A Markov Chain Model-based Method for Cancer Classification

Ding Li
Department of Automation,
University of Science and Technology of China,
Hefei 230027, PR China
Intelligent Computation Lab,
Hefei Institute of Intelligent Machines, Chinese Academy of Science, P.O.Box 1130,
Hefei, Anhui 230031, PR China

Hong-Qiang Wang*
Intelligent Computation Lab,
Hefei Institute of Intelligent Machines, Chinese Academy of Science, P.O.Box 1130,
Hefei, Anhui 230031, PR China

***Abstract*—In this paper, we propose a Markov chain model (MCM)-based method for cancer classification. By viewing a gene chain extracted from a gene pathway map as a gene Markov chain (GMC), the method can construct a MCM for different cancer classes. The resulted MCM captures the co-activity pattern of genes in terms of an initial state distribution and a state transition probability matrix. When used for cancer classification, the method first calculates the respective probabilities of a test sample belonging to different cancer classes according to the corresponding MCM, and then predicts the class of the sample as the one with the highest probability. We evaluate the proposed method on the publicly available leukemia dataset and compare it with several conventional methods.**

Keyword—***Cancer classification; Markov chain; Gene regulation; Gene pathway***

I. INTRODUCTION

Recently developed high-throughput technology can simultaneously measure expression levels of ten of thousands of genes in cells and makes it possible to study cancer pathology in a genomic level [1]. Recognizing cancer-related gene expression patterns can gain deep insight into gene activity resposible for cancer mechanism, and thus lead to more effective cancer classification than traditional histology-based methods. Currently, a number of gene expression data-based methods have been developed for cancer classification [2-5]. However, most of these methods classifies cancer like a black box and do not concern the pathology of cancer [6-9].

To exploit the pathway information for cancer classification, Guo et al. [10] proposed to predict the class of a sample using pathway expression levels consisting of the mean/median gene expression values in each pathway. Shin et al proposed to extract the complex functional-dependencies between genes to classify cancer subtypes [11]. Considering the correlation patterns between genes, we propose to apply Markov chain model to model a pathway for cancer classification in this paper. Simply speaking, we first build MCM for each cancer class based on a gene chain from a pathway, and then classify a test sample by comparing the occurrence probabilities of its gene state observation with respect to different cancer classes.

The rest of the paper is organized as follows. Related knowledge on MCM is first reviewed. In particular, two kinds of MCM models are introduced. Then a MCM-based cancer classification framework is presented. Finally, the proposed MCM method is evaluated on a real-world data set and is compared with several conventional methods including Support Vector Machine (SVM) and K-Nearest Neighbor (KNN).

II. METHODS

A. Markov Chain Model (MCM)

Markov chain model provides an approach for characterizing a kind of random processes, i.e., Markov chain (MC). MC is such a random process that meets the Markov property, i.e., the next state depends only on the current state but not on the sequence of events that preceded it. Assume a sequence set T= {1,2, ...,n} where n is a natural constant and a finite state space S of size m where m is a natural constant. Let $X(t)$ take values in S, a Markov chain can be represented as a sequence $\{X(t),\ t \in T\}$ that meets the following probability equation:

$$\begin{aligned} &P(X(t+1)=x_{t+1} \mid X(t)=x_t, X(t-1)=x_{t-1},\ldots,X(1)=x_1) \\ &=P\{X(t+1)=x_{t+1} \mid X(t)=x_t\} \end{aligned} \quad (1)$$

where $x_k \in T, k=1,2,\ldots t+1$, and P represents a conditional probability function.

MCM have two important parameters to characterize a MC: one is the distribution of the initial state, denoted by $p_0(x)$, and another is the l-step transition probability matrix at time t, denoted by M^t. The former reflects the probability of each possible state being present at time 1. For any $x \in S$, we have $p_0(x) \geq 0$ and $\sum p_0(x)=1$. The latter specifies the probabilities of any possible state transition from time t to time $t+l$. Specifically, the element of M^t at row i and column j, $p(x,y,t)$ represents the transition probability of the ith state x at time t to the jth state y at time $t+l$, i.e.,

*To whom correpondence should be addressed
This work was supported by the grants of the National Science Foundation of China, Nos. 30900321and 31071168.

$$p(x,y,t)=P(X(t+1)=x|X(t)=y) \quad (2)$$

For any $x \in S$, the following equation holds

$$\sum_{y \in S} p(x,y,t)=1 \quad (3)$$

For simplicity, we only consider the case of l=1 to model gene pathway for cancer classification in this present paper.

Given a MCM, the probability of an observation sample coming from it can be calculated as follows. Let $X=\{x_1,x_2,\ldots,x_n\}$represent an observation. Generally, the occurrence probability $P(x)$ can be represented as the joint probability of the observed states x_t, t=1,2,…,n, i.e.,

$$P(\mathrm{x})=P(X(1)=x_1,X(2)=x_2,...,X(n)=x_n) \quad (4)$$

According to the property of a Markov chain in (1), $P(x)$ can be further represented as

$$\begin{aligned} P(\mathrm{x}) &= p_0(x_1)\prod_{t=1}^{n-1} p(X(t+1)=x_{t+1} \mid X(t)=x_t) \\ &= p_0(x_1)\prod_{g=1}^{n-1} p(x_{g+1},x_g,g) \end{aligned} \quad (5)$$

where $p_0(x_1)$ and $p(x_{g+1}, x_g, g)$ are the initial state distribution and the state transition probabilities related to the MCM, respectively.

There are two types of Markov chains commonly used in practice: one is referred to as homogeneous MC (HMC) and another non-homogeneous MC (NMC). The major difference between them is that the former has a same k-step transition probability matrix at any time while the matrices change with time in the later. In what follows, we present the respective applications of the two MCM models to cancer classification.

B. Modeling Gene Pathway Based on MCM

Multiple gene chains constitute a gene pathway and play important roles to the operation of the gene pathway. In a gene chain, genetic signal can sequentially flows from the first gene to the last one in the form of regulation state transition. In light of this fact, we view such a gene chain as a gene Markov chain (GMC) to model gene coactivity in a gene pathway. Specifically, for each gene, consider a 3-state regulatory state space Sg={-1,0,1}, where -1,0 and 1 represent down-regulation, down-regulation and up-regulation respectively [12]. Let $G=\{g_1,g_2,\ldots,g_n\}$ represent a gene chain of length n, its GMC can be represented as $\{X(g), g \in G\}$ where $X(g)$ represents the regulation state of the gth gene. Based on the two types of MC, two models are generated to characterize the GMC, homogeneous MC (HMC) model and non-homogeneous MC (NMC) model. In what follows, the parameter learning processes for the two models are presented.

I) HMC Model

In the HMC model, we assume that a GMC is a homogeneous Markov chain. Consider a training sample set of size w, its initial state distribution $p_0(x)$ can be calculated through summarizing the frequency of each gene regulatory state at g=1 over all the training samples, i.e.,

$$p_0(x)=\frac{1}{w}\sum_{k=1}^{w} I(x_{k,1}=x) \quad (6)$$

where $x_{k,1}$ represents the state of the first gene of the GMC in sample k. For the calculation of the transition probability matrix M, since the transitions at all the steps in a HMC shares a same transition probability matrix, an accurate estimation of transition probability matrix can be obtained using the data over all the steps. Let x and y represent the ith and jth states, respectively, the element at row i and column j of M can be calculated as

$$p(x,y)=\frac{\sum_{k=1}^{w}\sum_{g=1}^{n-1} I(x_{k,g}=x \,\&\, x_{k,g+1}=y)}{\sum_{k=1}^{w}\sum_{g=1}^{n-1}\sum_{v \in Sg} I(x_{k,g}=x \,\&\, x_{k,g+1}=v)} \quad (7)$$

where $x_{k,g}$ and $x_{k,g+1}$ represent the states of gene g and g+1 in sample k, respectively, and I represents an indicator function whose value is equal to 1 if true and 0 otherwise. Likewise, other transition probabilities can be obtained in a similar way, and the transition probability matrix M can be obtained.

II) NMC Model

In the NMC model, we assume that a GMC is a non-homogeneous Markov chain. For the NMC model, the distribution of initial state $p_0(x)$ can be calculated using the same procedure as that in HMC. Since the transition steps are non-homogeneous, there are n transition probability matrices to be learned in the NMC model, each corresponding to a transition step. Specifically, for the transition probability matrix M^g at gene g, the element at row i and column j, i.e., the probability of the transition from the ith state x at gene g to the jth state y at gene g+1, is calculated over all the training samples as follows:

$$\begin{aligned} p_{ij}^{g} &= p(x,y,g) \\ &= \frac{\sum_{k=1}^{w} I(x_{k,g}=x \,\&\, x_{k,g+1}=y)}{\sum_{k=1}^{w}\sum_{v \in Sg} I(x_{k,g}=x \,\&\, x_{k,g+1}=v)} \end{aligned} \quad (8)$$

Additionally, to better represent the state transition patterns in the NMC, we integrate the resulted n-1 M^gs into a single matrix M as follows: by first reshape each M^g as a row vector, and then combineing them by row. Since each row of M corresponds to a one-step's regulation transition in a GMC, M provides a global visualization of the state transition patterns in the GMC.

C. A MCM-Based Framework for Cancer Classification

For an observation (sample), the probability that it comes from a given GMC model can be easily obtained by (5). In light of this, we devise a MCM-based framework for cancer classification as follows. For convenience, consider a binary

cancer classification problem with two classes labeled as $c1$ and $c2$, and let $P1$ and $P2$ represent the probability values of an unknown sample calculated by (5) based on the MCMs of the two classes, respectively. The framework will classify the sample into class $c1$ if $P1>P2$ and $c2$ otherwise. The framework can be easily interpreted biologically. Each MCM encapsulates the patterns of regulatory state transition specific to a corresponding cancer class. If the sample belongs to a class, the regulatory patterns represented in it should be closely consistent with those encapsulated in the MCM of the class and a sufficiently high probability will be resulted and vice versa.

III. RESULTS

We evaluated the proposed model on the leukemia dataset [10]. The dataset contains 72 samples, of which 47 are acute lymphoblastic leukemia (ALL) and 25 are acute myelogenous leukemia (AML). Each sample consists of the expression levels of 7129 genes.

To extract GMC for the data set, we downloaded two human pathways, *hsa04510* (focal adhesion) and *has04010* (MAPK Signaling Pathway), from the KEGG pathway database [http://www.genome.jp/kegg/]. The two pathways have been shown to be related to leukemia [11]. Table 1 shows related description of the two gene pathways. Two GMC were extracted, CH1 from hsa04510 and CH2 from hsa04010, as shown in Fig. 1. In the experiment, we applied the two MCM on the two GMC, and used them to classify the leukemia data. Note that before modeling, the expression values of each gene were manually discretized into 3 intervals, and the state of the gene in a sample was possibly one of the resulted 3 regulatory states.

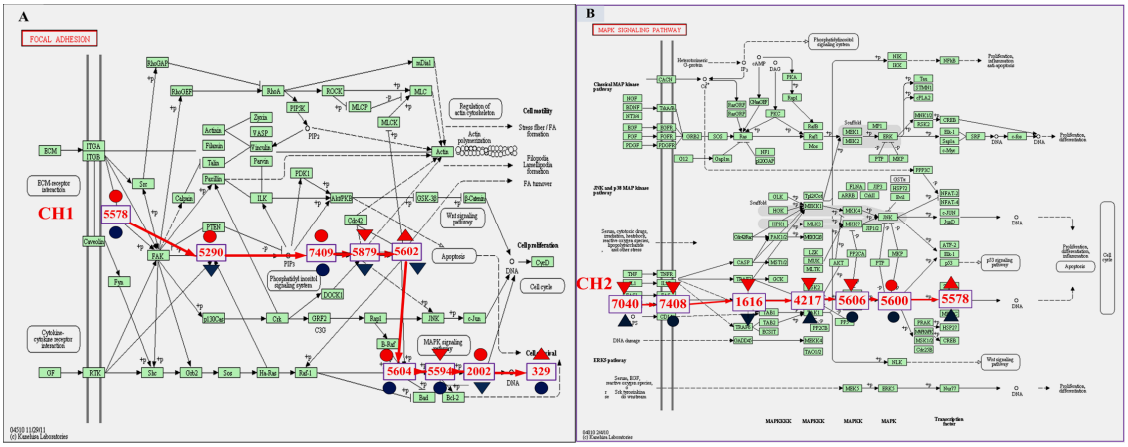

Figure 1. Maps of the two pathways used, has04510 (A) and has04010 (B), and PST of the two GMC, CH1 and CH2,for the two cancer classes, ALL (red) and AML (darkblue). Solid circle, down- and up-triangle represent non-regulated,down-regulated and up-regulated state, respectively. The numbers in red are the gene IDs of the corresponding genes.

TABLE 1. RELATED INFORMATION OF THE TWO PATHWAYS, has04510 AND has04010

Pathway ID	Pathway name	Description	#Genes
has04510	Focal adhesion – Homo sapiens (human)	Cell-matrix adhesions play essential roles in important biological processes including cell motility, cell proliferation, cell differentiation, regulation of gene expression and cell survival	201
has04010	MAPK signaling pathway – Homo sapiens (human)	The mitogen-activated protein kinase (MAPK) cascade is a highly conserved module that is involved in various cellular functions, including cell proliferation, differentiation and migration	268

We first applied the HMCM model to the first GMC, CH1. Figs. 2 and 3 shows the histograms of the initial state probability distributions and the heatmaps of the state transition probability matrices for the two classes, ALL and AML, respectively. From these figures, it can be found that the classes of leukemia exhibit very different gene activity patterns in CH1. In particular, the first gene (GeneID:5578) was non-regulated in AML with a significant higher probability relative to that in ALL (Fig. 2A), and the transition probability from up-regulation to up-regulation was higher in AML than in ALL(Fig. 2B). Fig. 3 shows that the heatmaps of the resulted state transition probability matrices for the two classes by NMCM. Theoretically, compared with HMCM, NMCM can gain deeper insight into the activity of a GMC by constructing transition matrices for each step. From Fig. 3, it can be found that the down-to-up transitions in ALL dominantly happened in the 4^{th} step of CH1, i.e., from gene 5879 to 5602, which is significantly different from that in AML where no transition is dominant. To clearly observe the difference of gene activity in the two classes, we collected the states with the highest transition probability in each step to form a principal state transition pattern (PST) for each class, as

shown in Fig. 1A. From Fig. 1A, it can be found that the two classes exhibit significantly different PST in the CH1, which could account for the distinction of the two classes. Similarly, we applied the two models to CH2, and the resulted PST for the two classes are shown in Fig.1B, suggesting the different gene activity patterns between the two classes.

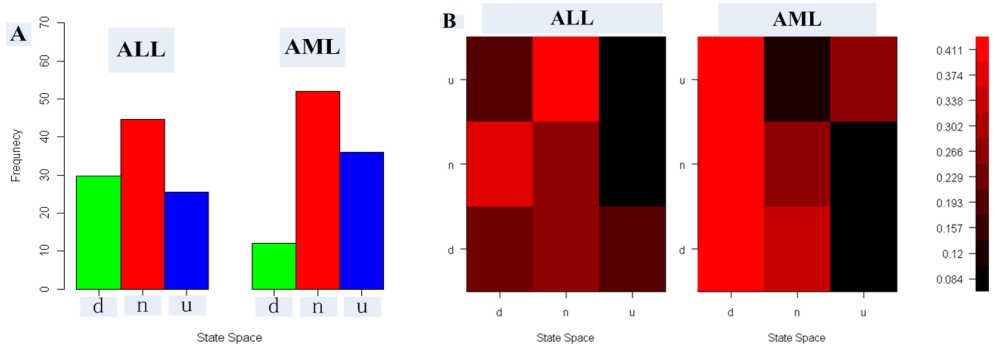

Figure 2. Histograms of the initial state probability distributions (A) and heatmaps of the state transition probability matrices(B) of HMCM on CH1 for ALL and AML. Symbols, u, d and n, represent up-, down- and non-regulation states.

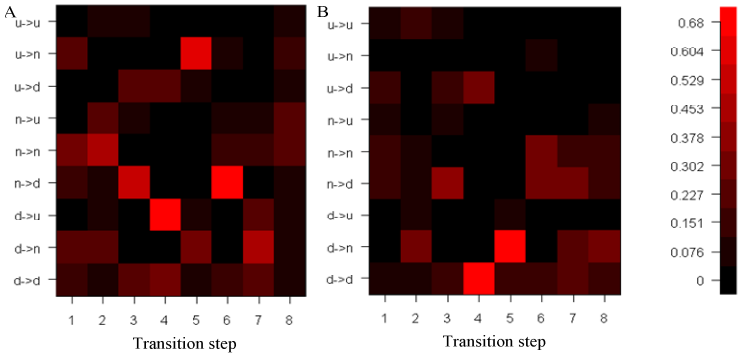

Figure 3. Heatmaps of the state transition probability matrices of NMCM on CH1 for ALL (A) and AML (B).

TABLE 2. COMPARISON OF THE CLASSIFICATION PERFORMANCE OF THE PROPOSED MCM MODELS WITH SEVERAL PREVIOUS METHODS ON THE LEUKEMIA DATA SET

Methods	**CH1**						**CH2**					
	ACC	***std***	***min ACC***	***max ACC***	***SPE***	***SEN***	***ACC***	***Std***	***max ACC***	***min ACC***	***SPE***	***SEN***
HMCM	74.54	0.82	50	90.91	76.4	70.57	71.68	0.67	90.91	50	74.47	65.71
NMCM	**80.55**	0.57	59.09	**95.45**	84.73	71.57	**74.36**	0.67	**90.91**	**54.55**	79.4	63.57
KNN (k =1)	73.59	0.54	54.55	90.91	80.87	58	71.27	0.62	**90.91**	50	81.07	50.29
KNN (k =2)	74.23	0.66	59.09	**95.45**	81.07	59.57	71.72	0.62	86.36	50	81.13	51.57
SVM(*Gaussian Kernel*)	79.36	0.40	63.62	90.91	85.8	65.57	73.41	0.67	**90.91**	45.45	88.2	41.71

Note: ACC-mean accuracy, minACC-min accuracy, maxACC-max accuracy, std-standard deivation of accuracies accuracy, SEN-sensitivity and SPE-specificity

To examine the cancer classification performance of the two MCM models, we next applied the MCM-based classification rule to the leukemia data. In particular, a 3-fold cross validation procedure was performed, where all samples were randomly evenly divided into 3 parts and each part was used as a test set in turn and the remaining as training set. To remove randomness, we repeated the procedure 100 times. Table 2 shows the maximum, minimum, mean and standard deviation of the resulted classification accuracy, and the average sensitivity and average specificity as well, for the two

MCM. From Table 2, it can be found that NMCM obtained better classification performance than HMCM, irrespective of the GMC used. This can be explained: NMCM can extract details in a GMC by separately constructing the transition probability matrices at each step, but HMCM represents all steps using one transition probability matrix. To further compare with previous methods, we applied KNN(k=1), KNN(k=2) and SVM(Gaussian kernel) to classify the leukemia data with the two GMC. As shown in Table 2, these conventional methods generally led to a average classification accuracy lower than those by our NMCM method. Furthermore, compared with the results by these previous methods, our method, HMCM or NMCM, achieved a more balanced values of specificity and sensitivity, irrespective of the GMC used.

IV. CONCLUSION

We have proposed two models based on Markov chain, HMCM and NMCM, for cancer classification. The models can capture gene co-activity patterns in a gene pathway using state transition probability matrices and thus help improve cancer classification. We evaluated the two models on leukemia data set and compared with several previous classification methods including KNN and SVM. Experiment results show the effectiveness and efficiency of the MCMs, especially NMCM, for cancer classification. Future work will focus on optimal generation of regulatory state space and extensive evaluations on more real-world data sets.

ACKNOWLEDGMENT

This work was supported by the grants of the National Science Foundation of China, Nos. 31071168, 30900321, 60975005, 61005010, 60873012, 60973153, 61133010 and 60905023.

REFERENCES

[1] A. Berns, "Cancer: gene expression diagnosis," Nature 403, pp. 491–492, 2000.

[2] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, p. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D . Botstein, P. O. Brown, "Systematic variation in gene expression patterns in human cancer cell lines," Nat Genet 24**,** pp. 227-235, 2000.

[3] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, L. M. Staudt , "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," Nature 403, pp. 503-511, 2000.

[4] A. Antoniadis, S. Lambert-Lacroix, F. Leblanc, "Effective dimension reduction methods for tumor classification using gene expression data," Bioinformatics 19, pp. 563-570, 2003.

[5] H. Zhang , C. Y. Yu, B. Singer, M. Xiong, "Recursive partitioning for tumor classification with gene expression microarray data," Proc Natl Acad Sci U S A 98, pp. 6730-6735, 2001.

[6] J. P. Novak, R.. Sladek, T. J. Hudson, "Characterization of variability in large-scale gene expression data: implications for study design," Genomics 79, pp. 104-113, 2002.

[7] Y. Tu, G. Stolovitzky, U. Klein, "Quantitative noise analysis for gene expression microarray experiments," Proc Natl Acad Sci U S A 99, pp. 14031-14036, 2002.

[8] H. Herzel, D. Beule, S. Kielbasa, J. Korbel, C. Sers, A. Malik, H. Eickhoff, H. Lehrach, J. Schuchhardt, "Extracting information from cDNA arrays," Chaos 11, pp. 98-107, 2001.

[9] E. P. Xing, M. I. Jordan, R. M. Karp, "Feature selection for high-dimensional genomic microarray data," (ICML2001), pp. 601-608, 2001.

[10] T.R. Golub, et al, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," Science 286(5439), pp. 531–537, 1999.

[11] E. Shin, et al, "TC-VGC: A Tumor Classification System using Variations in Genes' Correlation Comput Methods Programs," Biomed. doi:10.1016/j.cmpb, 2011.

[12] H. Q. Wang, D. S. Huang, "Regulation probability method for gene selection.," Pattern Recognition Letters 27, pp. 116–122, 2006.