A Gene Link-based Method for Identifying Differential Gene Pathways

Zirui Zhang
School of Mechanical and
Automotive Engineering
Hefei University of Technology
Hefei, Anhui, China
zhangzirui90@163.com

Ke Chen
School of Mechanical and
Automotive Engineering
Hefei University of Technology
Hefei, Anhui, China
k.chen@163.com

Hong-Qiang Wang*
Institute of Intelligent Machines
Chinese Academy of Science
P.O.Box 1130
Hefei 230031, Anhui, China
hqwang126@126.com

***Abstract*—pathway analysis plays an important role in exploring underlying connections between genomic data and complex diseases. In this paper, we propose a gene link–based method for identification of differentially expressed gene pathways. By viewing gene links in a pathway as a Markov chain, the proposed method first develops a gene link Markov chain model (MCM) and devises a Markov chain model-based classification rule to measure the biological importance of a gene link. Then, the expression difference of a pathway is estimated based on all the gene links in the pathway using the gene link MCM. The use of gene links, instead of individual genes, allows for exploring pathway topology that is crucial to pathway activity in cells. Results on two real-world gene expression data sets demonstrate that the effectiveness and efficiency of the proposed method in identifying differential gene pathways.**

Keywords—Pathway analysis; Markov chain Model; Gene link; Cancer clssification

I. Introduction

Molecular processes are responsible for the manifestation or development of caner or other diseases [1, 2]. So, it is crucial to connect molecular processes to cancerous state of cells for deciphering tumorigenesis [3]. Over the past decades, a tremendous amount of high-throughput biological data (HTBD), including deep sequencing data and microarray data, have been generated and accumulated. To mine cancer-related biological knowledge, a number of computational methods have been developed to analyze the genomic data. Many of them aimed to identify differentially expressed genes (DEG) between different conditions, for example, the approaches based on univariate or multivariate statistics [4-8]. However, these methods fail to take advantage of prior biological knowledge available in online databases, such as Gene Ontology (GO) or Kyoto Encyclopedia of Genes and Genomes (KEGG) [9, 10], thus making the results less reproducible and biologically interpretable [11-13]. Biologically, genes co-function together in cells. Prior knowledge about gene networks, *e.g.*, gene pathways, can benefit analyzing HTBD in a biologically meaningful way. Pathway analysis allows investigators to identify differential expressed pathways and are more explanatory than a list of differently expressed genes [14].

Although a number of pathway analysis methods have been developed, most of them equally treat genes in a pathway and ignore the structure information embedded in the pathway networks. Behind these methods, one of important philosophies commonly followed is counting differentially expressed genes and estimating the significance of a pathway against a list of differentially expressed genes using statistical hypothesis test methods, *e.g.*, Fisher’s exact test [15, 16]. Obviously, such kind of methods need to pre-define a list of DEGs and as a result, the performance heavily depends on the value of the cutoff parameter that is chosen for the selection of DEGs. Subramanian *et al* [17] creatively proposed a pathway-level statistic, GSEA, to overcome the shortcoming. The idea behind GSEA has been followed by many pathway analysis methods. Such pathway-level methods generally employed three main steps: 1) Calculating the association of each gene's expression pattern with phenotype by *t*-statistics or correlation measures; 2) Mapping pathway genes and computing pathway-level enrichment evidence scores based on the association scores of the pathway genes; 3) Estimating the significance of each pathway based on empirical distribution of the statistic or a permutation test. Because these methods do not consider the pathway topology, they are not perfect and especially, they will produce a same result for pathways that have a same gene set but different topological structures [18].

Recent studies showed that incorporating pathway topology can lead to better performance of pathway analysis. Pathway topological structure is basically composed of directed gene links. Some methods have been proposed to incorporate topological information into gene-level statistics for testing the significance of pathways. For example, Rahnenfuhrer *et al* [19] developed ScorePAGE by considering the similarity (e.g., correlation, covariance, etc.) between genes in a pathway. Gao and Wang [20] proposed to incorporate second-order information of gene expression to formulate a pathway connectivity index (PCI) of pathway activity. The TAPPA method was motivated by the molecular

This work was supported by the National Natural Science Foundation of China grant nos. 61374181, 61300058, 61272339, 61402010,30900321, and by the Anhui Province Natural Science Foundation grant no. 1408085MF133, and by K. C. Wong education foundation

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

connectivity concept in chemoinformatics. Gene connectivity over pairs of genes in a pathway was summarized to formulate the PCI statistic. TAPPA finally estimated the significance of association between a pathway and a phenotype based on PCI using Mann-Whitney test. Another representative method is the Bayesian Pathway Analysis (BPA) method recently proposed in [21, 22]. In BPA, a biological pathway was modeled as a Bayesian network (BN) after merging repeating entries and deleting cyclicity. BPA preserves gene dependencies entailed by the original pathway. The resulted BN, as a graphical representation of gene interactions rendered by the given pathway, was dealt with using non-informal, uniform belief priors. BPA can quantify the degree to which observed experimental data fit a given BN using Bayesian Dirichlet equivalent (BDe) score and estimate the statistical significance of pathways by testing it against randomization via bootstrapping.

Considering that pathway is in nature a set of gene links that represent various gene associations (binding, inhibition, activation, etc.), we propose to estimate the expression difference of a pathway based on gene links for pathway analysis. Biologically, gene links are directed and reflect cause-and-effect or time-dependent processes, and can dynamically change along with time and biological conditions. Inspired by this, we employed random process theory, Markov chain model, to model gene links for pathway analysis. Briefly, we model a gene link as a Markov chain, by viewing genes to be time tags, to make inference of the activity of pathway gene links in some cellular state. As integral parts of a pathway, gene links can contain subtle but consistent changes in pathway activity. To estimate the significance of a pathway, a permutation test is also devised based on random gene pairs. We evaluated the method on two publicly available gene expression data sets, liver cancer data [23] and ALL data [24].

II. Methods

A. *Markov chain-based modeling of gene link*

1) *Markov chain model*

In statistics, a Markov chain is defined as a stochastic process with Markov property, in which next state depends only on the current state but is not related with any previous event on the time sequence [25]. Such a Markov chain can be mathematically modeled by Markov chain model (MCM). Generally, a MCM consists of three quantities, *i.e.*, a state set, initial state distribution and state transition probability matrix. Given a MCM, one can estimate the occurring probability of an observed Markov chain. There are two types of Markov chain, stationary and non-stationary Markov chain. In stationary Markov chain, it assumes that the transition probability matrix is the same at any transition time, but not the case for non-stationary Markov chain. The division into stationary and non-stationary Markov chains only makes sense to Markov chains with 3 or more time tags.

2) *Gene link MCM*

By viewing gene sequence as a time sequence, we model gene links in a pathway as a Markov chain of length two, similar to that for gene chains in [26]. Given continuous gene expression values, we first use the biology-constrained discretization method [28] to discretize them into three states: down-regulated (-1), non-significantly regulated (0) and up-regulated (1), which constitute the state set S={-1,0,1} for a gene link MCM.

Now assume a gene link l with a starting gene g_1 and an ending gene g_2. Given a training set of size w, we estimate the initial state distribution of l, denoted by $P_0(x)$, by summarizing the proportion of each state x at the starting gene g_1 over the training set, *i.e.*,

$$P_0(x) = \frac{1}{w}\sum_{k=1}^{w} I(x_{k,1} = x) \qquad (1)$$

where $x_{k,1}$ represents the expression state of gene g_1 in sample k and I is an indicator function yielding 1 if the condition is true and 0 otherwise. Let x and y represent two any states, where $x,y \in S$, respectively, the probability of the transition from x to y, denoted by $P(y|x)$, can be estimated as follows:

$$P(y\,|\,x) = \frac{\sum_{k=1}^{w} I(x_{k,1} = x \,\&\, x_{k,2} = y)}{\sum_{k=1}^{w}\sum_{v\in S_g} I(x_{k,1} = x \,\&\, x_{k,2} = v)} \qquad (2)$$

where $x_{k,1}$ and $x_{k,2}$ are the states of genes g_1 and g_2 in the sample k. By varying x and y, we can obtain all 3×3=9 state transition probabilities and form a state transition probability matrix M [3×3] for the gene link MCM.

Consider that a sample s with expression state values (x_1, x_2) for the two genes, g_1 and g_2. According to the Markov property, the likelihood $P(s)$ that the sample comes from the gene link MCM can be estimated as a joint probability of the observed states x_i, i=1, 2, *i.e.* ,

$$\begin{aligned} P(s) &= P(x_1, x_2) \\ &= P_0(x_1)P(x_2\,|\,x_1) \end{aligned} \qquad (3)$$

where $P_0(x_1)$ and $P(x_2|x_1)$ are the initial state probability of x_1 and the probability of the state x_1 of gene g_1 transiting to x_2 of gene g_2, respectively, both of which can be obtained by Eqs. (1)-(2) respectively. Fig.1 shows the schematic framework of the gene link MCM.

3) *Gene link MCM -based cancer classification*

Consider K classes labeled as C_1, C_2…C_K respectively, and let MCM$_k$ represent the gene link MCM for class k by Eqs. (1) -(2). Given a test sample, the probability P_k, k=1,2,⋯,K, that the sample comes from class k can be estimated by Eq.(3), and then the test sample can be classified to class c with the maximum occurring probability, *i.e.*,

$$c = \arg\max_{k\in\{1,2,\cdots,K\}}(P_k) \qquad (4)$$

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

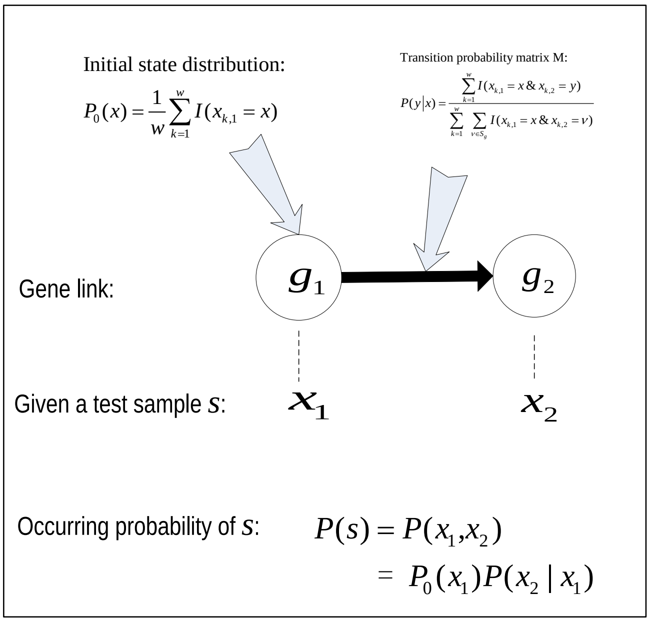
Initial state distribution:
$P_0(x) = \frac{1}{w}\sum_{k=1}^{w} I(x_{k,1} = x)$
Transition probability matrix M:
$P(y|x) = \frac{\sum_{k=1}^{w} I(x_{k,1} = x \,\&\, x_{k,2} = y)}{\sum_{k=1}^{w}\sum_{v \in S_g} I(x_{k,1} = x \,\&\, x_{k,2} = v)}$
Gene link:
g_1
g_2
Given a test sample s:
x_1
x_2
Occurring probability of s:
$P(s) = P(x_1, x_2) = P_0(x_1)P(x_2 \mid x_1)$

Fig. 1. Schematic illustration of the gene link MCM

B. *Classification power-based representation of gene links*

To overcome overfitting induced by small sample size of microarray data sets, we adopt 10-fold cross–validation (CV) to evaluate the classification performance of a gene link. In the 10-fold cross-validation, the whole dataset is randomly divided into ten folds, and each fold is in turn as testing set and the rest as training set for learning a MCM classifier by Eq.(4). We repeat the CV procedure 20 times and average the resulted accuraciesas an overall performance of the gene link, denoted by *ACC*.

To assess the statistical significance of an observed *ACC*, we devise the following permutation test to calculate a *p*-value: Randomly shuffling both the class labels of samples and the two genes of the link *B*=1000 times and calculating a 10-fold cross-validation accuracy based on MCM classifier for the permutated gene link. As a result, *B* permutated ACC (*rACC*) can be obtained. These *rACC*s represent the classification accuracies under null distribution. Based on the *B rACC*s, a *p*-value for an observed ACC can be calculated as

$$P = \frac{1}{B}\sum_{j=1}^{B} I(ACC < rACC_j) \qquad (5)$$

where $rACC_j$ is the *j*th *rACC*.

C. *Gene links-based identfication of differential pathways*

To measure the differential expression of a pathway, we formulate a pathway-level statistic based on gene links. Assuming *K* gene links in a pathway, a pathway score (PS) can be defined as the mean of ACCs over these links, *i.e.*,

$$PS = \frac{1}{K}\sum_{i=1}^{K} ACC_i \qquad (6)$$

The pathway score reflects the discriminative power of the whole pathways and can be taken as a pathway-level statistic for measuring the expression difference of the pathway. We then devise a permutation test to estimate the significance of an observed PS. We consider a null hypothesis that all the links within a pathway are non-discriminative. To simulate the null distribution, we randomly sample a same number (*K*) of gene pairs as that of gene links in the pathway to constitute a permutated pathway, and randomly shuffle the class labels of samples to calculate the PS for the permutated pathway by Eqs. (1)-(5). Assuming *Z*=1000 permutated pathways, the *p*-value for an observed pathway *ps* can be estimated as

$$P_{ps} = \frac{1}{Z}\sum_{i=1}^{Z} I(ps < PS_i) \qquad (7)$$

where PS_i, *i*=1, 2, ... *Z*, denotes the PS of the *i*th permutated pathway.

III. Results

A. *Datasets*

To evaluate the proposed method, we collected two benchmark gene expression data sets: liver cancer dataset [23] and ALL dataset [24]. In the liver cancer data, the samples are divided into two groups: one consisting of patients suffering from early intrahepatic recurrence (n=20, REC) and the other patients free from recurrence (n=40, NREC). Each sample consists of the expression levels of ~7129 genes. The ALL dataset characterizes acute lymphocytic leukemia (ALL) cells associated with known genotypic abnormalities in adult patients, and contains 37 observations from one experimental condition (n1=37, BCR; presence of BCR/ABL gene rearrangement) and 42 from another experimental condition (n2=42, NEG; absence of rearrangement). In the ALL data set, each sample consists of the expression levels of ~11556 genes. Table I gives the descriptions of the two data sets. To apply the proposed gene link MCM, we first used the biology-constrained discretization method (BCD) [28] to discretize the expression levels of genes into one of three regulatory states, down-regulated (-1), non-significantly regulated (0) and up-regulated (1).

TABLE I. Descriptions of the two datasets used

Dataset	#Gene	#Sample		
		Class 1	*Class 2*	*Total*
ALL	~11556	37(BCR)	42(NEG)	79
Liver cancer	~7129	20(REC)	40(NREC)	60

Pathway information was collected from the KEGG database [http://www.kegg.jp/kegg/pathway.html]. A total number of 220 KEGG pathways were considered in the experiment. To apply the proposed method, we first decomposed each pathway into a set of gene links. Given a pathway, not all genes in it will be present in a data set. After manual examination, only 213 of the 220 pathways were found to have more than one gene link present in the liver cancer dataset and 218 pathways in the ALL dataset. Only these pathways were used in the analyses for the two datasets.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

B. *Gene links signficantly classify cancer*

For the 213 pathways, there are totally 15717 gene links present in the liver cancer data. We first calculated the 10-fold CV accuracies of these gene links based on gene link MCM classifier and estimated the significance by the permutation test in Eq.(5). Fig. 2 shows the cumulative probability distribution (CPD) of the *p*-values for the pathway gene links. For comparison, we randomly sampled a same number (15717) of gene pairs from the total genes. With randomly shuffled class labels, the ACCs of these random gene pairs were calculated and then their *p*-values. Ideally, random classification accuracies follow a normal distribution around 0.5, and the CPD curve of *p*-values will approach the line y=x. From Fig. 2, it can be found that the observed CPD for the gene links is furthest away from the line y=x, showing that the pathway gene links tend to be more discriminative than by chance. The classification accuracies of the pathway links are significantly higher (*p*-value < 2.2e-16, according to *t*-test) than those of the random gene pairs. Similar results (24348 pathway gene links, *p*-value < 2.2e-16 by *t*-test) were also observed on the ALL data, as shown in Fig. 2, confirming the discriminative power of pathway gene links.

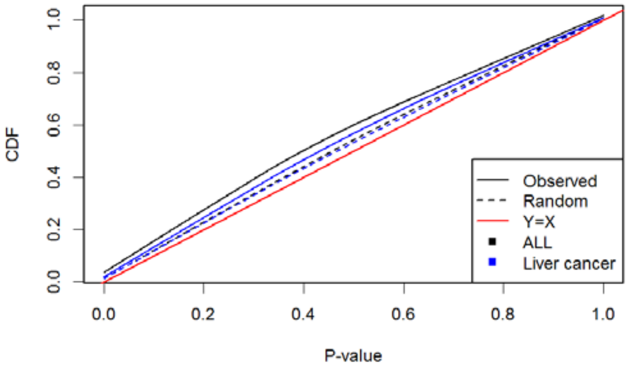

Fig. 2. The cumulative probability density (CPD) curve of the *p*-values of gene links in liver cancer dataset and ALL dataset.

C. *Identifying differentially expressed pathways based on gene links*

We next identified differentially expressed pathways for the liver cancer and the ALL data based on gene links.

1) Identification of liver cancer recurrence-related pathways

For the liver cancer data, totally 66 pathways were called significantly differentially expressed between the two liver cancer classes by our gene link method at an *ad hoc p*-value cutoff of 0.05. In contrast, at the same value of *p*-value cutoff, four previous methods, Global test [31, 32], WW test [33], LR[34] and TAPPA[20], identified very few significantly differentially expressed pathways, 9, 11, 8 and 36 for Global test, WW test, LR and TAPPA respectively, as listed in Table II. We further compared the CPD curves of *p*-values among our method and the four previous methods, as shown in Fig. 3, showing the stronger power of our method in identifying differentially expressed pathways.

TABLE II. NUMBERS OF SIGNIFICANTLY DIFFERENTIALLY EXPRESSED PATHWAYS (P-VALUE<0.05) BY OUR METHOD AND FOUR PREVIOUS METHODS

Method	**Liver cancer data**	**ALL data**
Our method	66	102
Global test	9	136
WW test	11	86
LR	8	21
TAPPA	36	12

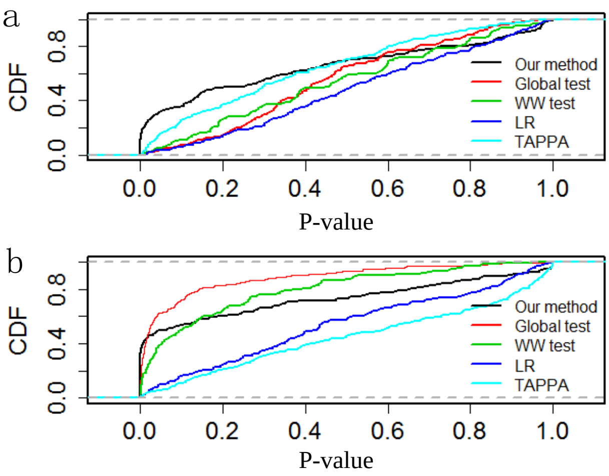

Fig. 3. Comparison of CPD curve of *p*-values among the four methods on the liver cancer (a) and ALL(b) data sets.

Among the significant pathways identified by our methods, most were previously reported to be related to liver cancer, for example, p53 signaling pathway (*p*-value=0.012), Transcriptional misregulation in cancer (*p*-value=0.003) and Hepatitis B (*p*-value=0). These pathways, however, were not called to be significantly differentially expressed by the four previous methods. p53 signaling pathway consists of 68 genes. Of the 68 genes, 38 are present in the liver cancer data, which constitute 55 gene links. Fig. 4(a) shows the distribution of the classification accuracies of the 55 gene links against that by chance, indicating that the gene links tend to be more discriminative. Similar results were obtained for the Transcriptional misregulation in cancer and Hepatitis B pathways, as shown in Fig.4 (b-c).

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

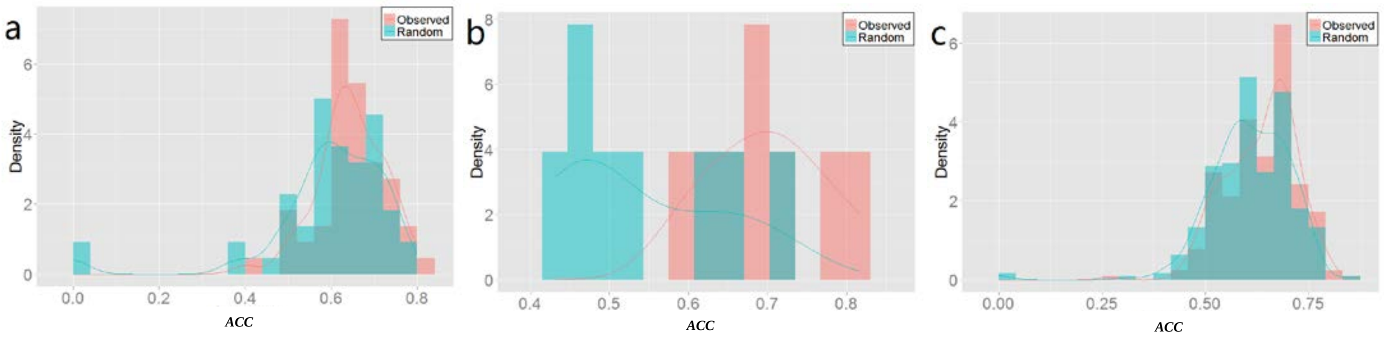

Fig. 4. Distributions of classification accuracies (ACC) of observed in p53 signaling pathway (a), Transcriptional misregulation in cancer (b) and Hepatitis B (c) against by chance.

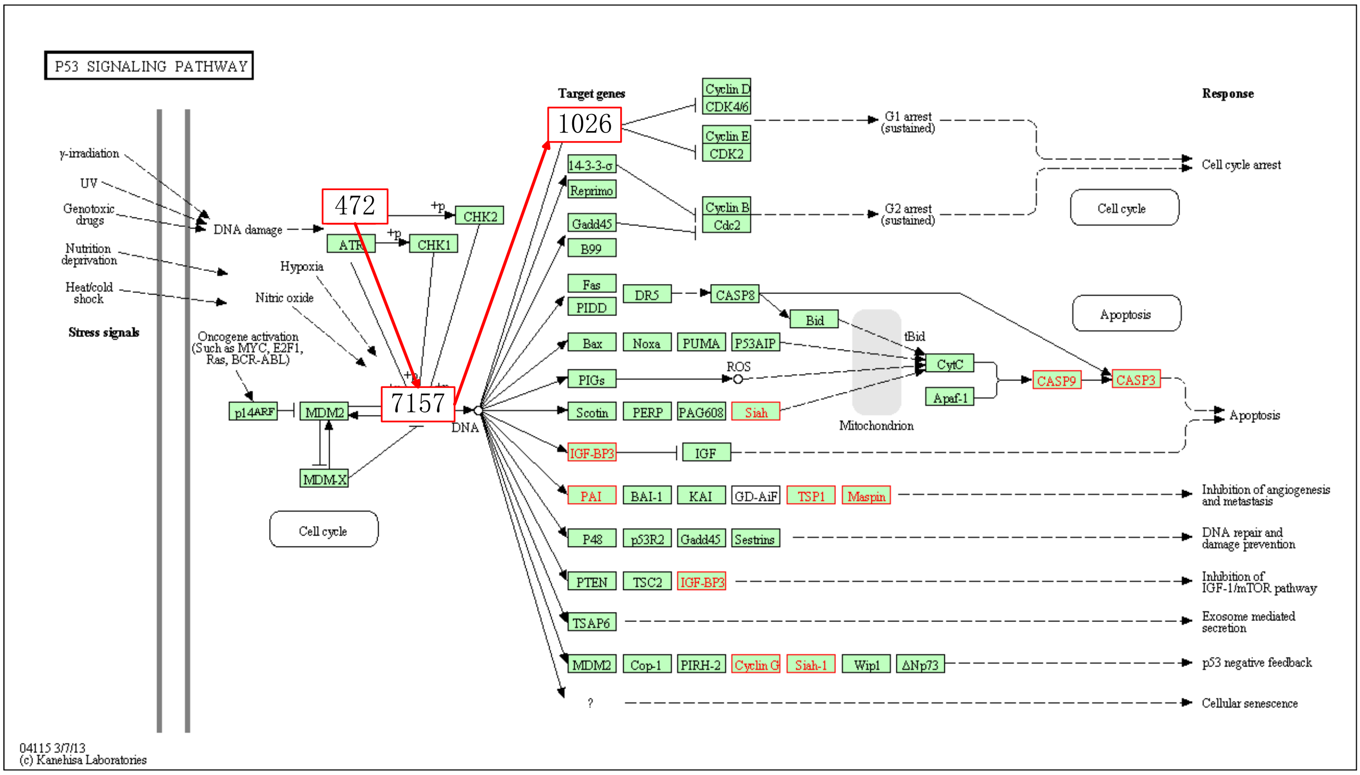

Fig. 5. Distributions of significant links in P53 pathway. Genes involved in significant links are highlighted in red.

Biologically, gene links that significantly classify the two liver classes tend to be important to the recurrence of liver cancer and potentially play crucial roles in pathway topological structure. We then overlaid the pathway gene links that were identified to be significantly discriminative (*p*-value<0.05) onto the KEGG pathway map, as shown in Fig. 5 for P53 pathway. From Fig. 5, it can be found that most of the significant links (10/11) involve p53 gene, which is in agreement with the importance of P53 in tumorigenesis. Especially, we can see that these significant links are closely associated with cancer-related responses, for example, "Apoptosis" that causes apoptosis and cell death of cells and "inhibition of angiogenesis and metastasis" , a cellular response that is closely related to cancer recurrence [35]. The 11 significant gene links constitute several paths of length 3 or more genes. Take one of the paths (highlighted in red lines in the figure) as example. We took the three genes (Entrez IDs 472, 7157 and 1026) involved in the path and examined the distributions of their expression states in each of the two classes of liver cancer, as shown in Fig. 6. In Fig.6, "U", "N", and "U" represent three gene expression states, down-regulated (-1), non-significantly regulated (0) and up-regulated (1), respectively, and the height of the letters represents the probability of the corresponding state occurring. This figure suggests the activity difference of the path in the two liver cancer subtypes.

2) Identification of BCR/ABL gene rearrangement -related pathways for the ALL data

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

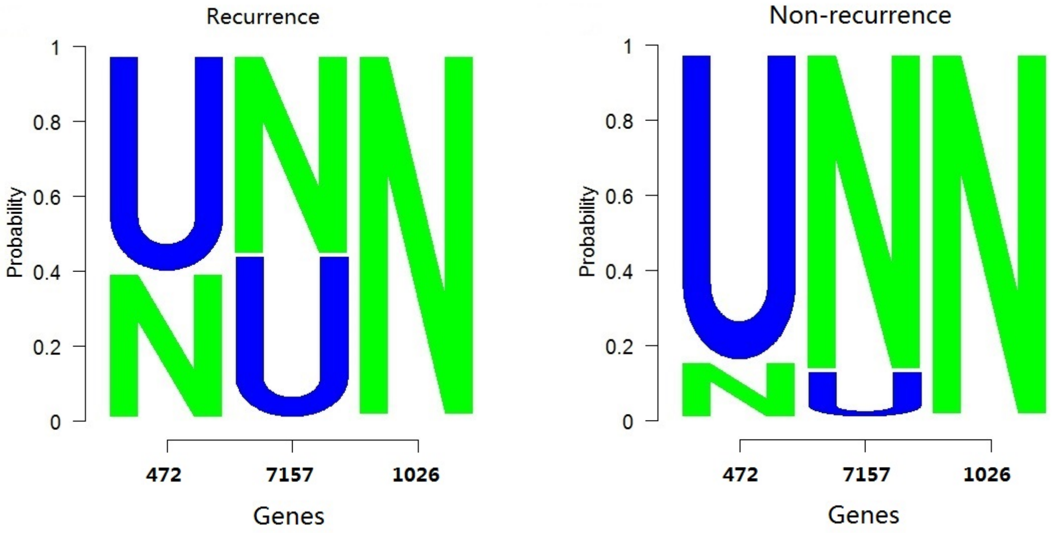

Fig. 6. Distributions of the expression state of the three genes with Entrez IDs 472, 7157 and 1026 in the two class of liver cancer.

TABLE III. IDENTIFICATION RESULTS (√) OF THE NINE PATHWAYS INCLUDING BCR AND/OR ABL1 BY DIFFERENT METHODS ($P - value < 0.05$)

PATHWAY	Our method	Global test	WW test	LR	TAPPA	clipper	BPA	SPIA	GSEA
ErbB signaling pathway									
Cell cycle	√	√	√	√		√	√		
Axon guidance	√	√	√	√		√			
Neurotrophin signaling pathway	√	√	√	√		√			
Pathogenic Escherichia coli infection	√	√	√			√		√	√
Shigellosis	√	√				√			
Pathways in cancer	√	√	√						√
Chronic myeloid leukemia	√	√	√			√			
Viral myocarditis	√	√	√	√		√		√	

For the ALL data, our methods and the four previous methods, Global test, WW test, LR and TAPPA, called 102, 136, 86, 21 and 12 significantly expressed pathways at a *p*-value cutoff of 0.05, respectively, as listed in Table II. It can be seen that the previous method, TAPPA, found very few differentially expressed pathways while Global test identified a very large number (136) of differentially expressed pathways. In contrast to these, our methods obtained a moderate result (102), which is close to the result of WW test (83). Fig.2 compares the CPDs of *p*-values among our method and the four previous methods for ALL dataset, confirming the competitive discovery power of the gene link method.

Given the presence of the BCR/ABL chimera, it is expected that pathways including BCR and/or ABL1 will be impacted biologically. Among the total 218 pathways present in the ALL data, there are 9 pathways found to be BCR and/or ABL1-involved. Recently, Martini *et al.*[36] reported the identification results by Clipper [36], GSEA[17], SPIA[37] and BPA[38]. With a relaxed *p*-value cutoff of 0.1, GSEA and SPIA identified 2 out of the 9 pathways and only one for BPA [38]. More surprisingly, none of the 9 pathways was identified by TAPPA. In contrast to these methods, our method as well as the global test method identified almost all the 9 related pathways (8), as shown in Table III, confirming the competitive power of our method in identifying differentially expressed pathways. What is worth of mentioning is that the global test method totally called far more (136) significant pathways (*p*-value<0.05) than that (102) by our method, as shown in Table II, which suggests that our method led to less false positives than those by the global test method.

IV. CONCLUSIONS

We have proposed a gene link-based approach for identifying differential pathways in microarray data analysis. The method takes advantage of gene links within a pathway to extract pathway topological information. To measure the importance of gene links, we modeled gene link using Markov chain model and then devised a gene link MCM classifier. Based on the discriminative performance of gene links, we then formulated a new statistic *PS* for measuring differential expression of a pathway, represented as the average accuracy over all the gene links in the pathway. The significance of differential expression of pathways is estimated using a permutation test. Experimental results on two real-world data sets, liver cancer and ALL data sets, demonstrated the effectiveness and efficiency of the proposed method.

We would like to note that the proposed method could suffer from the small sample problem inherent in microarray data and thus the learned gene link model be overfitted. In future, we will evaluate the method on simulation data sets and apply the method to more microarray or RNA-seq data sets for extensive evaluation.

References

[1] Emmert-Streib, F., The chronic fatigue syndrome: a comparative pathway analysis. J Comput Biol, 2007. 14(7): p. 961-72.

[2] Schadt, E.E., Molecular networks as sensors and drivers of common human diseases. Nature, 2009. 461(7261): p. 218-223.

[3] Liu, W., et al., Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. Bioinformatics, 2013. 29(17): p. 2169-77.

[4] Zheng, C.-H., Y.-W. Chong, and H.-Q. Wang, Gene selection using independent variable group analysis for tumor classification. Neural Computing and Applications, 2010. 20: p. 161-170.

[5] Ma, S. and J. Huang, Regularized gene selection in cancer microarray meta-analysis. BMC Bioinformatics, 2009. 10: p. 1.

[6] Wang, H.-Q. and D.-S. Huang, Regulation probability method for gene selection. Pattern Recognition Letter, 2006. 27(2): p. 116-122.

[7] Subramani, P., R. Sahu, and S. Verma, Feature selection using Haar wavelet power spectrum. BMC Bioinformatics, 2006. 7(1): p. 432.

[8] Bae, K. and B.K. Mallick, Gene selection using a two-level hierarchical Bayesian model. Binformatics, 2004. 20: p. 3423-3430.

[9] Kanehisa, M. and S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res, 2000. 28(1): p. 27-30.

[10] Kanehisa, M., et al., KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res, 2012. 40(Database issue): p. D109-14.

[11] Callow, M.J., et al., Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. Genome Res, 2000. 10(12): p. 2022-9.

[12] Tusher, V.G., R. Tibshirani, and G. Chu, Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A, 2001. 98(9): p. 5116-21.

[13] Dudoit, S., J.P. Shaffer, and J.C. Boldrick, Multiple hypothesis testing in microarray experiments. Statistical Science, 2003. 18(1): p. 71-103.

[14] Glazko, G.V. and F. Emmert-Streib, Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. Bioinformatics, 2009. 25(18): p. 2348-54.

[15] Khatri, P. and S. Draghici, Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics, 2005. 21(18): p. 3587-95.

[16] Drǎghici, S., et al., Global functional profiling of gene expression. Genomics, 2003. 81(2): p. 98-104.

[17] Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.

[18] Khatri, P., M. Sirota, and A.J. Butte, Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. 2012.

[19] Rahnenfuhrer, J., et al., Calculating the statistical significance of changes in pathway activity from gene expression data. Stat Appl Genet Mol Biol, 2004. 3: p. Article16.

[20] Gao, S. and X. Wang, TAPPA: topological analysis of pathway phenotype association. Bioinformatics, 2007. 23(22): p. 3100-2.

[21] Isci, S., et al., Pathway analysis of high-throughput biological data within a Bayesian network framework. Bioinformatics, 2011. 27(12): p. 1667-1674.

[22] Friedman, N., et al., Using Bayesian networks to analyze expression data, in Proceedings of the Fourth Annual International Conference on Computational Molecular Biology. 2000, ACM: Tokyo, Japan. p. 127-135.

[23] Iizuka, N., et al., Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. Lancet, 2003. 361(9361): p. 923-9.

[24] Chiaretti, S., et al., Gene expression profiles of B-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation. Clinical Cancer Research, 2005. 11(20): p. 7209-7219.

[25] Norris, J.R., Markov chains. 1998: Cambridge university press.

[26] Ding, L. and W. Hong-Qiang. A Markov chain model-based method for cancer classification. in Natural Computation (ICNC), 2012 Eighth International Conference on. 2012.

[27] Li, D., R. Li, and H.-Q. Wang, A Novel Discretization Method for Microarray-Based Cancer Classification, in Intelligent Computing Technology, D.-S. Huang, et al., Editors. 2012, Springer Berlin Heidelberg. p. 327-333.

[28] Wang, H.-Q., G.-J. Jing, and C.-H. Zheng, Biology-constrained gene expression discretization for cancer classification. Neurocomputing, 2014.

[29] SHAPIRO, S.S. and M.B. WILK, An analysis of variance test for normality (complete samples). Biometrika, 1965. 52(3-4): p. 591-611.

[30] Nornadiah, R. and W.Y. Bee, Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. Journal of Statistical Modeling and Analytics, 2011. 2(1): p. 21-33.

[31] Goeman JJ, et al., A global test for groups of genes: testing association with a clinical outcome. Bioinformatics, 2004. 20: p. 93-99.

[32] Mansmann U and M.R, Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. Methods of Information in Medicine, 2005. 44(3): p. 449-453.

[33] Rahmatallah, Y., F. Emmert-Streib, and G. Glazko, Gene set analysis for self-contained tests: complex null and specific alternative hypotheses. Bioinformatics, 2012. 28(23): p. 3073-80.

[34] Sartor, M.A., G.D. Leikauf, and M. Medvedovic, LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. Bioinformatics, 2009. 25(2): p. 211-7.

[35] Folkman, J. Role of angiogenesis in tumor growth and metastasis. in Seminars in oncology. 2002. Elsevier.

[36] Martini, P., et al., Along signal paths: an empirical gene set approach exploiting pathway topology. Nucleic Acids Res, 2013. 41(1): p. e19.

[37] Tarca, A.L., et al., A novel signaling pathway impact analysis. Bioinformatics, 2009. 25(1): p. 75-82.

[38] Isci, S., et al., Pathway analysis of high-throughput biological data within a Bayesian network framework. Bioinformatics, 2011. 27(12): p. 1667-74.