BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 17 2013, pages 2169–2177
doi:10.1093/bioinformatics/btt373

Gene expression

Advance Access publication July 10, 2013

Topologically inferring risk-active pathways toward precise cancer classification by directed random walk

Wei Liu[1,2,†], Chunquan Li[1,†], Yanjun Xu[1], Haixiu Yang[1], Qianlan Yao[1], Junwei Han[1], Desi Shang[1], Chunlong Zhang[1], Fei Su[1], Xiaoxi Li[3], Yun Xiao[1], Fan Zhang[1], Meng Dai[3,*] and Xia Li[1,*]

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China, [2]Department of Mathematics, Heilongjiang Institute of Technology, Harbin 150050, China and [3]Department of Health Management Center, Nanfang Hospital, Southern Medical University, Guangzhou 510515, China

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Motivation: The accurate prediction of disease status is a central challenge in clinical cancer research. Microarray-based gene biomarkers have been identified to predict outcome and outperform traditional clinical parameters. However, the robustness of the individual gene biomarkers is questioned because of their little reproducibility between different cohorts of patients. Substantial progress in treatment requires advances in methods to identify robust biomarkers. Several methods incorporating pathway information have been proposed to identify robust pathway markers and build classifiers at the level of functional categories rather than of individual genes. However, current methods consider the pathways as simple gene sets but ignore the pathway topological information, which is essential to infer a more robust pathway activity.

Results: Here, we propose a directed random walk (DRW)-based method to infer the pathway activity. DRW evaluates the topological importance of each gene by capturing the structure information embedded in the directed pathway network. The strategy of weighting genes by their topological importance greatly improved the reproducibility of pathway activities. Experiments on 18 cancer datasets showed that the proposed method yielded a more accurate and robust overall performance compared with several existing gene-based and pathway-based classification methods. The resulting risk-active pathways are more reliable in guiding therapeutic selection and the development of pathway-specific therapeutic strategies.

Availability: DRW is freely available at http://210.46.85.180:8080/DRWPClass/

Contact: lixia@hrbmu.edu.cn or dm42298@126.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 4, 2013; revised on May 29, 2013; accepted on June 23, 2013

1 INTRODUCTION

One important challenge in clinical cancer research is the accurate prediction of disease states and responses to treatment, which guide the choice of optimal therapy for the patients. Microarray-based gene biomarkers have been identified to predict outcome and outperformed traditional clinical parameters (Perez-Diez *et al.*, 2007; van 't Veer *et al.*, 2002; Wang *et al.*, 2005). However, the reproducibility of the individual gene biomarkers has been challenged (Dupuy and Simon, 2007; Ein-Dor *et al.*, 2006). The prediction performance of the gene signatures discovered in one dataset often decreased drastically in an independent dataset for the same disease phenotype. The main cause of this discrepancy may include the cellular heterogeneity within tissues, the inherent genetic heterogeneity across patients and the measurement error in microarray platforms. Moreover, the small number of samples that are available to build a classifier makes the problem more difficult (Ein-Dor *et al.*, 2006).

To address this problem, several methods have been proposed to find more robust biomarkers at the level of functional categories rather than of individual genes (Guo *et al.*, 2005; Kim *et al.*, 2012; Lee *et al.*, 2008; Su *et al.*, 2009). Because the gene products are well known to function coordinately by way of a functional module or signaling cascade, the high-level perturbed functional modules may be more consistent than individual genes. Thus, integrating the expressions of function-related genes and extracting classification features at the functional level may produce more reproducible biomarkers. The biomarkers at the functional level could alleviate the impact of heterogeneity across the tissues or samples and provide a better biological interpretation of the relationships between the disease and canonical pathways such as the Gene Ontology (Ashburner *et al.*, 2000) and KEGG databases (Kanehisa and Goto, 2000). A number of feature extraction methods have been used to integrate the expression value of the member genes in canonical pathways. Guo *et al.* used the mean and median of the expression values of member genes to infer the pathway activity (Guo *et al.*, 2005). Bild *et al.* used the first principal component of the expression values of the member genes to infer the pathway activity (Bild *et al.*, 2006). In contrast, Lee *et al.* inferred the pathway activity by using only a subset of genes [condition-responsive genes (CORGs)], whose combined expression delivers optimal discriminative power for the disease phenotype (Lee *et al.*, 2008). Other approaches estimated the pathway activity based on probabilistic inference (Efroni *et al.*, 2007; Su *et al.*, 2009). The existing methods successfully incorporated the

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

pathway information into the disease classification procedure and achieved better classification performance.

However, all these methods used the pathways as simple gene sets and treated all genes in the pathway equally but ignored the structure information embedded in the pathway network. It has been reported that the 'hub' genes with high connectivity are usually critical to the functionality of the whole pathway (Carter *et al*., 2004). The perturbation on a 'hub' gene may impact the pathway more greatly than that on a gene with little topological importance. However, the 'hub' genes tend to have low levels of change in gene expression (Lu *et al*., 2007). Their subtle but consistent changes could be easily overshadowed by differential non-hub genes, whose changes may be unstable across the samples. Thus, up-weighting the hub genes and down-weighting the non-hub genes according to their topological importance may contribute to a more robust pathway activity. A few methods have been proposed to mine the topological information of the network. Gao *et al*. designed a pathway score based on all connected gene pairs for the pathway enrichment analysis (Gao and Wang, 2007). Hung *et al*. improved the pathway enrichment analysis by weighting each gene based on the closest correlated neighbor genes (Hung *et al*., 2010). Recently, undirected random walk was used to mine the topological information of the protein–protein interaction network and showed superiority in prioritizing candidate disease genes or pathways (Can *et al*., 2005; Kohler *et al*., 2008; Li and Patra, 2010). However, these topological information mining methods consider the edges in the network as undirected; thus, they may lose some important information when used in the directed pathway networks, such as the position of the genes and the type of interactions, which are essential to measure the topological importance of the genes. For example, the genes located in the upstream are more important because they influence other genes downstream, such as the insulin receptor in the insulin signaling pathway, whose inactivation may shut off the entire pathway. To capture the topological properties in directed network and infer a more robust pathway activity, a flexible pathway topological information mining method is required.

In this work, we propose a directed random walk (DRW)-based method (Fig. 1) to mine the topological information and infer the pathway activity. DRW is performed on a merged global pathway network and attempts to evaluate the topological importance of each gene according to all the aforementioned topological properties. The strategy of weighting genes by their topological importance effectively enhances the reproducibility of the pathway activities. We apply the DRW method to the classification of six types of cancers and show that it can produce higher accuracy and stronger robustness compared with several existing pathway-based approaches, both within single datasets and in different independent datasets. The resulting risk-active pathways are more reliable in predicting clinical outcome and guiding therapeutic selection.

2 MATERIALS AND METHODS

2.1 Datasets

A total of 18 microarray datasets were used in this study. All were downloaded from the GEO database (Edgar *et al*., 2002). Our analysis focused on eight datasets. They are GSE10072 for lung cancer (Landi *et al*., 2008), GSE13911 for stomach cancer (D'Errico *et al*., 2009), GSE17856 for liver cancer (Tsuchiya *et al*., 2010), GSE5364 for thyroid cancer (Yu *et al*., 2008), GSE15641 and GSE17895 for kidney cancer (Dalgliesh *et al*., 2010; Jones *et al*., 2005) and GSE3494 and GSE1456 for breast cancer (Miller *et al*., 2005; Pawitan *et al*., 2005). For breast cancer, patients who died because of breast cancer within 5 years were defined as the 'poor' prognostic group, and the remaining patients were defined as the 'good' prognostic group (patients with a survival time <5 years without any reported event were excluded.). Each dataset contains gene expression profiles for ≥22 tumor/poor samples and ≥16 normal/good samples. To evaluate classification performance on independent dataset, we collected 10 additional datasets for validation (Supplementary Table S1).

The pathway information was obtained from the KEGG database (Kanehisa and Goto, 2000). This collection contains 150 metabolic and 150 non-metabolic pathways. The interactions in KEGG were manually drawn by summarizing experimental evidence in published literature.

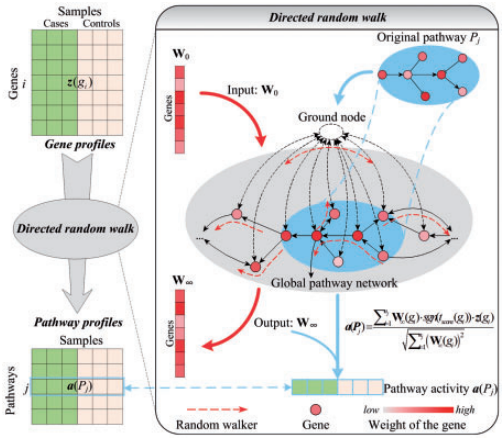

Fig. 1. Overview of the DRW-based pathway activity inference method. The gene profiles are translated into pathway profiles based on the DRW method. Row i of the gene profiles $[z(g_i)]$ is the expression value vector of gene g_i across all samples. Row j of the pathway profiles $[\boldsymbol{a}(P_j)]$ is the expression value vector (i.e. pathway activity) of pathway P_j across all samples. The diagram on the right is the detailed illustration of DRW-based pathway activity inference. The global pathway network is constructed on 150 metabolic and 150 non-metabolic pathways. A virtual ground node is added to balance the weights of the genes in the network. $\mathbf{W}_0$ is the initial weights of the genes. $\mathbf{W}_\infty$ is the output weight vector. The inference of the pathway activity of pathway P_j is presented as an example. The edge direction in the original pathway network P_j is reversed when we merge the pathway P_j into the global pathway network. The pathway activity $\boldsymbol{a}(P_j)$ of P_j is the integrated expression value vector of the significantly differentially expressed genes in the pathway weighting by $\mathbf{W}_\infty$

2.2 Constructing the global-directed pathway graph

We first converted each KEGG pathway into a directed graph using the SubpathwayMiner software package (Li *et al*., 2009, 2013). Then the 300 graphs were merged into a global-directed pathway graph, which covers 4113 nodes and 40 875 directed edges. Each node in the graph represents a gene, and each directed edge represents how genes interact and regulate each other. The direction of the edge is derived from the type of interaction between two genes, which is available from the KEGG. For example, if gene A activates (inhibits) gene B, then A points to B. The random walk (Lovasz, 1996) on the pathway graph is similar to the

PageRank algorithm used by the Google search engine (Brin and Page, 1998), but the direction of the random walk is opposite (a web page is important if other pages point to it, whereas a gene is important if it influences other genes (Draghici *et al.*, 2007). Thus, we reversed the direction of all edges on the global pathway graph (Fig. 1).

To perform random walk, two problems need to be solved. First, there are 938 dangling nodes (0 out-degree, such as the most upstream nodes in the pathway) on the directed pathway graph. The dangling nodes will absorb the weight of other nodes but never distribute their weight back to the graph. Besides, there are 396 nodes with 0 in-degree; they do not receive any weights from other nodes. This may be unsatisfactory from a biological perspective. To alleviate this problem, we introduced a virtual ground node connected to every node through bidirectional edges (Fig. 1). Thus, the graph became strongly connected and consisted of $4113+1$ nodes and $40\,875+2\times4113$ edges. We denoted the extended graph as $G=(V, E)$, where V is the set of nodes and E is the set of edges.

2.3 DRW on the global-directed pathway graph

The DRW with restart stimulates a random walker that starts on a source node s (or a set of source nodes simultaneously). At every time step, the walker transitions from its current node to a randomly selected neighbor (based on edge weights) or goes back to source node s with probability r. Formally, the DRW with restart is defined as

$$\mathbf{W}_{t+1} = (1-r)\mathbf{M}^T\mathbf{W}_t + r\mathbf{W}_0 \quad (1)$$

where $\mathbf{W}_t$ is a vector in which the i-th element holds the probability of being at node i at time step t, and $\mathbf{M}$ is the row-normalized adjacency matrix of the graph G.

To start the random walk, the initial probability vector $\mathbf{W}_0$ was constructed by assigning to each node (except for the ground node, whose initial probability was 0) its absolute t-test score, and normalized to a unit vector. The restart probability r was set as 0.7. $\mathbf{W}_t$ converges to a unique steady state in the presence of the ground node (Lü *et al.*, 2011). This was obtained by performing the iteration until the L_1-norm between $\mathbf{W}_t$ and $\mathbf{W}_{t+1}$ $<10^{-10}$. At the steady state, we evenly distributed the probability of the ground node back to all other nodes and obtained the final probability vector $\mathbf{W}_\infty$. $\mathbf{W}_\infty$ provided a measure of the topological importance of the genes in the global pathway graph and was used as the weight vector of genes at the step of pathway activity inference. Because we used the t-test scores as the initial probabilities, the magnitude of the t-test scores also contributed to weight adjustments. Thus, genes which are both topologically important and significantly differentially expressed will obtain higher weights.

2.4 Pathway activity inference

For each pathway, we calculated the t-test statistics of all member genes. Only those genes that are significantly differentially expressed ($P<0.05$, see Supplementary Table S2 for the corresponding fdr, Benjamin correction) are selected in constructing the pathway activity. Consider a pathway P_j that contains n_j differential genes $\{g_1, g_2, \ldots, g_{nj}\}$ after those genes that were not significantly differentially expressed in the training set have been removed. The pathway activity $\boldsymbol{a}_T(P_j)$ of the pathway P_j in the training dataset is calculated as follows:

$$\boldsymbol{a}_T(P_j) = \sum_{i=1}^{n_j} \mathbf{W}_\infty(g_i)\cdot sgn(t_{score}(g_i))\cdot \boldsymbol{z}_T(g_i) \Big/ \sqrt{\sum_{i=1}^{n_j}(\mathbf{W}_\infty(g_i))^2} \quad (2)$$

where $\mathbf{W}_\infty(g_i)$ is the weight of gene g_i, $t_{score}(g_i)$ is the t statistics of gene g_i from a two-tailed t-test on expression values between two phenotypes, $\boldsymbol{z}_T(g_i)$ is the normalized expression value vector of g_i across the samples in the training dataset and sgn() is the sign function that returns -1 for negative numbers and $+1$ for positive numbers. The pathway activity $\boldsymbol{a}_V(P_j)$ of the pathway P_j in the test dataset is given by

$$\boldsymbol{a}_V(P_j) = \sum_{i=1}^{n_j} \mathbf{W}_\infty(g_i)\cdot sgn(t_{score}(g_i))\cdot \boldsymbol{z}_V(g_i) \Big/ \sqrt{\sum_{i=1}^{n_j}(\mathbf{W}_\infty(g_i))^2}$$

where $\boldsymbol{z}_V(g_i)$ is the normalized expression value vector of g_i across the samples in the test dataset.

2.5 Reproducibility power

Yang *et al.* proposed that a pathway activity is reproducible if it provides similar discriminative power on the training-test pair datasets (Yang *et al.*, 2012). The reproducibility power is given by

$$C_{score}(N) = \frac{1}{N}\sum_{i=1}^{N} t_{score}(P_T^i)\cdot t_{score}(P_V^i) \quad (3)$$

where $t_{score}(P)$ is the t statistics of P from a two-tailed t-test on pathway activities between two phenotypes, P_T^i is the i-th pathway activity in descending order (ranked by absolute t-scores) in the training dataset, P_V^i is its corresponding pathway activity in the test dataset and N is the number of selected pathways. The reproducibility power reflects the discriminative power and the robustness of the pathway activity.

For *within-dataset* experiments, the samples in a dataset were randomly divided into five subsets of equal size. Four of these subsets were used as the training dataset, whereas the remaining subset was used as the test set. Each subset was used in turn as the test set to evaluate the reproducibility. For unbiased evaluation, we repeated these experiments for 100 random partitions for the entire dataset. The averaged C_{score} over 500 experiments was reported as the overall reproducibility performance. For *cross-dataset* experiments, the whole first dataset was used as the training dataset, and the second independent dataset was used as the test set.

2.6 Classification evaluation

For *within-dataset* experiments, we randomly divided the dataset so that four-fifths of the samples were used as the training set, and the remaining one-fifth was used as the test set (5-fold cross-validation). To select the best pathway marker set for classification, we further split the training set into three equal-sized subsets. Two subsets were used as the *marker evaluation dataset* to build the classifier and rank the pathway markers, and one subset was used as the *feature selection dataset* for assessing which pathway marker set produced the best classification performance. To build the classifier, we calculated the t-test statistics of the pathway activities on the *marker evaluation dataset* and ranked the pathways by their P-values in increasing order. The top 50 pathways were used as the candidate features to build the logistic regression model (or classifiers based on other classification algorithms). We first constructed the classifier with the pathway ranked first. Then the pathways were added sequentially to train the logistic regression model, and the performance of the classifier was measured by evaluating its area under the receiver operating characteristics curve (AUC) on the *feature selection dataset*. The added pathway marker was kept in the feature set if the AUC increased and was removed otherwise. We repeated the above process for the top 50 pathway markers to optimize the classifier and obtained the best feature set. The performance of the optimized classifier was evaluated on the test set by using the pathway markers in the best feature set. Each subset of the training set was used in turn as *feature selection dataset* to optimize the classifier. Therefore, each training set generated three optimized classifiers and generated three AUCs on the corresponding test set. Each of the five subsets in the whole dataset was evaluated in turn as the test set. Thus, one partition generated 15 AUCs. For unbiased evaluation and to estimate the variation of the AUC, we repeated the above process 100 times. The mean AUC across 1500 classifiers was reported as the overall performance of the classification method.

For *cross-dataset* experiments, the whole first dataset was used as the training set, and the second independent dataset was used as the test set. The first dataset was divided into five subsets of equal size. Four subsets

were used as the marker evaluation dataset to train the classifier, and one subset was used as the feature selection dataset to optimize the constructed classifier and select the best feature set. We repeated this experiment by using each subset as the feature selection dataset in turn and evaluated the optimized classifier on the test set. One hundred random partitions of the training set were generated for building the classifier and feature selection. The averaged AUC on the test set over 500 classifiers was reported as the final performance measure.

2.7 Other classification methods

For comparison with other pathway activity inference methods, we implemented four pathway-based classification methods, namely, the Mean and Median methods (Guo *et al.*, 2005), the Principle Component Analysis (PCA) method (Bild *et al.*, 2006) and the Pathway Activity inference using Condition-responsive genes (PAC) method (Lee *et al.*, 2008). The Mean and Median methods use the mean and median of the expression values of the member genes to infer the pathway activity. The PCA method uses the first principal component of the expressions of the member genes to represent the pathway activity. The PAC method performs a greedy search to identify CORGs and infers the pathway activity by using only the CORGs. We also compared the performance of a traditional gene-based classifier that uses individual genes as features (the Genes method). Additionally, we compared our method with another excellent gene-based method Reweighted Recursive Feature Elimination (RRFE) (Johannes *et al.*, 2010) implemented in pathClass (Johannes *et al.*, 2011). RRFE also incorporates topological information to improve AUC as well as the reproducibility and interpretability of selected gene features.

3 RESULTS

3.1 Inferring pathway activity by DRW

Figure 1 illustrates the overview of the DRW-based pathway activity inference method. DRW mines the topological information embedded in the global pathway network and weights the genes according to their topological importance. With these weights, we integrate the expression values of the differential member genes to topologically infer the activity for each pathway. The genes that are topologically important receive more weights and contribute more to pathway activity. Finally, the DRW method translates the gene profiles into pathway profiles. We applied the DRW method on the lung cancer dataset and inferred 245 pathway activities of lung cancer. The top 50 active pathways (ranked by absolute *t*-scores) are able to classify samples with no misclassifications by using two-way hierarchical clustering (Fig. 2A). The top 50 active pathways of seven other datasets also show good discriminative ability (Supplementary Figs S1–S7). The heat map in Figure 2B depicts the –log (*P*-value) of 200 common active pathways from a two-tailed *t*-test on pathway activities between the two phenotypes in eight cancer datasets. It provides a snapshot of the active levels of the pathways for each cancer. A few cancer-related pathways show high active levels in all datasets, such as apoptosis, p53 signaling pathway and acute myeloid leukemia. Compared with kidney cancer, the pathway activities of breast cancer are less discriminative and consistent between two breast cancer datasets, confirming the molecular heterogeneity of breast cancer.

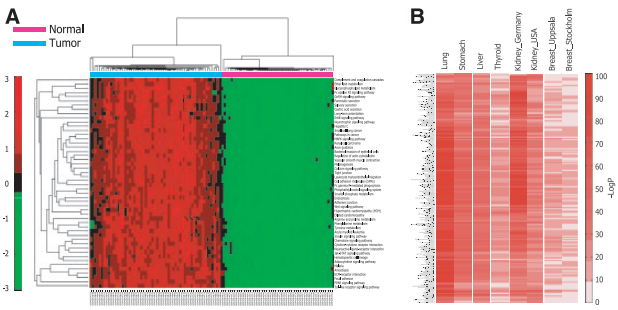

Fig. 2. Active levels of the pathway activities. (**A**) The top 50 active pathways of the lung cancer samples are clustered by complete linkage hierarchical clustering. (**B**) Heat map depicts the active levels of 200 common active pathways for each cancer

3.2 Pathway activities show high reproducibility

A major challenge in pathway activity inference is to find active pathways that have stronger discriminative power and robustness. We used the reproducibility power [Equation (3), see 'Materials and Methods' section] to evaluate the pathway activities. Greater reproducibility power indicates more discriminative power and stronger robustness of the pathway activity. We computed the reproducibility power of the active pathways and ranked the pathways according to their reproducibility power. The mean reproducibility power of the top pathways of five pathway activity inference methods, namely, the Mean and Median methods, the PCA method, the PAC method and the DRW method, was compared (Supplementary Fig. S8). Besides, we also evaluated the reproducibility power of the top single gene markers, which were chosen from the 4113 genes covered by the 300 pathways used in this study. The DRW-based pathway activities exhibited the greatest power to discriminate between tumor and normal samples for all eight datasets. They showed ~1.5-fold increased reproducibility over the PAC-based pathway activities, ~2-fold over individual gene markers and ~10-fold improvement over pathway activities inferred by other methods, such as the Mean, Median and PCA methods.

Another important measurement of the robustness of pathway activities is their reproducibility between independent datasets. We further analyzed the kidney and breast cancer datasets because two separate cohorts of patients were available for each cancer. The DRW method again obtained the largest reproducibility on average (Supplementary Fig. S9), further indicating the robustness of the DRW-based pathway activities. The Genes method performed worst in the 'Kidney-Germany->USA' case (Supplementary Fig. S9A). This is because the most discriminative markers in the Kidney-Germany dataset were not highly discriminative in the Kidney-USA dataset, indicating that the gene markers were not robust. The PAC method showed a similar phenomenon in that the pathway markers extracted from the 'Kidney-Germany' dataset were not highly discriminative in the 'Kidney-USA' dataset.

3.3 Risk-active pathways improve cancer classification

To test the usefulness of pathway activities in discriminating between different disease phenotypes, we performed classification on eight cancer datasets. For a fair and effective comparison with other methods, we carried out *within-dataset* experiments similar to those used in Lee *et al.* (2008) to evaluate the classification

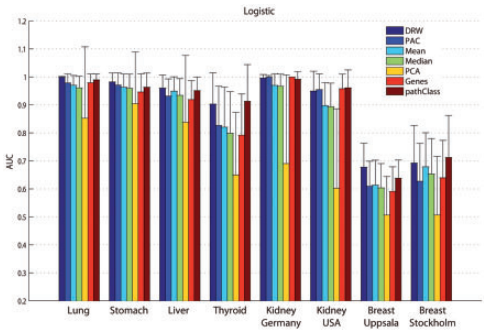

Fig. 3. Classification performances within datasets using logistic regression

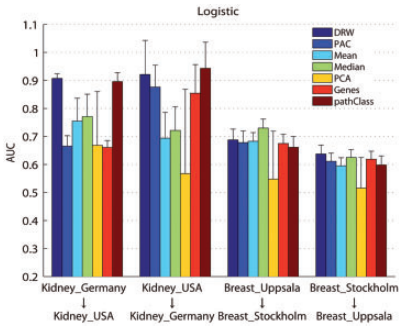

Fig. 4. Classification performances cross-datasets using logistic regression

performance (see 'Materials and Methods' section). To reduce the effect of sensitivity in feature selection when comparing different pathway-based methods, we also used the top 50 pathway activities ranked by their *P*-values in increasing order as the candidate features for feature selection. The classifier was built by using the selected pathway activities based on logistic regression. The Mean, Median, PAC and PCA methods followed the same procedure. The results of pathClass (RRFE) were obtained by a 10-times repeated 5-fold cross-validation by using R-package pathClass. For the Genes method, the top 50 discriminative gene markers were chosen as the candidate features to keep the maximum number of features identical across all classification methods (Su *et al*., 2009).

Figure 3 presents a summary of the AUC of the cross-validation experiments (Supplementary Fig. S10A shows the classification accuracy). The proposed DRW method obtained AUCs (accuracies) of 0.9996 (95.58%), 0.9820 (94.34%), 0.9602 (93.06%), 0.9029 (87.33%), 0.9961 (97.55%), 0.9496 (94.87%), 0.6776 (73.36%) and 0.6926 (82.21%) for eight cancer datasets, respectively, indicating that the DRW-based pathway markers were quite competent in discriminating between different disease phenotypes. The AUCs (accuracies) of the DRW method on the eight datasets were significantly greater than those of the PAC, Mean, Median, PCA and Genes methods at the 5% significance level [Wilcoxon signed-rank test: $P = 0.0391$ (0.0234), 0.0078 (0.0078), 0.0078 (0.0078), 0.0078 (0.0078) and 0.0391 (0.0391), respectively]. This indicates that the DRW method outperforms other methods not only in AUC but also in accuracy. The pathClass method also yielded favorable performance, but was slightly inferior to the DRW method. Of these eight datasets, the AUC of the DRW method was larger compared with pathClass in five datasets. Although the Genes method reached the highest AUC for kidney cancer, it did not perform well in five datasets, especially the thyroid and breast cancer datasets, suggesting that the individual gene markers are less stable than pathway markers.

To evaluate the generalization ability of the classifier, we carried out *cross-dataset* experiments on the kidney and breast cancer datasets. One dataset was used as the training set and another independent dataset as the test set (see 'Materials and Methods' section). To estimate the variation of the AUC of pathClass on independent dataset, we also randomly divided the whole training dataset into five subsets of equal size. Each subset was removed in turn, and the remaining four subsets were used as the training set to build the classifier. Each classifier was then used to predict the class of samples in the independent test set. The aforementioned process was repeated 10 times to obtain the average AUC and the standard deviation. Figure 4 shows the results of the *cross-dataset* experiments (Supplementary Fig. S10B shows the classification accuracy). The DRW method again outperformed other pathway inference methods, especially in the two kidney cancer datasets, which were detected on different microarray platforms. The AUC of the proposed method improved by 36.28, 20.14, 17.65, 35.67 and 37.30% in the 'Kidney-Germany->USA' case and 5.13, 32.76, 27.72, 62.28 and 7.81% in the 'Kidney-USA->Germany' case compared with the PAC, Mean, Median, PCA and Genes method. The pathClass also obtained good performance in the two kidney cancer datasets. A potential reason may be that both the DRW method and pathClass incorporate topological information to build the classifier. However, pathClass was inferior to the DRW method on average. The AUC of the DRW method was larger compared with pathClass in three of four paired training-test datasets. Compared with *within-dataset* experiments, although the PAC, Mean, Median and Genes methods showed good performances on the two kidney cancer datasets, respectively (Fig. 3), their performances decreased drastically in the independent datasets. This is not surprising if we consider the reproducibility power presented in Supplementary Figure S9A. The DRW method showed consistent performance and strong generalization ability, indicating that the DRW-based pathway activities are less sensitive to different cohorts of patients and microarray platforms and are more reliable in predicting clinical outcome in practice.

To further demonstrate the superior performance of DRW on independent dataset, we collected 10 additional datasets and performed *cross-dataset* experiments. Of the 10 experiments, the AUCs of DRW ranked first in six experiments (Table 1). They were significantly greater than those of the PAC, Mean, Median, Genes and DRW-NR methods at the 5% significance level (Wilcoxon signed- rank test: $P = 0.0137$, 0.0020, 0.0020, 0.0195, 0.0020, respectively). pathClass also obtained good performance, but it was inferior to DRW in seven experiments. In all, our DRW method outperformed all other methods on average.

In addition, to show that the superior performance of our DRW method was not dependent on the chosen classification

Table 1. Classification performance comparison on independent datasets

Cancer	Training set->Test set	DRW	PAC	Mean	Median	Genes	pathClass	DRW-NR[a]
Lung	GSE10072->GSE19804	**0.9788** ± 0.0063	0.9400 ± 0.0387	0.9360 ± 0.0213	0.9401 ± 0.0157	0.9600 ± 0.0143	0.9714 ± 0.0058	0.9752 ± 0.0271
Lung	GSE10072->GSE19188	0.9722 ± 0.0053	0.9348 ± 0.0192	0.9240 ± 0.0261	0.9332 ± 0.0150	0.9541 ± 0.0175	**0.9745** ± 0.0039	0.9717 ± 0.0057
Stomach	GSE13911->GSE19826	**0.9258** ± 0.0392	0.8887 ± 0.0360	0.7754 ± 0.0499	0.8226 ± 0.0439	0.8827 ± 0.0537	0.8907 ± 0.0514	0.8872 ± 0.0783
Stomach	GSE13911->GSE38940	**0.7250** ± 0.0707	0.5702 ± 0.0522	0.5773 ± 0.0119	0.5391 ± 0.0324	0.6552 ± 0.0933	0.6657 ± 0.0814	0.7210 ± 0.0831
Liver	GSE17856->GSE14520_1[b]	**0.9691** ± 0.0148	0.9505 ± 0.0342	0.9311 ± 0.0150	0.9245 ± 0.0151	0.8862 ± 0.0774	0.9555 ± 0.1487	0.9609 ± 0.0235
Liver	GSE17856->GSE14520_2	0.9646 ± 0.0274	0.9547 ± 0.0279	0.9264 ± 0.0113	0.9235 ± 0.0192	0.8808 ± 0.0724	**0.9791** ± 0.0112	0.9579 ± 0.0466
Thyroid	GSE5364->GSE33630	**0.9480** ± 0.0462	0.7762 ± 0.0883	0.6993 ± 0.0441	0.6963 ± 0.0686	0.9031 ± 0.0667	0.9456 ± 0.0643	0.9032 ± 0.0879
Thyroid	GSE5364->GSE29265	**0.9063** ± 0.0606	0.7806 ± 0.1119	0.7556 ± 0.0325	0.7447 ± 0.0452	0.7820 ± 0.0635	0.9020 ± 0.0553	0.8890 ± 0.1151
Kidney	GSE17895->GSE36895	0.9874 ± 0.0559	0.9705 ± 0.0688	0.9009 ± 0.0594	0.9500 ± 0.0333	**0.9882** ± 0.0470	0.9799 ± 0.0760	0.9637 ± 0.0999
Breast	GSE3494->GSE7390	0.6035 ± 0.0480	0.6330 ± 0.0430	0.5142 ± 0.0312	0.5538 ± 0.0382	0.6238 ± 0.0353	**0.6373** ± 0.0309	0.5954 ± 0.0537

Shown are the average AUC and the standard deviation. The evaluation was performed according to *cross-dataset* experiments. The AUC shown in bold is the best AUC for the corresponding paired training-test dataset. The DRW method outperformed all other methods on average.

[a]DRW-NR: DRW algorithm does not reverse the direction of the edges in the KEGG pathways.

[b]Roessler *et al.* provided tumor and paired non-tumor samples of patients from two independent cohorts in GSE14520. We used them as two independent test sets.

algorithm, we repeated the *within-dataset* and the *cross-dataset* experiments by using three additional classification algorithms: naïve Bayes (John and Langley, 1995), LibLINEAR (Fan *et al.*, 2008) and LibSVM (Chang and Lin, 2011). The results exhibited a similar tendency to that in logistic regression (Supplementary Figs S11–S13).

3.4 Robustness of risk-active pathways

The classification experiments showed that the DRW-based pathway activities had reliable predictive accuracy. The discriminative pathway activities that are frequently selected to build the classifier may reveal new, robust risk-active pathways for cancers. We provided the 20 most frequently selected risk-active pathways in Supplementary Table S3 (the genes used to infer these pathway activities are provided in Supplementary Table S4). Among these pathways, the regulation of actin cytoskeleton is identified as a risk-active pathway in most cancer datasets (Table 2). It has been reported that key proteins involved in the actin cytoskeleton, including WASP family proteins, Arp2/3 complex, LIM-kinase, cofilin and cortactin, are overexpressed in many cancers, such as breast (Wang *et al.*, 2004) and liver (Chuma *et al.*, 2004). These proteins coordinately regulate the formation of invasive protrusions in tumor cells and generate the driving force for cancer cell migration, invasion and metastasis (Yamaguchi and Condeelis, 2007). Focal adhesion was identified in four datasets. This pathway functions in the link between the actin cytoskeleton and the extracellular matrix (ECM). The ECM-receptor interaction is the most frequently selected active pathway in the lung cancer dataset (Supplementary Table S3), suggesting its important role in lung cancer, consistent with the study by Cho *et al.* (Cho *et al.*, 2011). These related pathways are all identified as risk-active pathways in multiple cancers, indicating their coactive effect on cancers. Besides, several widely studied pathways that are associated with cancer are identified as risk-active pathways in multiple cancers, such as pathways in cancer, MAPK, Wnt and calcium signaling pathway. Deregulated calcium signaling has been shown in many cancers and is used as a sensitive therapeutic target in some cancers, such as melanoma, lung cancer, prostate cancer (Haverstick *et al.*, 2000) and so forth.

Pathways specific to the phenotype of classification were identified. For example, ErbB signaling pathway was reported in the Breast-Uppsala dataset. From early studies of the prognostic indicator and therapeutic target of breast cancer (Hayes and Thor, 2002; Wright *et al.*, 1989) until recent reports on this topic (Weigel and Dowsett, 2010), there is an agreement that ErbB2, a critical component involved in ErbB signaling pathway, plays an important role in breast cancer prognosis and therapy. Furthermore, HER2/ c-erbB-2 is highly associated with progression and metastasis of breast cancer via promotion of cancer's important processes, such as tumor cell motility and apoptosis (Johnson *et al.*, 2010). In addition, calcium signaling pathway is the most frequently selected in both the Kidney-Germany and Breast-Uppsala datasets (Supplementary Table S3), suggesting that it may be a new biomarker for kidney and breast cancer. Tight junction pathway was associated with the occurrence and progression of lung cancer. Global gene expression analysis showed that the specific patterns of cell junction were implicated

Table 2. Risk-active pathways identified in multiple cancers

Active pathway	1[a]	2	3	4	5	6	7	8
Regulation of actin cytoskeleton	√	√	√	√	—	√	√	√
Pathways in cancer	—	√	√	√	—	√	√	√
Calcium signaling pathway	√	—	—	—	√	√	√	√
Focal adhesion	√	—	√	√	—	√	—	—
Tight junction	√	—	√	√	—	—	√	—
PPAR signaling pathway	√	√	—	—	—	√	√	—
MAPK signaling pathway	—	—	—	√	√	—	√	√
Endocytosis	—	—	—	√	—	√	√	√
ECM-receptor interaction	√	—	√	√	—	—	—	—
Wnt signaling pathway	—	√	—	√	—	—	√	—
Axon guidance	—	√	√	√	—	—	—	—
Adherens junction	√	√	√	—	—	—	—	—

[a]Cancer datasets: 1-Lung; 2-Stomach; 3-Liver; 4-Thyroid; 5-Kidney-Germany; 6-Kidney-USA; 7-Breast-Uppsala; 8-Breast-Stockholm.

in non–small cell lung cancer (Kuner *et al.*, 2009). Moreover, the disruption of tight junction has also been shown to be associated with lung cancer cell apoptosis (Zhang *et al.*, 2011) and to play an important role in cancer metastasis (Martin and Jiang, 2009). The detailed evidence from previous reports concerning these risk pathways in the classification of each disease is provided in Supplementary Table S5. These risk-active pathways could be useful in cancer diagnosis and could provide novel hypotheses in cancer therapy.

4 DISCUSSION

Identifying reproducible disease biomarkers is important for clinical utility in diagnostic and prognostic applications. Pathway markers have been proven to be more reproducible than individual gene markers and provide better biological interpretation at the functional level (Guo *et al.*, 2005; Kim *et al.*, 2012; Lee *et al.*, 2008; Su *et al.*, 2009). Pathway markers are becoming a useful tool for discriminating the disease status toward clinical prediction and treatment. In this context, we proposed a novel DRW method for inference of reproducible pathway activities and robust disease classification. We showed that DRW-based pathway activities are more discriminative and consistent than individual gene markers and pathway activities inferred by other methods. Based on their superior reproducibility, we used the DRW-based pathway activities to build the classifier and tested the classification performance on eight datasets. The results showed that the AUCs and classification accuracies of the DRW method were significantly higher than those of the Genes method and other pathway-based methods, both within single datasets and in different independent datasets.

The reliable performance of the DRW-based pathway activities could be attributed to the strategy of weighting genes according to their topological importance. By weighting the genes, the DRW method could amplify the signals of the key genes, whose variations in their expression levels may greatly impact the pathway but weaken the differential signals of the genes, which only appear somewhere downstream or does not affect the given pathway as much. Thus, the effectiveness of the topologically important genes associated with the phenotype of interest is up-weighted, whereas other genes that have little topological importance are weakened even if they are significantly differentially expressed. Therefore, the DRW method may alleviate the noise due to the heterogeneity of samples or the technical measurements, leading to more reproducible pathway activities. We take the calcium signaling pathway, which is the most frequently selected in the Kidney-Germany dataset, as an example for analysis. Figure 5 provides a snapshot of the perturbed nodes in the calcium signaling pathway (Supplementary Figs S14 and S15 show the perturbed nodes in each dataset, respectively). The topologically important genes located at the center of the pathway, such as those residing in CALM, PLCγ, PLCβ and PLCδ, are up-weighted (Supplementary Table S6 and S7 provides the full gene list). For example, the weights of CALM member genes (*CALM1*, *CALM2*, *CALM3*, *CALML3* and *CALML5*) exceed those of genes residing in GPCR (*AGTR1*, *CYSLTR2*, *TACR3* and *LHCGR*), although the *P*-values of the latter genes are much smaller ($P < 8.96\text{E-}13$). In contrast, the member genes of nodes

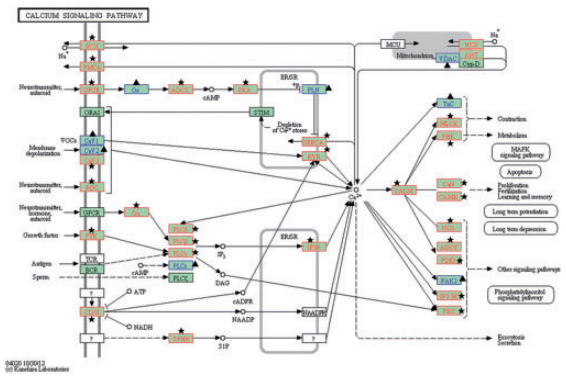

Fig. 5. A snapshot of the perturbed nodes in the calcium signaling pathway. A node is considered as disrupted if one of its member genes was altered (at the 5% significance level). The nodes marked with a star are disrupted in both datasets, whereas the nodes marked with a triangle symbol are disrupted in only one dataset

that have less topological importance (e.g. GPCR, PTK, FAK2, IP3 3K) are down-weighted. Interestingly, most inconsistent genes (shown in bold italics in Supplementary Table S6) are down-weighted. For example, *AGTR1*, *PTK2B* and *ITPKA* are differentially expressed only in the Kidney-Germany dataset, whereas *ADORA2A* and *GNA15* are differentially expressed only in the Kidney-USA dataset. The aforementioned phenomenon is in line with Lu *et al*.'s study (Lu *et al*., 2007); i.e. the hub genes (such as those residing in CALM) tend to exhibit low but consistent changes in gene expression. The highly differentially expressed genes may contribute greatly to the Genes method. For example, the most differential gene *AGTR1* ($P = 4.70E-33$) may be selected as a gene marker based on the Kidney-Germany dataset. However, it has little discriminative power ($P = 7.58E-02$) in the Kidney-USA dataset. Obviously, it will reduce the averaged reproducibility of the gene markers (Supplementary Fig. S9A) and thus the classification performance (Fig. 4) between independent datasets. The inconsistent genes may also impact the pathway-based methods, which treat all genes in the pathway equally. Considering the instability of these genes between independent cohorts of patients, it is not hard to speculate about the deteriorated performance of most methods in *cross-dataset* experiments. The DRW method, however, could alleviate the impact of inconsistent genes by reducing their weights; thus, it is more stable than other methods. This further indicates the importance of incorporating pathway topological information in building a robust cancer classifier.

DRW captures the topological properties of genes, including the position of the genes in the pathway, how many genes interact with the given gene and the type of interactions between them. Because we use the absolute *t*-test score as the initial probability in DRW, a gene connected to other differentially expressed genes may receive more weight. Thus, the differential hub genes exhibit an increased effect on the classifier. We reran the DRW method by deleting the top 'hub' genes to evaluate this influence. The genes were sorted in decreasing order according to out-degree (Supplementary Table S8). The top 2, 4, 6 and 8% of genes were deleted in turn. The results demonstrated that the AUC obtained by deleting 'hub' genes decreased to a greater extent compared with deleting randomly selected genes of the same number (Supplementary Fig. S16), confirming the importance of 'hub' genes. We further investigated the functional roles of the 'hub' genes (Supplementary Table S9). Transcription factors and membrane receptors comprised a high proportion ($24.3 + 10.2 = 34.5\%$) of genes with out-degree ($51 \sim 100$). This indicates that the influences of differential transcription factors and membrane receptors are enhanced by DRW if they connect to other differential genes and potentially serve as a driver of pathway activity to a certain extent.

To test the influence of the directionality on the classifier, we evaluated the classification performance of a modified DRW algorithm (DRW-NR) that does not reverse the direction of the edges in the KEGG pathways. DRW-NR is inferior to DRW in almost all datasets (Table 1). Also, the average standard deviations of AUCs of DRW-NR are ~65.8% larger than that of DRW, indicating that DRW is more robust in predicting disease states.

There is one parameter in the DRW algorithm: the restart probability r [Equation (1)]. To investigate the performance of the DRW method for varying restart probabilities, we set r at 0.1, 0.3, 0.5, 0.7 and 0.9 and performed *cross-dataset* experiments based on logistic regression. The AUC and classification accuracy did not change much with the change in the value of restart probability r (Supplementary Fig. S17), which was in line with the results of Can *et al*. (2005) and Li and Patra(2010).

The performance of the DRW method depends on the coverage and quality of the human pathway information. A more complete pathway topology could help to clarify the roles that the genes played in the pathway and weight the genes more precisely. With the rapid development of human interaction databases, we believe that having more complete biological pathway information will enable a more accurate prediction of disease status and provide a better guide for patient treatment.

ACKNOWLEDGEMENTS

We thank all the research staff for their contributions to this project.

Funding: This work was supported in part by the Funds for Creative Research Groups of the National Natural Science Foundation of China [Grant Nos. 81121003], the National Natural Science Foundation of China [Grant Nos. 61170154, 61073136, 31200996 and 31200998] and the Specialized Research Fund for the Doctoral Program of Higher Education of China [Grant Nos. 20102307120027 and 20102307110022].

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bild,A.H. *et al.* (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353–357.

Brin,S. and Page,L. (1998) The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, **30**, 107–117.

Can,T. *et al.* (2005) Analysis of protein–protein interaction networks using random walks. In: *Proceedings of the 5th International Workshop on Bioinformatics*. Chicago, Illinois, pp. 61–68.

Carter,S.L. *et al.* (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, **20**, 2242–2250.

Chang,C. and Lin,C. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**.

Cho,J.H. *et al.* (2011) Systems biology of interstitial lung diseases: integration of mRNA and microRNA expression changes. *BMC Med. Genomics*, **4**, 8.

Chuma,M. *et al.* (2004) Overexpression of cortactin is involved in motility and metastasis of hepatocellular carcinoma. *J. Hepatol.*, **41**, 629–636.

D'Errico,M. *et al.* (2009) Genome-wide expression profile of sporadic gastric cancers with microsatellite instability. *Eur. J. Cancer*, **45**, 461–469.

Dalgliesh,G.L. *et al.* (2010) Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature*, **463**, 360–363.

Draghici,S. *et al.* (2007) A systems biology approach for pathway level analysis. *Genome Res.*, **17**, 1537–1545.

Dupuy,A. and Simon,R.M. (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J. Natl Cancer Inst.*, **99**, 147–157.

Edgar,R. *et al.* (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Efroni,S. *et al.* (2007) Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS One*, **2**, e425.

Ein-Dor,L. *et al.* (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl Acad Sci. USA*, **103**, 5923–5928.

Fan,R. *et al.* (2008) LIBLINEAR: a library for large linear classification. *J. Mach. Learn. Res.*, **9**, 1871–1874.
Gao,S. and Wang,X. (2007) TAPPA: topological analysis of pathway phenotype association. *Bioinformatics*, **23**, 3100–3102.
Guo,Z. *et al.* (2005) Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, **6**, 58.
Haverstick,D.M. *et al.* (2000) Inhibition of human prostate cancer proliferation in vitro and in a mouse model by a compound synthesized to block Ca2+ entry. *Cancer Res.*, **60**, 1002–1008.
Hayes,D.F. and Thor,A.D. (2002) c-erbB-2 in breast cancer: development of a clinically useful marker. *Semin. Oncol.*, **29**, 231–245.
Hung,J.H. *et al.* (2010) Identification of functional modules that correlate with phenotypic difference: the influence of network topology. *Genome Biol.*, **11**, R23.
Johannes,M. *et al.* (2010) Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. *Bioinformatics*, **26**, 2136–2144.
Johannes,M. *et al.* (2011) pathClass: an R-package for integration of pathway knowledge into support vector machines for biomarker discovery. *Bioinformatics*, **27**, 1442–1443.
John,G.H. and Langley,P. (1995) Estimating continuous distributions in Bayesian classifiers. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., Montréal, Canada, pp. 338–345.
Johnson,E. *et al.* (2010) HER2/ErbB2-induced breast cancer cell migration and invasion require p120 catenin activation of Rac1 and Cdc42. *J. Biol. Chem.*, **285**, 29491–29501.
Jones,J. *et al.* (2005) Gene signatures of progression and metastasis in renal cell cancer. *Clin. Cancer Res.*, **11**, 5730–5739.
Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
Kim,S. *et al.* (2012) Pathway-based classification of cancer subtypes. *Biol. Direct*, **7**, 21.
Kohler,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.
Kuner,R. *et al.* (2009) Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung Cancer*, **63**, 32–38.
Lü,L. *et al.* (2011) Leaders in social networks, the delicious case. *PLoS One*, **6**, e21202.
Landi,M.T. *et al.* (2008) Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS One*, **3**, e1651.
Lee,E. *et al.* (2008) Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.*, **4**, e1000217.
Li,C. *et al.* (2009) SubpathwayMiner: a software package for flexible identification of pathways. *Nucleic Acids Res.*, **37**, e131.
Li,C. *et al.* (2013) Subpathway-GM: identification of metabolic subpathways via joint power of interesting genes and metabolites and their topologies within pathways. *Nucleic Acids Res.*, **41**, e101.
Li,Y. and Patra,J.C. (2010) Integration of multiple data sources to prioritize candidate genes using discounted rating system. *BMC Bioinformatics*, **11** (**Suppl. 1**), S20.
Lovasz,L. (1996) Random walks on graphs: a survey. In: *Combinatorics, Paul Erdos is Eighty*. Vol. 2, János Bolyai Mathematical Society, pp. 253–398.
Lu,X. *et al.* (2007) Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. *Mol. Syst. Biol.*, **3**, 98.
Martin,T.A. and Jiang,W.G. (2009) Loss of tight junction barrier function and its role in cancer metastasis. *Biochim. Biophys. Acta*, **1788**, 872–891.
Miller,L.D. *et al.* (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl Acad. Sci. USA*, **102**, 13550–13555.
Pawitan,Y. *et al.* (2005) Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res.*, **7**, R953–R964.
Perez-Diez,A. *et al.* (2007) Microarrays for cancer diagnosis and classification. *Adv. Exp. Med. Biol.*, **593**, 74–85.
Su,J. *et al.* (2009) Accurate and reliable cancer classification based on probabilistic inference of pathway activity. *PLoS One*, **4**, e8161.
Tsuchiya,M. *et al.* (2010) Gene expression in nontumoral liver tissue and recurrence-free survival in hepatitis C virus-positive hepatocellular carcinoma. *Mol. Cancer*, **9**, 74.
van 't Veer,L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
Wang,W. *et al.* (2004) Identification and testing of a gene expression signature of invasive carcinoma cells within primary mammary tumors. *Cancer Res.*, **64**, 8585–8594.
Wang,Y. *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671–679.
Weigel,M.T. and Dowsett,M. (2010) Current and emerging biomarkers in breast cancer: prognosis and prediction. *Endocr. Relat. Cancer*, **17**, R245–R262.
Wright,C. *et al.* (1989) Expression of c-erbB-2 oncoprotein: a prognostic indicator in human breast cancer. *Cancer Res.*, **49**, 2087–2090.
Yamaguchi,H. and Condeelis,J. (2007) Regulation of the actin cytoskeleton in cancer cell migration and invasion. *Biochim. Biophys. Acta*, **1773**, 642–652.
Yang,R. *et al.* (2012) Core module biomarker identification with network exploration for breast cancer metastasis. *BMC Bioinformatics*, **13**, 12.
Yu,K. *et al.* (2008) A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers. *PLoS Genet.*, **4**, e1000129.
Zhang,G. *et al.* (2011) Hydroxycamptothecin-loaded Fe3O4 nanoparticles induce human lung cancer cell apoptosis through caspase-8 pathway activation and disrupt tight junctions. *Cancer Sci.*, **102**, 1216–1222.