doi:10.1093/jmcb/mjv025
Published online April 26, 2015

Article

Diagnosing phenotypes of single-sample individuals by edge biomarkers

Wanwei Zhang[1,†], Tao Zeng[1,†], Xiaoping Liu[1], and Luonan Chen[1,2,*]

[1] Key Laboratory of Systems Biology, Innovation Center for Cell Signaling Network, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[2] School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China
[†] These authors contributed equally to this work.
[*] Correspondence to: Luonan Chen, E-mail: lnchen@sibs.ac.cn

Network or edge biomarkers are a reliable form to characterize phenotypes or diseases. However, obtaining edges or correlations between molecules for an individual requires measurement of multiple samples of that individual, which are generally unavailable in clinical practice. Thus, it is strongly demanded to diagnose a disease by edge or network biomarkers in one-sample-for-one-individual context. Here, we developed a new computational framework, EdgeBiomarker, to integrate edge and node biomarkers to diagnose phenotype of each single test sample. By applying the method to datasets of lung and breast cancer, it reveals new marker genes/gene-pairs and related sub-networks for distinguishing earlier and advanced cancer stages. Our method shows advantages over traditional methods: (i) edge biomarkers extracted from non-differentially expressed genes achieve better cross-validation accuracy of diagnosis than molecule or node biomarkers from differentially expressed genes, suggesting that certain pathogenic information is only present at the level of network and under-estimated by traditional methods; (ii) edge biomarkers categorize patients into low/high survival rate in a more reliable manner; (iii) edge biomarkers are significantly enriched in relevant biological functions or pathways, implying that the association changes in a network, rather than expression changes in individual molecules, tend to be causally related to cancer development. The new framework of edge biomarkers paves the way for diagnosing diseases and analyzing their molecular mechanisms by edges or networks in one-sample-for-one-individual basis. This also provides a powerful tool for precision medicine or big-data medicine.

Keywords: edge biomarker, edge feature, progressive stages, disease diagnosis, big biological data

Introduction

A complex disease is generally a problem resulted from the failure of the relevant system, which should be investigated in a dynamic and network manner (Auffray et al., 2010; Hood and Friend, 2011; Hood and Flores, 2012). Therefore, compared with single molecules, networks or edges among molecules are considered to be a more stable and reliable form for characterizing complex diseases or phenotypes as biomarkers. From the systems viewpoint, a group of individual molecules is only a set of irrelevant nodes, but a group of edges represents a system or network. In a biological system, it is the interactions (regulations) or edges between molecules rather than individual molecules that facilitate a biological function or signal transduction involved in diseases (Zeng et al., 2014). Thus, the signatures of a network with interactive biological elements, e.g. network biomarkers or edge biomarkers, are essential to achieve the predictive and personalized medicine. Generally, an edge in a molecular network is represented by its correlation coefficient, e.g. Pearson correlation coefficient (PCC) between a pair of molecules, which can be numerically estimated provided that there are multiple samples. Recently, due to rapid advance on high throughput technologies, network-based biomarker discovery by exploiting omics data has become a hot topic in the study of complex diseases or personalized medicine (Hood and Friend, 2011).

An important evidence of edge or interaction signatures is the finding of 'edgetics' diseases. The notion of 'edgotype' has generally linked the genotype to phenotype (Sahni et al., 2013). Meanwhile, the study of 'edgetics' also revealed the malfunctions of interactions (Chen et al., 2009; Sahni et al., 2013) as the key molecular mechanisms relevant to complex diseases, in which the genes work interactively in a network/system manner. One example is that traditional biomarker discovery mainly focuses on differentially expressed genes (DGs), leaving large amount of non-differentially expressed genes (NDGs) unexamined (Mor et al., 2005; Listgarten et al., 2007). However, a considerable amount of evidence has shown that the 'edgotype' (or edgetics) of NDGs can play key roles in altering the states of biological systems, e.g. from normal to disease (Lai et al., 2004; Sun et al.,

Received December 21, 2014. Revised January 16, 2015. Accepted January 27, 2015.

2013; Wang et al., 2013; Zeng et al., 2013; Yu et al., 2014). Different from DGs, an interacting gene-pair from NDGs can exhibit positive or negative correlation in different conditions, even though the involved genes are not differentially expressed. Such changes of genes' correlations result in the different states of a biological system (e.g. normal or diseased). Several methods have been developed to identify the signatures of NDGs (Lai et al., 2004; He et al., 2012; Liu et al., 2012b; Sun et al., 2013) and achieved certain success in revealing potential mechanisms altering the biological states.

Although traditional network-based methods can identify differential edges (DEs) or networks by comparing multiple known control and case samples, those edge biomarkers cannot be applied to diagnose an unknown sample, because the respective edges or PCCs cannot be obtained from a single sample, which, however, represents the majority of clinic cases for each individual. In other words, computing correlations or edges requires multiple samples, but generally only one sample is available when diagnosing an individual in clinic practice. Usually, a gene group or set with certain criterion, e.g. dense connections of protein–protein interactions, is derived from DEs in network-based methods, but DGs of the gene group instead of DEs are used for diagnosis and prognosis. Thus, network-based biomarkers are not network biomarkers but essentially molecular biomarkers (Zeng et al., 2014), because the edge or network information is not exploited to diagnose an unknown sample. Figure 1 schematically shows the procedures of discovery and application of traditional molecular or node biomarkers (based on DGs) and traditional network-based biomarkers (based on DEs).

To overcome the difficulty in obtaining correlations or edges from one sample, an EdgeMarker approach (Zhang et al., 2014) was developed by decomposing each PCC into multiple elements that form a new vector embedding correlation-like information for one sample. Provided that average values and standard deviations for gene expressions of DEs between control and case samples are similar, the obtained edges can serve as biomarkers to distinguish the states of the sample. Thus, it is efficient to identify edge biomarkers mainly from NDGs. In this study, we further extend the work for edge biomarker discovery by considering the whole gene set including both DGs and NDGs. Our new method, EdgeBiomarker, identifies both edge and node biomarkers from omics data, which can directly diagnose the state of a disease even for a single sample. To illustrate the effectiveness, we employ the edge biomarkers to distinguish earlier and advanced stages of complex diseases, e.g. progressive development of cancer. Specifically, we briefly summarize various models of biomarkers, describe the theoretical framework of EdgeBiomarker to extract signatures of network information and to diagnose each individual test sample, and study the states of lung and breast cancers as well as their metastasis risk by EdgeBiomarker. Both theoretical results and computational experiments demonstrate that edge biomarkers not only accurately distinguish phenotypes of each single sample but also provide new insights into the pathogenesis of complex diseases.

Various models of biomarkers

Biomarker discovery generally includes two steps: firstly in the learning step, effective marker molecules are identified to

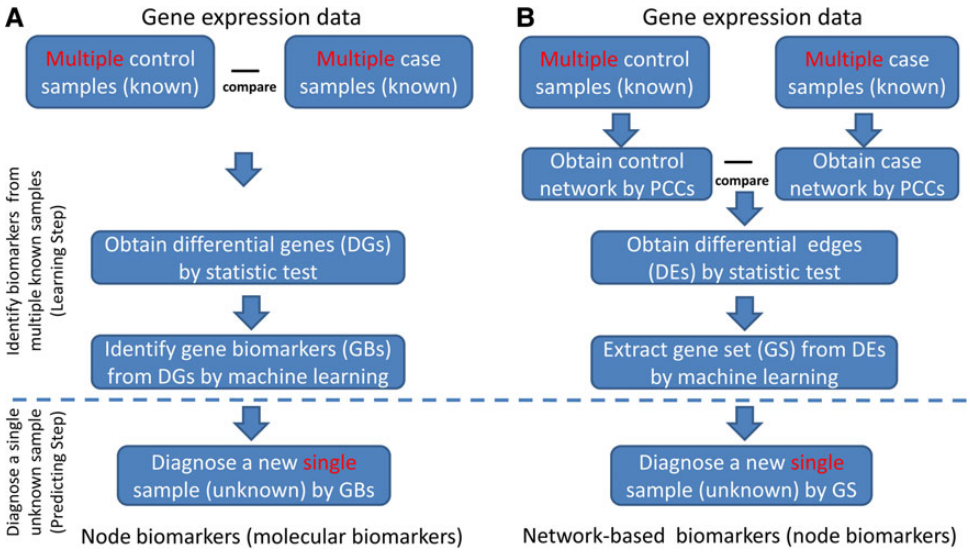

Figure 1 The procedures of discovery and application of traditional node biomarkers and traditional network-based biomarkers. (**A**) Traditional molecular or node biomarkers are obtained from differential genes (DGs) or nodes by machine learning techniques so as to accurately classify multiple case and control samples with known phenotypes. By the obtained node biomarkers, a new sample with an unknown phenotype can be classified into either case or control group. (**B**) On the other hand, from multiple control and case samples, we can obtain their respective networks and further differential edges (DEs). But traditional network-based approaches only use genes corresponding to DEs to diagnose or classify a new single sample (unknown phenotype) without considering network information in the new sample, because edges or correlations among molecules in a single sample cannot be calculated by the traditional methods. Thus, traditional network-based biomarkers are essentially node biomarkers.

discriminate different phenotypes (e.g. normal and disease groups of samples) by machine learning or optimization techniques; secondly in the predicting step, the phenotype of a test sample is evaluated by the marker molecules (Figure 1). According to the type of network information in such steps, biomarkers can be mainly classified as (Zeng et al., 2014) traditional node biomarkers (or molecular biomarkers), traditional network-based biomarkers, newly developed edge biomarkers (or network biomarkers), and dynamical network biomarkers (DNBs). Their main features are summarized as follows.

(A) Node biomarkers focus on differential expression levels of a number of individual molecules rather than differential expression associations/correlations among multiple molecules (Zeng et al., 2014) (Figure 1A).

(B) Although some network-based biomarkers and network-weighted biomarkers use network or correlation information to identify sub-networks or edges in the learning step, only molecules (or molecule set) related to those edges or sub-networks, without the edge or network information of the test sample, are used to diagnose the phenotype of a test sample in the predicting step (Zeng et al., 2014) (Figure 1B). Thus, essentially they are still node biomarkers.

(C) Recently, an approach, EdgeMarker, was developed to identify the differentially correlated gene pairs (DCPs) based on a new vector representation of an edge (Zhang et al., 2014). With the new edge representation, i.e. a correlation-like vector for each edge, this single-sample-based approach can make full use of edge information on an individual test sample. Both theoretical and computational experiments have shown that the edge biomarkers can distinguish the phenotype of a test sample in an accurate manner even though their genes may not have differential expressions. These NDGs ignored by conventional methods can be as informative as DGs for distinguishing different biological conditions or phenotypes (Zhang et al., 2014). However, EdgeMarker is not efficient for DGs due to its vector form representing correlations of gene pairs.

(D) By exploring dynamical information of data, the model for DNB was proposed to detect pre-disease state (or critical state) rather than disease state (Chen et al., 2012b). The pre-disease state is the limit of the normal state, or a normal state of an individual just before his/her critical transition to the disease state. Traditional biomarkers try to distinguish whether a sample is in a disease state or not, but they generally cannot diagnose the pre-disease state lacking significant difference between the normal and disease states, e.g. no significant differential expressions of molecules. DNB can be identified by satisfying the following conditions (Chen et al., 2012b; Liu et al., 2012a) when a biological system is approaching the pre-disease state: (i) the variance of the DNB molecules drastically increases; (ii) the correlation (PCC) between any two DNB molecules increases; (iii) the correlation (PCC) between any molecule in the DNB and another in the non-DNB decreases. Actually, DNB is a model-free approach for biomarker identification based on the observed big data. Furthermore, Yu et al. (2014) proposed edge-network to exploit higher-order statistics information among molecules inspired from DNB. On the other hand, Liu et al. (2014) further developed the DNB-S scoring method, as a single-sample-based approach, to indicate the pre-disease state on a single sample (given that there are a group of normal/control samples). However, DNB approach is designed to identify the pre-disease state rather than the disease state.

In this work, we aim to develop a new framework of edge biomarkers for distinguishing disease samples by further extending EdgeMarker to both DGs and NDGs.

EdgeBiomarker identifies both edge and node biomarkers to predict the phenotype of each single-sample individual

Given a number of control samples (m samples) and case samples (n samples), gene expression data with k genes can be represented as the data matrix shown in Figure 2A, where there are k genes, and the sample sizes of control and case groups are m and n, respectively, i.e. the dimensions of the matrix for node data are $k \times (m+n)$. Let $\boldsymbol{x}_i \in R^m$ and $\boldsymbol{y}_i \in R^n$ denote the expression vectors of the ith gene for control and case, respectively, i.e. x_{ij} is the expression of the ith gene for the jth sample in control group, while y_{ij} is the expression of the ith gene for the jth sample in case group. $\bar{\boldsymbol{x}}_u, \bar{\boldsymbol{x}}_v, \bar{\boldsymbol{y}}_u, \bar{\boldsymbol{y}}_v$ are the average expressions of genes u and v from control and case groups; Sx_u, Sx_v, Sy_u, Sy_v are the standard expression deviations of genes u and v from control and case groups; $k_1 = \sqrt{(m-1)/m}$ and $k_2 = \sqrt{(n-1)/n}$ are two adjusting factors. For each gene-pair $\langle u, v\rangle$, two coupled edge features $\langle u, v\rangle_N$ and $\langle u, v\rangle_D$ are constructed as in Figure 2B. Thus, the matrix for edge data in Figure 2B is constructed with the dimensions of $2k^2 \times (m+n)$. The average of each row vector in the gray box in Figure 2B is the PCC between genes u and v from control or case group, respectively. On the other hand, the average of the row vector in the gray box in Figure 2A is the average expression of gene u from control and case group, respectively.

Many traditional methods for detecting node biomarkers based on the matrix of Figure 2A are to select genes whose differential average expressions between control and case are significantly high, whereas the remaining NDGs are deleted from the matrix. Any test sample as a new column vector (Figure 2C) can be diagnosed based on the selected node biomarkers. To detect edge biomarkers from the matrix of Figure 2B, a similar scheme is to select edges whose differential correlations are significantly high, whereas the remaining non-differential edges are removed. Then, any test sample as a new column vector (Figure 2C) can be diagnosed by the selected edge biomarkers. As a novel approach to identify edge biomarkers for individual test samples, EdgeBiomarker combines the above traditional biomarker discovery procedure with the EdgeMarker approach (Zhang et al., 2014). The details are demonstrated in Figure 3. Given gene expression profiles under binary conditions (Figure 3A1), the DGs are selected (Figure 3A2), and the biomarkers are extracted from these DGs by any optimization algorithm, e.g. sequential forward floating selection (SFFS; Pudil et al.,

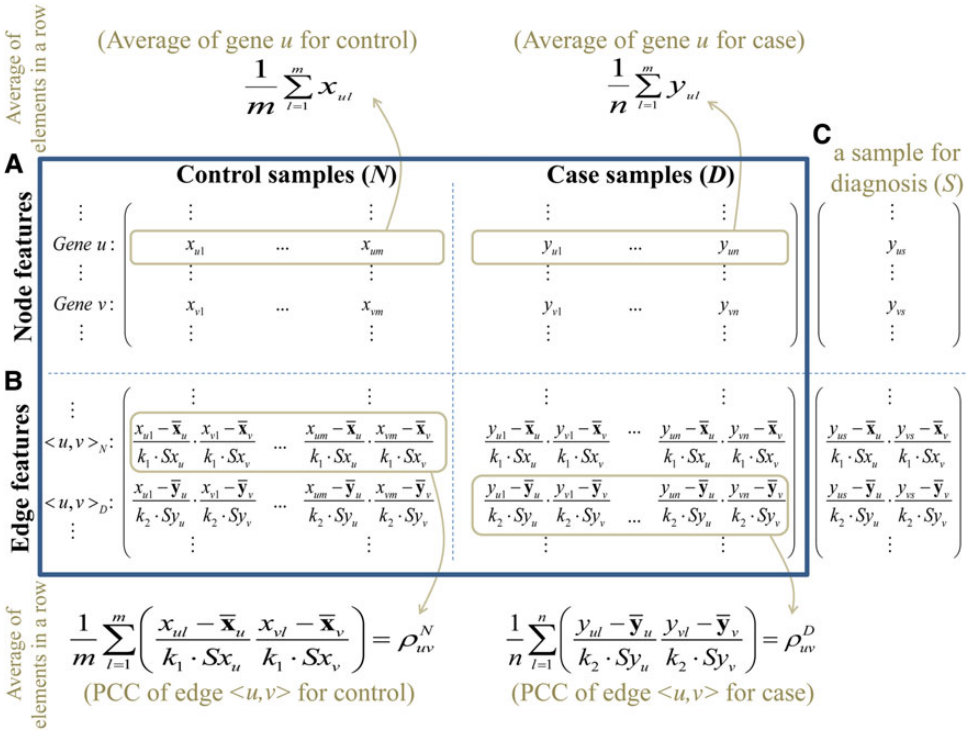

Figure 2 Data matrices for node features and edge features. (**A**) The data matrix for finding node biomarkers, which aims to identify effective DGs, whose differential expressions between control and case, in terms of average value, are large. Each column is one sample. Average value of each row represents the average expression of a gene for either control or case. (**B**) The data matrix for finding edge biomarkers, which aims to identify effective DEs or correlations mainly from NDGs. Each column is one sample. Average value of each row represents the PCC of a gene-pair or edge for either control or case. The combined matrix of node biomarkers (**A**) and edge biomarkers (**B**) is the data matrix of EdgeBiomarkers, which aims to identify effective DGs (node biomarkers) and DEs (edge biomarkers) from the information of both DGs and NDGs. (**C**) The column vector represents the test sample for diagnosis by node biomarkers and edge biomarkers.

1994) method (Figure 3A3). For those NDGs, we extracted the DCPs or DEs (Figure 3B1), where each DCP is defined as interactive two genes whose expression correlation has a significant change from one condition to another. Such DCPs are selected in the criterion of $\{\langle i,j\rangle||r_{ij}^N - r_{ij}^D| > \delta\}$, where r_{ij}^N and r_{ij}^D are the PCCs of two genes i and j in condition N (normal) and D (disease), and the threshold δ was empirically set. Based on the formula (2) as introduced below, the node data (i.e. expression values) of the selected gene pairs were then transformed into edge data, where gene–gene pairs instead of individual genes are features (Figure 3B2). Note that the matrix of edge data has $m + n$ columns as same as the matrix of node data, but has at most k^2 rows in contrast to at most k rows in the matrix of node data (Figure 2A and B). Subsequently, the edge biomarkers are selected by applying any optimization algorithm, e.g. SFFS algorithm (Pudil et al., 1994), to the edge data (Figure 3B3). The data for node biomarkers and the data for edge biomarkers were combined into one matrix and the SFFS algorithm was used again to obtain the optimal combined biomarkers, i.e. node biomarkers and edge biomarkers (Figure 3A4). To classify/predict a single test sample with unknown class label (Figure 3A5), the data of the combined biomarkers for this sample are used by formula (2) (Figure 3A6) and based on the trained classification model for phenotype prediction (e.g. disease or not). The detailed procedures of EdgeBiomarker are listed as follows (Figure 3). Note that we mainly state the procedure for detecting edges, and how to detect the nodes can be conducted in a similar way or refer to Figures 1A and 3A1–A3.

(i) *Pre-processing the datasets*. Genes with low expression level or with high coefficient of variance are likely to be affected by noise, and thus it is necessary to remove these genes. In practice, the genes are sorted by expression level or coefficient of variance from large value to small value, and removed when the average expression levels are in bottom 10% or the coefficients of variances are in top 10% among all genes.

(ii) *Selecting differentially correlated gene pairs*. In EdgeBiomarker, PCC is used to characterize the correlation between two genes (Figure 3B1). The DCPs or DEs are defined as follows:

$$\{\langle i,j\rangle||r_{ij}^N - r_{ij}^D| > \delta\}, \quad (1)$$

where i and j denote genes that are under the study. r_{ij}^N and r_{ij}^D are PCCs between gene i and gene j under normal (or control) and disease (or case) conditions, respectively. The threshold δ is set as 0.6 empirically in lung cancer dataset (~420 samples) and 0.45 in breast cancer dataset (~750 samples), although more sophisticated method can be applied to

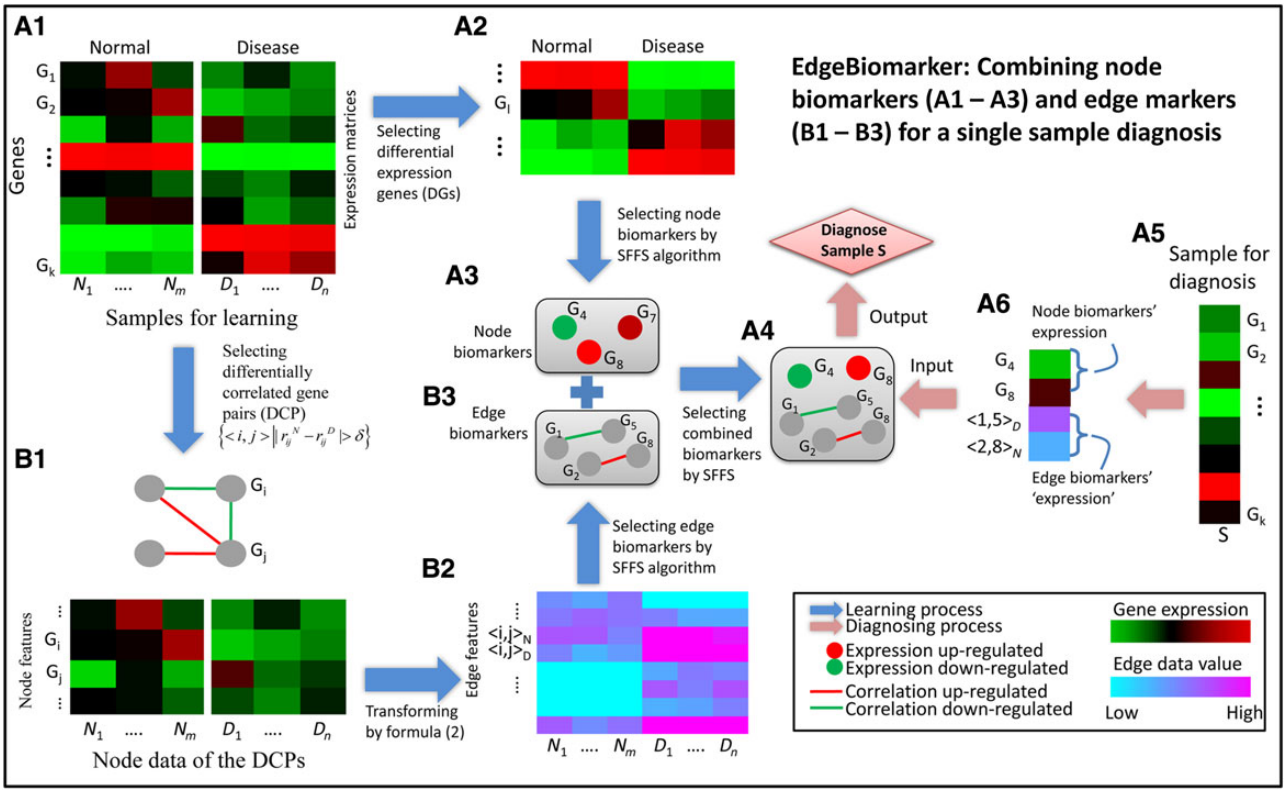

Figure 3 The procedure of EdgeBiomarker. Given gene expression profiles under binary conditions (**A1**), DGs (**A2**) and differentially correlated gene pairs (DCPs) (**B1**) are selected. Here, gene expressions are presented in green-red color map, where green/red represents low/high expression level. The red/green edges represent up-regulated/down-regulated correlations. The node data (i.e. expression values) of the selected gene pairs were then transformed into edge data where features are the gene–gene pairs (**B2**). The edge data are presented in light blue-purple color map, where blue/purple represents low/high value. Subsequently, the node biomarkers (**A3**) and edge biomarkers (**B3**) were selected from DGs and DCPs, respectively, by using sequential forward floating selection (SFFS) algorithm. The SFFS algorithm is used again to get the optimal combined biomarkers from node + edge biomarkers (**A4**). To classify the single test sample with unknown class label (**A5**), the data of the combined biomarkers for this sample are computed and input to the trained classification model for its phenotype prediction (**A6**).

determine the threshold. Note that the DCPs can be selected from all genes including DGs or NDGs. But the number of DGs is usually much smaller than that of NDGs, and the information of DGs is also used in finding node biomarkers. Thus, in the numerical computation of this study, we selected edge biomarkers from NDGs.

(iii) *Transforming node data into edge data*. From the viewpoint of network, gene expression profiles characterize the node feature, while the expression correlations represent the edge feature. Inspired by the Pearson's correlation coefficient, we designed the following matrix transformation from node data (e.g. genes u and v) to edge data (e.g. gene-pair u and v),

$$\text{Node features}: \quad \begin{array}{c} \\ \text{Gene } u: \\ \text{Gene } v: \end{array} \begin{array}{c} \text{Normal samples} \\ \begin{pmatrix} N_1 & \dots & N_m \\ x_{u1} & \dots & x_{um} \\ x_{v1} & \dots & x_{vm} \end{pmatrix} \end{array} \quad \begin{array}{c} \text{Disease samples} \\ \begin{pmatrix} D_1 & \dots & D_n \\ y_{u1} & \dots & y_{un} \\ y_{v1} & \dots & y_{vn} \end{pmatrix} \end{array}$$

$$\Downarrow \qquad (2)$$

$$\text{Edge features}: \quad \begin{array}{c} \\ \langle u,v\rangle_N: \\ \langle u,v\rangle_D: \end{array} \begin{pmatrix} \frac{x_{u1}-\bar{\mathbf{x}}_u}{k_1\cdot Sx_u}\cdot\frac{x_{v1}-\bar{\mathbf{x}}_v}{k_1\cdot Sx_v} & \cdots & \frac{x_{um}-\bar{\mathbf{x}}_u}{k_1\cdot Sx_u}\cdot\frac{x_{vm}-\bar{\mathbf{x}}_v}{k_1\cdot Sx_v} & \frac{y_{u1}-\bar{\mathbf{x}}_u}{k_1\cdot Sx_u}\cdot\frac{y_{v1}-\bar{\mathbf{x}}_v}{k_1\cdot Sx_v} & \cdots & \frac{y_{un}-\bar{\mathbf{x}}_u}{k_1\cdot Sx_u}\cdot\frac{y_{vn}-\bar{\mathbf{x}}_v}{k_1\cdot Sx_v} \\ \frac{x_{u1}-\bar{\mathbf{y}}_u}{k_2\cdot Sy_u}\cdot\frac{x_{v1}-\bar{\mathbf{y}}_v}{k_2\cdot Sy_v} & \cdots & \frac{x_{um}-\bar{\mathbf{y}}_u}{k_2\cdot Sy_u}\cdot\frac{x_{vm}-\bar{\mathbf{y}}_v}{k_2\cdot Sy_v} & \frac{y_{u1}-\bar{\mathbf{y}}_u}{k_2\cdot Sy_u}\cdot\frac{y_{v1}-\bar{\mathbf{y}}_v}{k_2\cdot Sy_v} & \cdots & \frac{y_{un}-\bar{\mathbf{y}}_u}{k_2\cdot Sy_u}\cdot\frac{y_{vn}-\bar{\mathbf{y}}_v}{k_2\cdot Sy_v} \end{pmatrix},$$

where $\bar{\boldsymbol{x}}_u, \bar{\boldsymbol{x}}_v$ and $\bar{\boldsymbol{y}}_u, \bar{\boldsymbol{y}}_v$ are the average expressions of genes u and v from control group and case group, respectively; Sx_u, Sx_v and Sy_u, Sy_v are the standard expression deviations of genes u and v from control group and case group, respectively; $k_1 = \sqrt{(m-1)/m}$ and $k_2 = \sqrt{(n-1)/n}$ are two adjusting factors. For each gene-pair $\langle u, v\rangle$, two coupled edge features $\langle u, v\rangle_N$ and $\langle u, v\rangle_D$ are constructed for control (normal) group and case (disease) group. It can be seen that the average values of each row vector of

$$\left(\frac{x_{u1}-\bar{\mathbf{x}}_u}{k_1\cdot Sx_u}\frac{x_{v1}-\bar{\mathbf{x}}_v}{k_1\cdot Sx_v},\ \ldots,\ \frac{x_{um}-\bar{\mathbf{x}}_u}{k_1\cdot Sx_u}\frac{x_{vm}-\bar{\mathbf{x}}_v}{k_1\cdot Sx_v}\right), \tag{3}$$

and

$$\left(\frac{y_{u1}-\bar{\mathbf{y}}_u}{k_2\cdot Sy_u}\frac{y_{v1}-\bar{\mathbf{y}}_v}{k_2\cdot Sy_v},\ \ldots,\ \frac{y_{un}-\bar{\mathbf{y}}_u}{k_2\cdot Sy_u}\frac{y_{vn}-\bar{\mathbf{y}}_v}{k_2\cdot Sy_v}\right), \tag{4}$$

are the Pearson's correlation coefficients between genes u and v under normal condition (formula (3)) and disease condition (formula (4)), respectively (see Figure 2).

(iv) *Filtering discriminative edge features as candidates of edge biomarkers*. For edge data of DCPs, the number of edge features is usually very large. Therefore it is necessary to filter a subset of features as biomarker candidates that are discriminative between two conditions. In this work, we select the features of which the *P*-values or adjusted *P*-values of Student's *t*-test between two conditions are <0.05. The representation of edge features is shown in formula (2).

(v) *Selecting discriminative node features as candidates of node biomarkers*. By standard *t*-test (*P*-value < 0.05) for case and control samples, we select the DGs as the node features or as the candidates of node biomarkers in EdgeBiomarker. The representation of node features is shown in formula (2).

(vi) *Selecting union biomarkers from the combined node and edge features*. After selecting the candidates of node and edge features from (v) and (iv), the corresponding matrices were combined into the matrix defined as formula (5).

$$\begin{array}{c} \\ \text{Edge features} \quad \text{Node features} \end{array}
\begin{array}{c}
\begin{array}{cccccccc} \multicolumn{4}{c}{\text{Normal samples}} & & \multicolumn{3}{c}{\text{Disease samples}} \end{array} \\
\begin{array}{cccccccc} N_1 & \ldots & N_m & & D_1 & \ldots & D_n \end{array} \\
\begin{pmatrix}
\vdots & & \vdots & & \vdots & & \vdots \\
x_{i1} & \ldots & x_{im} & & y_{i1} & \ldots & y_{in} \\
\vdots & & \vdots & & \vdots & & \vdots \\
\frac{x_{u1}-\bar{\mathbf{x}}_u}{k_1\cdot Sx_u}\frac{x_{v1}-\bar{\mathbf{x}}_v}{k_1\cdot Sx_v} & \cdots & \frac{x_{um}-\bar{\mathbf{x}}_u}{k_1\cdot Sx_u}\frac{x_{vm}-\bar{\mathbf{x}}_v}{k_1\cdot Sx_v} & & \frac{y_{u1}-\bar{\mathbf{x}}_u}{k_1\cdot Sx_u}\frac{y_{v1}-\bar{\mathbf{x}}_v}{k_1\cdot Sx_v} & \cdots & \frac{y_{un}-\bar{\mathbf{x}}_u}{k_1\cdot Sx_u}\frac{y_{vn}-\bar{\mathbf{x}}_v}{k_1\cdot Sx_v} \\
\vdots & & \vdots & & \vdots & & \vdots
\end{pmatrix}
\end{array}. \tag{5}$$

Any feature selection method can be applied on this matrix (5) to select the union biomarkers. In this work, the sequential forward floating selection algorithm (Pudil et al., 1994) is used. Here, we use the cross-validation accuracy of linear-SVM to characterize the goodness of a subset of features. This algorithm returns the optimal feature subsets with different sizes and the corresponding cross-validation accuracies. Based on the accuracy curve, the tradeoff for maximizing the accuracy and minimizing the feature number is made to select the final node, edge, or combined biomarkers.

(vii) *Predicting phenotypes by union biomarkers for an individual test sample*. After the combined biomarkers being selected, the classifier is built. The diagnosis or classification of each single sample can be conducted by transforming its expression values of genes of the biomarkers into the corresponding data form (node data + edge data) (Figure 3A6). The edge data as well as the node data are subsequently input to the classifier for predicting the classification of the test sample.

With the procedure above, we can diagnose a new test sample by the obtained node and edge biomarkers. Note that, for the traditional method, it is efficient to identify node biomarkers by exploiting DGs from the node matrix of Figure 2A, whereas for EdgeMarker, it is effective to identify edge biomarkers by exploiting NDGs from the edge matrix of Figure 2B. Thus, it is complementary to combine the two matrices of Figure 2A and B for biomarker discovery (including both node biomarkers and edge biomarkers) in EdgeBiomarker. Note that we can also use all gene pairs (including DG–DG, DG–NDG, and NDG–NDG pairs) as the candidates of edge biomarkers, instead of the gene pairs of NDGs in Figure 2.

Case study on metastasis risk of lung adenocarcinoma by EdgeBiomarker

For case studies, we investigated lung adenocarcinoma, which is one of the major histological subtypes of non-small-cell lung cancer (NSCLC) and is the most common type in patients of non-smokers (Herbst et al., 2008). Although being widely studied, the diagnosis and treatment of lung adenocarcinoma are still big challenges,

since the cancer pathogenesis and progressive mechanism is not completely clear. In this study, we divided the clinical stages of lung adenocarcinoma into two periods: earlier stage (including clinical stages IA, IB, IIA, IIB) and advanced stage (including clinical stages IIIA, IIIB, IV), where the patients in advanced stage have higher metastasis risks.

Lung adenocarcinoma datasets from TCGA and GEO databases

In this work, we collected the public RNA-seq data from the Cancer Genome Atlas database (http://cancergenome.nih.gov/). We also downloaded the microarray data from the Gene Expression Omnibus database (GSE13213; Tomida et al., 2009), which has many samples with matched clinical information, for independent validation. The clinical information for both datasets is available as survival time, stage, grade, sex, etc. Table 1 gives a brief summary of the two datasets. After data processing, there are in total 421 + 3 samples in TCGA dataset, where the additional three samples are known in stage I but without further information for 'IA' or 'IB'.

Edge biomarkers distinguish earlier and advanced lung adenocarcinoma with a higher accuracy than node biomarkers

EdgeBiomarker was applied to NDGs of TCGA lung cancer dataset (which are all ignored or deleted by traditional methods), and obtained marker sets with optimal cross-validation accuracies. By the traditional method, node biomarkers are selected from DGs.

Table 1 Summary on samples of two datasets from TCGA and GEO databases.

		Lung cancer stages							
		IA	IB	IIA	IIB	IIIA	IIIB	IV	Total
Dataset	TCGA	106	124	39	59	62	10	21	421 + 3
	GEO	42	37	4	9	20	5	0	117

Figure 4A shows the optimized accuracies achieved with different numbers of edge features (i.e. gene pairs, labeled in red) or node features (i.e. individual genes, labeled in blue). Obviously, the edge features (gene pairs) discriminate between earlier and advanced stages of lung cancer in a better performance than the node features (individual genes). Based on the accuracy curves, the tradeoff between maximizing accuracy and minimizing feature number is further tuned, and a few features are finally selected as final biomarkers. In Figure 4A, two arrows indicate the optimal selected edge biomarkers and node biomarkers, with cross-validation accuracies 96.98% ± 0.27% and 89.05% ± 0.39%, respectively. The selected 23 edge biomarkers and 24 node biomarkers are listed in Table 2, where the order of the biomarkers reflects the importance of these features by SFFS, such that important features tend to be at the top of the list. In Table 2, the arrows of edge biomarkers indicate the expression correlation change, while the arrows of node biomarkers indicate the expression level change from earlier to advanced disease (The heat maps of these two kinds of biomarkers are shown in Supplementary Figure S1). Note that the expression changes of all node biomarkers are significant; in contrast, edge biomarkers are identified from NDGs, thus the expression changes of genes for edge biomarkers are not significant. The independent validations of edge and node biomarkers have been performed on the GEO dataset, and the prediction accuracies are 72.24% and 70.62%, respectively.

To test whether the combination of edge and node biomarkers can improve the classification performance, we apply SFFS again on the union of edge biomarkers and node biomarkers, i.e. EdgeBiomarker. The result is shown in Figure 4B. The maximum accuracy of the combined biomarkers is slightly improved, and the corresponding union-features are almost the same as edge biomarkers, which indicate that edge biomarkers can indeed well classify earlier and advanced lung adenocarcinoma in this study.

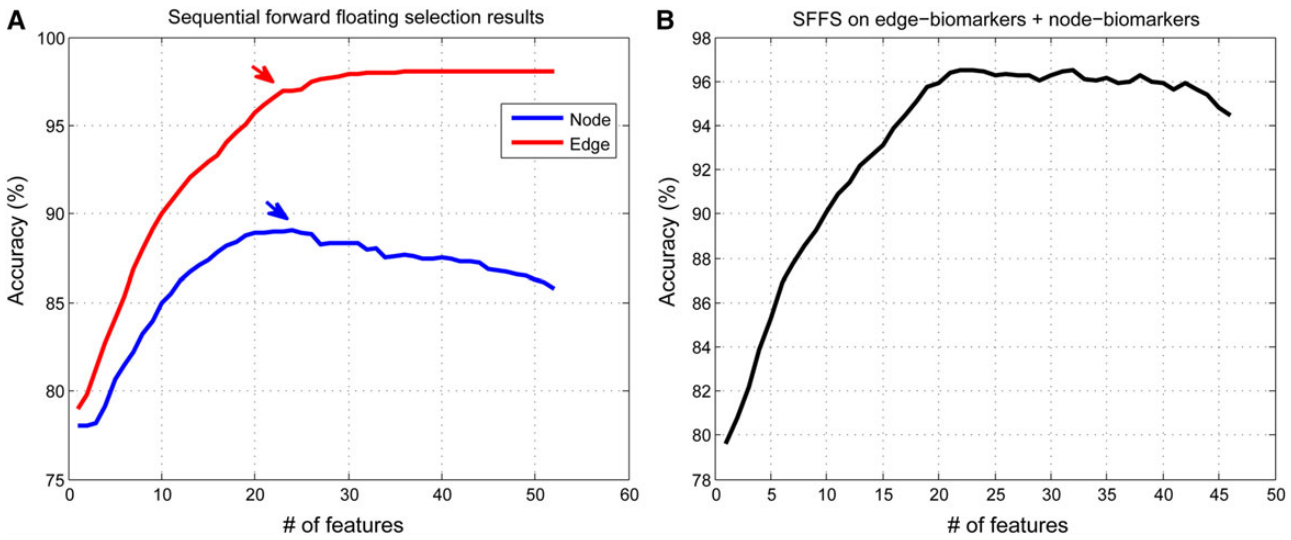

Figure 4 Optimized accuracy curves achieved with different numbers of features by EdgeBiomarker. (**A**) SFFS algorithm was applied to candidates of edge biomarkers (red curve) and node biomarkers (blue curve). (**B**) SFFS algorithm was applied to the union of edge biomarkers and node biomarkers.

Table 2 Selected edge features (gene pairs) and node features (individual genes) for classifying earlier and advanced lung adenocarcinoma.

Edge biomarkers	Correlation changing	Node biomarkers	Expression changing
ADORA1,ANKRD46	↗	ABCA6	↘
ATF6B,RPL26	↗	C1orf172	↘
DMGDH,SLC6A6	↗	FMNL2	↘
CLIC2,SLC4A1AP	↗	OR52E4	↘
DDX50,ZCCHC6	↗	C9orf64	↗
LOC283663,TMEM205	↗	GPSM3	↘
FAM38B,MMRN1	↘	C20orf107	↘
MACF1,TNNC1	↗	SIGLECP3	↘
CTF1,PGAP3	↗	NAPSB	↘
NUDT8,ULK4	↗	PSD4	↘
C18orf32,C8orf80	↗	ECD	↗
LOC100133612,PRDM8	↗	RNASE1	↘
ATP2B4,LOC134466	↗	TAS2R39	↘
MAMDC4,SPERT	↗	CLCNKB	↘
NEK3,PDCD7	↗	FAM65B	↘
NDEL1,TPCN2	↗	BDNFOS	↘
IMPA2,NLGN3	↗	FLJ33360	↘
DOHH,SLC27A3	↗	TRAFD1	↘
ETHE1,TMEM135	↗	PLEKHB1	↘
C14orf128,PIK3C3	↘	CRYGB	↘
PAPPA,SH2D2A	↗	GFRA1	↘
AGAP2,SP140	↘	ZNF167	↘
TRAP1,ZKSCAN1	↗	LPIN1	↘
		TXNDC11	↘

The symbols ↗ and ↘ mean the increase and decrease, respectively.

Edge biomarkers are enriched with transcriptional regulators and cancer/metastasis-related gene pairs

To investigate the biological significance of edge biomarkers, transcription factors (TFs) are mapped to marker gene-pairs. All predicted human TFs are downloaded from DBD, a transcriptional factor prediction database (Kummerfeld and Teichmann, 2006). There are in total 2078 TFs in DBD, among which 1258 TFs are detected in TCGA dataset. According to this TF list, 4 out of 46 genes of edge biomarkers are TFs (i.e. ATF6B, PRDM8, SP140, ZKSCAN1), whereas only 1 out of 24 node biomarkers is TF (i.e. ECD). This fact supports that edge biomarkers are more enriched with TFs, and thus related to the disease development or progression, e.g. cancer metastasis, as main regulators. Notably, a recent study has shown that the protein encoded by ZKSCAN1 can promote tumor metastasis by increasing the abilities of migration and invasion in gastric cancer cell (Li et al., 2014). The protein encoded by ATF6B is known to be a TF that participates in the unfolded protein response, which is known to protect tumor cells under hypoxia (Rouschop et al., 2010), and thus ATF6B may be a potential target in cancer therapy (Ma and Hendershot, 2004; Backer et al., 2011).

Since the edge biomarkers are a network of biomarkers, the network ontology analysis (NOA) was used to conduct enrichment analysis on their genes and gene associations, i.e. network structure (Wang et al., 2011). Supplementary Table S1 lists the annotated GO terms by using NOA tool. Remarkably, the association between TFs ATF6B and RPL26 was significantly annotated, and many works (Ofir-Rosenfeld et al., 2008; Chen et al., 2012a; Gazda et al., 2012; Li et al., 2012) have reported that RPL26 mediates the regulation of TP53, known as an important tumor-suppressor gene. In these functional annotations, actin binding and calcium ion binding are significantly enriched in our edge biomarker list. In fact, actin binding and calcium signaling are known to be involved in cancer progression and metastasis (Dos Remedios and Chhabra, 2008; Jiang et al., 2008; Yang et al., 2010; Prevarskaya et al., 2011), which suggests that the metastasis ability of cancer cell may be one of the important differences between earlier and advanced stages of diseases.

Edge biomarkers are significantly associated to KEGG pathways

To further systematically characterize the edge biomarkers at the biological pathway level, we carried out pathway enrichment analysis for edge biomarkers and node biomarkers, respectively. The pathway knowledge was downloaded from KEGG pathway database (http://www.genome.jp/kegg/pathway.html), which includes 284 human pathways. The hyper-geometric test was employed to test whether a pathway is significantly enriched by the edge or node biomarkers. Specifically, we mapped genes of the biomarkers to a given pathway to get the number of matched genes M. We denote the number of the background genes (i.e. the genes of the dataset that are left after data processing) G, the number of pathway genes detected by dataset N, and the number of genes of all biomarkers K. Then the probability of getting M match genes under null hypothesis that genes are randomly drawn is as follows:

$$P(M; G, N, K) = \frac{\binom{N}{M}\binom{G-N}{K-M}}{\binom{G}{K}}.$$

The P-value can be defined as

$$P\text{-value} = \sum_{m>M} P(m; G, N, K).$$

The significantly enriched pathways of edge biomarkers are listed in Table 3. Among these pathways, some are cancer-related or tumor metastasis-related according to other studies, e.g. sulfur metabolism (Ryu et al., 2011), cGMP-PKG signaling pathway (Li et al., 2013; Ren et al., 2014), inositol phosphate metabolism (Lee and Yuspa, 1991; Vucenik and Shamsuddin, 2003), phosphatidylinositol signaling system (Vivanco and Sawyers, 2002; Bunney and Katan, 2010), calcium signaling pathway (Parkash and Asotra, 2010; Yang et al., 2010), VEGF signaling pathway (Kowanetz and Ferrara, 2006; Waldner and Neurath, 2012), cell adhesion molecules (Paul et al., 1997; Weis and Cheresh, 2011), etc. The result of enrichment analysis on node biomarkers is listed in Supplementary Table S2 for comparison, which obviously contains less meaningful pathways for cancer. In addition, several genes from edge biomarkers are found to be located at the up-stream of pathways, e.g. phosphatidylinositol signaling system and cGMP-PKG signaling pathway as shown in Supplementary Figure S2. This would be the evidence that edge biomarkers are related to the causes or drivers of the disease development or progression.

Table 3 Significantly enriched pathways by KEGG pathway enrichment analysis on edge biomarkers.

ID	Pathway title	*P*-value	M	K	N	G	Genes included
hsa04261	Adrenergic signaling in cardiomyocytes	0.000374	3	46	103	13815	ATF6B, ATP2B4, TNNC1
hsa00920	Sulfur metabolism	0.00048	1	46	10	13815	ETHE1
hsa04972	Pancreatic secretion	0.000506	2	46	47	13815	ATP2B4, TPCN2
hsa04022	cGMP-PKG signaling pathway	0.000606	3	46	117	13815	ADORA1, ATF6B, ATP2B4
hsa00562	Inositol phosphate metabolism	0.000761	2	46	54	13815	IMPA2, PIK3C3
hsa04070	Phosphatidylinositol signaling system	0.001365	2	46	66	13815	IMPA2, PIK3C3
hsa04140	Regulation of autophagy	0.002625	1	46	23	13815	PIK3C3
hsa00260	Glycine, serine and threonine metabolism	0.00335	1	46	26	13815	DMGDH
hsa04020	Calcium signaling pathway	0.004458	2	46	100	13815	ATP2B4, TNNC1
hsa05030	Cocaine addiction	0.004743	1	46	31	13815	ATF6B
hsa04918	Thyroid hormone synthesis	0.007806	1	46	40	13815	ATF6B
hsa04024	cAMP signaling pathway	0.007912	2	46	123	13815	ADORA1, ATP2B4
hsa04970	Salivary secretion	0.008581	1	46	42	13815	ATP2B4
hsa05031	Amphetamine addiction	0.00898	1	46	43	13815	ATF6B
hsa04260	Cardiac muscle contraction	0.01066	1	46	47	13815	TNNC1
hsa04911	Insulin secretion	0.012004	1	46	50	13815	ATF6B
hsa04370	VEGF signaling pathway	0.012468	1	46	51	13815	SH2D2A
hsa05032	Morphine addiction	0.012468	1	46	51	13815	ADORA1
hsa05410	Hypertrophic cardiomyopathy (HCM)	0.0144	1	46	55	13815	TNNC1
hsa05414	Dilated cardiomyopathy	0.015412	1	46	57	13815	TNNC1
hsa04915	Estrogen signaling pathway	0.026446	1	46	76	13815	ATF6B
hsa04514	Cell adhesion molecules (CAMs)	0.03251	1	46	85	13815	NLGN3

Edge biomarkers reveal significant differential metastasis risk of the patients diagnosed in earlier and advanced stages

In TCGA datasets, patients' survival information was accessible. To examine the classification performance of edge biomarkers with respect to survival days, we evenly separate the dataset into training and prediction parts; classification model is learnt from training data and the prediction is performed on prediction data. Then the survival analysis is carried out on the predicted samples. The significance (*P*-value) of differences between survival curves of predicted earlier and advanced patients is also estimated. Figure 5 shows the survival curves for predicted earlier and advanced patients by using edge biomarkers (A), node biomarkers (B), and random assignment (C). The *P*-values of Cox proportional hazards regression model are 0.0041, 0.0170, and 0.6725, respectively. The results verify that edge biomarkers effectively identified patients in earlier or advanced stage, where the diagnosed patients in advanced stages exhibited worse prognosis.

Discussion

Essentially, EdgeBiomarker is a kernel-based method for detecting both edge biomarkers and node biomarkers. It may also suffer from 'over-fitting' problem and the 'curse of dimensionality', which are still two big challenges in machine learning field. Nevertheless, the proper pre-procession of data and careful selection of candidate features can overcome these problems to some degree. In this work, we applied EdgeBiomarker to the RNAseq datasets from TCGA database and tried to understand the pathogenic mechanism underlying earlier and advanced stages of cancer. For lung adenocarcinoma, we have identified edge biomarkers distinguishing patients in earlier stage from those in advanced stage. The classification performance of edge biomarkers is better than traditional node biomarkers (in both within-dataset cross-validation and independent-dataset validation). The functional analysis of edge biomarkers also reveals several relevant mechanisms underlying the metastasis risk of lung adenocarcinoma. Note that edge biomarkers and node biomarkers have weak prediction accuracy in the independent-dataset validations, which would be caused by the inconsistent types of datasets, where the TCGA dataset is from RNAseq data and the independent GEO dataset is from microarray data. Another case study on TCGA breast cancer dataset further supports the advantages of edge biomarkers, as shown in Supplementary material.

Particularly, focusing on the annotated gene-pair in NOA, i.e. ATF6B and RPL26, we have a hypothesis for lung cancer development and progression. L26, encoded by RPL26, is known to participate in the regulation of p53 by Mdm2. ATF6B encodes a TF that takes part in the unfolded protein response pathway. This gene-pair as one of the edge biomarkers indicates that there may be a changed crosstalk between two corresponding pathways. Our hypothesis is: p53 pathway and unfolded protein response pathway have no significant association in earlier stage of lung cancer, but they will have functional connection as pathway crosstalk when lung cancer is developed to the advanced stage (Figure 6). It means that the association between RPL26 and ATF6B would be potential 'drivers' of the lung adenocarcinoma development, whose appearance as biomarkers indicates the increasing risk of lung cancer metastasis.

The edge biomarkers (e.g. network biomarkers or DNBs; Zeng et al., 2014) explore network or edge information in both learning and predicting steps, which are networks of marker molecules and thus are essentially different from node biomarkers. With additional information on expression differences (e.g. differential expression variance/covariance), edge biomarkers can promote the prediction accuracy of phenotypes and reveal the markers-involved biological or pathogen mechanisms. By further integrating dynamic information of data, the dynamical edge biomarkers (e.g. DNB; Chen et al., 2012b) even have the ability to diagnose the 'pre-disease' or 'un-occurred disease' state before the

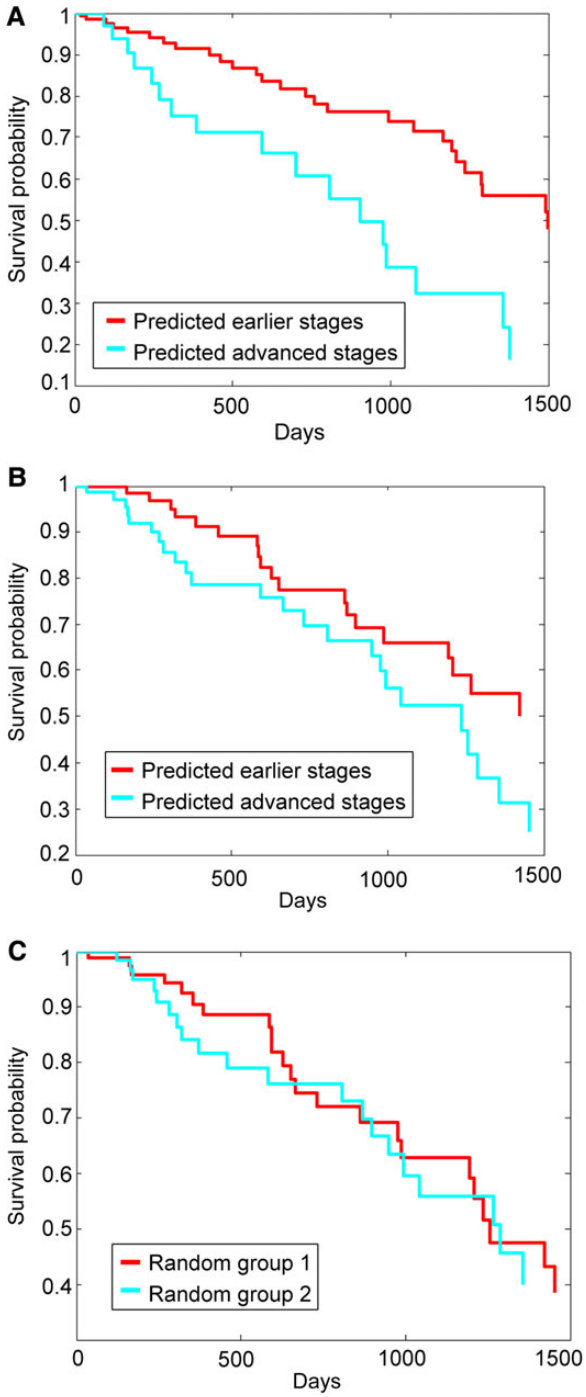

Figure 5 Survival analysis on the predicted patients using edge biomarkers (**A**), node biomarkers (**B**), and random assignment (**C**). Horizontal axes represent days elapsed from the diagnosis of the cancer, and vertical axes represent the proportion of patients that are still alive. Red curves are for patients predicted in earlier stages, and blue curves for those predicted in advanced stages.

occurrence of diseases. Actually, identifying and preventing 'un-occurred disease' state is also an important concept raised 2000 years ago in *Yellow Emperor's Medicine*, one of the earliest books for Traditional Chinese Medicine (Zeng et al., 2014). In

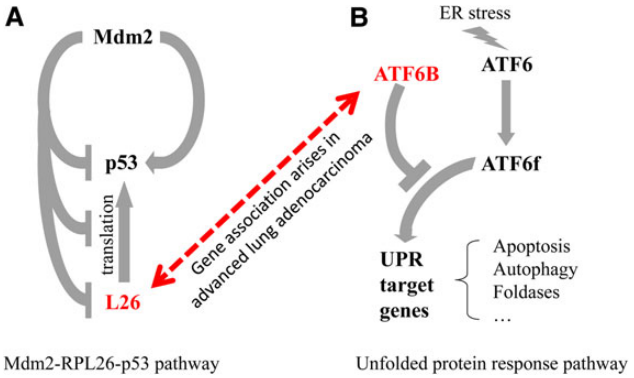

Figure 6 The hypothesis of crosstalk between Mdm2–RPL26–p53 pathway (**A**) and unfolded protein response (UPR) pathway (**B**) in advanced lung adenocarcinoma.

future, the edge biomarkers or dynamical edge biomarkers are expected to be extracted from big biological data for reliable characterization of complex diseases and biological processes.

In conclusion, edge biomarkers can extract differential correlations of gene expression changes, either from multiple samples or from single sample for an individual. Different from the conventional node biomarkers that representing genes with differential expressions, edge biomarkers further include genes with non-differential expressions but with differential correlations. In this study, EdgeBiomarker approach was applied to the TCGA cancer datasets to capture edge biomarkers from NDGs, in addition to node biomarkers. With either lung cancer or breast cancer datasets, edge biomarkers have shown a better performance in discrimination of progressive stages (i.e. earlier and advanced stages) of cancer than the conventional node biomarkers. Particularly, the EdgeBiomarker approach can measure the differential expression correlations in an individual sample, which enables the clinical application in individual patients, e.g. diagnosing diseases and the progressive stages for personalized medicine or precision medicine.

Supplementary material

Supplementary material is available at *Journal of Molecular Cell Biology* online.

Funding

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) (No. XDB13040700), the National Program on Key Basic Research Project (No. 2014CB910504), the National Natural Science Foundation of China (No. 91439103, 61134013, 31200987), and the Knowledge Innovation Program of SIBS of CAS (No. 2013KIP218).

Conflict of interest: none declared.

References

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

Backer, M.V., Backer, J.M., and Chinnaiyan, P. (2011). Targeting the unfolded protein response in cancer therapy. Methods Enzymol. *491*, 37–56.

Bunney, T.D., and Katan, M. (2010). Phosphoinositide signalling in cancer: beyond PI3K and PTEN. Nat. Rev. Cancer *10*, 342–352.

Chen, L., Wang, R.-S., and Zhang, X.-S. (2009). Biomolecular Networks: Methods and Applications in Systems Biology. Hoboken: Wiley.

Chen, J., Guo, K., and Kastan, M.B. (2012a). Interactions of nucleolin and ribosomal protein L26 (RPL26) in translational control of human p53 mRNA. J. Biol. Chem. *287*, 16467–16476.

Chen, L., Liu, R., Liu, Z.P., et al. (2012b). Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. Sci. Rep. *2*, 342.

Dos Remedios, C.G., and Chhabra, D. (2008). Actin-Binding Proteins and Disease. New York: Springer.

Gazda, H.T., Preti, M., Sheen, M.R., et al. (2012). Frameshift mutation in p53 regulator RPL26 is associated with multiple physical abnormalities and a specific pre-ribosomal RNA processing defect in diamond-blackfan anemia. Hum. Mutat. *33*, 1037–1044.

He, D., Liu, Z.P., Honda, M., et al. (2012). Coexpression network analysis in chronic hepatitis B and C hepatic lesions reveals distinct patterns of disease progression to hepatocellular carcinoma. J. Mol. Cell Biol. *4*, 140–152.

Herbst, R.S., Heymach, J.V., and Lippman, S.M. (2008). Lung cancer. N. Engl. J. Med. *359*, 1367–1380.

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

Hood, L., and Friend, S.H. (2011). Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat. Rev. Clin. Oncol. *8*, 184–187.

Jiang, P., Enomoto, A., Jijiwa, M., et al. (2008). An actin-binding protein Girdin regulates the motility of breast cancer cells. Cancer Res. *68*, 1310–1318.

Kowanetz, M., and Ferrara, N. (2006). Vascular endothelial growth factor signaling pathways: therapeutic perspective. Clin. Cancer Res. *12*, 5018–5022.

Kummerfeld, S.K., and Teichmann, S.A. (2006). DBD: a transcription factor prediction database. Nucleic Acids Res. *34*, D74–D81.

Lai, Y., Wu, B., Chen, L., et al. (2004). A statistical method for identifying differential gene-gene co-expression patterns. Bioinformatics *20*, 3146–3155.

Lee, E., and Yuspa, S.H. (1991). Changes in inositol phosphate metabolism are associated with terminal differentiation and neoplasia in mouse keratinocytes. Carcinogenesis *12*, 1651–1658.

Li, C., Ge, M., Yin, Y., et al. (2012). Silencing expression of ribosomal protein L26 and L29 by RNA interfering inhibits proliferation of human pancreatic cancer PANC-1 cells. Mol. Cell. Biochem. *370*, 127–139.

Li, N., Xi, Y., Tinsley, H.N., et al. (2013). Sulindac selectively inhibits colon tumor cell growth by activating the cGMP/PKG pathway to suppress Wnt/beta-catenin signaling. Mol. Cancer Ther. *12*, 1848–1859.

Li, Y., Tan, B.B., Zhao, Q., et al. (2014). ZNF139 promotes tumor metastasis by increasing migration and invasion in human gastric cancer cells. Neoplasma *61*, 291–298.

Listgarten, J., Neal, R.M., Roweis, S.T., et al. (2007). Difference detection in LC-MS data for protein biomarker discovery. Bioinformatics *23*, e198–e204.

Liu, R., Li, M., Liu, Z.P., et al. (2012a). Identifying critical transitions and their leading biomolecular networks in complex diseases. Sci. Rep. *2*, 813.

Liu, X., Liu, Z.P., Zhao, X.M., et al. (2012b). Identifying disease genes and module biomarkers by differential interactions. J. Am. Med. Inform. Assoc. *19*, 241–248.

Liu, R., Yu, X., Liu, X., et al. (2014). Identifying critical transitions of complex diseases based on a single sample. Bioinformatics *30*, 1579–1586.

Ma, Y., and Hendershot, L.M. (2004). The role of the unfolded protein response in tumour development: friend or foe? Nat. Rev. Cancer *4*, 966–977.

Mor, G., Visintin, I., Lai, Y., et al. (2005). Serum protein markers for early detection of ovarian cancer. Proc. Natl Acad. Sci. USA *102*, 7677–7682.

Ofir-Rosenfeld, Y., Boggs, K., Michael, D., et al. (2008). Mdm2 regulates p53 mRNA translation through inhibitory interactions with ribosomal protein L26. Mol. Cell *32*, 180–189.

Parkash, J., and Asotra, K. (2010). Calcium wave signaling in cancer cells. Life Sci. *87*, 587–595.

Paul, R., Ewing, C.M., Jarrard, D.F., et al. (1997). The cadherin cell-cell adhesion pathway in prostate cancer progression. Br. J. Urol. *79(Suppl 1)*, 37–43.

Prevarskaya, N., Skryma, R., and Shuba, Y. (2011). Calcium in tumour metastasis: new roles for known actors. Nat. Rev. Cancer *11*, 609–618.

Pudil, P., Novovicova, J., and Kittler, J. (1994). Floating search methods in feature-selection. Pattern Recognit. Lett. *15*, 1119–1125.

Ren, Y., Zheng, J., Yao, X., et al. (2014). Essential role of the cGMP/PKG signaling pathway in regulating the proliferation and survival of human renal carcinoma cells. Int. J. Mol. Med. *34*, 1430–1438.

Rouschop, K.M.A., van den Beucken, T., Dubois, L., et al. (2010). The unfolded protein response protects human tumor cells during hypoxia through regulation of the autophagy genes MAP1LC3B and ATG5. J. Clin. Invest. *120*, 127–141.

Ryu, C.S., Kwak, H.C., Lee, K.S., et al. (2011). Sulfur amino acid metabolism in doxorubicin-resistant breast cancer cells. Toxicol. Appl. Pharmacol. *255*, 94–102.

Sahni, N., Yi, S., Zhong, Q., et al. (2013). Edgotype: a fundamental link between genotype and phenotype. Curr. Opin. Genet. Dev. *23*, 649–657.

Sun, S.Y., Liu, Z.P., Zeng, T., et al. (2013). Spatio-temporal analysis of type 2 diabetes mellitus based on differential expression networks. Sci. Rep. *3*, 2268.

Tomida, S., Takeuchi, T., Shimada, Y., et al. (2009). Relapse-related molecular signature in lung adenocarcinomas identifies patients with dismal prognosis. J. Clin. Oncol. *27*, 2793–2799.

Vivanco, I., and Sawyers, C.L. (2002). The phosphatidylinositol 3-Kinase AKT pathway in human cancer. Nat. Rev. Cancer *2*, 489–501.

Vucenik, I., and Shamsuddin, A.M. (2003). Cancer inhibition by inositol hexaphosphate (IP6) and inositol: from laboratory to clinic. J. Nutr. *133*, 3778S–3784S.

Waldner, M.J., and Neurath, M.F. (2012). Targeting the VEGF signaling pathway in cancer therapy. Expert Opin. Ther. Targets *16*, 5–13.

Wang, J., Huang, Q., Liu, Z.P., et al. (2011). NOA: a novel Network Ontology Analysis method. Nucleic Acids Res. *39*, e187.

Wang, J., Sun, Y., Zheng, S., et al. (2013). APG: an Active Protein-Gene network model to quantify regulatory signals in complex biological systems. Sci. Rep. *3*, 1097.

Weis, S.M., and Cheresh, D.A. (2011). Tumor angiogenesis: molecular pathways and therapeutic targets. Nat. Med. *17*, 1359–1370.

Yang, H., Zhang, Q., He, J., et al. (2010). Regulation of calcium signaling in lung cancer. J. Thorac. Dis. *2*, 52–56.

Yu, X., Li, G., and Chen, L. (2014). Prediction and early diagnosis of complex diseases by edge-network. Bioinformatics *30*, 852–859.

Zeng, T., Sun, S.Y., Wang, Y., et al. (2013). Network biomarkers reveal dysfunctional gene regulations during disease progression. FEBS J. *280*, 5682–5695.

Zeng, T., Zhang, W., Yu, X., et al. (2014). Edge biomarkers for classification and prediction of phenotypes. Sci. China Life Sci. *57*, 1103–1114.

Zhang, W., Zeng, T., and Chen, L. (2014). EdgeMarker: identifying differentially correlated molecule pairs as edge-biomarkers. J. Theor. Biol. *362*, 35–43.