ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.1, 523–532
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.1.40

APPLICATION OF A HIDDEN MARKOV MODEL IN IDENTIFYING TARGET GENES IN BREAST CANCER

PARISA TORKAMAN

ABSTRACT. Breast cancer is one of the most common malignant cancers among women with increasing number of patients. Gene regulatory network and identifying target genes for cancer treatment, and reducing breast cancer death rates is of great importance medically. This study aims to model gene regulatory network of breast cancer using hidden Markov model which greatly aids doctors in early diagnosis and faster treatment of breast cancer using identification of target genes. In this study, gene expressions of 206 patients diagnosed with four subtypes of breast cancer including, Basal, Her2, LumA, LumB, were obtained from the Cancer Genome Atlas (TCGA). 8 genes with the verified interaction among them were investigated by hidden Markov model of gene regulatory network and target genes. with the results of transition probability matrix, FADD, TNFRSF10B, CASP8 are the target genes in the mentioned cancer subtypes so that genes that their transmit probabilities are more than an initial value of 0.125 are regulatory genes and transmit matrix identifies the probability of the mentioned cancers regarding gene expression level.

1. INTRODUCTION

Chromosome is inside the nucleus of every creature's body in which there is DNA made up of two regions of coding and non-coding. The coding region of DNA is the genes that contain information for making proteins. After a complex process, the final product of genes are proteins of big molecules based on

2010 *Mathematics Subject Classification.* 62M05, 62P10.
Key words and phrases. Breast Cancer, Hidden Markov Model, Target Gene.

which an organism is created. The existing information in genes is changed into proteins through two processes of transcription of genes to mRNA molecules and then translation in which the information leads to production of proteins. What occurs during these two processes is gene expression which needs various proteins for each job of the cell at different times. This can change based on different growing processes of the cell or environmental conditions permanently or temporarily. In other words, facing new conditions, such as infection, lack of nutrition, stress and anxiety, a cell responds with a new state like cell division by creating a new transcript, gene expression and protein production. In order to measure levels of gene expressions and cells, a new advancement called microarrays is created which allows measurement of thousands of genes simultaneously in one test. In fact, since presence of mRNA of a specific gene reflects its expression level, microarray tests are based on it. Gene expression is one of the most fundamental issues in genetics. To understand the function of an organism we need to know which genes, when and with what speed they are expressed. The final product of gene expression which is protein is produced via a complex process. In addition to specific features of genes, the interaction between proteins and genes and other materials create a complex system which plays an important role in proper performance of cell. Using modeling gene regulatory network, cell activities at molecular level can be interpreted. A descriptive system from interaction of constructing or modeling gene regulatory networks due to identification of diseases and function of unknown genes is vital. Nowadays, identification and understanding of such networks in human is of great importance since gene regulatory mechanism and interaction of genes can lead to complicated diseases. Therefore, identification and modeling gene regulatory network is influential in treating complicated diseases, selecting gene therapy candidates, and function of unknown genes. Moreover, the structure of this network allows comparison between different patterns of gene expressions. Providing a hypothesis for probable relationships between different genes and role of each gene and confirming these hypotheses are all applications of modeling gene regulatory network.

Takan et al., [8] formulated some hypotheses on probable relationships between different genes and the role of each of them in application of modeling of genetic networks. Rishi et al., [6] provide comparison between patterns of different gene expressions by examining function of unknown genes by which

the function of unknown genes can be understood. Numerous mathematical models in research on genes regulatory networks have been introduced including modeling gene regulatory networks, linear models, neural networks, differential equations, and Boolean network. Each method has its own weakness. For example, differential equations require numerous parameters that are hard to achieve, or linear models do not reflect nonlinear interaction between genes. Bayesian method was a practical and effective method that was introduced to examine gene regulatory network using data from gene expression of yeast. Hartmix et al., [2] and Ching [1, 11] carried out research on Bayesian and Markov networks. Bayesian network is a graph consisting of some nodes and links. Each edge of the graph shows a random variable that in gene network its value is the expression level of a given gene and it is a directed link from one point to the other showing that the first node produces a protein that influences the other node. Since data of gene expressions are noisy and probabilistic in nature, Bayesian networks are suitable because they are able to model the random state of these kinds of data. However, the problem in these kinds of networks is that understanding them is difficult so the number of possible structures of the network grows exponentially as number of nodes increase. Moreover, some of the features of gene expression data are fewer observations and high number of genes that make learning the Bayesian networks difficult. Since biological level of gene regulation indicates uncertainty, probable gene expression network has attracted great attention in recent years. This study, using hidden Markov model which can model the complex processes easily and it has been using widely, deal with the modeling of gene regulatory networks.

Hidden Markov model is the recommended model in this study because of two reasons; first, this is a model with strong mathematical structure which forms the theoretical foundation of many applications and, second if this model is implemented properly, it can be applied in various ways. Hidden Markov model is a kind of Markov model in which the modeled system is assumed as a Markov process with hidden states. Markov model was first introduced in 1960. This model is observable directly so the probability of transition between states is the only existing parameter; however, it is not observable directly in some states. Therefore, hidden Markov model is created and the output depending on the state is observable and each state has a probability distribution

on each output symbols. Thus, the produced sequence of symbols by a hidden Markov model produces some information about sequence of sates. Hidden Markov model as a powerful tool for classified learning, prediction and discovery of the hidden relationship - which could not be solved by other methods-aids in the process of problem solving and it is considered as one of the mining methods and machine learning. Hidden Markov model is more renowned due to its application in pattern recognition, voice and handwriting recognition, gesture recognition, bioinformatics etc.

As this model is a powerful tool in predicting random models, it has also been widely applied and investigated in medicine for predicting disease progression and discovering different states of disease diagnosis and so forth. It is also reduces laboratory costs and make the processes faster. For instance, Vimala et al., [9] to diagnose and classify ECG signals, H. Lee et al., [3] to predict the progression of lung cancer and Rafei et al., [5] to recognize Pulmonary Tuberculosis in Iran all used hidden Markov model. Woo Hung et al., [10] in a study called "Detecting Arrhythmia through identification of heart sounds" used hidden Markov model. Schemidth et al., [4, 7] carried out recognition and segmentation of recorded heart sounds with cystoscopy via hidden Markov model. Disease diagnosis through smart systems is fast, more precise and inexpensive which is crucial in health care making hidden Markov model extremely influential in medicine. This study aims to model gene regulatory network of breast cancer data and identify target genes using hidden Markov model as a powerful tool in modeling random processes.

2. Hidden Markov Model

The data used in this research was part of gene expression data from patients with 4 subtypes of cancer including Basal, Her2, LumA, LumB, which was obtained from TCGA. In Markov model, each state corresponding to an event is observable, but if the states are not directly observable, it means that observations are functions of probabilities so the model is a random model with a sub random process which is hidden and only a collection of random processes that produces a sequence of observations is observable. The hidden Markov model was first implemented in mid 1970s to describe speech and, in mid 1980s, it

continued to be utilized for biologic sequences and then spread to bioinformatics. The hidden Markov model tries to model complex Markov processes in which the states based on probability distribution of observations are produced. In fact, this a powerful model for processing and modeling random processes and among statistical methods it can model complex behaviors to predict and classify data. Assume that there are N states of gene expression level and in this study 8 different genes are taken into account and $S = \{S_1, S_2, ..., S_8\}$ is considered as distinct states and for time $t = 1, 2, ..$and the state at time t is shown with q_t. In each state $O = \{O_1, O_2, ...O_M\}$is the observed values that are 4 subtypes of cancer and v_t k th are observed value. In order to construct hidden Markov model in random processes,$\lambda = \{\pi, A, B\}$ parameters should be created.

Definition 2.1. *Probability of the initial state For each state, there is an initial state that Markov chain starts from that state with this initial probability*

$$\pi = [\pi_1, \pi_2, ..., \pi_N],$$

where $i = 1, ..., N$ *and* $\pi_1 = P(q_1 = S_1)$.

Definition 2.2. *Transition Probability matrix It is a matrix from probability of transitions between states* $A = [a_{ij}]$ *and because there are 8 states of gene expressions, the matrix is* 8×8 *and*

$$a_{ij} = P(q_t = S_j | q_{t-1} = S_i) \quad 1 < i, j < N,$$

where the probability of transitions between states has the following features:

$$\sum_{j=1}^{n} a_{ij} = 1, a_{ij} \geq 0.$$

Definition 2.3. *Probability matrix of observation Distribution It is a matrix in which each element shows the probability that if each observation belongs to each state,*

$$b_j(k) = P(O_t = V_k | q_t = S_j),$$

where $B = b_j(k)$ *is a matrix and* $\sum_{k=1}^{M} b_j(k) = 1, b_j(k) \geq 0$. 4 *subtypes of cancer and 8 gene expressions of an 8× 4 matrix are created.*

Three problems should be solved so that the hidden Markov model can be used in real world.

1. Problem of Evaluation. With a sequence of observations $O = \{O_1, O_2, ...O_M\}$ and model $\lambda = \{\pi, A, B\}$ evaluation means calculating the probability of producing a sequence of observations $O = \{O_1, O_2, ...O_M\}$ with $\lambda = \{\pi, A, B\}$ using the model. To solve the problem, prospective and retrospective methods are recommended.
2. Coding problem. With observation sequence of $O = \{O_1, O_2, ...O_M\}$ abd model $\lambda = \{\pi, A, B\}$, state sequence $\{q_1, q_2, ..., q_t\}$ is produced. In fact, generating a sequence of optimal states is done using Viterbi algorithm and it is used in recognition phase.
3. Problem of learning. It means achieving parameters of the model $\lambda = \{\pi, A, B\}$. In fact, it is the optimal estimate of parameters of the model or learning problem of Markov model which is carried out using Baum-Welch algorithm or Viterbi method. Generating the optimal parameters of the hidden Markov model via teaching algorithm can be done using a combination of the two mentioned methods. Since genes are related to biological functions of the body, a basic hypothesis is that genes with similar expression levels are usually regulated. Most of the gene regulatory networks are built based on clustering and making a network using hidden Markov model which divides genes into similar classes. The point is that in using hidden Markov model the model is the main subject of teaching which is possible through Baum-Welch algorithm.

This method offers an optimal value for the three parameters of the model.

First step: the initial parameter λ_0 of hidden Markov model shows that the number of states equals the number of genes and each initial value of state transition matrix equals $\frac{1}{N}$ where N show the number of states and $P(O|\lambda_0)$ can be calculated.

Second step: Reevaluating Markov model based on parameter λ_0 so that Baum-Welch algorithm used in the model can present another λ which is the estimate of the parameter.

Third step: Calculating $P(O|\lambda)$ based on the obtained $\lambda_0 = \lambda$ in the model which is done via prospective and retrospective algorithms.

Fourth step: if $P(O|\lambda) - P(O|\lambda_0) < \varepsilon$ then $\lambda_0 = \lambda$ and it returns to the second step. Otherwise, the teaching process is over and the final

TABLE 1. transition probability matrix of Genes .

	A	B	C	D	E	F	G	H
A	0.0000	0.2609	0.1304	0.1739	0.4350	0.0870	0.0217	0.0870
B	0.0172	0.0000	0.2759	0.2069	0.0690	0.0346	0.1034	0.1379
C	01667	0.0833	0.1250	0.1250	0.0417	0.0471	0.1667	0.2500
D	0.0833	0.1667	0.1250	0.0833	0.2083	0.1250	0.1250	0.0833
E	0.0526	0.1053	0.0010	0.1579	0.0526	0.2105	0.3684	0.0526
F	0.0870	0.0870	0.1304	0.1739	0.2609	0.0870	0.1304	0.0435
G	0.1667	0.1944	0.0833	0.0278	0.0278	0.1389	0.0833	0.2778
H	0.1111	0.2222	0.0370	0.0370	0.0741	0.1825	0.2963	0.0371

hidden Markov model is generated and it can determine a sequence of observations.

3. ANALYSING REAL DATA

In this study expression data of 206 patients diagnosed with 4 subtypes of breast cancer including 44 patients with basal cancer, 25 patients with Her2 cancer, 86 with Lum A and 51 with Lum B cancer were examined. From these genes, 8 genes including, TNFRSF10(A), RIPK1(B), TNFRSF10B(C), CASP10(D), IKBKB(E), FADD(F), CASP8(G) and TNFSF10(H) with a verified interaction between them were chosen. Using hidden Markov model, the optimal value for probability transition matrix between different gene expressions is as follows: These analyses were conducted using MATLAB, 2012.

The initial probability transition matrix has 1/8 elements. In Table 1, each row shows the transition probability corresponding to the target gene which is determined with the probability of regulatory genes of the target genes. For example, the probability of transition of genes TNFRF10B, TNFRSF10, FADD and TNFSF10 to gene CASP8 is greater than the initial value (0.125) so that these genes regulate the target gene CASP8. The Table 2 lists some regulatory genes of target genes.

The Table 3 is release matrix that indicates the investigated probability of gene expressions in creating kinds of breast cancer.

TABLE 2. some regulatory genes of target genes

Target Gene	Regulatory Genes
CASP8	TNFRF10B TNFRSF10 IKBKG FADD TNFSF10
TNFRSF10	TNFRF10 BRIPK1 FADD
FADD	CASP10 IKBKG CASP8 TNFSF10

TABLE 3. Probability matrix of cancer

	basal	Her2	lumA	LumB
TNFRF10B	0.2917	0.0833	0.4583	0.1667
RIPK1	0.2759	0.1379	0.4483	0.1379
TNFRSF10	0.1667	0.0833	0.4583	0.2917
CASP10	0.2500	0.1250	0.4167	0.2083
IKBKG	0.1579	0.1111	0.3889	0.0807
FADD	0.1739	0.1304	0.6087	0.3056
CASP8	0.1944	0.1111	0.3889	0.3056
TNFSF10	0.1852	0.3333	0.1111	0.3704

Using the Table 3, the probability that gene TNFRSF10 leading to cancer Lum A and Her2 is 0.45 and 0.0833, respectively, and this probability for gene expression of CASP10 for basal cancer is 0.2500. Thus, according the obtained results from release matrix, the probability of these four kinds of cancer concerning gene expression levels are determined.

Corollary 3.1. *Breast cancer as the most common cancer among women worldwide is considered as the second leading cause of cancer death and many significant methods have been offered to prevent and treat this disease. Breast cancer is a kind of cancer that develops in breast tissue with various signs including, a lump in the breast, a change in the breast shape, dimpling of the skin, fluid from the nipples and scaly patch of skin. Studies revealed that this cancer is the result of a step-by-step process and occurs because of different genetic changes. Annual estimate in the United States and other countries reveal that mortality rate due to this cancer is on the rise which has made this as a major problem in many countries, including Iran. Numerous studies are developing to discover new ways relating*

to molecular mechanism of cancer and finding a suitable solution for its treatment. Recent studies have presented new ways to treatment strategies against this disease by identifying the effective role of genes in beginning and development of breast cancer. Breast cancer based on markers Estrogen receptor, progesterone and Her2 in cancer cells are divided to four groups of Basal, Her2, LumA and LumB, each of which having different prognosis and treatment. Identifying target genes for breast cancer diagnosis is extremely vital. In other words, since breast cancer can be treated in early stages, studying target genes can aid early diagnosis and increase survival rate. However, getting information about diagnosis is an arduous and time-taking process. Hence, models that can do such thing precisely and efficiently in little time are highly required. Hidden Markov model recommended in this study, allows precise and fast prediction of these kinds of diseases which can be a useful tool to diagnose and treat a disease faster and provide prognosis in clinical research. One of the main findings of this study is probability transition matrix for 8 states of gene expressions which shows regulatory genes and target genes with probable expressions. There are also other issues such as how to choose the initial model or how to evaluate gene regulatory network or complex dynamic system.

REFERENCES

[1] W. CHING, E. FUNG, M. NG, T. AKUSTU: *On construction of stochastic genetic networks based on gene expression sequences,* International Journal of Neural Systems., **15** (2005), 297-310.

[2] A. HARTEMINK, D. GIFFORD, T. JAAKKOLA: *Bayesian methods for elucidating genetic regulatory network*, IEEE intelligent Systems, **17**(2) (2002), 37-43.

[3] HK. LEE , J. LEE, H. KIM , JK. HA, KJ. LEE: *Snoring detection using a piezo snoring sensor based on hidden Markov models*, PhysiolMeas., **35**(4) (2013), 41-52.

[4] S. SCHMIDT, E. TOFT, C. HOLST-HANSEN, C. GRAFF, J. STRUIJK: *Hidden Markov Models Based Research on Lung Cancer Progress Modeling,* Engineering and Technology, **6** (13) (2013), 2470-2473.

[5] R. ORAK : *Tuberculosis Surveillance Using a Hidden Markov Model* , IJPH., **41**(10) (2012), 87-97.

[6] R. RISHI, L. CUPTA, E. K ACHENIE: *A network model for gene regulation, computers and chemical Engineering*, 2007.

[7] S. SCHMIDT, E. TOFT, C. HOLST-HANSEN, C. GRAFF, J. STRUIJK: *Segmentation of heart sound recording from an electronic stethoscope by a Hidden-Markov model*, Comput-Cardiol., **35** (2008), 345-349.

[8] M. TAKAN: *Inference of Gene Regulatory Networks from Large Scale Gene Expression Data,* Montreal, degree of Master of Science, 2003.

[9] K. VIMALA : *Stress causing Arrhythmia Detection from ECG Signal using HMM,* IJIRCCE., **2**(10) (2014), 6079-6085.

[10] H. WU, S. KIM, K. BAE: *Hidden Markov Model with heart sound signals for identification of heart diseases,* Proceeding of 20th International Congress on Acoustics (ICA), Sydney, Australia, (2010), 23-27.

[11] S. Q. ZHANG, W. K. CHING, J. YUE: *Construction and control of genetic regulatory networks*, A multivariate Markov chain approach, Journal of Biomedical Science and Engineering, **1** (2008), 15-21.

DEPARTMENT OF STATISTICS
FACULTY OF MATHEMATICAL SCIENCES AND STATISTICS
MALAYER UNIVERSITY, MALAYER, IRAN
E-mail address: p.torkaman@malayeru.ac.ir