Published online 16 August 2006 *Nucleic Acids Research, 2006, Vol. 34, No. 14* ***4069–4080***
doi:10.1093/nar/gkl583

Topology-based cancer classification and related pathway mining using microarray data

Chun-Chi Liu[1,2], Wen-Shyen E. Chen[1], Chin-Chung Lin[2], Hsiang-Chuan Liu[3], Hsuan-Yu Chen[4], Pan-Chyr Yang[5], Pei-Chun Chang[3,*] and Jeremy J.W. Chen[2,5,*]

[1]Department of Computer Science, National Chung-Hsing University, Taichung, Taiwan, ROC, [2]Institutes of Biomedical Sciences and Molecular Biology, National Chung-Hsing University, Taichung, Taiwan, ROC, [3]Department of Biotechnology and Bioinformatics, Asia University, Taichung, Taiwan, ROC, [4]Graduate Institute of Epidemiology, National Taiwan University, Taipei, Taiwan, ROC and [5]NTU Center for Genomic Medicine, National Taiwan University College of Medicine, Taipei, Taiwan, ROC

Received May 29, 2006; Revised July 24, 2006; Accepted July 26, 2006

ABSTRACT

Cancer classification is the critical basis for patient-tailored therapy, while pathway analysis is a promising method to discover the underlying molecular mechanisms related to cancer development by using microarray data. However, linking the molecular classification and pathway analysis with gene network approach has not been discussed yet. In this study, we developed a novel framework based on cancer class-specific gene networks for classification and pathway analysis. This framework involves a novel gene network construction, named ordering network, which exhibits the power-law node-degree distribution as seen in correlation networks. The results obtained from five public cancer datasets showed that the gene networks with ordering relationship are better than those with correlation relationship in terms of accuracy and stability of the classification performance. Furthermore, we integrated the ordering networks, classification information and pathway database to develop the topology-based pathway analysis for identifying cancer class-specific pathways, which might be essential in the biological significance of cancer. Our results suggest that the topology-based classification technology can precisely distinguish cancer subclasses and the topology-based pathway analysis can characterize the correspondent biochemical pathways even if there are subtle, but consistent, changes in gene expression, which may provide new insights into the underlying molecular mechanisms of tumorigenesis.

INTRODUCTION

The advance of molecular diagnosis offers a systematic and precise prospect for cancer classification. One of the typical methods is DNA microarray technology, which is a powerful tool in functional genome studies (1–3). Recently, the gene expression data derived from such analyses have been employed to many cancer classification studies (4–6).

The studies of gene network have been used in drug discovery (7,8), identification of the signature of disease mechanism (9), analysis of acute systemic inflammation in human leukocyte (10), investigation of cellular regulatory processes (11), hub gene analysis (12,13), active pathway extraction (14), molecular characterization of the cellular state (15) and so on. In these studies, Segal *et al*. (16) developed a module-network approach to identify modules of coregulated genes by using microarray data, enrichment analysis and promoter analysis, which is further applied to discover the signature of the mechanisms underlying tumorigenesis (9). Calvano *et al*. (10) integrated the structured network knowledge-base approach, pathway analysis and microarray data analysis to develop an analytic method of systemic inflammation. In summary, these gene networks can be constructed from the correlation between gene expressions (so-called correlation or relevance network approach) (15,17,18), module-network approach (9,16,19), structured network knowledge-based approach (10), Boolean network approach (20), Bayesian network approach (21), gene-specific perturbations (7), the analysis and integration of biological databases (22), promoter element detection (23), time-series gene expression data (24) and many more.

The correlation networks are based on the linear correlation between the expressions of gene pairs (15,17,18) and exhibit the power-law node-degree distributions (15). The power-law distribution is a topology structure of the concentration of links within a few hubs, and is a scale-free network.

*To whom correspondence should be addressed. Tel: 886 4 22840485; ext. 226; Fax: 886 4 22853469; Email: jwchen@dragon.nchu.edu.tw
*Correspondence may also be addressed to P.C. Chang. Tel: 886 4 23323456; ext. 1868; Fax: 886 4 23316699; Email: pcchang@asia.edu.tw

This is in contrast to a random network where each node has the same scale of links (25). The various biological networks also revealed the power-law node-degree distributions in the previous studies (15,26,27). In addition, the global topological properties of the correlation networks are consistent with the previously characterized biological networks (15). However, cells may maintain some conservative ordering relationships among the genes' expressions in some situations, such as flower differentiation (28) and human tumor progression (29). So far, the gene networks based on ordering relationship, which remain steady in nonlinear gene regulation system, have not been studied.

Besides the applications described above, the gene networks have not been used for systematic cancer classification. The reason might be that the gene network can be constructed by the numerous training samples (microarray data), but it is difficult to construct a new gene network by using a single test sample and apply to determine its class. In addition to cancer classification, functional enrichment analysis regarding pathways is one of the advanced methods using microarray data to characterize biological processes and becoming popular (30). How to link the cancer classification and pathway enrichment analysis together by gene network approach has not been reported. The objective of this study was to develop a novel framework that can construct the gene networks in addition to overcome the above problems and to perform the topology-based classification and pathway analysis using microarray data.

MATERIALS AND METHODS

The software programs, source code and datasets used in this work are available to download at our web site http://biochip.nchu.edu.tw/supl/top/.

Public expression datasets

To demonstrate this novel framework, five publicly available datasets of gene expression profiles were used in this study, including the ALL-subtype (4), GCM (31), Lung-cancer (5), Lung-subtype (32) and MLL-leukemia (6) datasets. These datasets included training and test datasets, except the Lung-subtype. Therefore, the Lung-subtype dataset was randomly separated into the training and test datasets, which was named as Lung-subtype-2 (the original Lung-subtype dataset was named as Lung-subtype-1). The detailed information about the datasets is described in Supplementary Data and Supplementary Table S1. All of the microarray datasets, without respect to training and test datasets, were simply processed by clipping the values to a minimum of 0 U and a maximum of 30 000 U.

Gene selection procedure

All of the genes on the arrays of training dataset were sorted according to their signal to noise ratio (S2N) (33) values that correlate with each tumor subclass. The S2N values can be imagined as the significant level of marker genes. In addition, 1000 times of permutation test were performed on the training dataset (Figure 1a), and the S2N value of each gene was re-calculated. Those genes with S2N values less than the significant level 0.05 by permutation test were further filtered out. The top ranking genes were used for topology-based classification. Gene selection methods can improve the classification performance (34), which is also the common procedure for most classification applications (4). It is conceivable that gene selection can increase the accuracy since it may reduce the noise (35). We applied gene selection on training dataset to reduce the noise and investigate the influence of various gene numbers on classification performance.

Networks construction

To construct the gene networks used in topology-based classification, the correlation and ordering networks were created. The correlation network construction is based on the Pearson correlation of all pairwise gene expressions in the dataset. The detailed procedures of the correlation network construction are described in Supplementary Data and can be found in the previous reports (15,36). In this study, the definition of ordering coefficient of gene pair is modified from our previous study that developed the generalized index of ordering coefficient for the semantic structure analysis in the field of educational measurement and statistics (37). The ordering network construction is described as follows. The paired genes would be linked if the level of observed ordering relationship exceeded the significance level, which could assess whether the paired genes have consistent ordering relationship of expression intensity in most samples. We defined the ordering coefficient r_{ij} between gene i and j as follows to measure the ordering relationship level:

$$\gamma_{ij} = 1 - \sum_{s=1}^{N} \frac{[x_{s_i} - x_{s_j}]^+}{N(x_{\max} - x_{\min})}$$
$$[x_{s_i} - x_{s_j}]^+ = \begin{cases} x_{s_i} - x_{s_j} & x_{s_i} > x_{s_j} \\ 0 & x_{s_i} \leqslant x_{s_j} \end{cases}, \quad \mathbf{1}$$

where x_{si} is the expression intensity of gene i in sample s; N is the total number of samples; $x_{\max}$ and $x_{\min}$ are the maximum and minimum expression values in the microarray data, respectively. If $x_{si} \leqslant x_{sj}$ for s = 1, 2, . . . , N (i.e. gene i is less than gene j in all the samples), then $r_{ij} = 1$ that means gene i and gene j have consensus ordering relationship in all the samples.

To determine whether the r_{ij} is significant, the ordering coefficient threshold (r_{th}) was obtained by using permutation method. The procedure is described as follows: (i) randomly select two genes and assign to gene i and gene j from microarray data; (ii) compute the ordering coefficient r_{ij}; (iii) repeat above step 1 and step 2 for 5000 rounds to obtain a distribution of the ordering coefficient r_{ij} and determine the r_{th} at the significant level 0.05.

Once the ordering coefficient r_{ij} between the pair of gene i and j exceeded the threshold ($r_{ij} \geqslant r_{th}$), the edge $i \rightarrow j$ (from i to j) would be created. Therefore, the ordering networks were directional graphs. An edge $i \rightarrow j$ meant the expression intensity of gene j was larger than gene i in most situations (samples). So the high input degree of a gene node implied the relatively high expression level and vice versa. The ordering networks possessed conservative topological properties when the networks were created from the samples of the same subclass of cancer.

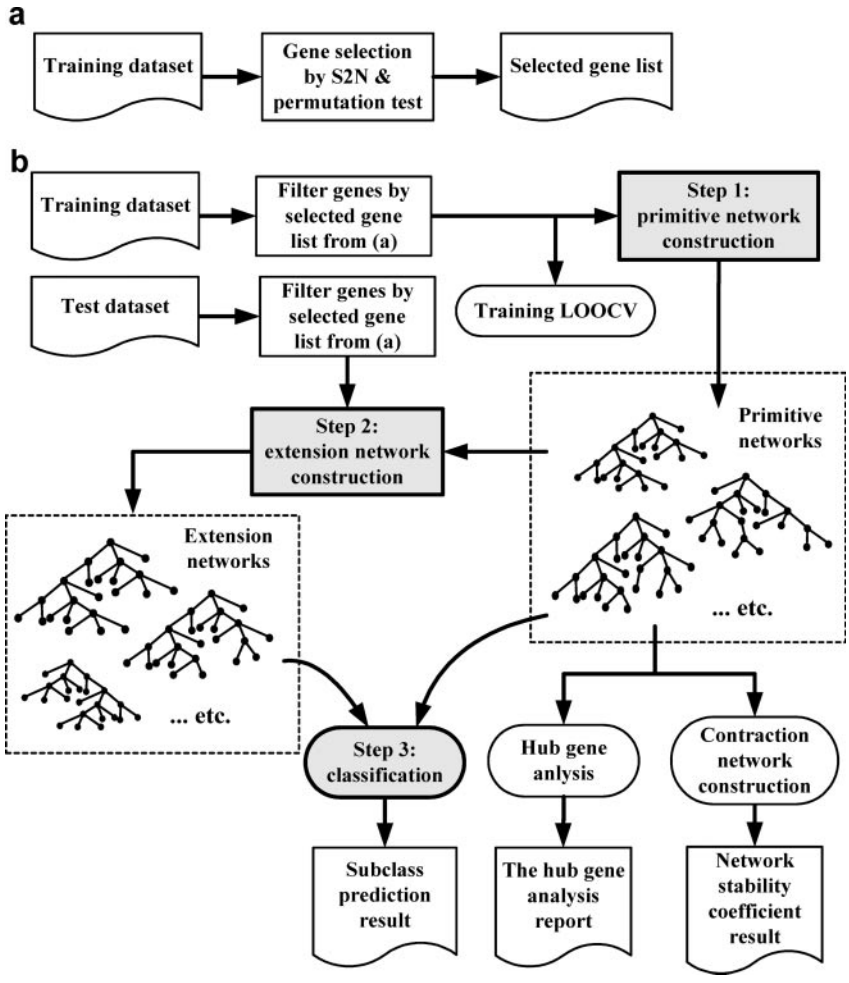

Figure 1. Flowchart of topology-based cancer classification framework. (**a**) Gene selection by S2N and permutation test is performed in the training dataset. (**b**) In order to reduce the noise and investigate the impact of various gene numbers on the classification performance, the microarray data are filtered by the selected gene list that is derived from the above (a). The topological classification framework includes three major steps: (i) the primitive gene network construction; (ii) the extension gene network construction and (iii) the computed similarity between the primitive and extension networks for cancer classification. The primitive networks are also applied to perform hub gene analysis and the calculation of network stability coefficient. A leave-one-out cross-validation (LOOCV) of the training dataset is performed to obtain a training accuracy before using the test dataset, and the training accuracy can indicate the quality of the training dataset.

Elimination of the common edges

To improve the classification accuracy and obtain the class-specific networks, we developed the method to eliminate the common edges (non-class-specific edges). For each subclass, a primitive network is constructed from the samples of this subclass in the training dataset, and each primitive network corresponds to a specific subclass. If an edge (a connection between two genes) of the network was concurrent on many various primitive networks, it was called 'common edge', which might decrease the discriminating capacity of class-specific networks. We also defined the common edge of a primitive network to occur in 80% of all primitive networks (the great majority of the primitive networks). We established the common edge list by examining all of the edges of all the primitive networks, and then reconstructed all the primitive networks by eliminating all of the common edges. The common edge list would continue to be used in the further network construction (extension and contraction networks) in the following procedures.

Topology-based cancer classification framework

A novel topology-based cancer classification framework is illustrated as Figure 1. The details of classification procedure are shown in Figure 2 and the pseudo code of the topology-based classification algorithm can be found in Supplementary Table S2. In addition, to measure the similarity of network topology, we investigated the classification performance with three topological quantities, including degree vector (DV), clustering coefficient vector (CCV) and weighted adjacency distribution (WAD), which are described in Supplementary Data. Due to the well performance of classification, the DV was chosen for the measurement of the network similarity in the following description. There are three major steps in the algorithm and described as follows:

Step 1: primitive gene network construction. First, the network construction method and the elimination of the common edges, as described above, were utilized to construct the primitive network of each subclass in the training dataset.

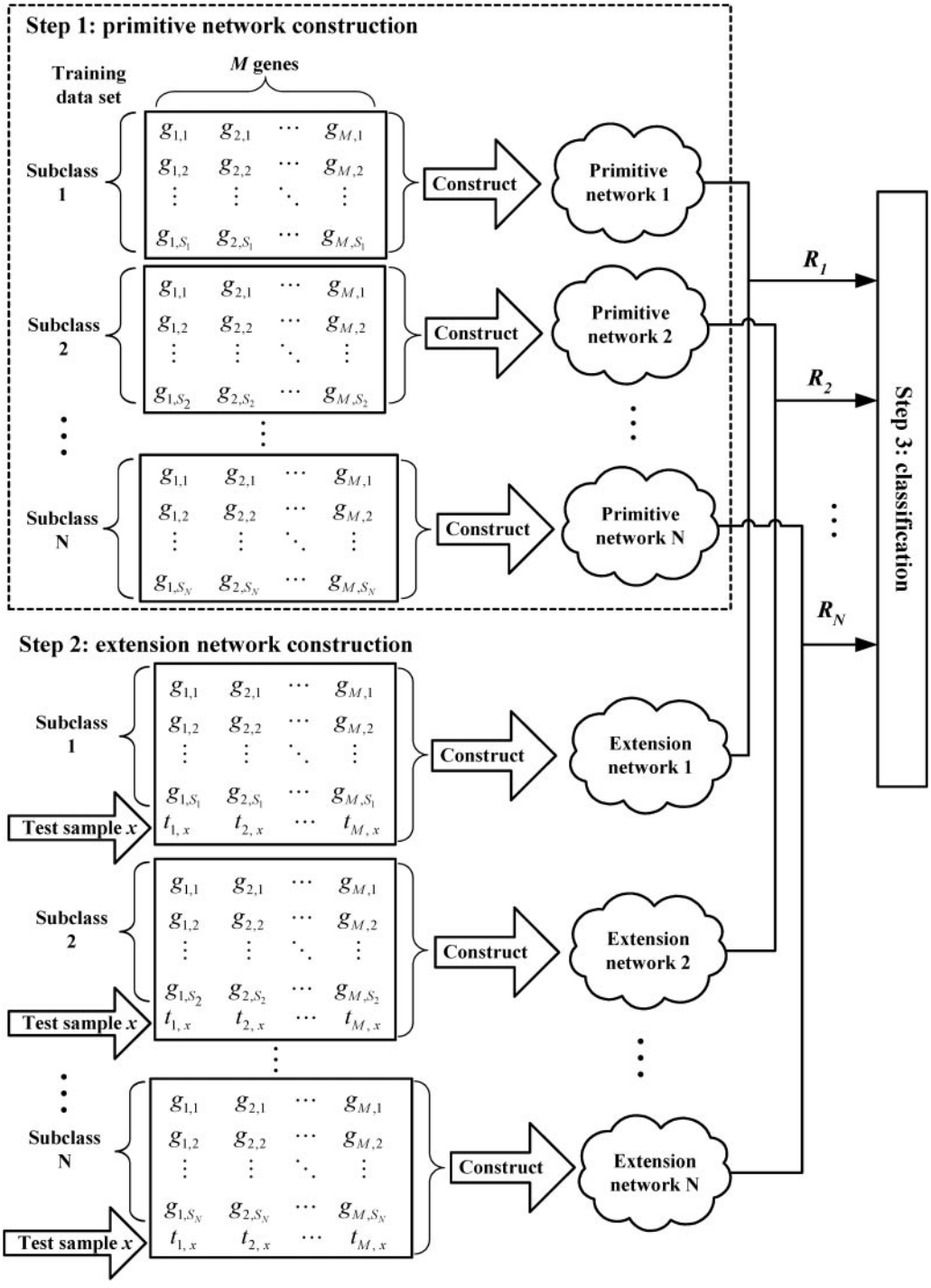

Figure 2. The diagram of the topology-based classification algorithm. If the sample number of subclass i is S_i , then in the matrix of the training dataset, each column represents a gene (total M genes), and each row represents a sample. $g_{i,j}$ is the gene expression level of the gene i in sample j. Given that the gene expression data of test sample x is from the test dataset, x takes the subclass with the largest correlation coefficient of $R_1, R_2, \ldots, R_N$. This diagram is just a representative of the classification procedures for one test sample, since others follow the same way.

Each subclass of the training dataset can represent a matrix (each column represents a gene and each row represents a sample) (see Figure 2). One primitive network corresponded to one cancer subclass as well as a matrix. The primitive networks had the fundamental characters for the corresponding subclasses of cancer.

Step 2: extension gene network construction. For each test sample (microarray gene expression data) in the test dataset that went to classification, all of the matrices in the primitive networks would be extended for this test sample (see Figure 2). Thus all of the primitive networks were reconstructed after adding the test sample into the matrices of all

the cancer subclasses. These reconstructed networks were named as the extension networks. Figure 2 shows the method of the extension network construction. Suppose the test sample is x, there are N subclasses, and all the primitive networks are: $P_1, P_2, \ldots, P_N$. Then, we constructed the extension networks: $E_1, E_2, \ldots, E_N$ for the sample x, where E_i is constructed from the primitive network P_i and test sample x.

Step 3: cancer classification. The extension networks ($E_1, E_2, \ldots, E_N$) were constructed for test sample x in step 2. Since the topological quantity for characterizing the network was a DV, the Pearson correlations of the DV between the primitive networks and the extension networks were calculated to estimate the network similarity. The pairs of primitive networks and extension networks were: (P_1, E_1), (P_2, E_2), ..., (P_N, E_N), and then a series of correlation coefficients $R_1, R_2, \ldots, R_N$ were generated, where R_i is the correlation coefficients of the DV between the primitive network P_i and the extension network E_i. The test sample x then takes the subclass with the largest correlation coefficient as the predicted subclass. If the sample x belongs to the subclass k, the extension network E_k should have minimum topology change from P_k. Therefore, the correlation coefficient R_k of (P_k, E_k) should have the maximum value.

Topology-based pathway analysis

The pathway analyses are the advanced microarray data analyses, in which the gene lists usually selected from the stringent criterion. The previous pathway analyses might ignore the genes without significant differential expression. We developed a novel method, named topology-based pathway analysis, which utilizes the topological quantity of the ordering networks without gene selection for pathway analysis. In addition, it is a way to demonstrate the biomedical significance of the ordering networks. The comprehensive web server CRSD (our previous work) (38) is used to provide the pathway information in this study (http://biochip.nchu.edu.tw/crsd1/), which included the Kyoto Encyclopedia of Genes and Genomes (KEGG) (39,40) and BioCarta (http://www.biocarta.com) pathways. The method of topology-based pathway analysis is as follows: (i) analyze all pathways (KEGG and BioCarta) and extract the gene symbols of each pathway; (ii) extract the microarray data associated to the gene symbols of each pathway; (iii) the ordering networks based on the gene symbols of each pathway are constructed; (iv) eliminate the common edges using the procedures described above and (v) compute the topology impact score by the common and remained edges for each pathway as follows:

$$\text{topology impact score} = \frac{e_1 e_2 \cdots e_N}{(e_1 + e_{\text{comm}})(e_2 + e_{\text{comm}}) \cdots (e_N + e_{\text{comm}})}, \quad \mathbf{2}$$

where N is the number of subclass, e_i is the remained edge number of the network of ith subclass, and e_{comm} is the number of the common edges.

RESULTS

Selection of topological quantity: the DV

Three topological quantities (DV, CCV and WAD) were applied to all of the tested datasets for the evaluation of classification accuracies (see Supplementary Figures S1–S6). Both the correlation and the ordering network constructions are included for the performance comparison. Classification experiments with various selected gene number per subclass (10, 20, ..., 100) were performed. The average accuracy, the best accuracy and the standard deviation of accuracies for all datasets are shown in Table 1. The average accuracy

Table 1. The comparison of three topological quantities between the correlation and ordering networks

Dataset	Statistic[c]	Correlation[a] DV	CCV	WAD	Ordering[b] DV	CCV	WAD
ALL-subtype	Average	0.9247	0.9012	0.8788	0.9659	0.8706	0.7647
	Maximum	0.9294	0.9294	0.9647	0.9882	0.9765	0.8235
	SD	0.0149	0.0381	0.0392	0.0151	0.0620	0.0283
GCM	Average	0.3957	0.3978	0.5500	0.6783	0.3978	0.6326
	Maximum	0.4348	0.5000	0.6304	0.7174	0.4565	0.6522
	SD	0.0367	0.0492	0.0562	0.0321	0.0411	0.0216
Lung-cancer	Average	0.9906	0.9805	0.9946	0.9960	0.9423	0.9859
	Maximum	0.9933	0.9933	1.0000	1.0000	0.9933	0.9933
	SD	0.0085	0.0124	0.0053	0.0035	0.0278	0.0080
Lung-subtype-1	Average	0.8985	0.8361	0.8238	0.9178	0.8718	0.3104
	Maximum	0.9257	0.8812	0.9109	0.9307	0.9059	0.3168
	SD	0.0150	0.0264	0.0893	0.0097	0.0208	0.0078
Lung-subtype-2	Average	0.9721	0.9519	0.9473	0.9667	0.9279	0.8612
	Maximum	0.9845	0.9767	0.9535	0.9690	0.9457	0.9767
	SD	0.0138	0.0219	0.0080	0.0037	0.0110	0.2638
MLL-leukemia	Average	1.0000	0.9467	0.9533	1.0000	0.9467	1.0000
	Maximum	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	SD	0.0000	0.1080	0.0549	0.0000	0.0878	0.0000

The topology-based classification analyses were performed to compare the three topological quantities (DV; CCV; WAD) in all the tested datasets. The accuracies of the classification experiments are listed in this table.

[a]The classification is based on the correlation network construction with the permutation significance threshold ($P < 0.05$).

[b]The classification is based on the ordering network construction with the permutation significance threshold ($P < 0.05$).

[c]The classification experiments are performed by using various gene number per subclass, such as 10, 20, ..., 100 genes. The statistic quantities are the average accuracy, the best accuracy (Maximum) and the SD of accuracies.

can assess the classification accuracy while the standard deviation of accuracy can assess the stability of the classification method.

The results in Table 1 revealed several constructive information: (i) average accuracy: the best values occurred on the DV in all six datasets (consider the maximum value of each row for different topological quantities and networks in the average accuracies); (ii) standard deviation: there are four datasets, except for GCM and Lung-subtype-1, having the minimum values of standard deviation on the DV; (iii) ordering networks: there are five datasets, except for Lung-subtype-2, achieving the best average accuracy, and there are also five datasets, except for the ALL-subtype, having the minimum standard deviation and (iv) correlation networks: there are only two datasets achieving the best average accuracy, and there are also two datasets having the minimum standard deviation. Hence, the DV is the best choice and the ordering networks are obviously much better than the correlation networks on both accuracy and stability.

Comparison of classification accuracies

A previous report compared the performance of several machine learning classifiers, including support vector machines (SVM), Naive Bayes, K-Nearest Neighbor (KNN) and Decision Tree, and indicated that SVM is a better classifier than others (35). To compare the accuracies of the topology-based classification framework with other approaches, aside from the classification algorithms previously published in the tested datasets, the SVM multiclass (41) was included in this study. The classification accuracies of the original papers, the SVM multiclass and the topology-based approach in all of the tested datasets were shown in Table 2. Top 40 and 80 genes per subclass were selected based on S2N score with significant level 0.05 in the permutation test. Except for the GCM dataset, the accuracies with the ordering networks (top 40 genes per subclass) were better than the originally reported accuracies, and 93% (98% on average). The ordering networks, without respect to 40 or 80 genes, had better accuracies than the correlation networks in the ALL-subtype, GCM and Lung-subtype-1 datasets, and had the same accuracies as the correlation networks in the Lung-cancer and MLL-leukemia datasets. It was evident that the ordering network construction is the better approach.

Table 2. Classification accuracies of all the tested datasets

Dataset	Original[c]	SVM[d]	Correlation[a] 40	Correlation[a] 80	Ordering[b] NSC[e]	Ordering[b] 40	Ordering[b] 80	Best[f]
ALL-subtype	0.96	1.00	0.93	0.93	0.96	0.99	0.96	0.99
GCM	0.78	0.70	0.37	0.41	0.90	0.72	0.67	0.72
Lung-cancer	0.96	0.99	0.99	0.99	0.95	0.99	0.99	1.00
Lung-subtype-1	0.87	0.94	0.91	0.88	0.95	0.93	0.91	0.93
Lung-subtype-2	—	0.93	0.98	0.98	0.94	0.97	0.97	0.97
MLL-leukemia	0.95	1.00	1.00	1.00	0.96	1.00	1.00	1.00

Top 40 and 80 genes per subclass are selected based on S2N score and permutation test ($P < 0.05$).

[a]The classification is based on the correlation network construction with the permutation significance threshold ($P < 0.05$).

[b]The classification is based on the ordering network construction with the permutation significance threshold ($P < 0.05$).

[c]The classification methods of the original papers: ALL-subtype, SVM; GCM, SVM; Lung-cancer, gene expression ratios; Lung-subtype-1, KNN; MLL-leukemia, KNN.

[d]For the purpose of comparison, SVM multiclass is performed for the classification.

[e]The average of the NSC (network stability coefficient) of all subclasses with ordering network constructed by the top 40 genes per subclass.

[f]The classification experiments are performed by using various gene number per subclass, such as 10, 20, . . . , 100 genes. The best accuracy among these cases was reported.

The ordering networks constructed by the top 40 genes also had the better accuracies than the SVM multiclass in the GCM and Lung-subtype-2 datasets, and had equal accuracies in the Lung-cancer and MLL-leukemia datasets. In all the tested datasets, the accuracies of the ordering networks (top 40 genes per subclass) were between 72 and 100% (93.3% on average), and better than the accuracies of SVM that were between 70 and 100% (92.7% on average). Generally speaking, the classification accuracies of the SVM multiclass and topology-based ordering network were better than those of previously published results using the same datasets. However, both algorithms did not achieve the originally reported accuracy in the GCM dataset (78%). Interestingly, our results indicated that the GCM dataset had the minimum network stability coefficient (NSC, see Supplementary Data), which imply that accuracy is hard to increase when NSC is low. Our results also showed that the best classification accuracy may not occur at the condition of top 40 or 80 genes, but their performances approximate the best accuracy. Therefore, classification using the topology-based ordering network is a stable approach.

To evaluate the true accuracy, sensitivity and specificity of classification prediction in all of the datasets, as well as the function of the NSC, the ordering networks and top 40 genes were applied to this study (Supplementary Tables S3–S8). The average specificities were between 97 and 100% in all the subclasses of all the tested datasets. Except for the GCM dataset, the true accuracies were between 93 and 100% as well as the average sensitivities were between 96 and 100%. In the ALL-subtype and MLL-leukemia datasets, the majority of subclasses had 100% accuracy, sensitivity and specificity. For example, there were four subclasses that had 100% true accuracy and five subclasses with 100% sensitivity and specificity in the ALL-subtype dataset.

Large-scale gene networks and power-law node-degree distribution

We further investigated the properties of the topology-based framework with either the ordering or correlation networks on the construction of large-scale gene networks and the power-law node-degree distribution. As illustrated in Figure 3a, the results showed that the classification accuracies of the ordering networks were better than those of the correlation networks in all of the tested datasets while the networks were large-scale. Furthermore, the ALL-subtype dataset was used to verify the power-law node-degree distribution property in the large-scale ordering and correlation gene networks (Figure 3b). The ordering networks of the ALL-subtype dataset had the degree exponent r between 1.72 and 1.96, and determination coefficient $R^2 \geqslant 0.84$ (0.91 on average), and the correlation networks had the degree exponent r between 2.76 and 3.10 and $R^2 \geqslant 0.68$ (0.79 on average). In contrast, R^2 were between 0.32 and

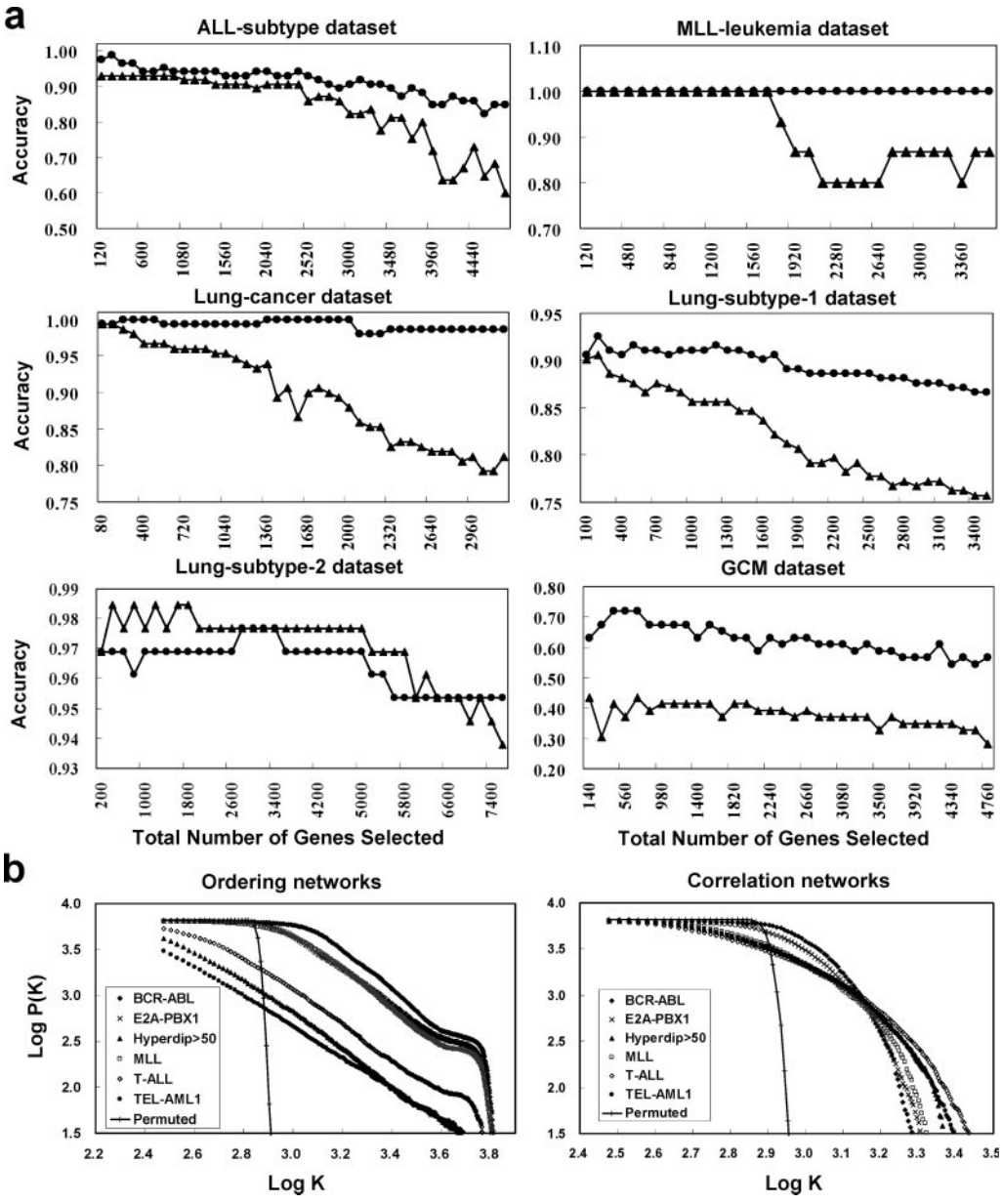

Figure 3. Large-scale gene networks and the power-law node-degree distribution. (**a**) Classification accuracy profiles of large networks with the ordering and correlation networks. All of the tested datasets exhibit the classification accuracy profiles of the large-scale networks. The solid circle represents the ordering network, while the solid triangle represents the correlation network. The accuracy of the correlation networks decreased and became high variance while the network size increased and the gene number was in excess of some threshold. (**b**) The ALL-subtype dataset is used to verify the power-law node-degree distribution property in the ordering and correlation networks. These are the log-log plots of degree K versus the number of nodes with degree $\geqslant$ K i.e. P(K), and there are 6602 genes in the both ordering and correlation networks. The linear regression measures the linearity between log[P(K)] and log(K), which is a condition of the power-law node-degree distribution, where P(K) can be represented by power-law with a degree exponent r : $P(K) \approx K^{-r}$. The determination coefficient R^2 ranges from 0 to 1, with 1 representing perfect linearity (i.e. a perfect power-law distribution). The permuted network is constructed by randomly permuting all of the edges from the T-ALL subclass network. The degree exponent r and determination coefficient R^2 of each subclass-specific network are shown in Supplementary Table S9.

0.36 in the permuted networks. The results indicated that the topology-based ordering networks also possessed power-law node-degree distribution and were better than the correlation networks.

Hub gene and topology-based pathway analysis

The details of hub gene analysis method are described in Supplementary Data. The hub gene annotations and the adjacency matrices of the ordering networks in the ALL-subtype, Lung-subtype-1 and Lung-cancer datasets are available at our web site (http://biochip.nchu.edu.tw/supl/top).

To demonstrate the availability of topology-based pathway analysis, the Lung-cancer dataset was selected for verification. There are two classes in the Lung-cancer dataset: malignant pleural mesothelioma (MPM) and lung adenocarcinoma (AD). These two contained 181 tissue samples (31 MPM and 150 AD), of which 31 MPM and 31 AD samples were

selected to construct the ordering networks and correlation networks. We analyzed 115 KEGG and 317 BioCarta pathways, and there were remained 102 KEGG and 208 BioCarta pathways after filtering the pathways with the gene number less than 8. We sketched the ordering networks for 310 pathways (see our web site http://biochip.nchu.edu.tw/supl/top), and the ordering networks of the top two pathways, by the topology impact score, were shown in Figure 4a and c, in which the networks include MPM- and AD- specific networks. Also, the correlation networks of these two pathways were also shown in Figure 4b and d. MPM- and AD-specific networks in each pathway, eicosanoid and ascorbate/aldarate metabolisms, showed that the network topology had apparent differences between two class-specific networks, as well as between the hub genes.

DISCUSSION

Both the molecular classification and the pathway analysis based on microarray data are the advanced approaches to

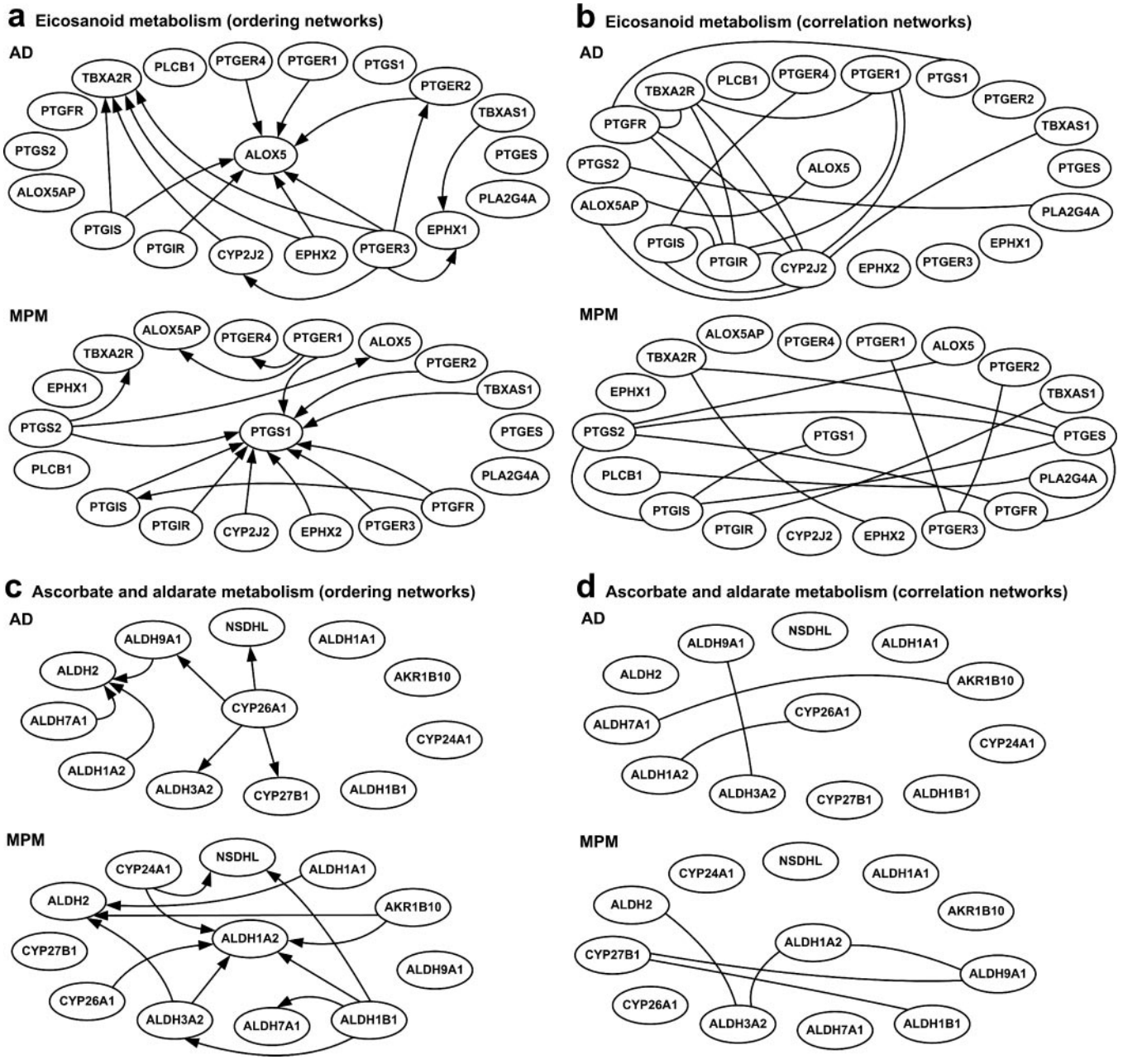

Figure 4. The ordering and correlation network diagrams for the significant pathways. The Lung-cancer dataset was used for topology-based pathway analysis, and the two significant pathways, derived from the BioCarta and KEGG pathway databases, respectively. (**a**) Ordering networks of eicosanoid metabolism; (**b**) correlation networks of eicosanoid metabolism; (**c**) ordering networks of ascorbate and aldarate metabolism; (**d**) correlation networks of ascorbate and aldarate metabolism. Two class-specific networks (MPM and AD) are shown for each pathway. The network topology reveals the apparent difference between the two class-specific networks, as well as the change of the hub gene. In the ordering networks, the hub gene is *ALOX5* in AD but is *PTGS1* in MPM (a); the hub gene is *CYP26A1* in AD but is *ALDH1A2* in MPM (c).

characterize the biological properties and processes. In this report, we presented a novel and highly efficient computational framework for topology-based classification by using microarray data, which is a simple and stable approach. The classification accuracies of the original papers, the SVM multiclass, and the topology-based approach were reported in this study. The SVM classifier constructs a complicated mapping between samples and their class labels, which functions as a black box (42,43); therefore the inside information of SVM classifier cannot interpret biological significance. Although the SVM (for gene selection) and enrichment analysis can be integrated to perform pathway analysis, which might ignore the genes that play important role in certain pathways but without significant differential expression in different classes. The results of this type of pathway analysis might be difficult to interpret the biological significance. In contrast, the ordering network construction was not only utilized to the classification, but also can benefit the biological pathway analysis, hub gene analysis and visualization.

In this study, we first utilized the correlation network for developing the topology-based classification. The classification performance of the correlation networks was impressive in some conditions. However, their accuracy and stability in all of the tested datasets are not very good, particularly in large-scale gene networks (see Figure 3a). The correlation networks are linear relationship but the gene regulations are almost nonlinear (44,45), which might be a reason to affect the accuracy and stability. Therefore, we further developed the ordering network construction for classification. The ordering relationship developed in this study is a nonlinear measurement, which may add some information in the analysis of complex gene regulation system. Our results showed that the topology-based ordering networks are suitable for a classification study based on gene expression profiles and may help to interpret biological meanings through pathway analysis.

It is undoubted that cells may maintain some conservative ordering relationships among the essential genes in some situations, which may be more important than the correlation relationships. For example, flower identities are tightly modulated by three groups of genes, which are well known as ABC model in *Arabidopsis thaliana* and many other plant species. The ordering relationships of the three groups of gene expressions can dramatically affect flower differentiation and development. The tissue will differentiate into sepals and petals when the expression level of group A is higher than group C; in contrast, if the ordering relationship is reversed as group C > A, the same tissue can develop into stamens and carpels (28). Another example is the ordering relationship of *CDH1* (*E-cadherin*) and *CDH2* (*N-cadherin*), which has biomedical significance in human tumor progression (29). *CDH1* is expressed abundantly but the expression of *CDH2* is not detected in prostate glandular epithelium (*CDH1* > *CDH2*). However, human prostate carcinoma cell lines show loss of *CDH1* and expression of *CDH2* (*CDH1* < *CDH2*) (29). Thus, the ordering relationship of gene expression levels between *CDH1* and *CDH2* is reversed in different class. It is evident that the ordering networks could play some roles in interpreting the biological significance.

Currently, many methods on gene network constructions do not allow establishing large-scale networks because of the limitation of biological databases (27), the high time complex (36) and the noise accumulation (35) during the large-scale network construction. Real gene networks are very complex and enormous, so the capability of large-scale network analysis is very important. It is evident that the topology-based framework with ordering relationship can handle large-scale networks very well (see Figure 3a). The accuracy of the correlation networks decreased and was highly variable when the network size increased. Therefore, the ordering networks may tolerate the noise accumulation efficiently in large-scale networks. The performances of the ordering networks are better in accuracy and stability in all of the tested datasets when the networks are large-scale.

To further investigate the topological properties of the ordering networks, we analyzed the power-law node-degree distribution property of the ordering and correlation gene networks. The log-log plots of degree K versus the number of nodes with degree $\geqslant$ K are shown in Figure 3b. The topological properties of interactome maps usually possess the degree exponent r between 1.59 and 2.75 and determination coefficient $R^2 \geqslant 0.84$, as seen in several experimental datasets (27). In the ALL-subtype dataset, the ordering networks have the degree exponent r between 1.72 and 1.96 and $R^2 \geqslant 0.84$, and the correlation networks have the degree exponent r between 2.76 and 3.10 and $R^2 \geqslant 0.68$ (see Figure 3b and Supplementary Table S9). Our results suggested that the ordering networks have stronger power-law distribution property and are consistent with the previous study (27). Although, the various biological networks revealed the power-law node-degree distributions in the previous studies (15,26,27), it is unproven that this property could be as a criterion to assess the quality of gene network construction.

The ordering network construction is also low in time complexity. We compared the execution time of topology-based classification between ordering networks and correlation networks by using the top 40 and 200 genes per subclass (see Supplementary Table S10). The results revealed that the average execution time of the ordering networks constructed by the top 40 and 200 genes decreased 6 and 24% as compared with those of the correlation networks, respectively. Therefore, the larger the network size, the better the performance of the ordering network.

Genes with a high degree of connections with other genes may dominate the network topology and are the hub genes of the network, which are the critical genes in a network (15). Most genes are connected to the hub genes by a relatively short path (12,13). In this study, we analyzed the node-degree distributions of the ordering networks to demonstrate its biological significance by the previous literature reports in three datasets, including the ALL-subtype, Lung-subtype-1 and Lung-cancer. A total of 52 genes were selected as the hub genes of the ALL-subtype dataset, 42 as the hub genes of the Lung-subtype-1 dataset, and 24 as the hub genes of the Lung-cancer dataset (see Supplementary Tables S11–S19). These three datasets were derived from the microarray experiments using Affymetrix GeneChip array HG_U95Av2, U95Av2, and the gene annotation was obtained from the Affymetrix web site. The adjacency matrices of these networks were available at our web site http://biochip.nchu.edu.tw/supl/top, while the annotation of the hub genes, the

input and output degree distributions, and the corresponding subclass were shown in Supplementary Tables S11–S19. Furthermore, the network diagrams for the hub genes are shown in Supplementary Figure S7 and S8. The results showed that some previous reports can support the biological significance of the hub genes in the ordering networks, especially the genes that have been previously characterized as the subclass-specific markers, such as *CD3D* (4), *CBFA2T3* (46), *TRB@* (47) and so on. The detailed information about the hub gene analysis is described in Supplementary Data.

Functional enrichment analysis with pathways is one of the advanced microarray data analyses that is becoming popular, in which the gene lists usually selected from the stringent criterion, e.g. P-value < 0.05 or fold-change > 2 (30). Then the pathway analysis may identify the underlying biological mechanisms of these gene lists. There are several statistical methods that can be utilized to analyze the intersection of a gene list and a pathway or functional annotation, e.g. the fisher exact test (48), hypergeometric distribution (49) and binomial distribution (50). These analyses might ignore the genes that play important role in certain pathways but without significant differential expression in different classes. Gene set enrichment analysis (GSEA) is a statistical method to detect the subtle difference of individual genes but coordinate changes in the groups of functionally related genes (51). The results of GSEA may be difficult to interpret (30), and there have also been some statistical debate regarding GSEA (52). Therefore, we developed topology-based pathway analysis, which integrates the pathway database, the classification information, the topological quantity and all microarray data without significant differential gene selection. The results can be visualized and be employed to interpret the biological meanings easily.

It is a challenge to characterize the biological process in which the genes' expressions are subtle but consistent changes. The topology-based pathway analysis utilizes the topology change among the ordering gene networks of the different classes to discover class-specific pathways. To demonstrate the biomedical significance of the ordering networks applied to topology-based pathway analysis, the Lung-cancer dataset containing MPM and AD was employed to this study. The top 10 significant pathways were selected from KEGG pathways (see Supplementary Table S20) and the top 20 significant pathways were selected from BioCarta pathways by the topology impact score (see Supplementary Table S21). High topology impact score of pathway indicated high variance between subclasses. The top significant pathway in BioCarta, by the topology impact score, was eicosanoid metabolism, and the top significant pathway in KEGG was ascorbate and aldarate metabolism. To illustrate the significant different network topology, the ordering and correlation networks of these two pathways were shown in Figure 4. The ordering networks are directional graphs and illustrate the relative expression levels and ordering structures in the networks, which have not been discussed yet. On the other hand, the correlation networks are non-directional graphs, which have been discussed in various aspects (15–18). We believe both the ordering networks and the correlation networks have important biological implications.

The eicosanoid metabolism is the important pathway in non-small cell lung-cancer (NSCLC) (53,54). In previous studies, several genes in the eicosanoid metabolism have been investigated in AD or MPM, such as *PTGIS* (5,54), *PTGS1* (55), *PTGS2* (53,55) and *ALOX5* (53). The previous report demonstrated increased expression of *PTGS2* (*Cox-2*) and simultaneous down-regulation of *PTGS1* (Cox-1) in NSCLC (55). Our results showed that *PTGS2* has 0 degree in AD and 3 output degrees in MPM, and *PTGS1* has 0 degree in AD and 10 input degrees in MPM (Figure 4a), which meant the relatively high expression level of *PTGS2* and relatively low expression level of *PTGS1* in AD. On the other hand, both *CYP24A1* and *AKR1B10* are involved in the ascorbate and aldarate metabolism and have been demonstrated to be highly expressed in AD (56,57). Our results showed that both *CYP24A1* and *AKR1B10* have 0 degree in AD and 2 output degrees in MPM (Figure 4c), which meant the relatively high expression level in AD. The detailed information of other significant pathways is described in Supplementary Data.

The traditional analysis for microarray data usually filters out the genes with no significant differential expression in different classes. In order to understand whether the significant pathways contained any significant genes, we investigated the relationship between the 200 significant genes derived from S2N selection with permutation test (see our web site http://biochip.nchu.edu.tw/supl/top) and the 30 significant pathways (Supplementary Tables S20 and S21) in the Lung-cancer dataset. There are nine significant pathways that do not contain any significant gene, and the complete list is described in Supplementary Data. Although the genes of these nine pathways are not significant, our results revealed that the local network structures of the pathways are significantly changed. For example, some of these pathways have been proved their biological significance by the previous reports, such as the CBL mediated ligand-induced down-regulation of EGF receptors and the erythrocyte differentiation pathway (see Supplementary Data). It is possible that a gene with subtle change can still contribute to the complex regulatory mechanism, and impact on the local network structure. The topology-based pathway analysis using the ordering network is the novel approach to observe the behavior of the local networks, and it can discover the impact on the local network structure.

In summary, we have developed the topology-based classification framework and pathway analysis by using the ordering networks. The accuracy and stability of the classification performance were demonstrated in the topology-based classification. Because of literature reports, the hub genes and significant pathways based on the ordering networks have the potential biological significance. In addition, according to the significant difference of the hub genes and significant pathways between subclasses, they can help to discover new insights in the underlying molecular mechanisms related to disease development.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by the National Chung-Hsing University Biotechnology Center through the Teaching Core

Facility Project grant, as well as partially supported by the National Science Council grant (NSC 95-2314-B-005-005-MY3). Funding to pay the Open Access publication charges for this article was provided by National Chung-Hsing University.

Conflict of interest statement. None declared.

REFERENCES

1. Morley,M., Molony,C.M., Weber,T.M., Devlin,J.L., Ewens,K.G., Spielman,R.S. and Cheung,V.G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.
2. Chen,J.J., Peck,K., Hong,T.M., Yang,S.C., Sher,Y.P., Shih,J.Y., Wu,R., Cheng,J.L., Roffler,S.R., Wu,C.W. *et al.* (2001) Global analysis of gene expression in invasion by a lung cancer model. *Cancer Res.*, **61**, 5223–5230.
3. Chen,J.J., Lin,Y.C., Yao,P.L., Yuan,A., Chen,H.Y., Shun,C.T., Tsai,M.F., Chen,C.H. and Yang,P.C. (2005) Tumor-associated macrophages: the double-edged sword in cancer progression. *J. Clin. Oncol.*, **23**, 953–964.
4. Yeoh,E.J., Ross,M.E., Shurtleff,S.A., Williams,W.K., Patel,D., Mahfouz,R., Behm,F.G., Raimondi,S.C., Relling,M.V., Patel,A. *et al.* (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.
5. Gordon,G.J., Jensen,R.V., Hsiao,L.L., Gullans,S.R., Blumenstock,J.E., Ramaswamy,S., Richards,W.G., Sugarbaker,D.J. and Bueno,R. (2002) Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res.*, **62**, 4963–4967.
6. Armstrong,S.A., Staunton,J.E., Silverman,L.B., Pieters,R., den Boer,M.L., Minden,M.D., Sallan,S.E., Lander,E.S., Golub,T.R. and Korsmeyer,S.J. (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genet.*, **30**, 41–47.
7. di Bernardo,D., Thompson,M.J., Gardner,T.S., Chobot,S.E., Eastwood,E.L., Wojtovich,A.P., Elliott,S.J., Schaus,S.E. and Collins,J.J. (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.*, **23**, 377–383.
8. Gardner,T.S., di Bernardo,D., Lorenz,D. and Collins,J.J. (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102–105.
9. Segal,E., Friedman,N., Kaminski,N., Regev,A. and Koller,D. (2005) From signatures to models: understanding cancer using microarrays. *Nature Genet.*, **37**, S38–S45.
10. Calvano,S.E., Xiao,W., Richards,D.R., Felciano,R.M., Baker,H.V., Cho,R.J., Chen,R.O., Brownstein,B.H., Cobb,J.P., Tschoeke,S.K. *et al.* (2005) A network-based analysis of systemic inflammation in humans. *Nature*, **437**, 1032–1037.
11. Guido,N.J., Wang,X., Adalsteinsson,D., McMillen,D., Hasty,J., Cantor,C.R., Elston,T.C. and Collins,J.J. (2006) A bottom-up approach to gene regulation. *Nature*, **439**, 856–860.
12. Strogatz,S.H. (2001) Exploring complex networks. *Nature*, **410**, 268–276.
13. Reka,A. and Albert-Laszlo,B. (2002) Statistical mechanics of complex networks. *Rev. Modern Phys.*, **74**, 47.
14. Parsons,A.B., Brost,R.L., Ding,H., Li,Z., Zhang,C., Sheikh,B., Brown,G.W., Kane,P.M., Hughes,T.R. and Boone,C. (2004) Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nat. Biotechnol.*, **22**, 62–69.
15. Carter,S.L., Brechbuhler,C.M., Griffin,M. and Bond,A.T. (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, **20**, 2242–2250.
16. Segal,E., Shapira,M., Regev,A., Pe'er,D., Botstein,D., Koller,D. and Friedman,N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genet.*, **34**, 166–176.
17. Steuer,R., Kurths,J., Fiehn,O. and Weckwerth,W. (2003) Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, **19**, 1019–1026.
18. Butte,A.J., Tamayo,P., Slonim,D., Golub,T.R. and Kohane,I.S. (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl Acad. Sci. USA*, **97**, 12182–12186.
19. Segal,E., Friedman,N., Koller,D. and Regev,A. (2004) A module map showing conditional activity of expression modules in cancer. *Nature Genet.*, **36**, 1090–1098.
20. Ramo,P., Kesseli,J. and Yli-Harja,O. (2005) Stability of functions in Boolean models of gene regulatory networks. *Chaos*, **15**, 34101.
21. Helman,P., Veroff,R., Atlas,S.R. and Willman,C. (2004) A Bayesian network classification methodology for gene expression data. *J. Comput. Biol.*, **11**, 581–615.
22. Kelley,R. and Ideker,T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.*, **23**, 561–566.
23. Tamada,Y., Kim,S., Bannai,H., Imoto,S., Tashiro,K., Kuhara,S. and Miyano,S. (2003) Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, **19**, II227–II236.
24. MacCarthy,T., Pomiankowski,A. and Seymour,R. (2005) Using large-scale perturbations in gene network reconstruction. *BMC Bioinformatics*, **6**, 11.
25. Barabasi,A.L. (2002) *Linked: The New Science of Networks*. Perseus Books Group. Cambridge, MA, pp. 55–65.
26. Jeong,H., Mason,S.P., Barabasi,A.L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
27. Han,J.D., Dupuy,D., Bertin,N., Cusick,M.E. and Vidal,M. (2005) Effect of sampling on topology predictions of protein–protein interaction networks. *Nat. Biotechnol.*, **23**, 839–844.
28. Espinosa-Soto,C., Padilla-Longoria,P. and Alvarez-Buylla,E.R. (2004) A gene regulatory network model for cell-fate determination during *Arabidopsis thaliana* flower development that is robust and recovers experimental gene expression profiles. *Plant Cell*, **16**, 2923–2939.
29. Tran,N.L., Nagle,R.B., Cress,A.E. and Heimark,R.L. (1999) N-Cadherin expression in human prostate carcinoma cell lines. An epithelial-mesenchymal transformation mediating adhesion with Stromal cells. *Am. J. Pathol.*, **155**, 787–798.
30. Curtis,R.K., Oresic,M. and Vidal-Puig,A. (2005) Pathways to the analysis of microarray data. *Trends Biotechnol.*, **23**, 429–435.
31. Ramaswamy,S., Tamayo,P., Rifkin,R., Mukherjee,S., Yeang,C.H., Angelo,M., Ladd,C., Reich,M., Latulippe,E., Mesirov,J.P. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.
32. Bhattacharjee,A., Richards,W.G., Staunton,J., Li,C., Monti,S., Vasa,P., Ladd,C., Beheshti,J., Bueno,R., Gillette,M. *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.
33. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
34. Statnikov,A., Aliferis,C.F., Tsamardinos,I., Hardin,D. and Levy,S. (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, **21**, 631–643.
35. Li,T., Zhang,C. and Ogihara,M. (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, **20**, 2429–2437.
36. de la Fuente,A., Bing,N., Hoeschele,I. and Mendes,P. (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, **20**, 3565–3574.
37. Liu,H.C. (2003) A study on mixing semantic structure analysis. *J. Educational Measurement Stat.*, **11**, 1–16(in Chinese).
38. Liu,C.C., Lin,C.C., Chen,W.S.E., Chen,H.Y., Chang,P.C., Chen,J.J.W. and Yang,P.C. (2006) CRSD: a comprehensive web server for composite regulatory signature discovery. *Nucleic Acids Res.*, **34**, W571–W577.
39. Kanehisa,M. (1997) A database for post-genome analysis. *Trends Genet.*, **13**, 375–376.
40. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
41. Crammer,K. and Singer,Y. (2001) On the Algorithmic Implementation of Multi-class SVMs. *Journal of Machine Learning Research*, **2**, 265–292.

42. Byvatov,E. and Schneider,G. (2004) SVM-based feature selection for characterization of focused compound collections. *J. Chem. Inf. Comput. Sci.*, **44**, 993–999.
43. Li,J., Liu,H., Downing,J.R., Yeoh,A.E. and Wong,L. (2003) Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. *Bioinformatics*, **19**, 71–78.
44. Thomas,R. (1998) Laws for the dynamics of regulatory networks. *Int. J. Dev. Biol.*, **42**, 479–485.
45. Goutsias,J. and Kim,S. (2004) A nonlinear discrete dynamical model for transcriptional regulation: construction and properties. *Biophys. J.*, **86**, 1922–1945.
46. Lindberg,S.R., Olsson,A., Persson,A.M. and Olsson,I. (2005) The Leukemia-associated ETO homologues are differently expressed during hematopoietic differentiation. *Exp. Hematol.*, **33**, 189–198.
47. Soulier,J., Clappier,E., Cayuela,J.M., Regnault,A., Garcia-Peydro,M., Dombret,H., Baruchel,A., Toribio,M.L. and Sigaux,F. (2005) HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). *Blood*, **106**, 274–286.
48. Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.
49. Yap,Y.L., Lam,D.C., Luc,G., Zhang,X.W., Hernandez,D., Gras,R., Wang,E., Chiu,S.W., Chung,L.P., Lam,W.K. *et al.* (2005) Conserved transcription factor binding sites of cancer markers derived from primary lung adenocarcinoma microarrays. *Nucleic Acids Res.*, **33**, 409–421.
50. Rhodes,D.R., Kalyana-Sundaram,S., Mahavisno,V., Barrette,T.R., Ghosh,D. and Chinnaiyan,A.M. (2005) Mining for regulatory programs in the cancer transcriptome. *Nature Genet.*, **37**, 579–583.
51. Mootha,V.K., Lindgren,C.M., Eriksson,K.F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstrale,M., Laurila,E. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.*, **34**, 267–273.
52. Damian,D. and Gorfine,M. (2004) Statistical concerns about the GSEA procedure. *Nature Genet.*, **36**, 663.
53. Laskin,J.J. and Sandler,A.B. (2003) The importance of the eicosanoid pathway in lung cancer. *Lung Cancer*, **41**, S73–S79.
54. Stearman,R.S., Dwyer-Nield,L., Zerbe,L., Blaine,S.A., Chan,Z., Bunn,P.A.,Jr, Johnson,G.L., Hirsch,F.R., Merrick,D.T., Franklin,W.A. *et al.* (2005) Analysis of orthologous gene expression between human pulmonary adenocarcinoma and a carcinogen-induced murine model. *Am. J. Pathol.*, **167**, 1763–1775.
55. Ermert,L., Dierkes,C. and Ermert,M. (2003) Immunohistochemical expression of cyclooxygenase isoenzymes and downstream enzymes in human lung tumors. *Clin. Cancer Res.*, **9**, 1604–1610.
56. Beer,D.G., Kardia,S.L., Huang,C.C., Giordano,T.J., Levin,A.M., Misek,D.E., Lin,L., Chen,G., Gharib,T.G., Thomas,D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Med.*, **8**, 816–824.
57. Penning,T.M. (2005) AKR1B10: a new diagnostic marker of non-small cell lung carcinoma in smokers. *Clin. Cancer Res.*, **11**, 1687–1690.