substantially from those produced by drugs with established mechanisms of action[33,34] and whether expensive research for improving target selectivity of experimental medicines should be undertaken. In addition, analysis of clinical symptoms and drug-effect patterns and associated signal transduction network topologies may provide clues on how the router-level, signal transduction network topology is altered in disease and how this information flow is rerouted through administration of medicines. In this respect, drugs and disease have the same target: the router-level connectivity of the signal transduction network. Understanding how compensatory mechanisms in biological systems work may open up new avenues for the discovery of medicines.

Experimental Section

Medicine−Protein Coinvestigation Frequency Data. More than 5000 biomedical journals containing over 15 million citations from Medline 2006 were scanned for co-occurrence of query compounds (Supporting Information) and proteins (Supporting Information), resulting in 10.9 million compound−protein associations across more than 1 million abstracts.[35] A full matrix of 1320 compounds with coinvestigation counts against 1179 proteins was created using the Python coding language (www.python.org). This data set was then normalized, i.e., wherein all coinvestigation counts >100 were set to 100. Bootstrapping experiments on the coinvestigation matrix were conducted as previously described.[36]

Medicine−COSTART Cocitation Frequency Data. The COSTART cocitation matrix creation was based on previously published text mining work,[12] whereby cocitation was defined as the occurrence of both a compound name and COSTART medical terminology (Supporting Information) within the same Medline abstract. A full matrix of 1320 compounds with cocitation counts against 1082 COSTART terms was created using the Python coding language (www.python.org). This data set was normalized as described above.

Sorting Drug-Effect Spectra. Spotfire Decision Site 8.1 software was used for hierarchical clustering and profile similarity determinations.[9−12]

Constructing Drug-Induced Protein Network Topology Maps. The vertical displayed dendrogram sections derived from the clustering of protein−medicine coinvestigation frequency information identifies clusters (associations) of proteins with unique confidence in cluster similarity values (CCS) (scoring wherein 0 = lowest to 1 = highest). Network fragments were constructed by first connecting the proteins in individual dendrogram sections that are most closely aligned (having the highest CCS values) and reported to be directly connected (corroborated by Ingenuity's direct protein-interaction information). For those

highly associated dendrogram proteins that cannot be directly connected, nearest neighbor (filler) proteins were identified using the Ingenuity platform. Network fragments were generated by starting with the cluster proteins with highest CCS values and adding sufficient neighbor proteins so that all cluster proteins are connected. For example, using the dendrogram-derived protein associations in clusters I−IX (Supporting Information) and the minimum number of proteins required to connect all the dendrogram proteins in I−IX, protein network fragments I^n−IX^n (Supporting Information) were generated.

Acknowledgment. The authors thank Joel Morris and Bert Chenard for stimulating discussions.

Supporting Information Available: A list of the 1320 medicines, a list of the 1179 proteins (Accession Numbers), a list of the 1082 COSTART terms, a list of the proteins (Accession Numbers) in network fragments I−IX, protein network fragments I^n−IX^n, and a list of the 11 medicines with effect/protein spectra similarity to rosiglitazone. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Castiglioni, A. *A History of Medicine*; Alfred A. Knopf: New York, 1958.
(2) Schadt, E. E.; Friend, S. H.; Shaywitz, D. A. A network view of disease and compound screening. *Nat. Rev. Drug Discovery* **2009**, *8*, 286–295.
(3) Petti, A. A.; Church, G. M. A network of transcriptionally coordinated functional modules in *Saccharomyces cerevisiae*. *Genome Res.* **2005**, *15*, 1298–1306.
(4) 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.
(5) Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T. A.; Judson, R. S.; Knight, J. R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; Qureshi-Emili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.; Johnston, M.; Fields, S.; Rothberg, J. M. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **2000**, *403*, 623–627.
(6) Li, S.; Armstrong, C. M.; Bertin, N.; Ge, H.; Milstein, S.; Boxem, M.; Vidalain, P. O.; Han, J. D.; Chesneau, A.; Hao, T.; Goldberg, D. S.; Li, N.; Martinez, M.; Rual, J. F.; Lamesch, P.; Xu, L.; Tewari, M.; Wong, S. L.; Zhang, L. V.; Berriz, G. F.; Jacotot, L.; Vaglio, P.; Reboul, J.; Hirozane-Kishikawa, T.; Li, Q.; Gabel, H. W.; Elewa, A.; Baumgartner, B.; Rose, D. J.; Yu, H.; Bosak, S.; Sequerra, R.; Fraser, A.; Mango, S. E.; Saxton, W. M.; Strome, S.; Van Den Heuvel, S.; Piano, F.; Vandenhaute, J.; Sardet, C.; Gerstein, M.; Doucette-Stamm, L.; Gunsalus, K. C.; Harper, J. W.; Cusick, M. E.; Roth, F. P.; Hill, D. E.; Vidal, M. A map of the interactome network of the metazoan *C. elegans*. *Science* **2004**, *303*, 540–543.
(7) Ambesi-Impiombato, A.; di Bernardo, D. Computational biology and drug discovery: from single-target to network drugs. *Curr. Bioinf.* **2006**, *1*, 3–13.
(8) Ekins, S. Systems-ADME/Tox: Resources and network approaches. *J. Pharmacol. Toxicol. Methods* **2006**, *53*, 38–66.
(9) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biospectra analysis: model proteome characterizations for linking molecular structure and biological response. *J. Med. Chem.* **2005**, *48*, 6918–6925.
(10) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biological spectra analysis: linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 261–266.
(11) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nat. Chem. Biol.* **2005**, *1*, 389–397.
(12) Fliri, A. F.; Loging, W. T.; Volkmann, R. A. Analysis of system structure−function relationships. *ChemMedChem* **2007**, *2*, 1774–1782.
(13) Campillos, M.; Kuhn, M.; Gavin, A.-C.; Jensen, L. J.; Bork, P. Drug target identification using side-effect similarity. *Science* **2008**, *321*, 263–266.
(14) McDaniel, R.; Weiss, R. Advances in synthetic biology: on the path from prototypes to applications. *Curr. Opin. Biotechnol.* **2005**, *16*, 476–483.
(15) Watts, D. J.; Strogatz, S. H. Collective dynamics of "small-world" networks. *Nature* **1998**, *393*, 440–442.
(16) Crutchfield, J. P.; Mitchell, M. The evolution of emergent computation. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 10742–10746.
(17) Lim, J.; Hao, T.; Shaw, C.; Patel, A. J.; Szabó, G.; Rual, J. F.; Fisk, C. J.; Li, N.; Smolyar, A.; Hill, D. E.; Barabási, A. L.; Vidal, M.; Zoghbi, H. Y. A protein−protein interaction network for human inherited ataxias and disorders of purkinje cell degeneration. *Cell* **2006**, *125*, 801–814.
(18) Calvano, S. E.; Xiao, W.; Richards, D. R.; Felciano, R. M.; Baker, H. V.; Cho, R. J.; Chen, R. O.; Brownstein, B. H.; Cobb, J. P.; Tschoeke, S. K.; Miller-Graziano, C.; Moldawer, L. L.; Mindrinos, M. N.; Davis, R. W.; Tompkins, R. G.; Lowry, S. F. A network-based analysis of systemic inflammation in humans. *Nature* **2005**, *437*, 1032–1037.
(19) Sharan, R.; Ulitsky, I.; Shamir, R. Network-based prediction of protein function. *Mol. Syst. Biol.* **2007**, *3*, 88.
(20) Tan, T.; Shlomi, T.; Feizi, H.; Ideker, T.; Sharan, R. Transcriptional regulation of protein complexes within and across species. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1283–1288.
(21) Medina, M. Genomes, phylogeny, and evolutionary systems biology. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (Suppl. 1), 6630–6635.
(22) Kim, P. M.; Lu, J. L.; Xia, Y.; Gerstein, M. B. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **2006**, *314*, 1938–1941.
(23) Clarke, R.; Ressom, H. W.; Wang, A.; Xuan, J.; Liu, M. C.; Gehan, E. A.; Wang, Y. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat. Rev. Cancer* **2008**, *8*, 37–49.
(24) Song, C.; Havlin, S.; Makse, H. A. Self-similarity of complex networks. *Nature* **2005**, *433*, 392–395.
(25) Pansiot, J.-J.; Grad, D. On routes and multicast trees in the internet. *ACM SIGCOMM Comput. Commun. Rev.* **1998**, *28*, 41–50.
(26) Gross, D.; Harris, C. M. *Fundamentals of Queueing Theory*, 3rd ed.; John Wiley & Sons: New York, 1998.
(27) Bestavros, A.; Byers, J. W.; Harfoush, K. A. Inference and labeling of metric-induced network topologies. *IEEE Trans. Parallel Distributed Syst.* **2005**, *16*, 1–13.
(28) Racunas, S.; Shah, N.; Fedoroff, N. V. A contradiction-based framework for testing gene regulation hypotheses. Proceedings of the IEEE Bioinformatics Conference , 2nd, Stanford, CA, Aug 11−14, 2003.
(29) Dunn, R.; Dudbridge, F.; Sanderson, C. M. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinf.* **2005**, *6*, 39.
(30) Becquet, C.; Blachon, S.; Jeudy, B.; Boulicaut, J.-F.; Gandrillon, O. Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. *Genome Biol.* **2002**, *3*, RESEARCH0067.
(31) Rabbat, M. G.; Figueiredo, M. A. T.; Nowak, R. D. Network inference from co-occurrences. *IEEE Trans. Inf. Theory* **2008**, *54*, 4053–4068.
(32) Jensen, L. J.; Saric, J.; Bork, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.* **2006**, *7*, 119–129.
(33) Hopkins, A. L. Network Pharmacology. *Nat. Biotechnol.* **2007**, *25*, 1110–1111.
(34) Yildirim, M. A.; Goh, K.-I.; Cusick, M. E.; Barabasi, A.-L.; Vidal, M. Drug-target network. *Nat. Biotechnol.* **2007**, *25*, 1119–1126.
(35) Srinivasan, P.; Hristovski, D. Distilling conceptual connections from MeSH co-occurrences. *Medinfo* **2004**, *11*, 808–812.
(36) Davison, A. C.; Hinkley, D. V. *Bootstrap methods and their application*; Cambridge University Press: Cambridge, 1997.