SCIENCE CHINA
Life Sciences

THEMATIC ISSUE: Computational life sciences
• REVIEW •

November 2014 Vol.57 No.11: 1103–1114
doi: 10.1007/s11427-014-4757-4

Edge biomarkers for classification and prediction of phenotypes

ZENG Tao[1], ZHANG WanWei[1], YU XiangTian[1,2], LIU XiaoPing[1], LI MeiYi[1], LIU Rui[3] & CHEN LuoNan[1*]

[1]*Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China;*
[2]*School of Mathematics, Shandong University, Jinan 250100, China;*
[3]*School of Mathematics, South China University of Technology, Guangzhou 510640, China*

Received June 9, 2014; accepted August 7, 2014; published online October 14, 2014

In general, a disease manifests not from malfunction of individual molecules but from failure of the relevant system or network, which can be considered as a set of interactions or edges among molecules. Thus, instead of individual molecules, networks or edges are stable forms to reliably characterize complex diseases. This paper reviews both traditional node biomarkers and edge biomarkers, which have been newly proposed. These biomarkers are classified in terms of their contained information. In particular, we show that edge and network biomarkers provide novel ways of stably and reliably diagnosing the disease state of a sample. First, we categorize the biomarkers based on the information used in the learning and prediction steps. We then briefly introduce conventional node biomarkers, or molecular biomarkers without network information, and their computational approaches. The main focus of this paper is edge and network biomarkers, which exploit network information to improve the accuracy of diagnosis and prognosis. Moreover, by extracting both network and dynamic information from the data, we can develop dynamical network and edge biomarkers. These biomarkers not only diagnose the immediate pre-disease state but also detect the critical molecules or networks by which the biological system progresses from the healthy to the disease state. The identified critical molecules can be used as drug targets, and the critical state indicates the critical point of disease control. The paper also discusses representative biomarker-based methods.

biomarker, edge biomarker, dynamical network biomarker, classification, prediction, phenotype, disease diagnosis, disease prognosis

Citation: Zeng T, Zhang WW, Yu XT, Liu XP, Li MY, Liu R, Chen LN. Edge biomarkers for classification and prediction of phenotypes. Sci China Life Sci, 2014, 57: 1103−1114, doi: 10.1007/s11427-014-4757-4

Most complex diseases are caused by multiple factors, which must be investigated in a systematic and dynamical manner [1–3]. From a systems biology viewpoint, a disease results not from the malfunction of individual molecules but from failure of the relevant system or network, which can be considered as a set of interactions or edges among molecules. Thus, complex diseases can be more reliably characterized by networks or edges than by single molecules (i.e., genes, proteins, metabolites and other measurable biological elements). Network features can be used to indicate the state of a disease (diagnosis), estimate the effect of treatment or predict the survival time (prognosis). Networks and edges of biomolecules also play important roles in the phenotypic variations among plants and animals, which can be similarly used for the classification and prediction of phenotypes in evolutionary studies.

Traditionally, researchers have used genetic or biomolecular biomarkers, also known as gene signatures or molecular biomarkers respectively, to distinguish particular biological phenotypes (e.g., states of complex diseases). In conventional clinical study, molecular biomarkers are

*Corresponding author (email: lnchen@sibs.ac.cn) widely used for (i) diagnosing a disease state and (ii) predicting the outcome of a disease state (prognosis). The growing popularity of "P4 medicine", i.e., predictive, preventive, personalized and participatory medicine has diverted our attention from passive treatment of disease to active prevention [2]. To reach the lofty targets of modern biomedical studies, early-warning signals should be detected and disease states should be predicted before the occurrence or serious deterioration of disease. Such early intervention can halt the disease before it progresses to an irreversible state. Obviously, conventional molecular biomarkers are not directly applicable to "P4 medicine" because they reflect biological elements rather than the predictors/causes of a disease. On the other hand, biological functions and signal transductions are facilitated by interactions (regulations) between molecules, which constitute the edges in biological networks. Thus, P4 medicine relies on the networks and interactions among biological elements (such as genes, proteins and patients) rather than the biological elements themselves. Network analysis captures previously unobserved features in both the network (edges) and dynamics. Therefore, as our theoretical and clinical thinking has advanced, biomarkers have evolved from single molecules (e.g., individual genes) to multiple molecules (e.g., gene sets), associated molecules (e.g., molecular networks) and finally to dynamical interactive molecules (e.g., dynamical molecule networks). From a network perspective, biomarkers are classified as node biomarkers, network-based biomarkers, network-weighted biomarkers, network (edge) biomarkers, and dynamical network biomarkers (DNBs) or dynamical edge biomarkers, as shown in Table 1. In particular, as components of general network biomarkers [4], edge biomarkers exploit information of interactions or associations among molecules, rather than individual molecules. Many researchers have investigated the influence of perturbations on interaction networks. Perturbations fall into two categories; loss of gene products ('node removal', the removal of a node from the interaction network), and loss of protein or gene interactions ('edge removal', the removal of an edge from the interaction network). Recently, 'edgetic' disease has been identified as the result of an interaction removal (edge removal) rather than a gene removal (node removal) [5]. On the other hand, our recent EdgeMarker analysis [6] demonstrated that non-differentially expressed genes, which are traditionally ignored, can be as informative as differentially expressed genes for classifying biological conditions and sample phenotypes. All of the above studies indicate that edge biomarkers can provide new insights into the pathogenesis of complex diseases at the network level. The strength of an edge between two molecules is frequently determined by the molecules' co-expression, measured by the Pearson's correlation coefficient (PCC) [6].

Next, we give a comprehensive review of the traditional node biomarker and the contemporary edge biomarker, and demonstrate the importance of edge biomarkers in translational biomedicine study. In Section 1, we first classify biomarkers based on the network information used in the learning and predicting steps. Section 2 briefly introduces conventional node biomarkers without network information, while Section 3 describes the biological background and motivation of node biomarkers with network information in the learning procedure. Section 4 presents the major topic of this paper: the use of edge biomarkers with network information in both learning and predicting procedures. Particular emphasis is placed on dynamical edge biomarkers, which exploit both network and dynamical information, and thus can detect pre-disease states that are missed by traditional biomarkers. Sections 2, 3 and 4 are accompanied by the computational approaches used to analyze the respective biomarkers. The paper concludes with several general remarks on biomarkers.

1 Categories of biomarkers

Biomarkers can be identified from observed samples or data (in a machine learning context) by a two-step process of learning and predicting. In the learning step, effective marker molecules that distinguish different phenotypes in samples are identified. Phenotypic examples are control and case samples, or normal and disease samples. The prediction step decides the candidate phenotype of a test sample based on the identified marker molecules. Depending on the data used for learning and predicting, biomarkers [4] are grouped into several broad categories as shown in Table 1. Besides traditional node (molecular) biomarkers, these categories include network-based biomarkers and network-weighted biomarkers, which use molecular network to identify sub-networks or edges from sample data in the learning step, but use only those molecules (or molecule set) related to the edges or sub-networks in the prediction step. Although sub-networks and edges are identified by network or correlation information, the edge or network information is excluded from the phenotypic prediction; thus, network-based and network-weighted biomarkers are essentially node biomarkers. In contrast, edge biomarkers, which include network biomarkers and dynamical network biomarkers (DNBs), use network or edge information in both learning and predicting steps. As such, they form networks of marker molecules and are essentially different from node biomarkers. Rather than merely diagnosing a disease state, DNBs exploit dynamical information in the data [7–11], and thus can detect pre-disease states (or early-warning disease signals), which cannot be determined by traditional biomarkers. Note that the pre-disease state can be viewed as a special normal state immediately before the major deterioration or critical transition of a disease [4]. In Table 1, biomarkers are also categorized by whether one or multiple samples are used in the predicting step.

Table 1 Categories of biomarkers

Biomarker categories	Input data in learning step	Output data in learning step	Input data in predicting step	Output data in predicting step	Type of biomarker
Molecular bio- marker, or gene-set biomarker	Omics data for samples of two groups, e.g. gene expressions, protein expressions	Identifying differential expression molecules (marker molecules) for classification of two groups	Marker molecules with their expressions for one test sample	Normal or disease state (control or case)	Node biomarker
Network-based biomarker, or gene-set biomarker	Omics and interactome data, e.g. gene expressions and Protein-Protein Interactions (PPIs) or other network	Identifying differential expression molecules (marker molecules) constrained by static network structure	Marker molecules with their expressions for one test sample		Node biomarker
Network-weighted biomarker-I (static network)		Identifying differential expression molecules (marker molecules) or edges constrained by static network structure	Marker molecules with their expressions weighted by network topology for one test sample		Node biomarker
Network-weighted biomarker-II (correlation network)		Identifying differential expression molecules (marker molecules) or edges constrained by correlation network	Marker molecules with their expressions weighted by network topology for one test sample		Node biomarker
Network biomarker-I, or edge biomarker-I		Identifying differential expression edges (marker edges) constrained by correlation network	Marker edges with their correlation information for multiple test samples		Edge biomarker
Network biomarker-II, or edge biomarker-II		Identifying differential expression edges (marker edges) constrained by correlation network	Marker edges with their correlation-like information for one test sample		Edge biomarker
Dynamical network biomarker (DNB)-I or dynamical edge biomarker-I		Identifying differential expression edges (marker edges) constrained by network, based on deviation, covariance, and distribution of their molecule expressions	Marker edges with deviation, covariance of their molecule expressions for a number of test samples	Normal or pre-disease state (control or case)	Edge biomarker
Dynamical network biomarker (DNB)-II or dynamical edge biomarker-II		Identifying differential expression edges (marker edges) constrained by network, based on deviation, covariance, and distribution of their molecule expressions	Marker edges with deviation, covariance of their molecule expressions and their distributions for one test sample		Edge biomarker

In contrast to node biomarkers, which are typically represented by differential expressions (i.e., the first-order statistics) of individual molecules or a molecular set, edge biomarkers exploit the association or correlation information among molecules to predict the phenotype of a test sample. These associations are represented by differential correlations and differential deviations (i.e., the second-order statistics) of molecules. The additional information provided by correlations or interactions not only improves the accuracy of phenotype prediction, but also reveals the biological or pathogenic mechanism underlying the marker molecules (e.g., the driver genes). Furthermore, by including the dynamic information of data, dynamical network biomarkers (or dynamical edge biomarkers) can diagnose a "pre-disease" or "un-occurred disease" state at an early stage. This concept is not new; it has been mentioned in *Yellow Emperor's Medicine* (one of the earliest books on traditional Chinese medicine) 2000 years ago.

2 Node biomarkers for classification and prediction without network information

Node (or molecular) biomarkers have been widely studied, especially for detecting and combating complex diseases. For instance, (i) the tumor suppressor genes BRCA1 and BRCA2, mainly involved in DNA damage repair and transcriptional regulation, are well-known causal genes of breast, ovarian and similar cancers [12]. The chromosomal stability of these genes is altered by mutations. Somatic alterations weaken the genes' suppressor ability, loosening their control on proliferation and ultimately leading to tumor occurrence and development. (ii) Pre-diabetic individuals show significant metabolic variations of a few important metabolites [13], which are expected to become novel markers for distinguishing impaired glucose tolerant (IGT) individuals from those with normal glucose tolerance. If pre-diabetic individuals with impaired glucose tolerance are treated early, their metabolisms can recover to the normal state, avoiding the development of Type 2 diabetes (T2D). Thus, this small number of IGT-specific metabolites and their associated T2D-related genes/proteins could be newly exploited in T2D prevention. (iii) In the case of pain, neurological and psychiatric diseases, the acid-sensing ion channels (ASICs) play important roles in the central and peripheral nervous systems [14]. Investigations of these ASICs have revealed the molecular mechanism of extracellular acid sensing by neurons. Potential inhibiting or potentiating ASICs cause physiological and behavioral disorders, leading to disease occurrence and development. Collectively, these studies have shown that one or a few molecules (e.g., master regulators) are important molecular indicators of the establishment and progression of complex diseases.

In addition to conventional low-throughput technologies, high-throughput approaches (such as microarrays) have provided big data for identifying molecular biomarkers, enabling cheap effective methods for predicting human disease risk. At the level of conventional node biomarkers, these methods demonstrate the effectiveness of classifying samples by individual molecules at different scales, i.e., single genes and sets of genes. Several single genes have been already identified as disease biomarkers, especially in cancers. An example is the IL28B gene implicated in liver cancer [15,16]. More importantly, the driver mutations of genes have been identified in a genome-wide association study [17] using next-generation sequencing technology [18]. Genes inducing pathogenic physical changes are usually drivers rather than passengers of disease consequents [19]. However, gene sets [20] are known to improve the classification accuracy of marking complex disease phenotypes, for the following reason. Unlike hereditary diseases, which always occur by mutations on a few genes, cancer and metabolic diseases (and many other complex diseases) [21–23] are usually caused by numerous associated genes. There are different strategies for predicting disease states from sets of marker genes. First, the normal and disease states can be distinguished by differential expression of the identified marker genes [24]. Second, a common discriminative gene group can be mined from multiple datasets [25], or multiple classifiers can be integrated from sets of marker genes [26,27]. These approaches enhance the robustness or consensus of sets of expression-dependent marker genes in heterogeneous disease cohorts. Finally, disease phenotypes can be characterized by differential enrichment rather than by differential expression [28–31]; the so-called enrichment-based approach or activity-based method [31,32]. According to this approach, pathogenic genotypes are characterized by significant extent and number of differentially regulated genes in a particular gene set.

Although conventional node biomarkers have greatly advanced the study of disease diagnosis, several formidable problems remain, including improved diagnostic accuracy, early diagnosis before the disease occurs or clinical symptoms appear, and therapy prediction. While these problems have been identified in conventional node biomarker studies, their solution requires biomarkers that integrate network information and even dynamical information, such as (dynamical) edge biomarkers. Such biomarkers have been extensively proposed and investigated in recent years [4,33,34], and are described in the next sections.

3 Node biomarkers for classification and prediction with network information limited to the learning step

Biomarkers that integrate network information have received much attention in recent translational biomedical research. Many studies have shown that signaling pathways, protein complexes, and sub-networks have more discriminative power for identifying disease phenotypes than individual gene and protein expression. For instance, condition-responsive genes (CORGs) integrate pathway information or biological network into a new measurement of gene signatures [32], providing more robust indicators for characterizing disease samples. Similarly, a newly proposed multi-pathway-based method called Pathifier [35] transforms gene-level information into pathway-level information by a pathway deregulation score. Pathifier has revealed highly reproducible, well-preserved and significant biological features of complex diseases. Many researchers have explored the causal associations between gene pairs and phenotypes to understand how perturbations influence interactome networks. As mentioned in the Introduction, perturbations are of two types; loss of gene products (‘node removal’, or removal of a node from the interactome network), and loss of gene interactions (‘edge removal’, or removal of an edge from the interactome network). Interaction (edge) removals, rather than gene (node) removals, are responsible for the so-called ‘edgetic’ diseases [5]. Thus, differentially expressed interactions or networks can provide more details about human pathogenic states and realize better molecular therapeutic strategies than conventional differentially expressed genes.

Because network-associated biomarkers integrate differential gene expression and differential expression correlations among genes, they characterize diseases in a reliable manner. Depending on the input and output data included in the learning and predicting steps (Table 1), biomarkers with network information are roughly divided into network-based biomarkers, network-weighted biomarkers, network (or edge) biomarkers, and DNBs or dynamical edge biomarkers. Unlike other types of biomarkers, edge biomarkers use network information to enhance the power of marker genes and their associations in both the learning and prediction steps.

Network-based biomarkers use the network information only for selecting a set of molecules (or marker molecules) from a static background network in the learning step.

These marker molecules and their expressions are then used to identify the phenotype (e.g., normal or disease) of a test sample in the prediction step. Similarly, network-weighted biomarkers select marker molecules based on the network information. These markers, together with their expressions, are then weighted by the network topology before being input to the prediction step for phenotypic identification. Clearly, although network information (e.g., correlation) is used in the learning step, it is disregarded in the prediction step. Therefore, both network-based and network-weighted biomarkers are essentially node biomarkers. However, network information is difficult to apply on a single test sample in the prediction step because, unlike gene expression, pairs of molecules or edges cannot be easily correlated in a single sample.

On the other hand, network and edge biomarkers use static network information (e.g., background protein interaction networks) or dynamical network information (e.g., co-expression networks) to identify integrative marker molecules and their marker molecule-pairs (or marker edges). These marker edges are then used in both learning and predicting steps. In particular, since DNBs or dynamical edge biomarkers extend the dynamical information to the prediction step, they can detect the pre-disease (or pre-transition) state immediately prior to disease occurrence. Clearly, because they exploit edges or networks, these biomarkers can exist as simple network or edge biomarkers or can form a network of biomarkers. In addition, the difficulty of representing an edge or network using a single sample has been recently overcome [6], which greatly improves the feasibility of edge biomarkers in clinical practice. Representative biomarkers with network information are summarized in Table 2. Network-associated biomarkers are detailed in the following subsections.

3.1 Network-based biomarkers

Chuang et al. [36] proposed the PinnacleZ approach for identifying so-called network markers. Network markers are not individual genes but sub-networks extracted from a protein interaction network. In PinnacleZ, the activity of each sub-network is defined as the regularized mean expression of the genes within it, and cancers are classified by the module activity profile rather than by the gene expression profile. Dao et al. [37] adopted to identify discriminative sub-network markers by a color coding technique implemented in a network-based classification algorithm called OptDis. In this approach, features are sub-network markers rather than single genes, and feature values (marker expressions) are the average expression levels of genes in the sub-network. He et al. [38] proposed a network-based approach that identifies dysfunctional modules (DM) in context-specific protein-protein interaction networks, and classifies phenotypes by the aggregated expression activity of gene modules (e.g., the regularized mean expression of genes in a module). Jin et al. [40] developed an integrative pipeline for biomarker discovery (IPBD) that combines disease information and expression profiles for proteins and genes with protein-protein interactions. Indeed, Jin et al.'s paper coined the term "network biomarker". For each sample, the intensity (expression) of a network biomarker is the weighted sum of intensities of genes in the network, where the weights are the *P*-values of the differential expression of each gene. Winter et al. [41] implemented a network-based approach called NetRank, which is similar to Google's PageRank. This method ranks genes by their prognostic relevance based on both expression and network information. The expression levels of the identified predictive marker genes are directly learned by a support vector machine, and the prognoses of tumor samples are categorized as poor or good. Eddy et al. [39] designed Differential Rank Conservation (DIRAC), a rank-based algorithm that considers the combinatorial effect of gene interactions within a network (the pathways) and evaluate the differential network rankings by quantitative measures such as rank difference scores. It should be noted that these quantitative measures are based on the ordering of gene expressions within the network but disregard the topological structure of the network. Obviously, all of the above methods extract discriminative gene sets constrained by the prior network in the learning step, but exclude the network or edge information in the prediction step. Therefore, these biomarkers are essentially node biomarkers, at least from the viewpoint of biomarker application.

3.2 Network-weighted biomarkers

Alternatively, marker molecules can be identified from the prior network (with static network topology) in the learning step (see Class I network-weighted biomarkers in Table 2). The CORGs based classification method proposed by Lee et al. [32] infers the activity level of a given pathway by summarizing the gene expression levels of the CORGs. Although Lee et al. do not consider the topological structure of the pathway, this information could be accommodated in the method. The DART (denoising algorithm based on relevance network topology) algorithm, recently designed by Jiao et al. [42], improves pathway activity estimates by filtering out noise in advance. The algorithm first builds a pruned expression relevance network (as a co-expression network of genes), in which the noiseless pathway activity is computed as a metric on the largest connected component (e.g., the maximal connected sub-network of relevance network). Similar to the previous sub-network expression metric [32], the activity metric may be sign-weighted, which registers whether the genes involved in pathway regulation are up- or down-regulated, or topologically weighted, which considers the number of neighbors of a single gene on the relevance network. Recently, Wen et al. [43] developed a network-based approach that identifies the putative

Table 2 Examples of biomarkers with network information

Methods	Biomarker categories	Biomarkers identified in learning step	Indicator of biomarkers in predicting step
PinnacleZ [36]	Network-based biomarker	Gene sub-network greedily identified in background network	Activity of interactive genes
OptDis [37]	Network-based biomarker	Optimally discriminative sub-network identified by color-coding technique	Activity of interactive genes
DIRAC [39]	Network-based biomarker	Variably expressed networks among phenotypes identified by rank difference score	Rank difference score according to the ordering of expressions of genes
IPBD [40]	Network-based biomarker	Single-, paired-, triple-, square biomarkers selected by SVM	Significance P-value weighted sum of expression intensity
NetRank [41]	Network-based biomarker	Predictive marker genes identified by network-based approach, e.g., Google's PageRank	The expression level of each marker gene, and no kind of aggregation was used
DM [38]	Network-based biomarker	Dysfunctional modules identified based on information flow and mutual information	Aggregated expression activity of interactive genes in module
CORGs [32]	Network-weighted biomarker-I	A new classification method based on condition-responsive genes	Summarizing the gene expression levels of pathway's CORGs
DART [42]	Network-weighted biomarker-I	Noise-less pathway activity as a metric on the largest connected component of pruned expression relevance network	Weighted expression activity
PCM [43]	Network-weighted biomarker-I	Putative causal module identified by MCL based on PPI and integrating epigenomic data, gene expression data	Activity as degree-weighted mean expression of genes in one module
NBS [44]	Network-weighted biomarker-I	Subtypes identified by a network-based stratification method to integrate tumor somatic mutations with gene networks	'Network-smoothed' profiles or mRNA expression signature
NGF [45]	Network-weighted biomarker-I	Logic functions (e.g., decision trees) as sub-networks in prior network identified by Network-Guided Forests	Expressions of genes in identified sub-networks and their decision values
CDN [46]	Network-weighted biomarker-II	Module biomarker as a core of differential networks corresponding to different cancer developmental stages	Expression of marker genes
DHP [47]	Network-weighted biomarker-II	Inter-modular hub and intra-modular hub proteins investigated in the changed interactome and their predictive relation with patient outcome	Expression differences of significant hubs and their interactors
CRV [48]	Network-weighted biomarker-II	Carcinogenesis relevance proteins identified by protein association models and protein association networks mapping to normal and disease	Mapping errors of gene expressions on two protein association networks
CTN [49]	Network biomarker-I or edge biomarker-I	Extracting patient-specific temporal gene biclusters	Similarity score based on bicluster similarity and its PPIScore
EdgeMarker [6]	Network biomarkers-II or edge biomarker-II	Extract discriminative information from non-differentially expressed genes by differentially correlated gene pairs	Co-expression values of differentially correlated gene pairs
DNB [7–11]	Dynamical network biomarker-I	Pre-disease state and critical point detected by dynamical network biomarker on multiple samples	DNB index determined by high expression variance, high intra-correlation and low inter-correlation
ENA [50]	Dynamical network biomarker-I	Network of molecular pairs rebuilt based on higher-order statistics information among molecules	Predictive index determined by high expression variance, high intra-correlation and low inter-correlation
DNB-S [10]	Dynamical network biomarker-II	Pre-disease state and critical point detected and predicted by dynamical network biomarker for one test sample	DNB-S index determined by differential expression distributions compared to a group of control samples

causal module (PCM) biomarkers of complex diseases by integrating epigenomic data, gene expression data, and protein–protein interaction networks. In this method, the modules are extracted from the protein interaction network by a Markov clustering algorithm (MCL), and their activities are defined as the degree-weighted mean expression of genes in each module. Dutkowski et al. used the network-guided forests (NGF) algorithm to identify logic functions (e.g., decision trees) with the same topological structure as the sub-networks in a prior network [45]. NGF is expected to connect the activity of each predictive module to the activity of its component genes. Hofree et al. [44] implemented a novel network-based stratification method (NBS) that integrates tumor somatic mutations with gene networks. In several steps, this method builds a robust stratification of patients (subtypes). First, it prepares the binary mutation profiles, then projects these profiles onto a human gene interaction network. A 'network-smoothed' profile is generated by network propagation and entered into the NetNMF algorithm, which clusters patients into subtypes. Patient assignments are then consolidated by consensus clustering. This method also classifies mutation-derived subtypes, based on which an mRNA expression signature can also be applied for subtype prediction.

Other methods identify molecular markers from extracted discriminative gene sets constrained by the correlation network in the learning step (see Class II network-weighted biomarkers in Table 2). Taylor et al. [47] investigated changes in the organization of the interactome and their predictive relations with patient outcomes. They identified tissue-specific inter-modular hub proteins and tissue-common intra-modular hub proteins. Each hub protein and its co-expressed interacting partners constitute one module (DHP: dynamical hub and partner). Individual patients are assessed not by their gene expressions, but by their differing expressions of significant hubs and their interactors. Meanwhile, using protein association models, Wang et al. [48] constructed a network-based biomarker (CRV) that identifies proteins implicated in carcinogenesis and performs diagnostic evaluation. The CRV provides two protein association networks corresponding to the normal and disease states. In particular, a new sample can be classified as normal or diseased based on the mapping errors of its gene expressions onto the two protein association networks. Liu et al. [46] presented a differential network based approach (CDN: core of differential networks) for identifying disease genes and dysfunctional sub-networks during multi-stage cancer progression. In CDN, the module biomarkers of disease risk are sets of genes (where risk is evaluated from overlapping genes among differential networks corresponding to different cancer developmental stages). However, similarly to their network-based counterparts, network-weighted biomarker methods extract the discriminative gene sets constrained by the network in the learning step, but ignore the network or edge information in the prediction step. Thus, weighted biomarkers are also essentially node biomarkers. In fact, network-weighted biomarkers and follow-up network (or edge) biomarkers are very similar; the main difference is that network-weighted biomarkers (such as DHP [47] and CRV [48]) recognize discriminative molecules (nodes) in the learning step, whereas network biomarkers identify discriminative molecule-pairs (edges).

We reiterate that all of the above network-based and network-weighted biomarkers are node biomarkers rather than edge or network biomarkers. Both categories examine the expressions of individual marker molecules in the prediction step. Network-weighted biomarkers must also estimate the expression weights in the learning step, since the weights must balance the trade-off between the learning and prediction accuracies. Network-weighted biomarkers-I (of Class I in Table 2) and network-weighted biomarkers-II (of Class II in Table 2) differ chiefly in that the former uses a static network structure while the latter uses a dynamical network structure. Thus, Class II network-weighted biomarkers can be expected to not only identify discriminative molecules as disease markers but also reveal condition-specific interactive maps among molecules, by which researchers could unravel disease mechanisms.

4 Edge biomarkers for classification and prediction with network information in both learning and predicting steps

The main advantage of edge biomarkers (or dynamical edge biomarkers) is their inclusion of network information in the prediction step. Note that predicting human phenotype generally requires several samples for calculating the correlations among nodes/genes (e.g., Pearson's correlation coefficient) or edge/gene pairs (higher-order statistics or correlations). In clinical practice, however, diagnosis or prognosis is evaluated on individuals. Constructing and interpreting correlation-like information in the prediction step of phenotypic diagnosis is difficult when very few test samples are available. To address this problem, several studies have investigated the edge biomarkers of single samples. Mathematical and computational approaches for phenotype prediction are broadly classifiable into multi-sample-based approaches (if multiple samples are available for use in the prediction step) [50] and single-sample-based approaches (if a single sample is available) [10]. Note that a group of nodes constitutes a node set, but a group of edges is a network. This implies that edge biomarkers are equivalent to biomarker networks. Representative analysis methods for network and edge biomarkers and their dynamical equivalents are discussed in the following subsections.

4.1 Network and edge biomarkers

Qian et al. [49] proposed a multi-sample-based approach for analyzing Class I network and edge biomarkers. Their method, called CTN (classifying time series gene expression via integration of biological networks) improves the classification and realizes reliable and sound predictions [49]. CTN hybridizes a hidden Markov model (HMM) and Gaussian mixture models (GMMs) to explore the time-dependence of the expression data, and thereby improve the prediction results. The learning step first infers the gene states by the HMM/GMM hybrid model, which converts the original gene expression level into a discrete gene state. Gene biclusters of each patient are then extracted from the temporal gene state matrix by a QL-biclustering algorithm, based on the suffix string and longest common prefix extracts. Each bicluster is assigned a PPI score based on the connection of its contained genes within the protein–protein interaction network (PPI). In the prediction step, the phenotype of each test patient is predicted by PPI-SVM-KNN (where KNN denotes *k*-nearest neighbors algorithm) according to the bicluster similarity (Jaccard index) of the patient and the PPI score of the bicluster. In experiments, learning of the early-stage data improved the performance of the later-stage phenotype prediction.

The EdgeMaker approach of Zhang et al. [6] is a single-sample-based approach for Class II network and edge biomarkers. EdgeMaker identifies edge biomarkers (differentially correlated gene pairs) by representing the edges in a new vector format. This method optimizes the classification ability and overcomes the difficulty of using edge information from a single sample in the prediction step, because each edge is represented as a correlation-like vector. EdgeMarker can predict or diagnose the phenotype of individual samples from the edge information, which is a major advantage of the method. In particular, theoretical and computational experiments show that the obtained edge biomarkers accurately distinguish phenotypes even when their genes are not differentially expressed. The results also support that non-differentially expressed genes, which are usually ignored by traditional methods, can be as informative as differentially expressed genes when assigning a biological condition or phenotype to a sample [6]. This finding provides new insight into the pathogenesis of complex diseases.

4.2 Dynamical network biomarkers and dynamical edge biomarkers

Chen et al. [7] developed a multi-sample-based approach for Class I dynamical network and edge biomarkers. This approach fully utilizes the second-order statistics, i.e., the expression deviations and inter-molecular correlations, to detect pre-disease rather than conventional disease states [7]. Here, the pre-disease state is defined as the normal state of an individual immediately before critical transition to the disease state, i.e., the limit of the normal state. Traditional node and edge biomarkers can distinguish between disease and normal states, but usually cannot diagnose the pre-disease state [7] because the molecular expressions of normal and early disease states are not significantly different. DNBs do distinguish between the pre-disease and normal states because they detect the early-warning signals of complex diseases regardless of sample size [7,8]. Identification of DNB molecules (a group of molecules) is based on the following theory: as the system approaches the pre-disease state or the critical transition, (i) the expression of the DNB molecules dramatically deviates from that of the normal state; (ii) the expression correlation between any two DNB molecules increases; (iii) the expression correlation between any molecule in the DNB and any molecule in the non-DNB decreases. Unlike the critical slowing-down theory [51], DNB theory does not require a large number of samples from each individual, and detects the early-warning signals of critical transitions directly from the high-dimensional data. Thus DNB is a model-free approach for observing big data.

Inspired by the DNB, Yu et al. [50] proposed an edge network that exploits the higher-order statistical information among molecules. In this approach, the traditional node-network representing the first-order statistics distinguishes the disease from the normal state, whereas the edge-network representing the higher-order statistics (ENA: edge-network analysis) distinguishes the pre-disease state from the normal state. At the molecular level, a biological system can be described by a stochastic dynamics model comprising a master equation or stochastic differential equations [52]. By linearizing the system and assuming a Gaussian distribution of the components, the system can be exactly expressed by two sets of equations; one evolving the mean vector of molecules (first-order statistical representation), and the other evolving the covariance matrix of molecules (second-order statistical representation). If the equations involving the covariance matrix are omitted from this model, the model reduces to the traditional node (molecular) networks (such as gene and protein interaction networks) [50], which cannot represent the stochastic dynamics of the original system. In other words, a node-network represents a biological system in the absence of stochastic fluctuation or noise. By contrast, since an edge-network reflects the second-order statistical information of a dynamical system, it captures the stochastic dynamics of the original biological system as well as the node network (assuming that the expression levels of each molecule are Gaussian-distributed). Theoretically, the first-order statistical information (node network) distinguishes between diseased and normal samples by identifying molecular biomarkers (e.g., node biomarkers). The higher-order statistical information (edge-network) additionally distinguishes between pre-diseased and normal samples by identifying edge biomarkers (such as dynamical network biomarkers) [50]. This additional information is crucial for early diagnosis or prediction of a disease.

A single-sample-based approach for Class II dynamical network and edge biomarkers is the DNB-S scoring method of Liu et al. [10]. This method was developed to identify the pre-disease states of single samples (given a group of normal control samples) by exploring the distributions of differential expressions of the pre-detected DNB and non-DNB molecules [10]. Intuitively, the expressions of DNB molecules should display a double-peak (bimodal) distribution in the pre-disease state but a single-peak distribution in the normal state. By contrast, the expressions of non-DNB molecules display a single-peak distribution in both normal and pre-disease states. Thus, if multiple control samples are available, the differential expression distributions between the pre-detected DNB molecules and non-DNB molecules can be estimated for any single sample. These differential distributions, recorded as the DNB-S score, can predict the pre-disease state of a test sample. When a biological system progresses from the normal to the pre-disease state, the DNB-S score alters as follows: (i) The K–L divergence of the test and control samples with regard to molecular DNB expression increases, because the expression distributions of the DNB molecules widely differ between the two samples; (ii) The K–L divergence regarding the expression distributions of the DNB and non-DNB molecules increases, because the differential distributions of the DNB molecules differ from those of non-DNB molecules in the pre-disease test sample; (iii) The K–L divergence of the test and control samples regarding the expression distributions of non-DNB molecules exhibits no significant change, because the expressions of non-DNB molecules are unaltered in the normal and pre-disease states. Collectively, these behaviors imply that if the DNB-S score of a test sample is much higher than a trained threshold, this sample is in the pre-disease state. Like its DNB predecessor, the DNB-S scoring method is a model-free approach for identifying an at-risk individual before the disease appears or seriously degrades the individual's health. Furthermore, the DNB-S scoring method fully utilizes high-dimensional data information to compensate for insufficient samples, and thus detects the pre-disease state from a single sample based on high-throughput data.

5 Conclusion

The "big data" era [53–58] has provided exciting opportunities for biomarker study and "P4" medicine. To achieve predictive, preventive, personalized and participatory medicine [2], we must identify accurate and reliable biomarkers for classifying and predicting the state of a test sample. This paper has reviewed the features and computational methods of various biomarkers from static nodes to dynamical edges biomarkers. Below, we highlight the special considerations of edge biomarker study.

(i) Many traditional approaches identify the discriminative networks or edges by network-based or edge-based methods. Test samples are then classified based on the genes or proteins (or gene set) involved in the identified networks or edges, rather than on the networks or edges (or their correlations) themselves. Such approaches are essentially node-biomarker based methods because the networks or edges are used only to identify discriminative genes or node-biomarkers. Zhang et al. [6] developed a novel method that represents edges in a correlation-like vector format. Although this approach identifies the discriminatively differentially correlated gene pairs as edge biomarkers, and thereby predicts the phenotype of individual test samples, more effective methods that exploit edges and their correlations are required for determining normal or diseased states of single samples.

(ii) In contrast to the first order statistics (i.e., average gene expressions) used in traditional networks or biomarkers, higher-order statistics (e.g., deviations and covariances of gene pairs as second order statistics, skewness as a third-order statistic, and kurtosis as a fourth-order statistic) provide rich information of the original systems. This additional information is excluded in traditional network analysis. In theory, the original stochastic dynamics can be fully recovered by including sufficiently high-order statistics. The edge network developed in [50] represents network nodes by their second-order statistics, and thereby significantly improves disease prediction. High-order statistical information is expected to become widely explored in future biomarker study.

(iii) Both the network and its dynamics are important facets of living organisms, and collectively characterize the states of biological systems in a reliable and stable manner. Thus, how to incorporate dynamical information into biomarkers is a key future direction. DNB theory, which merges nonlinear dynamics and statistics besides exploring the network information, detects early-warning signals of critical transitions to the disease state [7]. Although DNB can in principle determine the critical state and the leading molecules prior to drastic transition from the observed data alone, eliminating the large amount of noise and developing an efficient algorithm for accurately detecting the DNB molecules is required in future.

(iv) In addition to the "bad" molecules (e.g., disease genes), we are also interested in biomarker identification of "good" molecules (e.g., health and longevity genes) and human networks. Quantifying "health" and "wellness" states of humans is an important task requiring both network and dynamical information. This problem is closely related to resilience and robustness evaluation in systems, and can be measured by DNB theory [7]. The concept is illustrated in Figure 1. Health eventually reaches a critical state due to gradual changes in internal or external factors, which can be measured by DNB if big dynamical data are available (Figure 1A and C). By detecting critical factors (including "good" molecules) and critical networks (i.e., DNB) and appropriately attenuating the "bad" molecules while enhancing the "good" molecules, we can expect to significantly improve human health (that is, we can significantly delay the critical state). These scenarios are illustrated in Figure 1A→B and C→D. In particular, DNB can detect the leading molecules and network by which the entire system begins to progress from one state (e.g., health state) to another (e.g., unhealthy state). Thus, the drastic deterioration of an individual's health can be prevented by appropriately treating or perturbing the leading molecules [7]. In other words, health and wellness can be encouraged by enhancing the "good" genes before the critical transition to disease, rather than applying treatment once the disease has appeared (Figure 1). Such a strategy of treating "unoccurred" disease (equivalent to the pre-disease or critical state) is endorsed in the ancient Chinese text *Yellow Emperor's Medicine*, which states that the best doctor treats un-occurred disease, the better doctor treats occurring disease, and the inferior doctor treats occurred disease. By offering a means of detecting incipient disease, dynamical network biomarkers may significantly benefit preventative medicine and health welfare.

Similar to disease study, networks or edges could be exploited in the evolutionary study of plants and animals,

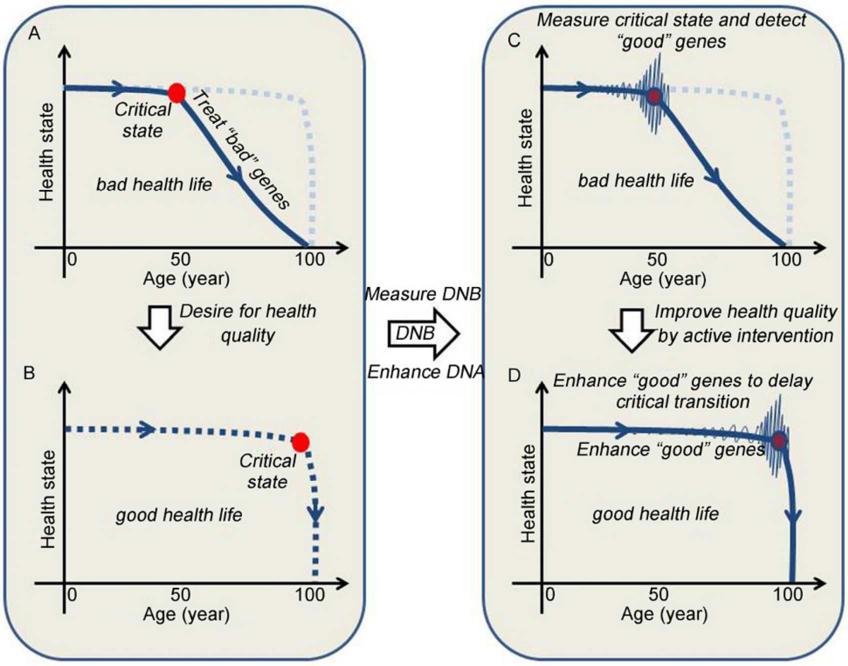

Figure 1 (color online) Quantifying human health conditions and "good" genes and networks by dynamical network biomarkers. DNB can quantify the health or wellness state of a person by measuring its "distance" from the limit or critical state based on collective fluctuations of big data. The critical state of a healthy person is eventually reached by gradual changes in internal or external factors (A), which can be measured by DNB analysis of big dynamical data (C). In contrast to traditional passive treatment of "bad" molecules after disease occurrence, human health could be improved by detecting and enhancing crucial health factors (including "good" molecules) and networks (note that the members of DNB are closely related to these important molecules) before the critical transition to disease. That is, we can significantly delay the critical state or critical transition (B and D). In particular, DNB detects the leading molecules and network that initially drive the system from one state (e.g., health) to another (e.g., disease). Drastic deteriorations in health could be prevented by appropriately treating or perturbing the precursors of the diseased state.

since they largely characterize the variations used for phenotypic classification and prediction.

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) (XDB13040700), the National Basic Research Program of China (2014CB910504), the National Natural Science Foundation of China (61134013, 91029301, 31200987 and 91130033) and the Knowledge Innovation Program of SIBS of CAS (2013KIP218).

1 Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. N Biotechnol, 2012, 29: 613–624

2 Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol, 2011, 8: 184–187

3 Auffray C, Charron D, Hood L. Predictive, preventive, personalized and participatory medicine: back to the future. Genome Med, 2010, 2: 57

4 Zeng T, Sun SY, Wang Y, Zhu H, Chen L. Network biomarkers reveal dysfunctional gene regulations during disease progression. FEBS J, 2013, doi:10.1111/febs.12536

5 Zhong Q, Simonis N, Li QR, Charloteaux B, Heuze F, Klitgord N, Tam S, Yu H, Venkatesan K, Mou D, Swearingen V, Yildirim MA, Yan H, Dricot A, Szeto D, Lin C, Hao T, Fan C, Milstein S, Dupuy D, Brasseur R, Hill DE, Cusick ME, Vidal M. Edgetic perturbation models of human inherited disorders. Mol Syst Biol, 2009, 5: 321

6 Zhang W, Zeng T, Chen L. EdgeMarker: identifying differentially correlated molecule pairs as edge-biomarkers. J Theor Biol, 2014, doi:10.1016/j.jtbi.2014.05.041

7 Chen L, Liu R, Liu ZP, Li M, Aihara K. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. Sci Rep, 2012, 2: 342

8 Liu R, Li M, Liu ZP, Wu J, Chen L, Aihara K. Identifying critical transitions and their leading biomolecular networks in complex diseases. Sci Rep, 2012, 2: 813

9 Liu R, Wang X, Aihara K, Chen L. Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. Med Res Rev, 2013, doi:10.1002/med.21293

10 Liu R, Yu X, Liu X, Xu D, Aihara K, Chen L. Identifying critical transitions of complex diseases based on a single sample. Bioinformatics, 2014, doi:10.1093/bioinformatics/btu084

11 Li M, Zeng T, Liu R, Chen L. Detecting tissue-specific early warning signals for complex diseases based on dynamical network biomarkers: study of type 2 diabetes by cross-tissue analysis. Brief Bioinform, 2013, doi: 10.1093/bib/bbt1027

12 Welcsh PL, King MC. BRCA1 and BRCA2 and the genetics of breast and ovarian cancer. Hum Mol Genet, 2001, 10: 705–713

13 Wang-Sattler R, Yu Z, Herder C, Messias AC, Floegel A, He Y, Heim K, Campillos M, Holzapfel C, Thorand B, Grallert H, Xu T, Bader E, Huth C, Mittelstrass K, Döring A, Meisinger C, Gieger C, Prehn C, Roemisch-Margl W, Carstensen M, Xie L, Yamanaka-Okumura H, Xing G, Ceglarek U, Thiery J, Giani G, Lickert H, Lin X, Li Y, Boeing H, Joost HG, de Angelis MH, Rathmann W, Suhre K, Prokisch H, Peters A, Meitinger T, Roden M, Wichmann HE, Pischon T, Adamski J, Illig T. Novel biomarkers for pre-diabetes identified by metabolomics. Mol Syst Biol, 2012, 8: 615

14 Wemmie JA, Taugher RJ, Kreple CJ. Acid-sensing ion channels in pain and disease. Nat Rev Neurosci, 2013, 14: 461–471

15 Lampertico P, Vigano M, Cheroni C, Facchetti F, Invernizzi F, Valveri V, Soffredini R, Abrignani S, de Francesco R, Colombo M. IL28B polymorphisms predict interferon-related hepatitis B surface

antigen seroclearance in genotype D hepatitis B e antigen-negative patients with chronic hepatitis B. Hepatology, 2013, 57: 890–896

16 Joshita S, Umemura T, Katsuyama Y, Ichikawa Y, Kimura T, Morita S, Kamijo A, Komatsu M, Ichijo T, Matsumoto A, Yoshizawa K, Kamijo N, Ota M, Tanaka E. Association of IL28B gene polymorphism with development of hepatocellular carcinoma in Japanese patients with chronic hepatitis C virus infection. Hum Immunol, 2012, 73: 298–300

17 Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway LA, Pe'er D. An integrated approach to uncover drivers of cancer. Cell, 2010, 143: 1005–1017

18 Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet, 2010, 11: 31–46

19 Schadt EE. Molecular networks as sensors and drivers of common human diseases. Nature, 2009, 461: 218–223

20 Ren X, Wang Y, Chen L, Zhang XS, Jin Q. ellipsoidFN: a tool for identifying a heterogeneous set of cancer biomarkers based on gene expressions. Nucleic Acids Res, 2013, 41: e53

21 Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet, 2001, 69: 124–137

22 Jorde LB. Linkage disequilibrium and the search for complex disease genes. Genome Res, 2000, 10: 1435–1444

23 Stower H. Complex disease: family history versus SNPs for disease predictions. Nat Rev Genet, 2012, 13: 827

24 van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 2002, 415: 530–536

25 Li J, Lenferink AE, Deng Y, Collins C, Cui Q, Purisima EO, O'Connor-McCourt MD, Wang E. Identification of high-quality cancer prognostic markers and metastasis network modules. Nat Commun, 2010, 1: 34

26 Gormley M, Dampier W, Ertel A, Karacali B, Tozeren A. Prediction potential of candidate biomarker sets identified and validated on gene expression data from multiple datasets. BMC Bioinformatics, 2007, 8: 415

27 Koziol JA, Feng AC, Jia Z, Wang Y, Goodison S, McClelland M, Mercola D. The wisdom of the commons: ensemble tree classifiers for prostate cancer prognosis. Bioinformatics, 2009, 25: 54–60

28 Holec M, Klema J, Zelezny F, Tolar J. Comparative evaluation of set-level techniques in predictive classification of gene expression samples. BMC Bioinformatics, 2012, 13(Suppl 10): S15

29 Lee S, Kim J. A comparative study on gene-set analysis methods for assessing differential expression associated with the survival phenotype. BMC Bioinformatics, 2011, 12: 377

30 Abraham G, Kowalczyk A, Loi S, Haviv I, Zobel J. Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. BMC Bioinformatics, 2010, 11: 277

31 Levine DM, Haynor DR, Castle JC, Stepaniants SB, Pellegrini M, Mao M, Johnson JM. Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. Genome Biol, 2006, 7: R93

32 Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Comput Biol, 2008, 4: e1000217

33 Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet, 2011, 12: 56–68

34 Kreeger PK, Lauffenburger DA. Cancer systems biology: a network modeling perspective. Carcinogenesis, 2010, 31: 2–8

35 Drier Y, Sheffer M, Domany E. Pathway-based personalized analysis of cancer. Proc Natl Acad Sci USA, 2013, 110: 6388–6393

36 Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol, 2007, 3: 140

37 Dao P, Wang K, Collins C, Ester M, Lapuk A, Sahinalp SC. Optimally discriminative subnetwork markers predict response to chemotherapy. Bioinformatics, 2011, 27: i205–213

38 He D, Liu ZP, Chen L. Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach. BMC Genomics, 2011, 12: 592

39 Eddy JA, Hood L, Price ND, Geman D. Identifying tightly regulated and variably expressed networks by Differential Rank Conservation (DIRAC). PLoS Comput Biol, 2010, 6: e1000792

40 Jin G, Zhou X, Cui K, Zhang XS, Chen L, Wong ST. Cross-platform method for identifying candidate network biomarkers for prostate cancer. IET Syst Biol, 2009, 3: 505–512

41 Winter C, Kristiansen G, Kersting S, Roy J, Aust D, Knosel T, Rummele P, Jahnke B, Hentrich V, Ruckert F, Niedergethmann M, Weichert W, Bahra M, Schlitt HJ, Settmacher U, Friess H, Büchler M, Saeger HD, Schroeder M, Pilarsky C, Grützmann R. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. PLoS Comput Biol, 2012, 8: e1002511

42 Jiao Y, Lawler K, Patel GS, Purushotham A, Jones AF, Grigoriadis A, Tutt A, Ng T, Teschendorff AE. DART: denoising algorithm based on relevance network topology improves molecular pathway activity inference. BMC Bioinformatics, 2011, 12: 403

43 Wen Z, Liu ZP, Liu Z, Zhang Y, Chen L. An integrated approach to identify causal network modules of complex diseases with application to colorectal cancer. J Am Med Inform Assoc, 2013, 20: 659–667

44 Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods, 2013, 10: 1108–1115

45 Dutkowski J, Ideker T. Protein networks as logic functions in development and cancer. PLoS Comput Biol, 2011, 7: e1002180

46 Liu X, Liu ZP, Zhao XM, Chen L. Identifying disease genes and module biomarkers by differential interactions. J Am Med Inform Assoc, 2012, 19: 241–248

47 Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol, 2009, 27: 199–204

48 Wang YC, Chen BS. A network-based biomarker approach for molecular investigation and diagnosis of lung cancer. BMC Med Genomics, 2011, 4: 2

49 Qian L, Zheng H, Zhou H, Qin R, Li J. classification of time series gene expression in clinical studies via integration of biological network. PLoS One, 2013, 8: e58383

50 Yu X, Li G, Chen L. Prediction and early diagnosis of complex diseases by edge-network. Bioinformatics, 2013, doi: 10.1093/bioinformatics/btt620

51 van de Leemput IA, Wichers M, Cramer AO, Borsboom D, Tuerlinckx F, Kuppens P, van Nes EH, Viechtbauer W, Giltay EJ, Aggen SH, Derom C, Jacobs N, Kendler KS, van der Maas HL, Neale MC, Peeters F, Thiery E, Zachar P, Scheffer M. Critical slowing down as early warning for the onset and termination of depression. Proc Natl Acad Sci USA, 2014, 111: 87–92

52 Ichikawa S, Ito Y, Uchida K. Periodic Lyapunov differential equation for noise evaluation in oscillatory genetic networks. IEEE Control Applications (CCA) & Intelligent Control (ISIC), 2009. 83–88

53 Sejdic E. Medicine: adapt current tools for handling big data. Nature, 2014, 507: 306

54 Restifo NP. A “big data” view of the tumor "immunome". Immunity, 2013, 39: 631–632

55 Gijzen H. Development: big data for a sustainable future. Nature, 2013, 502: 38

56 Boyle J. Biology must develop its own big-data systems. Nature, 2013, 499: 7

57 Marx V. Biology: the big challenges of big data. Nature, 2013, 498: 255–260

58 Buxton B, Hayward V, Pearson I, Karkkainen L, Greiner H, Dyson E, Ito J, Chung A, Kelly K, Schillace S. Big data: the next Google. Interview by Duncan Graham-Rowe. Nature, 2008, 455: 8–9

59 Ideker T, Krogan NJ. Differential network biology. Mol Syst Biol, 2012, 8: 565

60 Liu M, Liberzon A, Kong SW, Lai WR, Park PJ, Kohane IS, Kasif S. Network-based analysis of affected biological processes in type 2 diabetes models. PLoS Genet, 2007, 3: e96

61 Ulitsky I, Krishnamurthy A, Karp RM, Shamir R. DEGAS: *de novo* discovery of dysregulated pathways in human diseases. PLoS One, 2010, 5: e13367

62 Zhang B, Li H, Riggins RB, Zhan M, Xuan J, Zhang Z, Hoffman EP, Clarke R, Wang Y. Differential dependency network analysis to identify condition-specific topological changes in biological networks. Bioinformatics, 2009, 25: 526–532

63 Zhang B, Tian Y, Jin L, Li H, Shih Ie M, Madhavan S, Clarke R, Hoffman EP, Xuan J, Hilakivi-Clarke L, Wang Y. DDN: a caBIG(R) analytical tool for differential network analysis. Bioinformatics, 2011, 27: 1036–1038

64 Kim Y, Kim TK, Yoo J, You S, Lee I, Carlson G, Hood L, Choi S, Hwang D. Principal network analysis: identification of subnetworks representing major dynamics using gene expression data. Bioinformatics, 2011, 27: 391–398

65 West J, Bianconi G, Severini S, Teschendorff AE. Differential network entropy reveals cancer system hallmarks. Sci Rep, 2012, 2: 802

66 Sun SY, Liu ZP, Zeng T, Wang Y, Chen L. Spatio-temporal analysis of type 2 diabetes mellitus based on differential expression networks. Sci Rep, 2013, 3: 2268

67 Zeng T, Zhang CC, Zhang W, Liu R, Liu J, Chen L. Deciphering early development of complex diseases by progressive module network. Methods, 2014, http://dx.doi.org/10.1016/j.ymeth.2014.01.021

Appendix

Differential-network-based methods

Network biomarkers discriminate among samples with different phenotypes. Many alternative network-related studies infer groups of significantly differentially expressed genes interacting in context-specific or differential networks. These differential networks are considered to represent different disease phenotypes (Table S1). In fact, interactome architectures are known to be massively rewired during a cellular or adaptive response. Thus, identifying differential networks by exploring interaction spaces rather than gene spaces has become a standard network analysis technique [59]. For example, gene network enrichment analysis (GNEA) is a simple methodology that searches for significant aberrations in the collective expression of a set of interactive genes (whose protein products form a connected protein complex or sub-network) rather than aberrant expression of individual genes [60]. Another method (dysregulated gene set analysis via subnetworks; DEGAS) identifies sub-networks that are significantly enriched by dysregulated genes during disease [61]. Differential dependency network (DDN) analysis exploits the differential topological changes in biological networks. Based on a local dependency model, DDN detects significant topological changes in the transcriptional networks between two biological conditions, rather than changes in their expression levels [62,63]. By contrast, principal network analysis (PNA) captures the major dynamic activation patterns and their associated protein and metabolic sub-networks under multiple conditions [64]. Meanwhile, West et al. [65] found that cancer cells can be characterized by differential network entropy (DNE). Focusing on the differential functional changes of biological networks, they found that differences in gene expression in normal and cancer tissues anti-correlate with changes in local network entropy. Specifically, genes driving cell-proliferation in cancer cells or encoding oncogenes are usually associated with reduced network entropy. Another framework, differential expression network (DEN), is based on the 'dysfunctional interaction' concept. DEN characterizes the differential information involved in different disease stages by integrating the differential gene and differential network paradigms [66]. A dysfunctional interaction results from either node perturbations (i.e., interaction perturbations caused by the differentially expressed individual genes) or edge perturbations (i.e., interaction perturbations caused by the differentially co-expressed genes). A new model, called progressive module network (PMN), uses DNBs to identify the modules presenting at the pre-disease stage. It then detects the modules in the advanced stage by cross-tissue gene expression analysis; and finally finds the modules related to early disease development by a progressive module network [67].

Table S1 Examples of differential-network-based methods for biomarkers

Methods	Differential gene expression	Differential gene co-expression	Scale of differential network
GNEA [60]	Yes		A set of interactive genes
DEGAS [61]	Yes		Dys-regulated genes enriched on sub-networks
DDN [62,63]		Yes	Local dependency network
PNA [64]	Yes	Yes	Major dynamic activation patterns within sub-networks
DNE [65]	Yes	Yes	Local change of differential network entropy
DEN [66]	Yes	Yes	Global differential expression network
PMN [67]	Yes	Yes	Network modules, module network and its re-organizations