J. Biomedical Science and Engineering, 2010, 3, 633-637 **JBiSE**
doi:10.4236/jbise.2010.36086 Published Online June 2010 (http://www.SciRP.org/journal/jbise/).

The application of hidden markov model in building genetic regulatory network

Rui-Rui Ji, Ding Liu, Wen Zhang

Department of Automation and Information Engineering, Xi'an University of Technology, Xi'an, China.
Email: jirui@xaut.edu.cn; liud@xaut.edu.cn; zhangwendec@yahoo.com.cn

Received 6 March 2010; revised 15 April 2009; accepted 25 April 2009.

ABSTRACT

The research hotspot in post-genomic era is from sequence to function. Building genetic regulatory network (GRN) can help to understand the regulatory mechanism between genes and the function of organisms. Probabilistic GRN has been paid more attention recently. This paper discusses the Hidden Markov Model (HMM) approach served as a tool to build GRN. Different genes with similar expression levels are considered as different states during training HMM. The probable regulatory genes of target genes can be found out through the resulting states transition matrix and the determinate regulatory functions can be predicted using nonlinear regression algorithm. The experiments on artificial and real-life datasets show the effectiveness of HMM in building GRN.

Keywords: Genetic Regulatory Network; Hidden Markov Model; States Transition; Gene Expression Data

1. INTRODUCTION

In order to understand the functioning in living organisms, it should be known which genes are expressed, when and where, and to which extent. The regulation of gene expression is achieved through the interactions between DNA, RNA, proteins, and small molecules. This regulatory system can be described by the structure of network called genetic regulatory network (GRN). Building GRN is a reverse-engineering of real gene expression data to help studying the relationship between genes systematically, understanding the essential rule of biological phenomena and provide valuable idea to treat some complex diseases [1,2]. Many mathematics models have been proposed during GRN research, such as Boolean model, Linear combination model, Weighted matrix model, Differential equations model, and so on [3,4]. Resent years, probabilistic GRN model has been paid more attention for the real biological system is stochastic and the determinate model cannot infer the complex process. Bayesian network and Markov Chain have been studied [5-7]. The final result of probabilistic GRN is represented using a graph consisting vertices (genes) and edges (relationships). The relationships between genes are described through probability. Considering the current probabilistic GRN is simple and cannot give the dynamics behavior. This paper studies the application of Hidden Markov Model (HMM) and nonlinear regression algorithm to GRN building.

This paper is organized as follows: Section 2 gives a theoretical representation of HMM. Section 3 focuses on the application of HMM in building GRN. Section 4 shows the experimental result and discussion. Section 5 is the conclusion.

2. HIDDEN MARKOV MODEL

2.1. Basic Theory

Classical Hidden Markov Model is a kind of stochastic state machine based on statistical signal, which is a double stochastic process where the sequences of states can not to be observed directly. The fundamental stochastic process in this model is a Markov chain which describes the state transition; another stochastic process describes the statistical corresponding relationship between states and the observed value. Different states generate particular sequences of observation according to different probability. The observer can know the feature of the state only through the stochastic process. So this model is called Hidden Markov Model.

A parameter set $\lambda = (\pi, A, B)$ is used to describe a HMM. Where π means the initial state probability, A is the states transition matrix and B is the observed probability. Suppose the number of state is N, $S = [S_1, S_2, ..., S_N]$ is the state set. q_t means the state of the model at time t, $1 \le t \le T$, $O = (O_1, O_2, ..., O_T)$ is the observed value sequence and v_k is the observed value, the three parameters in a HMM model are defined as follows,

$$\pi = [\pi_1, \pi_2, \ldots, \pi_N], \quad \pi_i = P_r(q_1 = S_i) \tag{1}$$

where π_i is the probability of the initial state q_1 equals S_i.

$$A=(a_{ij}),\ a_{ij}=P_r(S_j/S_i),\ \sum_{j=1}^{N}a_{ij}=1 \quad (2)$$

where a_{ij} means the transition probability from state S_i to state S_j. A is a $N\times N$ matrix.

$$B=(b_j(k)),\ b_j(k)=P_r(v_k/s_j),\ \sum_{k=1}^{M}b_j(k)=1 \quad (3)$$

$b_j(k)$ means the observed probability that generating v_k under state s_j. B is a M × N matrix.

2.2. Fundamental Problems in HMM

There are three problems needed to be solved when building a HMM.

1) Evaluation

For a given HMM $\lambda=(\pi,A,B)$ and the observed sequence $O=(O_1,O_2,...,O_T)$, evaluation means calculating the probability $P(O|\lambda)$ corresponding the observed sequence generated by the model, which can evaluate the similarity of the observed sequence with the given model.

Forward-Backward Algorithm proposed by Baum is used to solve the evaluation problem. In practical application, the results are very small and usually normalized or carried on logarithmic operation in calculation process.

2) Decoding

Ascertaining an optimal state transition sequence $Q^*=(q_1^*,q_2^*,...q_T^*)$ on the given HMM $\lambda=(\pi,A,B)$ and observed sequence $O=(O_1,O_2,...,O_T)$ is called decoding.

Viterbi Algorithm can solve the decoding problem. Here, the optimal state transition sequence $Q^*=(q_1^*,q_2^*,...q_T^*)$ is the Q which can make the value of $P(Q,O|\lambda)$ maximum.

3) Learning

Learning can obtain the optimal parameters λ^* of HMM through training algorithm, where λ^* make the value of $P(O|\lambda)$ maximum, for the given HMM $\lambda=(\pi,A,B)$ and observed sequence $O=(O_1,O_2,...,O_T)$.

There are two solutions to solve the learning problem. One uses gradient technique, another is based on iteration or recursion like Baum-Welch algorithm which is often used to train parameters of HMM.

3. MODELING GRN BASED ON HMM

3.1. Constructing HMM

A fundamental assumption is that genes sharing similar expression levels are commonly regulated, and the genes are involved in related biological functions. Most of GRNs are built on the basis of clustering. The process of constructing HMM in this paper is carried on the genes clustered into the same class.

1) States:

Considering the genes clustered into same class, different genes are considered as different states. So the size of state transition matrix is just the same as the number of genes. State transition probability corresponds with the regulatory probability between genes. One gene may be regulated by any other genes, even itself, so the wholly connected connection structure is used.

2) Observed sequence:

The expression profiles of genes are considered as the observed sequence. Since these data is easily affected by noise, smoothing is used firstly to reduce the influence of noise.

3) Training steps:

Step 1: initializing parameters λ_0 of HMM. The number of states is equal to the number of genes and each value of state transition matrix is initialized as average value $1/N$, N is the number of states, and $P(O|\lambda_0)$ can be computed;

Step 2: revaluating HMM's parameter λ_0. Baum-Welch algorithm is used to train HMM model to acquire λ;

Step 3: computing $P(O|\lambda)$ under the obtained model λ using Forward-Backward algorithm;

Step 4: judging the convergence criterion. If $|P(O|\lambda)-P(O|\lambda_0)|\le\varepsilon$ is not satisfied, then $\lambda_0=\lambda$ and return to step 2. Else, training process is finished and final HMM model close to the observed sequence can be acquired.

3.2. Building Probabilistic GRN

Regulatory genes for each target gene can be found out based on the state transition matrix after training HMM. Then the structure of regulatory network can be built according to the following steps:

Step 1: For one target gene $x_i(i=1,2,...,N)$, the genes whose transition probability in the trained state transition matrix A is bigger than the initial average probability are found out and these genes are regarded as the parental regulatory genes of each target gene;

Step 2: repeating step 1 until finding out the global information for each target gene;

Step 3: predicting the determinate regulatory function f_i between target gene and its parental regulatory genes using multiple nonlinear regression and least squares algorithm.

4. EXPERIMENTS AND DISCUSSIONS

4.1. Experiment with Artificial Data

In order to evaluate the efficiency of our algorithms, a group of networks are required whose structure had been known. However, the real structures of GRN are unknown completely because the research about GRN is

still in an early stage. So artificial data reported in paper [8] are used in this study. Here, the adopted ALARM network contains 37 discrete variables, 46 edges and the value of every variable ranges from 2 to 4.

The network with known structure is called target network N_t and the result of our algorithm is called deduced network N_d. Three index sensitivity, specificity and F-factor are used to evaluate our algorithm. Sensitivity is used to test the inference ability, specificity reflects the degree of accuracy and F-factor is the balance of above two indicators. The bigger F-factor means the higher accuracy.

$$Sensitivity = \frac{s_1}{s_t}, \quad Specificity = \frac{s_1}{s_d} \tag{4}$$

where, s_1 means the number of the same edges in both and N_d and N_t; s_t means the total number of the edges in N_t; s_t means the total number of the edges in N_d.

Table 1 shows the comparison of standard Simulated Annealing algorithm, called BANJO developed by Hartemink and our algorithm. It can be seen that our result is better than general Simulated Annealing algorithm in all items.

4.2. Experiment with Real-Life Data

The real-life experimental data in this paper comes from yeast cell cycle expression datasets created by Spellman [9], which imply the regulatory information about genetic property in yeast cell cycle. However, the above mentioned index cannot be used to evaluate the result because the real biological regulatory network is unknown completely. So the existing regulatory relationships which had been already proved are used to evaluate our results.

As listed in Futcher's paper [10], 15 main transcriptional factors: MBP1, SWI4, SWI6, FKS1, MCM1, FKH1, NDD1, SWI5, ACE2, CDC28, CLN3, CLB2, SIC1, CLN2 and HHT1 are discussed. It had been verified that there exist interactions among these genes. For convenience, these genes are marked No. 1, 2,…15.

Table 1. The comparison of BANJO and HMM.

algorithm	BANJO	HMM
right side number	22	31
reverse side number	2	5
sensitivity	0.48	0.67
specificity	0.29	0.76
F-factor	0.36	0.71

1) Experimental results

Initial elements of state transition matrix are assigned as 1/15, **Table 2** is the final state transition matrix trained, and each row shows the transition probability corresponding with target gene, through which the probable regulatory genes of target genes can be found out.

Considering the transition probability of MBP1, SWI6, MCM1 and CLN3 to gene SWI4 are bigger than initial probability, so these genes can be regarded as the regulatory genes of target gene SWI4. **Table 3** lists the regulatory genes of target genes SWI4, SWI5, CLN2, and CLB2.

2) Determinate regulatory relationships

Predicting the determinate regulatory relationships can offset the drawbacks of the probabilistic GRN, which can not describe the specific dynamics behavior. The obtained regulatory relationships between target gene CLB2 (x_{12}) and its regulatory genes MCM1 (x_5) and SIC1 (x_{13}) is:

$$x_{12} = -0.1503 + 0.0943x_5 + 0.0075x_5^2 + 0.0039x_{13} - 0.0026x_{13}^2 - 0.17x_5x_{13}$$

Figure 1 compares the real expression profile of gene CLB2 marked in blue and its predicted value marked in black.

Figure 2 gives a local structure of our resulting GRN. Where, black connecting lines represent the verified existing edges [11]. Blue connecting lines represent the reversed direction with known relationships. Red con-necting lines represent the regulatory relationships pre-

Table 2. The states transition matrix.

Gene	MBP1	SWI4	SWI6	MCM1	SWI5	ACE2	CLN3	CLB2	SIC1	CLN2
MBP1	0.0542	0.0832	0.0623	0.0583	0.1014	0.0501	0.1357	0.0602	0.0532	0.0536
SWI4	0.0348	0.0523	0.0401	0.0347	0.0434	0.0482	0.1523	0.0531	0.0379	0.1768
SWI6	0.0629	0.0961	0.0658	0.0659	0.0653	0.0662	0.0631	0.0661	0.0598	0.0652
MCM1	0.0435	0.0759	0.0403	0.0431	0.1108	0.0463	0.103	0.1204	0.0446	0.1462
SWI5	0.0622	0.0635	0.0608	0.0592	0.0642	0.0631	0.0629	0.0633	0.1261	0.0642
ACE2	0.0582	0.0596	0.0574	0.1137	0.1302	0.0602	0.0588	0.058	0.0569	0.0611
CLN3	0.0578	0.1138	0.0584	0.0603	0.0536	0.0548	0.0637	0.0581	0.0592	0.0514
CLB2	0.0571	0.0562	0.0583	0.0615	0.0592	0.1583	0.0583	0.0633	0.0544	0.1019
SIC1	0.0585	0.0599	0.0592	0.0623	0.0582	0.0592	0.0552	0.0538	0.0653	0.0916
CLN2	0.0598	0.0635	0.066	0.0576	0.0658	0.0649	0.0538	0.0498	0.0531	0.0653

Table 3. The regulatory genes of several target genes.

target gene	regulatory genes
SWI4	MBP1、CLN3、MCM1、SWI6
SWI5	ACE2、MCM1、MBP1
CLN2	MCM1、SWI4、CLB2、SIC1
CLB2	MCM1、SIC1

dicted by our algorithm which is remained to be verified further.

3) Discussions

It had been verified in Reference [12] that gene CLN3 activates gene SWI4, gene SWI4 regulates gene CLN2 meanwhile, gene CLN1 and CLN2 are both influenced by gene CLB2, and CLB2 regulates CLN2. The expression pattern of gene SWI5 is similar with SIC1 and it has been proved that SWI5 regulates SIC1. These conclusions confirm that our algorithm is effective.

Moreover, it had been verified that CLN3 is regulated by gene SWI4 and CLB2 is regulated by gene SIC1, which is identical with the predicted results by our algorithm.

5. CONCLUSIONS

This paper discusses the application of HMM in building GRN. The regulatory genes for each target gene can be found out through the state transition matrix and then the global structure of GRN can be determined. Simulative experiment proves that this algorithm is more effective. The results in real-life data also show its rationality. Compared with the determinate model, HMM is more scientific because it describes the transcriptional regulatory degree between genes through probability. Especially, the present algorithms can find out self-regulatory relationships of genes.

There are still many problems should also be considered during the research of GRN using HMM, for example, how to choose the initial model. Since the biological GRN is a time-continuous and complicated dynamic system and haven't been completely known, as a result, how to evaluate the GRN integrated with biological meanings effectively is the next research.

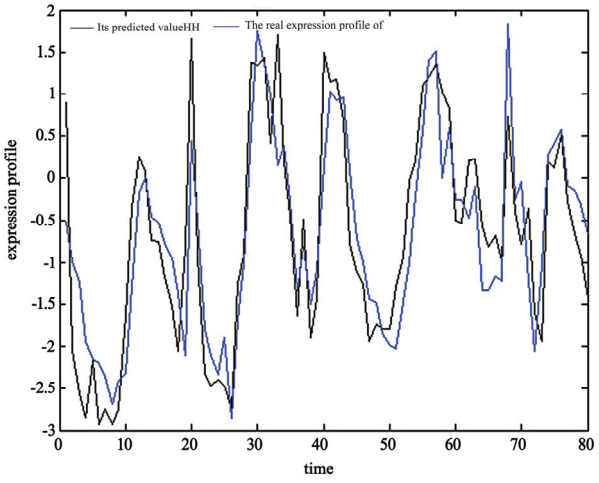

Figure 1. The real expression profile and regression result of gene CLB2.

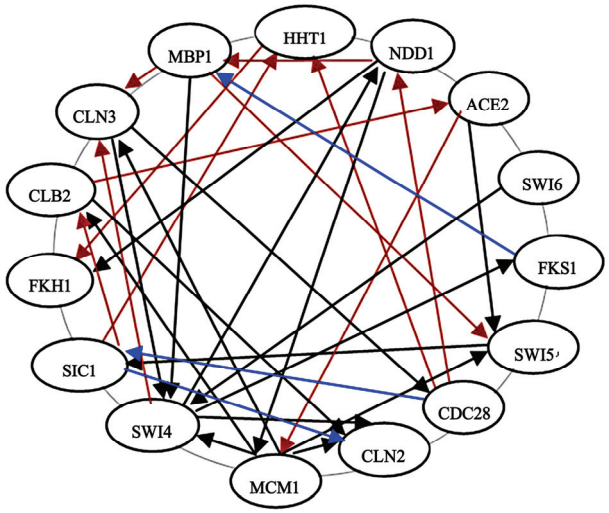

Figure 2. The local structure of GRN.

REFERENCES

[1] Zhao, G.P. (2002) Bioinformatics. Science Press, Beijing.

[2] Hidde, D.J. (2002) Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, **9(9)**, 67-103.

[3] Akustu, T., Miyano, S. and Kuhara, S. (2000) Inferring Qualitative relations in genetic networks and metabolic arrays. *Bioinformatics*, **16(8)**, 727-734.

[4] Bower, J. (2001) Computational modeling of genetic and biochemical networks. MIT Press, Cambridge.

[5] Hartemink, A., Gifford, D., Jaakkola, T., *et al*. (2002) Bayesian methods for elucidating genetic regulatory networks. *IEEE Intelligent Systems*, **17(2)**, 37-43.

[6] Ching, W., Fung, E., Ng, M. and Akustu, T. (2005) On construction of stochastic genetic networks based on gene expression sequences. *International Journal of Neural Systems*, **15(4)**, 297-310.

[7] Zhang, S.-Q., Ching, W.-K. and Yue, J. (2008) Construction and control og genetic regulatory networks: A multivariate Markov chain approach. *Journal of Biomedical Science and Engineering*, **1**, 15-21.

[8] Tsamardinos, I., Brown, L.E. and Aliferis, C.F. (2006) The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, **65(1)**, 31-78.

[9] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Amders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1997) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, **9(12)**, 3273-3297.

[10] Futcher B. (2002) Transcriptional regulatory networks and yeast cell cycle. *Current Opinion in Cell Biology*,

14(6), 676-683.

[11] Paul A.F. M.D. (2004) IncyteDB/OL. http://www.i-ncyte.com/proteome/YPD

[12] Zhang, Z.-F. (2004) Constructing and predicting gene regulatory network using micro-array data. National Central University, Taiwan.

