



EdgeMarker: Identifying differentially correlated molecule pairs as edge-biomarkers



Wanwei Zhang^a, Tao Zeng^{a,*}, Luonan Chen^{a,b,**}

^a Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

^b Collaborative Research Center for Innovative Mathematical Modelling, Institute of Industrial Science, The University of Tokyo, Tokyo 153-8505, Japan

ARTICLE INFO

Article history:

Received 15 January 2014

Received in revised form

22 May 2014

Accepted 27 May 2014

Available online 12 June 2014

Keywords:

Edge-biomarker

Differentially correlated molecular pair

Differential network

Single sample

Network biomarker

ABSTRACT

Biomarker discovery is one of the major topics in translational biomedicine study based on high-throughput biological data analysis. Traditional methods focus on differentially expressed genes (or node-biomarkers) but ignore non-differentials. However, non-differentially expressed genes also play important roles in the biological processes and the rewired interactions / edges among non-differential genes may reveal fundamental difference between variable conditions. Therefore, it is necessary to identify relevant interactions or gene pairs to elucidate the molecular mechanism of complex biological phenomena, e.g. distinguish different phenotypes. To address this issue, we proposed a new method based on a new vector representation of an edge, *EdgeMarker*, to (1) identify edge-biomarkers, i.e. the differentially correlated molecular pairs (e.g., gene pairs) with optimal classification ability, and (2) transform the 'node expression' data in node space into the 'edge expression' data in edge space and classify the phenotype of each single sample in edge space, which generally cannot be achieved in traditional methods. Unlike the traditional methods which analyze the node space (i.e. molecular expression space) or higher dimensional space using arbitrary kernel methods, this study provides a mathematical model to explore the edge space (i.e. correlation space) for classification of a single sample. In this work, we show that the identified edge-biomarkers indeed have strong ability in distinguishing normal and disease samples even when all involved genes are not significantly differentially expressed. The analysis of human cholangio-carcinoma dataset and diabetes dataset also suggested that the identified edge-biomarkers may cast new biological insights into the pathogenesis of human complex diseases.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Biomarker discovery is one of the major topics in the comparative high-throughput biomedical experiments of binary conditions (e.g. normal and disease) or multi-conditions (e.g. multiple stages of diseases). Traditional biomarker discovery methods mainly focus on the differentially expressed genes (DEGs) or node-biomarkers, leaving large amount of the non-differentially expressed genes (NDEGs) unexamined (Listgarten et al., 2007; Mor et al., 2005). However, it has been widely recognized that genes are working interactively in a network manner and the study of 'edgetics' can reveal the differences of interactions (Chen et al., 2009; Sahni et al., 2013; Yu et al., 2013) and also elucidate the molecular mechanism of complex biological phenomena at the

system level. In a recent review, (Sahni et al., 2013) raise the notion of 'edgetype' to link the genotype to phenotype. Actually, a considerable amount of evidence shows that 'edgetype' (or edgetics) of NDEGs can play key roles in altering the state of biological systems (Lai et al., 2004; Sun et al., 2013; Wang et al., 2013; Yu et al., 2013; Zeng et al., 2013). As a building element of a biological network, an interacting gene pair can exhibit positive or negative correlation observed in the gene expression profile. Changing of these gene correlations may result in the different states of a biological system (e.g. normal or disease state). Therefore, the differentially correlated gene pairs (DCPs) may well indicate the alteration of biological states even though the involved genes are not differentially expressed. In fact, some methods have been developed to identify the DCPs (He et al., 2012; Lai et al., 2004; Liu et al., 2012; Sun et al., 2013). They achieve some success to reveal potential mechanisms altering the biological states. However, these methods are unable to use the DCPs directly to distinguish or predict the state of a new single sample mainly because computing correlation requires multiple samples (or observations) and it is expensive or usually impractical to get multiple

* Corresponding author.

** Corresponding author at: Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China.

E-mail addresses: zengtao@sibs.ac.cn (T. Zeng), lnchen@sibs.ac.cn (L. Chen).

observations of high-throughput data for each individual. Some of these works avoid this problem by using the differently expressed genes of the identified DCPs to do the prediction or diagnosis, while others by using all genes of DCPs with arbitrary kernels. In any case, when referred to diagnosing, these methods are still node-based (node-biomarkers), and not specific to the identified DCPs or edges (edge-biomarkers), which would miss-lead the diagnosis. In other words, the edge or correlation information of each tested sample is not used in the prediction or diagnosis. Notably, Top-scoring pairs (TSP) and TSP-based machine learning algorithms had been proposed (Geman et al., 2004; Shi et al., 2011). However they simply considered the rank difference of gene pairs, of which the biological meaning was implicit, which hampered their application.

On the other hand, in the field of machine learning, one of the major topics is variable and feature selection (Guyon et al., 2003), i.e., to select a subset of features that has the optimal classification ability. Based on different learning machines, various feature selection methods were developed. Generally, there are three kinds: filter, wrapper and embedded methods. Filters assess the relevance of each feature individually to the target label (or response) and select the most relevant features. Wrappers use the classification performance to score the subset of the features and select the one with optimal classification ability. Embedded methods such as LASSO select features while training classifier. As illustrated in (Guyon et al., 2003), these methods have their own advantages and disadvantages. Filters are computation-efficient but do not take into account the combined effects of features. Wrappers are subset-based but computation-inefficient. Embedded methods consider the combined effects of features and some of them can be computation-efficient but they suffer from overfitting problems in the domains with large amount of input variables. For a selected feature set, the prediction performance is mainly affected by the kernel that maps the original data into higher dimensional space. Mostly, the kernel is chosen by blind searching and the physical meaning is implicit. Therefore, the reliability of prediction or diagnosis is weakened. To address these issues in this paper, we proposed a new vector representation of an edge and adopted the following strategy: first map the original data into a proper higher dimensional space with each feature corresponding to a vector of components of Pearson correlation coefficient (PCC), and then the edge-biomarker discovery is conducted on this new space by a machine learning method. Such a representation for features not only has strong physical meaning (i.e., correlation) but also can achieve the prediction for a single sample with the information of edges. Specifically, we map the 'node' data (molecule expression profile) in the 'node' space into the 'edge' space where each value represents a component of the 'correlation expression' of molecule pairs. Thus, all of those variables as a whole represent the information of Pearson correlation coefficient (PCC) or edge. If we take the mean value of the 'correlation expression' values of a given gene pair in a given class (e.g., control or case group), we can get the corresponding correlation coefficient. In such a way, the prediction or diagnosis of phenotypes on a single sample can be achieved by directly using the correlation or edge information, rather than traditional node information.

In this study, we mainly develop a method based on a new representation of an edge, *EdgeMarker*, to identify a subset of differentially correlated molecular pairs (DCPs) whose expression correlation information can be directly used to diagnose the phenotype of a single sample. These differentially correlated molecular pairs are called edge-biomarkers. Although each pair of genes connected to an edge-biomarker can be differentially expressed genes (DEGs) or non-differentially expressed genes (NDEGs), in this work, we particularly focus on NDEGs which are

ignored by the traditional methods. In the analysis of human cholangiocarcinoma dataset and human type 1 diabetes dataset, we found that the NDEGs indeed contain rich information, not in terms of their individual gene expressions but in terms of their interactions. The identified edge-biomarkers not only distinguish the phenotypes of each single sample accurately but also provide new insights into the pathogenesis of those complex diseases.

2. Method

Since a particular biological function can be considered to be facilitated by a cascade of biochemical reactions or a network, it is straightforward to identify the biomarkers (markers that indicate the difference between biological states) from differentially expressed molecules in a high-throughput comparative biomedical study. However, it is well-known that the genetic backgrounds of within-class samples are heterogeneous and the biochemical reactions are stochastic (Raser and O'Shea, 2005). Therefore, the identified biomarkers are often highly false-discovered and the true markers are poorly recalled. These node-based biomarkers (or node-biomarkers) are usually disconnected individuals. Yet on the other hand, a biological function is generally achieved by a group of cooperative molecules or a network of interactively-working molecules. Thus it is reasonable to introduce the edge-based biomarkers (or, edge-biomarkers, network biomarkers) which utilize the interaction rewiring or differential correlations to distinguish disease samples from normal samples. Although many network-based or edge-based methods have been developed to identify the discriminative networks or edges, they only use the genes or proteins (or gene set) involved in those networks or edges for the prediction or classification of each single sample, rather than using networks or edges (or correlations) themselves. Thus, they are basically node-biomarker based methods. There are no effective methods to directly use edges or correlations for classifying a single sample for normal and disease groups. One major difficulty is that the computation of correlation coefficient requires several samples. To overcome this problem, in this paper, we develop a novel method based on a new vector representation of an edge (1)–(2), *EdgeMarker*, to identify the differentially correlated molecular pairs as the edge-biomarkers, which can in turn be directly used for the diagnosis and prediction of a single tested sample. The method is described in the following sections.

The following abbreviations are used throughout this article: DEG (statistically differentially expressed genes), NDEG (statistically non-differentially expressed genes), DCP (differentially correlated gene pairs), SD (standard deviation), PCC (Pearson's correlation coefficient), SFFS (sequential forward floating selection), TSP (top-scoring pairs), and SVM (support vector machine)

2.1. Pre-excluding the genes with low expression variation among within-class samples

With the premise that gene expression bears noise of unknown magnitude, the reliability of PCCs that are computed from genes of different magnitude of within-class variation is fairly different. The lower the magnitude of within-class variation, the more likely the computed correlation coefficient is affected by noise. Therefore, to get the high confident correlations, it is necessary to pre-exclude the genes of low within-class variation. Because the magnitude of the expression noise is basically unknown, we simply exclude the genes with lowest 20% of standard deviation (SD) in any classes (e.g., normal samples or disease samples). Specifically, for normal and disease groups, we sort the genes decreasingly by their within-class SD respectively; select the top 80% of genes for both classes and take the overlapped genes for next analysis.

2.2. Selecting statistically non-differentially expressed genes (NDEGs) and differentially correlated gene pairs (DCPs)

Although *EdgeMarker* can be applied to both DEGs and NDEGs, to show the capability of edges on diagnosis, this work particularly focuses on NDEGs which were thought to be informationless and were discarded in most biomarker studies. To get NDEGs, the p -values of t -test between disease and normal samples are calculated for each of the genes filtered. Those genes with p -value > 0.6 are collected as NDEGs. Then the PCCs of all possible pairs of NDEGs are computed for normal and disease groups respectively. Those gene pairs with absolute difference of $|PCC| > 0.9$ are chosen as differentially correlated gene pairs (DCPs).

Suppose there are k NDEGs, and the sample sizes of normal and disease groups are m and n respectively. Let $x_i \in R^m$ and $y_i \in R^n$ denote the expression vectors of i^{th} NDEG for all normal samples (x) and all disease samples (y), respectively, i.e., x_{ik} is the expression of i^{th} NDEG for the k^{th} sample in normal group while y_{ik} is the expression of i^{th} NDEG for the k^{th} sample in disease group. A gene pair $\langle u, v \rangle$ is said to be differentially correlated if and only if $|\rho_{uv}^x - \rho_{uv}^y| > \delta$ where ρ_{uv}^x and ρ_{uv}^y are the PCCs between genes u and v of normal (x) and disease (y) respectively, and δ is set to 0.9 in our work.

2.3. Transforming data from node space (gene expressions) into edge space (correlation components)

In this section, we propose a new vector form to represent an edge, which can make the classification of a single sample possible with correlation information. After selecting the DCPs, we can map the original expression data into the higher-order space where the correlation of each gene pair can be depicted by two coupled features. For convenience, we call the data before and after this transformation as node data and edge data, respectively. Suppose gene pair $\langle u, v \rangle$ is a DCP. Then the transformation of gene expressions of u and v by (1) into edge space by (2) is as follows:

$$\begin{matrix} \text{Gene } u: & (x_{u1} & \dots & x_{um} & | & y_{u1} & \dots & y_{un}) \\ \text{Gene } v: & (x_{v1} & \dots & x_{vm} & | & y_{v1} & \dots & y_{vn}) \end{matrix} \rightarrow \quad (1)$$

$$\begin{matrix} \langle u, v \rangle_N: & \left(\frac{x_{u1} - \bar{x}_u}{k_1 \cdot Sx_u}, \frac{x_{v1} - \bar{x}_v}{k_1 \cdot Sx_v}, \dots, \frac{x_{um} - \bar{x}_u}{k_1 \cdot Sx_u}, \frac{x_{vm} - \bar{x}_v}{k_1 \cdot Sx_v}, \dots, \frac{y_{u1} - \bar{y}_u}{k_2 \cdot Sy_u}, \frac{y_{v1} - \bar{y}_v}{k_2 \cdot Sy_v}, \dots, \frac{y_{un} - \bar{y}_u}{k_2 \cdot Sy_u}, \frac{y_{vn} - \bar{y}_v}{k_2 \cdot Sy_v} \right) \\ \langle u, v \rangle_D: & \left(\frac{x_{u1} - \bar{y}_u}{k_2 \cdot Sy_u}, \frac{x_{v1} - \bar{y}_v}{k_2 \cdot Sy_v}, \dots, \frac{x_{um} - \bar{y}_u}{k_2 \cdot Sy_u}, \frac{x_{vm} - \bar{y}_v}{k_2 \cdot Sy_v}, \dots, \frac{y_{u1} - \bar{y}_u}{k_2 \cdot Sy_u}, \frac{y_{v1} - \bar{y}_v}{k_2 \cdot Sy_v}, \dots, \frac{y_{un} - \bar{y}_u}{k_2 \cdot Sy_u}, \frac{y_{vn} - \bar{y}_v}{k_2 \cdot Sy_v} \right) \end{matrix} \quad (2)$$

where m and n are the numbers of samples for normal and disease groups respectively; $\bar{x}_u, \bar{x}_v, \bar{y}_u, \bar{y}_v$ are the mean expressions of genes u and v from normal and disease groups; Sx_u, Sx_v, Sy_u, Sy_v are the standard expression deviations of genes u and v from normal and disease groups; $k_1 = \sqrt{(m-1)/m}$ and $k_2 = \sqrt{(n-1)/n}$ are two adjusting factors. By considering both normal and disease groups, (1) is data of each pair of genes in node space while (2) represents each pair of genes in edge space. Here, in contrast to the traditional method, we represent an edge by a vector form (2), which enables us to classify a single sample by using edge information.

Note that for each DCP $\langle u, v \rangle$, there are two coupled edge features $\langle u, v \rangle_N$ and $\langle u, v \rangle_D$ corresponding to it. The difference between these two features lies in that different means and standard deviations are used to center and scale the expression values. Thus from (2), we have a matrix M of all gene pairs for the normal and the disease groups with each two rows as $\langle u, v \rangle_N$ and $\langle u, v \rangle_D$, and the dimension of M is $2d(m+n)$ where d is the total number of DCPs.

Through simple calculation of (2), the summations for the left row vector $\langle u, v \rangle_N$ of the normal group

$$\left(\frac{x_{u1} - \bar{x}_u}{k_1 Sx_u}, \frac{x_{v1} - \bar{x}_v}{k_1 Sx_v}, \dots, \frac{x_{um} - \bar{x}_u}{k_1 Sx_u}, \frac{x_{vm} - \bar{x}_v}{k_1 Sx_v} \right)$$

and for the right row vector $\langle u, v \rangle_D$ of the disease group

$$\left(\frac{y_{u1} - \bar{y}_u}{k_2 Sy_u}, \frac{y_{v1} - \bar{y}_v}{k_2 Sy_v}, \dots, \frac{y_{un} - \bar{y}_u}{k_2 Sy_u}, \frac{y_{vn} - \bar{y}_v}{k_2 Sy_v} \right)$$

are equal to the PCCs of gene pair $\langle u, v \rangle$ in normal and disease groups respectively, i.e.

$$\frac{1}{m} \sum_{l=1}^m \left(\frac{x_{ul} - \bar{x}_u}{k_1 Sx_u} \frac{x_{vl} - \bar{x}_v}{k_1 Sx_v} \right) = \rho_{uv}^x$$

and

$$\frac{1}{n} \sum_{l=1}^n \left(\frac{y_{ul} - \bar{y}_u}{k_2 Sy_u} \frac{y_{vl} - \bar{y}_v}{k_2 Sy_v} \right) = \rho_{uv}^y$$

Similarly, it can be seen that the summations for the upper right and the lower left vectors for all samples of (2) are just shifting and scaling of PCCs of $\langle u, v \rangle$, i.e.

$$\frac{1}{n} \sum_{l=1}^n \left(\frac{y_{ul} - \bar{x}_u}{k_1 Sx_u} \frac{y_{vl} - \bar{x}_v}{k_1 Sx_v} \right) = A \rho_{uv}^y + B$$

and

$$\frac{1}{m} \sum_{l=1}^m \left(\frac{x_{ul} - \bar{y}_u}{k_2 Sy_u} \frac{x_{vl} - \bar{y}_v}{k_2 Sy_v} \right) = C \rho_{uv}^x + D$$

where

$$A = \frac{k_2^2 Sy_u Sy_v}{k_1^2 Sx_u Sx_v}, \quad B = \frac{(\bar{y}_u - \bar{x}_u)(\bar{y}_v - \bar{x}_v)}{k_1^2 Sx_u Sx_v}, \quad C = \frac{k_1^2 Sx_u Sx_v}{k_2^2 Sy_u Sy_v},$$

$$D = \frac{(\bar{x}_u - \bar{y}_u)(\bar{x}_v - \bar{y}_v)}{k_2^2 Sy_u Sy_v}.$$

In this sense, the node data of (1) is mapped to the coupled edge data of (2). Then we can perform feature selection methods on this edge data and use learning machines to do classification or prediction. It has advantages over traditional methods in that the features selected on edge space are exactly specific to the correlation information, whereas traditional methods select features on node space or other space with implicit physical meaning by blind searching of the kernels.

2.4. Selecting discriminative edge features with top Fisher's discriminant score

After transforming the expression data in node space (1) into the correlation data in edge space (2), we apply any machine learning method for feature selection from (2). In this work, we use the feature selection pipeline which combines filter and wrapper methods to explore their respective advantages. Specifically we first utilize a filter method to narrow down the candidate set of edge-biomarkers and then use a wrapper method to select the edge-biomarkers from the filtered candidate set to achieve the best prediction power. For filter method, we choose the Fisher's discriminant score (Bishop, 1995) to rank the edge features decreasingly and get n_{top} (100 in this work) features on top of the rank list. The Fisher's discriminant score of feature i between disease (+) and normal (−) is given as follows,

$$f_i = \frac{(\mu_i(+)-\mu_i(-))^2}{\sigma_i(+)^2 + \sigma_i(-)^2}$$

where $\mu_i(+)$ and $\mu_i(-)$ are the means of the i^{th} feature in 2 different sample groups (i.e., disease group + and normal group −).

-); $\sigma_i(+)$ and $\sigma_i(-)$ are the standard deviations in 2 different sample groups. For wrapper method, we employ sequential forward floating selection (SFFS) (Pudil et al., 1994), which is described in the following section.

2.5. Identifying edge biomarkers by sequential forward floating selection (SFFS)

Suppose the complete set of edge features is $F = \{f_i | i = 1, \dots, D\}$. $F_k \subset F$ is a subset (or selection) of k features. $J(F_k) \in R$ is the criterion function to evaluate efficiency of a selection. Here we define $J(F_k)$ as the leave-one-out cross validation accuracy when using F_k to train the linear support vector machine. The cross validation accuracy is measured by area under the ROC curve (AUC). We used linear SVM because the edge data are of higher-dimension and linear SVM is less sensitive to overfitting. Suppose f^+ and f^- are features to be added and removed, respectively. Sequence of selections, $S = \{F_k | k = 1, \dots, K\}$, are stored in S .

Then, from all of the edge features, we can use any machine learning method to classify the two groups of samples by identifying the best edge-biomarkers. For instance, the forward floating method can be used to conduct such a procedure, which mainly includes two steps: inclusion step and exclusion step. In the inclusion step, add edge features to the set of candidate edge-biomarkers one by one until $J(F_k)$ cannot be improved anymore, while in the exclusion step, delete edge features from the set of candidate edge-biomarkers one by one until $J(F_k)$ cannot be improved anymore, and repeat the inclusion and exclusion steps until the computation for $J(F_k)$ is converged, and then we have the set of the best edge-biomarkers.

In this paper, we particularly use the SFFS algorithm shown below to train the classifier. Intuitively, the SFFS implements elimination every time when adding a new feature into the set of candidate edge-biomarkers. For more details, one can refer to (Pudil et al., 1994). The set of the edge features or candidate edge-biomarkers with maximum classification accuracy is selected as edge-biomarkers. If the optimal solutions are not unique, select the smallest feature set based on the principle of sparseness.

SFFS Algorithm for identifying edge biomarkers:

Step 1 (Initiation). Set $F = \{f_i | i = 1, \dots, D\}$, $F_k = \phi$, $k = 0$.

Repeat {Step 2; $k = k + 1$; } until $k = D$.

Step 2 (Inclusion). If $k = D$, terminate the algorithm and output $\{F_k | k = 1, \dots, D\}$, else find the best feature f^+ from $F - F_k$, i.e. $f^+ = \arg \max_{f_i \in F - F_k} J(F_k + f_i)$, and set $F_{k+1} = F_k + f^+$.

Step 3 (Conditional exclusion). Find the worst feature f^- from F_{k+1} , i.e. $f^- = \arg \max_{f_i \in F_{k+1}} J(F_{k+1} - f_i)$. If $f^- = f^+$ then set $k = k + 1$

and return to Step 2, else set $F'_k = F_{k+1} - f^-$. Now if $k = 2$ then set $F_k = F'_k$ and return to Step 2, else go to Step 4.

Step 4 (Continuation of exclusion). Find the worst feature f^- from F'_k , i.e. $f^- = \arg \max_{f_i \in F'_k} J(F'_k - f_i)$. If $J(F'_k - f^-) \leq J(F_{k-1})$, set

$F_k = F'_k$ and return to Step 2, else set $F'_{k-1} = F'_k - f^-$, $k = k - 1$. Now if $k = 2$, set $F_k = F'_k$ and return to Step 2, else repeat Step 4.

2.6. Diagnosing phenotype for each single sample

After the edge-biomarkers are identified and the classifier is built, we can predict the phenotype of a new sample by transforming the involved genes' expression of any new sample into the edge data by formula (2). Specifically, suppose that the two edge-biomarkers are selected as $\{<g_1, g_2>_N, <g_2, g_3>_D\}$, and the expression values of the three genes g_1 , g_2 and g_3 for a new single sample are z_1 , z_2 and z_3 . Then, the edge data of the new sample for classification with the

two edge-biomarkers are as follows

$$\begin{aligned} <g_1, g_2>_N : \begin{pmatrix} z_1 - \bar{x}_1 & z_2 - \bar{x}_2 \\ k_1 Sx_1 & k_1 Sx_2 \end{pmatrix} \\ <g_2, g_3>_D : \begin{pmatrix} z_2 - \bar{y}_2 & z_3 - \bar{y}_3 \\ k_2 Sy_2 & k_2 Sy_3 \end{pmatrix} \end{aligned}$$

where $\bar{x}_1$, Sx_1 , $\bar{x}_2$, Sx_2 are the mean expression values and standard deviations of g_1 and g_2 in the normal group of training data, $\bar{y}_2$, Sy_2 , $\bar{y}_3$, Sy_3 are the mean expression values and standard deviations of g_2 and g_3 in the disease group of training data. By inputting the above edge data (i.e., (2)) into the trained classifier, we are able to predict the phenotype of the new sample by EdgeMarker. The computational procedure is shown in Fig. 1.

3. Results and discussion

3.1. Detecting differentially correlated pairs as biomarkers

It has been widely accepted that genes are performing their functions by working interactively with other molecules and these interactions or relationships can be abstracted as edges of a biomolecular network. In translational biomedical area, there were a large number of research works in finding biomarkers of human diseases. Many of these works considered biomarkers as isolated molecules while others considered the combinatorial effects of biomarkers by using blind-searched kernels of which biological meanings are not explicit. A typical way of those works is to adopt network concept to identify the edges by comparing disease and normal samples, but only the gene set involved in those edges without the interaction information is used to distinguish or diagnose the disease samples rather than edges themselves, and thus those works are actually network-based biomarkers not network biomarkers (Chen et al., 2012; Liu et al., 2013). They achieved great success in the studies of monogenic diseases but suffered a setback in the analysis of complex human diseases whose pathogenesis is multi-genetic. To face the challenge of complex human diseases, we turn to edge-biomarkers that characterize the rewiring of gene-gene interactions (or differential correlations) and develop a new method, *EdgeMarker* by using differential correlations to diagnose each sample. Fig. 1 shows the differences of traditional biomarker discovery methods (A1–A4) and our *EdgeMarker* (B1–B7). In comparative high-throughput experiments, ten thousands of genes are measured in parallel for a collection of normal samples and a collection of disease samples. A typical expression profile is generated as in (A1) where rows are genes and columns are biological samples of normal and disease. For traditional methods, statistically differentially expressed gene (DEG) set is usually selected (A2). Then biomarkers are chosen from the DEG set by applying feature selection methods (A3) and further a classifier is built. To classify a new sample, the expression values of marker genes are input to the classifier and the predicted label is output (A4). For our *EdgeMarker*, differentially correlated gene pairs (DCPs) are selected by the process described in Section 2.2 (B1, B2). The expression table of all genes that are involved in the DCPs is (B3) which is a small part of (A1). This expression table is transformed, as described in Section 2.3, into edge space (B4) where rows are gene pairs and columns are samples by (1)–(2). Then the discriminative DCPs (DDCPs) are selected by filter method of Fisher's discriminant score (B5). After DDCPs are selected, the edge-biomarkers can be identified by sequential forward floating selection (SFFS) (B6) and further the classifier can be trained. For classifying a new single sample, the edge data are generated by the method described in Section 2.6 and then are input to the classifier (B7) for diagnosis.

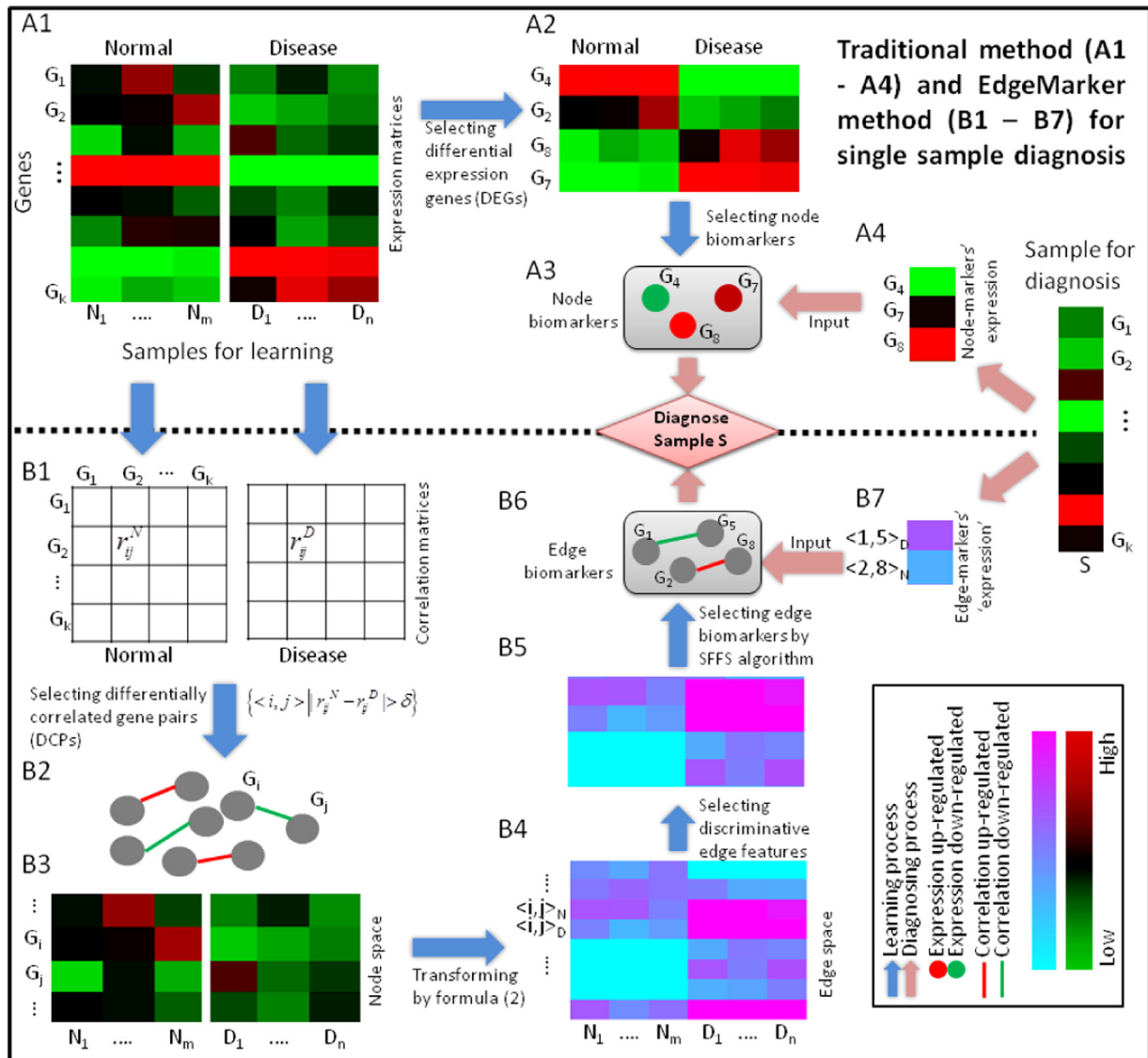


Fig. 1. Flow of traditional biomarker discovery methods (A1–A4) and *EdgeMarker* method (B1–B7) for single sample diagnosis. (A1) Comparative gene expression profile of binary conditions was generated from microarray or other high-throughput technologies. Rows are genes and columns samples. (A2) differentially expressed genes are selected by two-sample *T*-test or other statistical tests. (A3) Feature selection methods were utilized to refine the biomarkers with optimal prediction power. Red nodes represent up-regulated genes when comparing disease to normal samples, and green nodes are down-regulated genes. (A4) The expression values of marker-genes were extracted from a new sample to predict or diagnose the state of that sample. (B1) Gene–gene correlation matrices of normal and disease state were computed from expression profile. (B2) Differentially correlated gene pairs were selected by $\{ \langle i, j \rangle \mid |r_{ij}^N - r_{ij}^D| > \delta \}$, where r_{ij}^N and r_{ij}^D are Pearson's correlation coefficients of gene i and gene j under normal and disease state respectively. We set $\delta = 0.9$ in this work. Red edges represent up-regulated correlations when comparing disease to normal samples, while green edges are down-regulated correlations. (B3) Extract the involved expressions of all genes in the differentially correlated pairs. (B4) Transform the involved gene expressions into edge space by Eqs. (1),(2). In edge space, rows are gene pairs (edge-feature) and columns are samples. (B5) Select significant DCPs by statistically testing the difference of edge-data between normal and disease samples. (B6) Sequential forward floating (SFF) algorithm was employed to select the edge-biomarkers with optimal prediction power. (B7) Edge-markers 'expressions' were extracted from a new single sample to predict the phenotype of that sample.

3.2. Analysis on human cholangiocarcinoma dataset

We applied our method to analyze a human cholangiocarcinoma dataset with accession ID GSE26566 from GEO DataSets (Andersen et al., 2012). The downloaded expression table was processed by Genome Studio v2010, the background noise was subtracted, and the data was quantile normalized. There are 163 samples in total, where 104 samples are tumors and 59 samples matched non-cancerous surrounding liver samples. 1312 NDEGs were selected with the criterion of p -value > 0.6 . For these NDEGs, 538 gene pairs are differentially correlated under the criterion of $|\Delta PCC| > 0.9$ (Fig. 2A). Top 100 discriminative edge features were selected from the 1076 coupled edge features of 538 DCPs. For these features, the expression profile of the involved NDEGs (Fig. 2B) and the edge data

(Fig. 2C) are shown in heat maps which were hierarchically clustered using correlation distance. The opposite patterns between normal and disease are clear in edge data but not in the gene expression data due to NDEGs, which suggests that the edge features we selected are reasonable. Then we performed SFFS on the top 100 discriminative edge features and on the involved 88 NDEGs. The optimizing curves are shown in (Fig. 2A, 'Edge-NDEG' and 'Node-NDEG'). Fig. 2B is the network view of the top 100 edge features. The edge-biomarkers are highlighted in bold lines. Interestingly, many of the edge-biomarkers are connected to hub genes, which indicate that the reversion of some backbone functions may account for the canceration. Note that in this network, a coupled edge features are presented as one edge (see Supplementary Information for the full list of gene pairs). For method comparison and evaluation to traditional molecular biomarkers (or

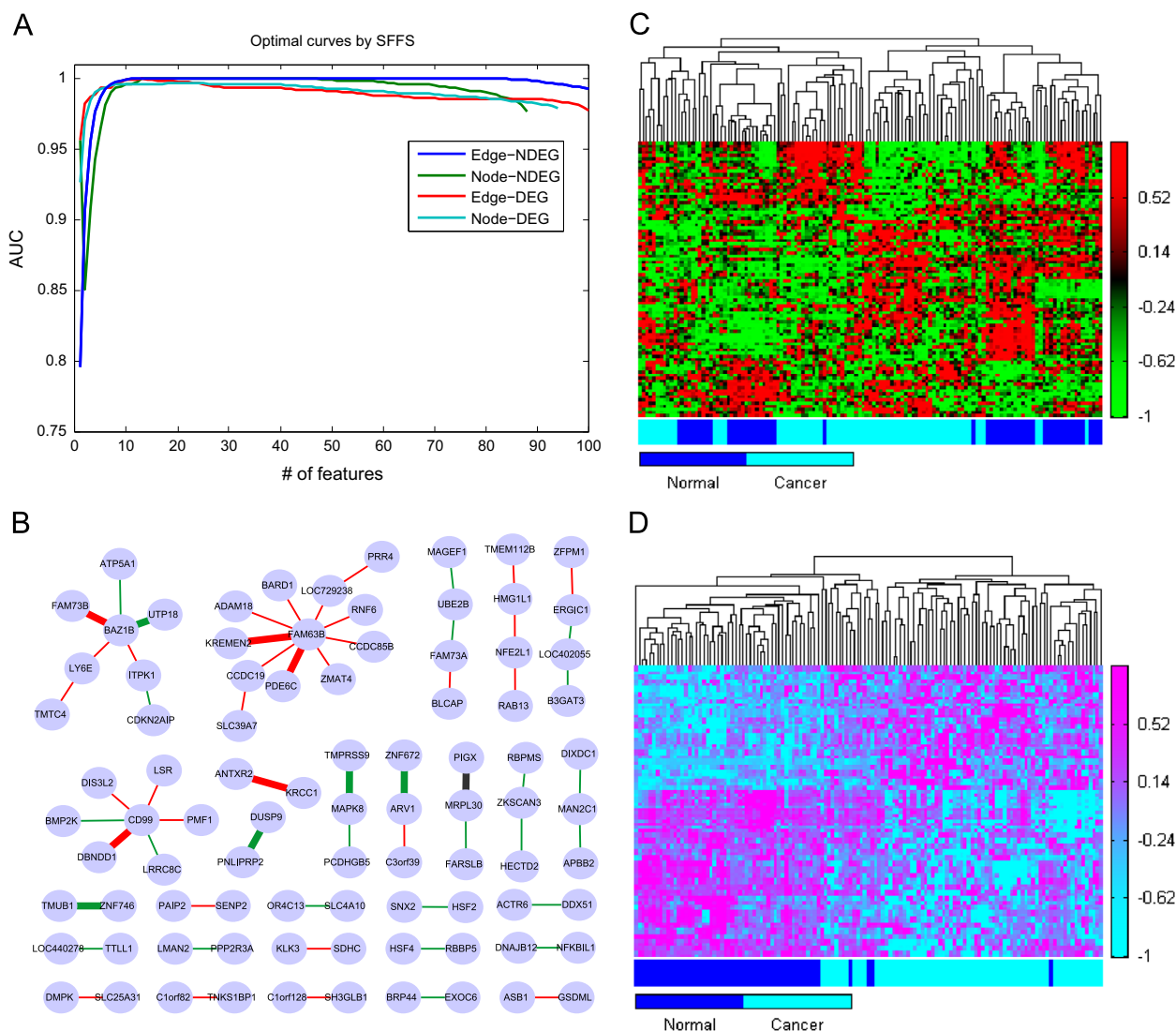


Fig. 2. Results of the cholangiocarcinoma dataset analysis. (A) For NDEG and DEG, the optimizing paths by SFFS methods on the top 100 edge features and the involved genes are plotted. (B) The discriminative edge features are presented, where bold lines are chosen as edge-biomarkers. Green lines represent down-regulated correlations, while red lines up-regulated correlations. (C) Heatmap (expression heatmap) of the genes involved in the discriminative edge features. (D) Heatmap of the edge data transformed by formula (2).

node-biomarkers), the same pipelines are used to analyze DEGs. The optimizing curves are also shown in Fig. 2A ('Edge-DEG' and 'Node-DEG').

Seeing Fig. 2A, due to the good quality of the dataset, the difference between these optimizing curves is significant. But according to the trend of the tails, the overall performance of *EdgeMarker* on NDEGs is better, while *EdgeMarker* on DEGs has the advantage in selecting sparse features (the early part of the 'Edge-DEG' curve raises most rapidly than other curves). Considering that the edge-biomarkers chosen from NDEGs by *EdgeMarker* are performed relatively well, it is an impressive result because those genes or edges are generally ignored by traditional methods and were thought to have no information to distinguish disease and normal samples.

Among the top 100 edge features, three genes (BLCAP, LY6E and RBBP5) are cholangiocarcinoma related according to *GeneCards* database. Notably, the hub gene CD99 was reported as cancer related gene (Edlund et al., 2012; Hong et al., 2012; Rajagopalan et al., 2013). The hub gene BAZ1B was reported as a versatile transcription factor whose functions include chromatin assembly, DNA repair, and RNA polymerase I and III gene regulation (Barnett and Krebs, 2011). By literature search, The hub gene FAM63B

(Coulouarn et al., 2012), DBNDD1 (Andersen et al., 2012), MAPK8 (Borad et al., 2014), ARV1 (Calin et al., 2008), and KREMEN2 (Maehata et al., 2008) were reported as cholangiocarcinoma or cancer related. These suggest that the edge biomarkers we found are highly related to the disease.

3.3. Analysis on human diabetes dataset

A human diabetes dataset (GSE9006, GEO DataSets) (Kaizer et al., 2007) was also analyzed. From this dataset, we extracted 24 healthy samples and 43 newly diagnosed type I diabetes (T1D) samples. About 1600 NDEGs were selected with the criterion of $P\text{-value} > 0.8$. Among these NDEGs, there are 365 DCPs whose absolute differences of PCC between healthy and T1D groups are > 1.0 . We selected top 100 edge features based on the rank of Fisher's discriminant score as discriminative edge features (see *Supplementary Information* for the full list of gene pairs). Then the SFFS was employed to select the edge-biomarkers. The results are shown in Fig. 3, where edge data or edge expressions in Fig. 3D are the matrix represented by (2). Basically they are consistent with the result of Fig. 2 except that the network niche of the edge-biomarkers are quite different: i.e., for cholangiocarcinoma, the

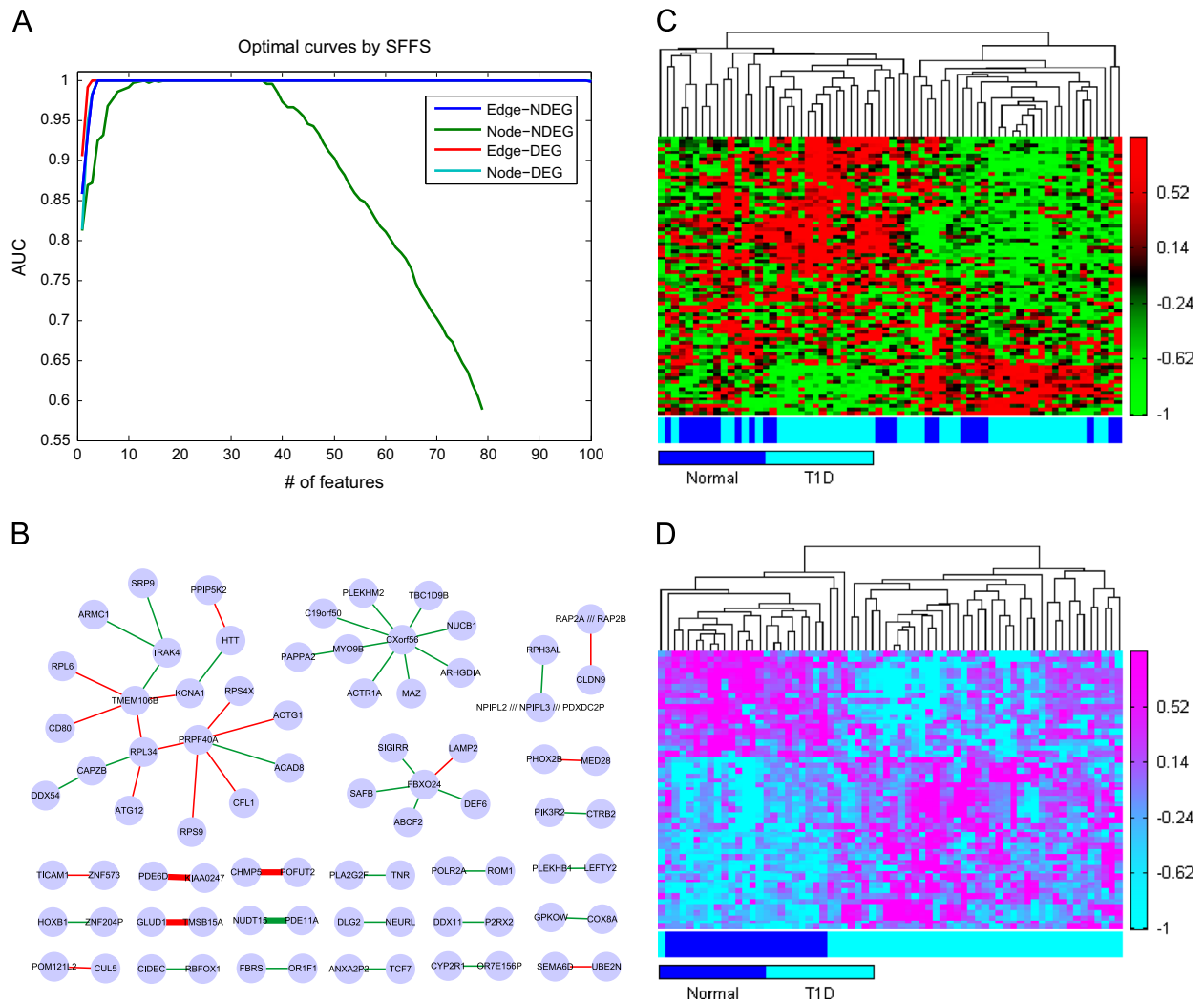


Fig. 3. Results of the diabetes dataset analysis. (A) For NDEGs and DEGs, the optimizing curves by SFFS method on the top 100 edge features and the involved genes are plotted. (B) The discriminative edge features are presented, where bold lines are edge-biomarkers. Green lines represent down-regulated correlations, while red lines up-regulated correlations. (C) Expression heatmap of the genes involved in the discriminative edge features. (D) Heatmap of the edge data transformed by formula (2).

edge-biomarkers tend to connect to hub while for type I diabetes the edge-biomarkers tend to be marginal. This suggests the two distinct modes of pathogenesis of human complex diseases. This example also demonstrated that those NDEGs ignored by the traditional analysis actually have rich information on the phenotypes, and can achieve almost the similar or better classification or prediction effect to those DEGs provided that we consider their edge information as EdgeMarker.

Among these biomarkers, 19 genes (e.g. CD80, CYP2R1, HTT, MYO9B, TCF7, PHOX2B) are related to T1D according to *GeneCards* database. Through literature search, we found that among the genes of 4 edge-biomarkers, GLUD1 (Kapoor et al., 2009; Tanizawa et al., 2002), NUDT15 (Chen et al., 2013b; Ito et al., 2011), PDE6D (Heiliger et al., 2012), and POFUT2 (Chen et al., 2013a) were reported to be associated with diabetes. Note that none of the edge-biomarkers are connected to hub genes, which is quite different from the cholangiocarcinoma dataset where many edge-biomarkers are connected to hubs.

4. Conclusions

Many traditional approaches use network-based or edge-based methods to identify the discriminative networks or edges, and

then use the genes or proteins (or gene set) involved in those networks or edges for the classification of each test sample, rather than using networks or edges (or correlations) themselves. Thus, they are basically node-biomarker based methods because the networks or edges in those methods are only used to identify discriminative genes or node-biomarkers. There are no effective methods to directly use edges or correlations for classifying a single sample for normal and disease groups. One major difficulty is that the computation of correlation coefficient requires several samples. In contrast to this problem, in this paper, we developed a novel method based on a new vector representation of an edge (1)–(2) to identify the discriminatively differentially correlated gene pairs as edge biomarkers for distinguishing the binary or even multiple conditions (control/case, normal/disease, etc.). We show that our edge biomarkers can distinguish phenotypes in an accurate manner although their genes may have no differential expressions. Our method also overcame the difficulty for using edge information on single samples, i.e., with the proposed new edge matrix (2), we can predict or diagnose the phenotype of each single sample by using edge information, which is one of major contributions in this paper. This method was applied to analyze human cholangiocarcinoma and diabetes datasets. The results imply two distinct modes of pathogenesis of human complex diseases. The results also support that non-differentially expressed

genes, which are usually ignored by traditional methods, can be as informative as differentially expressed genes in terms of classifying different biological conditions or phenotypes of samples, which can also provide us network-based new insight into the pathogenesis of complex diseases. As a future topic, we will extend this work to network biomarkers (He et al., 2012; Liu et al., 2012; Wen et al., 2013) or dynamical network biomarkers (DNBs) (Chen et al., 2012; Li et al., 2013; Liu et al., 2013a; Liu et al., 2013b) by considering available information of networks (Zhang et al., 2013; Zhang et al., 2012) and dynamics. In addition to the study of complex diseases, this work can be directly used to the analysis of other biological problems with different phenotypes in a similar way.

Acknowledgements

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040700), and the National Program on Key Basic Research Project [Grant No. 2014CB910504]. This work was supported by the National Natural Science Foundation of China (NSFC) with Nos. 61134013, 91029301, 31200987 and 91130033. This work was also supported by the Knowledge Innovation Program of SIBS of CAS with Grant No. 2013KIP218.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at <http://dx.doi.org/10.1016/j.jtbi.2014.05.041>.

References

- Andersen, J.B., Spee, B., Blechacz, B.R., Avital, I., Komuta, M., Barbour, A., Conner, E. A., Gillen, M.C., Roskams, T., Roberts, L.R., Factor, V.M., Thorgerisson, S.S., 2012. Genomic and genetic characterization of cholangiocarcinoma identifies therapeutic targets for tyrosine kinase inhibitors. *Gastroenterology* 142, 1021–1031 e15, doi: 10.1053/j.gastro.2011.12.005.
- Barnett, C., Krebs, J.E., 2011. WSTF does it all: a multifunctional protein in transcription, repair, and replication. *Biochem. Cell Biol.* 89, 12–23, <http://dx.doi.org/10.1139/O10-114>.
- Bishop, C.M., 1995. *Neural Networks for Pattern Recognition*. Clarendon Press; Oxford University Press, Oxford New York.
- Borad, M.J., Champion, M.D., Egan, J.B., Liang, W.S., Fonseca, R., Bryce, A.H., McCullough, A.E., Barrett, M.T., Hunt, K., Patel, M.D., Young, S.W., Collins, J.M., Silva, A.C., Condjella, R.M., Block, M., McWilliams, R.R., Lazaridis, K.N., Klee, E. W., Bible, K.C., Harris, P., Oliver, G.R., Bhavsar, J.D., Nair, A.A., Middha, S., Asmann, Y., Kocher, J.P., Schahl, K., Kipp, B.R., Barr Fritcher, E.G., Baker, A., Aldrich, J., Kurdoglu, A., Izatt, T., Christoforides, A., Cherni, I., Nasser, S., Reiman, R., Phillips, L., McDonald, J., Adkins, J., Mastrian, S.D., Placek, P., Watanabe, A.T., Lobello, J., Han, H., Von Hoff, D., Craig, D.W., Stewart, A.K., Carpten, J.D., 2014. Integrated genomic characterization reveals novel, therapeutically relevant drug targets in FGFR and EGFR pathways in sporadic intrahepatic cholangiocarcinoma. *PLoS Genet.* 10, e1004135, <http://dx.doi.org/10.1371/journal.pgen.1004135>.
- Calin, G.A., Cimmino, A., Fabbri, M., Ferracin, M., Wojcik, S.E., Shimizu, M., Taccioli, C., Zanesi, N., Garzon, R., Azeilán, R.I., Alder, H., Volinia, S., Rassenti, L., Liu, X., Liu, C.G., Kipps, T.J., Negrini, M., Croce, C.M., 2008. MiR-15a and miR-16-1 cluster functions in human leukemia. *Proc. Natl. Acad. Sci. USA* 105, 5166–5171, <http://dx.doi.org/10.1073/pnas.0800121105>.
- Chen, C.Y., Jan, Y.H., Juan, Y.H., Yang, C.J., Huang, M.S., Yu, C.J., Yang, P.C., Hsiao, M., Hsu, T.L., Wong, C.H., 2013a. Fucosyltransferase 8 as a functional regulator of non-small cell lung cancer. *Proc. Natl. Acad. Sci. USA* 110, 630–635, <http://dx.doi.org/10.1073/pnas.1220425110>.
- Chen, K., Wu, D., Zhu, X., Ni, H., Wei, X., Mao, N., Xie, Y., Niu, Y., Li, M., 2013b. Gene expression profile analysis of human intervertebral disc degeneration. *Genet. Mol. Biol.* 36, 448–454, <http://dx.doi.org/10.1590/S1415-47572013000300021>.
- Chen, L., Wang, R.-S., Zhang, X.-S., 2009. *Biomolecular Networks : Methods and Applications in Systems Biology*. Wiley, Hoboken, NJ.
- Chen, L., Liu, R., Liu, Z.P., Li, M., Aihara, K., 2012. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci. Rep.* 2, 342, <http://dx.doi.org/10.1038/srep00342>.
- Coulouarn, C., Cavard, C., Rubbia-Brandt, L., Audebourg, A., Dumont, F., Jacques, S., Just, P.A., Clement, B., Gilgenkrantz, H., Perret, C., Terris, B., 2012. Combined hepatocellular-cholangiocarcinoma exhibit progenitor features and activation of Wnt and TGFbeta signaling pathways. *Carcinogenesis* 33, 1791–1796, <http://dx.doi.org/10.1093/carcin/bgs208>.
- Edlund, K., Lindskog, C., Saito, A., Berglund, A., Ponten, F., Goransson-Kultima, H., Isaksson, A., Jirstrom, K., Planck, M., Johansson, L., Lambe, M., Holmberg, L., Nyberg, F., Ekman, S., Bergqvist, M., Landelius, P., Lamberg, K., Botling, J., Ostman, A., Micke, P., 2012. CD99 is a novel prognostic stromal marker in non-small cell lung cancer. *Int. J. Cancer* 131, 2264–2273, <http://dx.doi.org/10.1002/ijc.27518>.
- Geman, D., d'Avignon, C., Naiman, D.Q., Winslow, R.L., 2004. Classifying gene expression profiles from pairwise mRNA comparisons. *Stat. Appl. Genet. Mol. Biol.* 3, <http://dx.doi.org/10.2202/1544-6115.1071> (Article19).
- Guyon, I., Andr, #233, Elisseeff, 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182.
- He, D., Liu, Z.P., Honda, M., Kaneko, S., Chen, L., 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, <http://dx.doi.org/10.1093/jmcb/mjs011>.
- Heiliger, K.J., Hess, J., Vitagliano, D., Salerno, P., Braselmann, H., Salvatore, G., Ugolini, C., Summerer, I., Bogdanova, T., Unger, K., Thomas, G., Santoro, M., Zitzelsberger, H., 2012. Novel candidate genes of thyroid tumorigenesis identified in Trk-T1 transgenic mice. *Endocr. Relat. Cancer* 19, 409–421, <http://dx.doi.org/10.1530/ERC-11-0387>.
- Hong, J., Park, S., Park, J., Jang, S.J., Ahn, H.K., Sym, S.J., Cho, E.K., Shin, D.B., Lee, J.H., 2012. CD99 expression and newly diagnosed diffuse large B-cell lymphoma treated with rituximab-CHOP immunochemotherapy. *Ann. Hematol.* 91, 1897–1906, <http://dx.doi.org/10.1007/s00277-012-1533-z>.
- Ito, R., Sekiguchi, M., Setoyama, D., Nakatsu, Y., Yamagata, Y., Hayakawa, H., 2011. Cleavage of oxidized guanine nucleotide and ADP sugar by human NUDT5 protein. *J. Biochem.* 149, 731–738, <http://dx.doi.org/10.1093/jb/mvr028>.
- Kaizer, E. C., Glaser, C. L., Chaussabel, D., Banchereau, J., Pascual, V., White, P. C., 2007. Gene expression in peripheral blood mononuclear cells from children with diabetes. *J. Clin. Endocrinol. Metab.* 92, 3705–3711, <http://dx.doi.org/10.1210/jc.2007-0979>.
- Kapoor, R.R., Flanagan, S.E., Fulton, P., Chakrapani, A., Chadeaux, B., Ben-Omran, T., Banerjee, I., Shield, J.P., Ellard, S., Hussain, K., 2009. Hyperinsulinism-hyperammonemia syndrome: novel mutations in the GLUD1 gene and genotype-phenotype correlations. *Eur. J. Endocrinol.* 161, 731–735, <http://dx.doi.org/10.1530/EJE-09-0615>.
- Lai, Y., Wu, B., Chen, L., Zhao, H., 2004. A statistical method for identifying differential gene-gene co-expression patterns. *Bioinformatics* 20, 3146–3155, <http://dx.doi.org/10.1093/bioinformatics/bth379>.
- Li, M., Zeng, T., Liu, R., Chen, L., 2013. 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.*, <http://dx.doi.org/10.1093/bib/bbt027>.
- Listgarten, J., Neal, R.M., Roweis, S.T., Wong, P., Emili, A., 2007. Difference detection in LC-MS data for protein biomarker discovery. *Bioinformatics* 23, e198–e204, <http://dx.doi.org/10.1093/bioinformatics/btl326>.
- Liu, R., Aihara, K., Chen, L., 2013a. Dynamical network biomarkers for identifying critical transitions and their driving networks of biologic processes. *Quant. Biol.* 1, 105–114, <http://dx.doi.org/10.1007/s40484-013-0008-0>.
- Liu, R., Wang, X., Aihara, K., Chen, L., 2013b. Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. *Med. Res. Rev.*, <http://dx.doi.org/10.1002/med.21293>.
- Liu, X., Liu, Z.P., Zhao, X.M., Chen, L., 2012. Identifying disease genes and module biomarkers by differential interactions. *J. Am. Med. Inform. Assoc.* 19, 241–248, <http://dx.doi.org/10.1136/amiajnl-2011-000658>.
- Liu, X., Wang, J., Chen, L., 2013. Whole-exome sequencing reveals recurrent somatic mutation networks in cancer. *Cancer Lett* 340, 270–276, <http://dx.doi.org/10.1016/j.canlet.2012.11.002>.
- Maehata, T., Taniguchi, H., Yamamoto, H., Noshio, K., Adachi, Y., Miyamoto, N., Miyamoto, C., Akutsu, N., Yamaoka, S., Itoh, F., 2008. Transcriptional silencing of Dickkopf gene family by CpG island hypermethylation in human gastrointestinal cancer. *World J. Gastroenterol.* 14, 2702–2714.
- Mor, G., Visintin, I., Lai, Y., Zhao, H., Schwartz, P., Rutherford, T., Yue, L., Bray-Ward, P., Ward, D.C., 2005. Serum protein markers for early detection of ovarian cancer. *Proc. Natl. Acad. Sci. USA* 102, 7677–7682, <http://dx.doi.org/10.1073/pnas.0502178102>.
- Pudil, P., Novovicova, J., Kittler, J., 1994. Floating Search Methods in Feature-Selection. *Pattern Recogn. Lett.* 15, 1119–1125 (doi:10.1016/0167-8655(94)90127-9).
- Rajagopalan, A., Browning, D., Salama, S., 2013. CD99 expression in Merkel cell carcinoma: a case series with an unusual paranuclear dot-like staining pattern. *J. Cutan. Pathol.* 40, 19–24, <http://dx.doi.org/10.1111/cup.12049>.
- Raser, J.M., O'Shea, E.K., 2005. Noise in gene expression: origins, consequences, and control. *Science* 309, 2010–2013, <http://dx.doi.org/10.1126/science.1105891>.
- Sahni, N., Yi, S., Zhong, Q., Jaikhan, N., Charleaux, B., Cusick, M.E., Vidal, M., 2013. Edgotype: a fundamental link between genotype and phenotype. *Curr. Opin. Genet. Dev.* 23, 649–657, <http://dx.doi.org/10.1016/j.gde.2013.11.002>.
- Shi, P., Ray, S., Zhu, Q., Kon, M.A., 2011. Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction. *BMC Bioinformatics* 12, 375, <http://dx.doi.org/10.1186/1471-2105-12-375>.
- Sun, S.Y., Liu, Z.P., Zeng, T., Wang, Y., Chen, L., 2013. Spatio-temporal analysis of type 2 diabetes mellitus based on differential expression networks. *Sci Rep* 3, 2268, <http://dx.doi.org/10.1038/srep02268>.
- Tanizawa, Y., Nakai, K., Sasaki, T., Anno, T., Ohta, Y., Inoue, H., Matsuo, K., Koga, M., Furukawa, S., Oka, Y., 2002. Unregulated elevation of glutamate dehydrogenase

- activity induces glutamine-stimulated insulin secretion: identification and characterization of a GLUD1 gene mutation and insulin secretion studies with MIN6 cells overexpressing the mutant glutamate dehydrogenase. *Diabetes* 51, 712–717.
- Wang, J., Sun, Y., Zheng, S., Zhang, X.S., Zhou, H., Chen, L., 2013. APG: an Active Protein–Gene network model to quantify regulatory signals in complex biological systems. *Sci Rep* 3, 1097, <http://dx.doi.org/10.1038/srep01097>.
- Wen, Z., Liu, Z.P., Liu, Z., Zhang, Y., Chen, L., 2013. An integrated approach to identify causal network modules of complex diseases with application to colorectal cancer. *J Am Med Inform Assoc* 20, 659–667, <http://dx.doi.org/10.1136/amiajnl-2012-001168>.
- Yu, X., Li, G., Chen, L., 2013. Prediction and early diagnosis of complex diseases by edge-network. *Bioinformatics*, <http://dx.doi.org/10.1093/bioinformatics/btt620>.
- Zeng, T., Sun, S.Y., Wang, Y., Zhu, H., Chen, L., 2013. Network biomarkers reveal dysfunctional gene regulations during disease progression. *FEBS J* 280, 5682–5695, <http://dx.doi.org/10.1111/febs.12536>.
- Zhang, X., Liu, K., Liu, Z.P., Duval, B., Richer, J.M., Zhao, X.M., Hao, J.K., Chen, L., 2013. NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics* 29, 106–113, <http://dx.doi.org/10.1093/bioinformatics/bts619>.
- Zhang, X., Zhao, X.M., He, K., Lu, L., Cao, Y., Liu, J., Hao, J.K., Liu, Z.P., Chen, L., 2012. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics* 28, 98–104, <http://dx.doi.org/10.1093/bioinformatics/btr626>.