Analyzing Time-Course Gene Expression Data Using Profile-State Hidden Markov Model

Qiang Huang, Ling-Yun Wu*, Ji-Bin Qu and Xiang-Sun Zhang
National Center for Mathematics and Interdisciplinary Sciences
Institute of Applied Mathematics
Academy of Mathematics and Systems Science, CAS, Beijing 100190
* Corresponding author. Email: lywu@amt.ac.cn

***Abstract*—More and more gene expression data are available due to the rapid development of high-throughput experimental techniques such as microarray and next generation sequencing (NGS). The gene expression data analysis is still one of the fundamental tasks in bioinformatics. In this paper, we propose a new profile-state hidden Markov model (HMM) for analyzing time-course gene expression data, which gives a new point of view to explain the variation of gene expression and regulation in different time. This model addresses the bicluster problem in time-course data efficiently and can identify the irregular shape and overlapping biclusters. The comprehensive computational experiments on simulated and real data show that the new method is effective and useful.**

I. Introduction

High-throughput experimental techniques such as microarray and next generation sequencing (NGS) have been widely used for measuring expression levels of thousands of genes simultaneously. The gene expression data analysis is one of the fundamental tasks in bioinformatics in the post-genomic era. For example, clustering gene expression data has been applied for predicting gene functions [1], discovering transcription regulations [2], revealing cell populations [3], understanding disease processes [4]. The underlying assumption is that the expression levels of genes with similar functions, on the same pathway, or regulated by same transcriptional factors may be highly correlated. On the other hand, using the clusters instead of individual genes can greatly reduce the curse of dimensionality in downstream studies. Many approaches have been proposed for clustering gene expression data in the past decades.

However, a cellular process is active only under some conditions and a single gene may participate in multiple pathways [5]. Therefore, it may be difficult to find the genes that correlate with each other in all conditions or time points. Instead of clustering genes only, we need simultaneously find out the genes and the conditions under which these genes have similar expression profiles. This problem is called biclustering. There are already many biclustering methods such as spectral biclustering [6], Cheng and Church's algorithm [7], iterative signature algorithm [8], high-dimensional linear geometries method [9], non-negative matrix factorization [10], nonparametric bayesian biclustering [11].

In this paper, we propose a novel temporal biclustering algorithm based on profile-state hidden Markov model (HMM) for analyzing time-course gene expression data. As an important and useful tool for bioinformatics, HMM has been widely adapted to analyze gene expression data [12], [13], [14], [15]. However, to the best of our knowledge, there is no method based on HMM for biclustering problem. Using a hidden state to represent an expression profile, we can deal with any type of profile in a bicluster. Several computational experiments on the simulated and real data are conducted to show that the new method can find biclusters in time-course data efficiently and accurately.

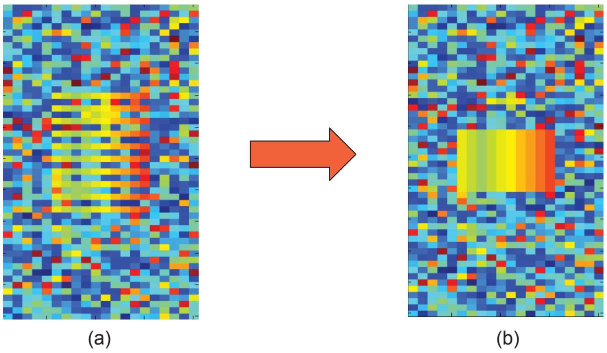


Fig. 1. (a) A gene expression matrix with ten genes that have the same expression profiles under six continuous time points. The blue cells are other irrelevant expression values. (b) The rows of gene expression matrix are reordered to show the bicluster clearly.

II. Method

A. Biclustering Problem

Here we briefly give a description of the biclustering problem on time-course gene expression data. Given a gene expression matrix with n genes (rows) and m time points (columns), the temporal biclustering problem is to find a subset of genes I and a continuous segment of time points J, such that the expression values of genes I follow a desired profile under time points J, as shown in Figure 1. More detail of the biclustering problem can be found in [16].

B. Profile-State HMM

A standard HMM is characterized by the following elements [17]: 1) N, the number of hidden states. 2) M, the number of distinct observation symbols, which is only valid for the

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

discrete HMM and not used here since the gene expression values are continuous values. 3) $A = \{a_{ij}\}$, the state transition probability distribution, where a_{ij} is the transition probability from state i to j. 4) $B = \{b_j(l)\}$, the emission probability distribution, where $b_j(l)$ is the emission probability to observe l in state j. In gene expression data analysis, l is the gene expression value. 5) $\pi = \{\pi_i\}$, the initial state distribution, where π_i is the start point probability of the state i. An HMM is often simply notated as $\lambda = (A, B, \pi)$.

In order to model the time-dependence of the time-course gene expression data, we assume that expression values of each gene are from a Markov process, which is widely used by many existing clustering methods [14]. Each gene stays in one state at a time point. The gene can stay in the same state or switch to other states in the next time point. The genes in the same state have similar expression pattern which is modeled by the state profile. For example, the genes in the same state may participate in the same pathway, and the transition to other states may indicate the gene no longer takes part in the pathway associated with the old state. Generally, the transitions between different states only happen at a few of time points. In other words, the probability of staying in the same state is often larger than that of transiting to other states.

The profile of state k can be parameterized as: $P_k = \{(\mu_{k1}, \sigma_{k1}), \cdots, (\mu_{kj}, \sigma_{kj}), \cdots, (\mu_{km}, \sigma_{km})\}$, where m is the number of time points in the gene expression data, μ_{kj} is the mean expression value and σ_{kj} is the standard deviation of the genes in state k at time j. The expression value of each gene in state k at time j follows a normal distribution defined by the corresponding profile P_k, that is, $x_{ij} \sim N(\mu_{kj}, \sigma_{kj}) + \varepsilon$, $\varepsilon \sim N(0, \sigma)$, where σ is unknown. In order to reduce the model complexity and risk of overfitting, in this study we assume that the Markov chain is time-homogeneous, that is, $A = \{a_{ij}\}$ is same during all time points. To model the genes which are different from all state profiles, we add an background (outlier) state.

C. Biclustering with Profile-State HMM

Given the gene expression data with n genes and m time points, the profile-state HMM is firstly trained by using EM algorithm [18]. Then Viterbi algorithm [17] is applied to infer the hidden state sequences for all genes. Finally, the genes are reordered to show the biclusters clearly.

Initialization Given the gene expression matrix G, the number of states N in profile-state HMM is set as $K + 1$, that is, K biclusters with one background state. K should be less than the upper bound of possible biclusters and larger than the actual number of biclusters. The transition probabilities are initialized uniformly except the self transition, i.e. $a_{ij} = (1 - a_{ii})/(N - 1)$ for $j \neq i$. The initial state probabilities are set as $\pi_i = 1/N$. The initial state profiles are generated by hierarchical clustering genes into N clusters. Then the means and standard deviations of N clusters are assigned to the state profiles. The emission probability for the gene with expression value l in the state k at time j is computed as $b_{kj}(l) = \Pr(x = l | N(\mu_{kj}, \sigma_{kj}))$, while the emission probability for the gene in background state is set as $b_0(l) = \sigma_0$, where σ_0 is a predetermined parameter.

Parameters update The parameters are updated by EM algorithm as follows.

$$\begin{aligned} a_{ij} =& \frac{\text{the number of transitions from state } i \text{ to state } j \text{ in all genes}}{\text{the number of transitions from state } i \text{ in all genes}}; \\ \mu_{kj} =& \text{the mean of expression values of genes in state } k \text{ at time } j; \\ \sigma_{kj} =& \text{the standard deviation of expression values of genes in} \\ & \text{state } k \text{ at time } j. \end{aligned}$$

Post process The hierarchical clustering algorithm is applied to show the bicluser clearly. For each state k of HMM (i.e. bicluster), we construct the state index matrix S^k as follows. The matrix has the same dimension as the gene expression matrix, i.e. n rows and m columns. Each row denotes a gene and each column represents a time point. If the gene i is in the state k at time point j, S^k_{ij} is set as 1, otherwise 0. Then the rows of state index matrix are reordered by using hierarchical clustering. By this way, we get a reordered state index matrix for each bicluster, which can clearly show the structure of the bicluster.

III. Results

The proposed profile-state HMM method was implemented by using Matlab. The Matlab toolbox HMMall[1] is used to train the HMM model and infer the state labels. We set all the initial self transition probability $a_{ii} = 0.9$.

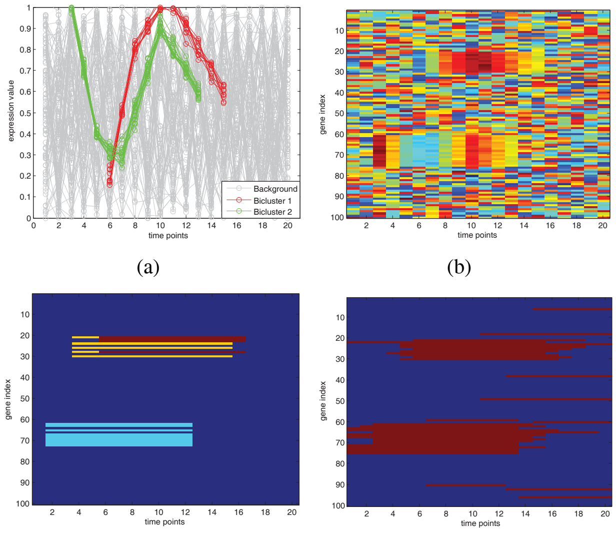
The computational complexity of HMM method is $O(InmK^2)$ where I is the iteration number, and the data matrix is n rows by m columns. In all experiments, HMM method stops in less than 50 iterations.

A. Simulated data

In order to investigate the performance of profile-state HMM, we run the computational experiments on the simulated data with different patterns, different noise levels and different number of biclusters.

The first example contains 100 genes and 20 time points with two simulated biclusters. The first bicluster is 10 gens in time points from 6 to 15 with mean expression values $[0.2, 0.5, 0.8, 0.9, 1, 1, 0.9, 0.8, 0.7, 0.6]$. The second bicluster is 15 genes in time points from 3 to 13 with mean expression values $[1, 0.7, 0.4, 0.3, 0.3, 0.5, 0.7, 0.9, 0.8, 0.7, 0.6]$. Other gene expression values are randomly drawn from the uniform distribution on the interval $[0, 1]$. Noise is imitated by random values drawn from a zero-mean normal distribution with different standard deviations 0.01, 0.02, 0.03, respectively. Figure 2 (a) and (b) shows the simulated gene expression values with noise level 0.03. The HMM method with $K = 4$ and $\sigma_0 = 0.27$ get the correct results for all the noise levels, as shown in Figure 2 (d).

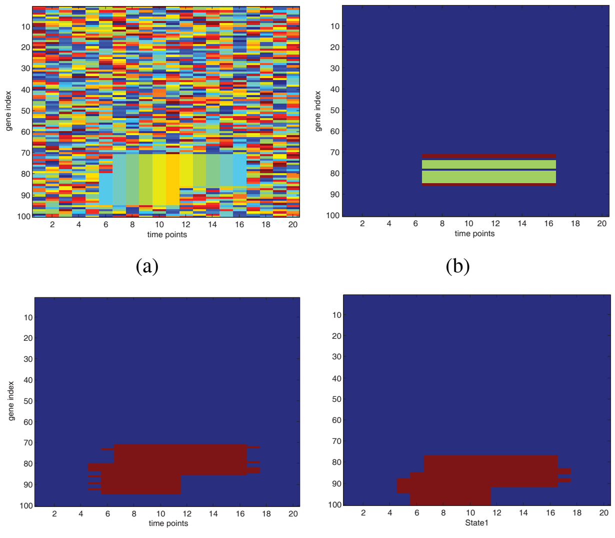
[1] http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html



(a) (b)

(c) (d)

Fig. 2. The simulated data with two biclusters. (a) The expression values of all genes. (b) The heat map of expression data with noise level 0.03. (c) The state labels output by CCC biclustering. (d) The state labels output by HMM.



(a) (b)

(c) (d)

Fig. 3. The simulated data with irregular shape biclusters. (a) The heat map of expression data. (b) The state labels output by CCC biclustering. (c) The state labels output by HMM. (d) The bicluster identified after post-processing.

We also test five existing clustering methods with the default parameters based on the software platform BiCAT [19] for clustering-based data analysis on the same data. All methods fail to find two biclusters while the noise is increasing.

The second example contains one bicluster with irregular shape (non-rectangle) as shown in Figure 3 (a). The result of HMM with $K = 7$ and $\sigma_0 = 0.05$ shows that the bicluster is clearly identified (Figure 3 (c–d)).

The third example is similar as in [9], which contains four overlapping biclusters with a cos curve in 7 time points, a sin curve in 10 time points, a linear increasing curve in 12 time points and a linear decreasing curve in 9 time points, as shown in Figure 4 (a). The noise is drawn from a normal distribution with standard deviation 0.1. The result of HMM with $K = 7$ and $\sigma_0 = 0.05$ is shown in Figure 4 (c–h).

To compare with the temporal biclustering methods, we test the CCC biclustering [20] on the same simulated examples. We set the parameter of overlapping as 0.25. The CCC biclusters are firstly sorted by the statistical significance p-value and then filtered by the parameter of overlapping. The results are shown in Figure 2 (c) (Only the result with noise level 0.03 is shown here), Figure 3 (b) and Figure 4 (b), respectively. These results show that our method is more accurate for identifying biclusters.

B. Real data

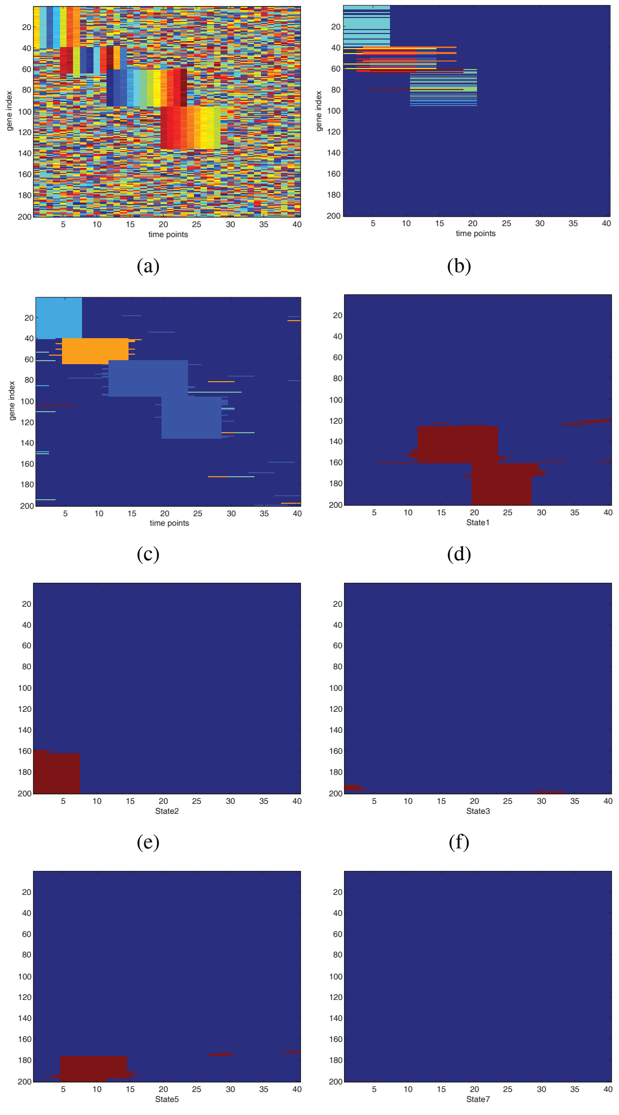
The profile-state HMM is applied to the time-course gene expression data obtained from a well-established model of T-cell activation process in [21] with 58 genes and 10 time points, as shown in Figure 5 (a). Firstly, the expression values of each gene are normalized to zero median. By applying HMM method with $K = 20$, three time-lapse biclusters are found with different expression profiles. The result is shown in Figure 5 (b). We use the g:profiler [22] tool to analysis the enrichment of GO terms for the three biclusters. The enrichment terms are corresponding to three phases: macrophage activation, G1 phase, activation of immune response, respectively, which are clearly described in the T-cell activation process. This result shows that our method can find the dynamic functional variation in actual biological processes.

IV. Discussion and Conclusion

In this paper, we developed a profile-state HMM for analyzing time-course gene expression data. The profile-state HMM addresses a special case of biclustering problem in which the columns are not interchangeable. Each state of HMM is a profile that represents a bicluster center. The major advantage of new method is that it can find the irregular shape and overlapping biclusters, while most of existing biclustering methods focus on discovering regular shape biclusters. For biological data such as gene expression, the bicluster means that the genes have similar expression pattern in some time interval. But this time interval may not be exactly same for all genes in the bicluster. In other words, the genes in the same bicluster may not enter or leave the mode at the same time, so that the bicluster does not necessary have regular rectangle shape. Therefore, the new method is more suited for analyzing time-course gene expression data.

There are two important parameters in the new method. The first is the number of biclusters K. A large K can reveal more small biclusters, but may bring the risk that the large bicluster is divided into several parts. The second is the emission probability of background state σ_0. Increasing σ_0 will let more gene expression values are filtered out from any bicluster. That is, the sizes and standard deviations of biclusters will decrease, and the biclusters will become more tight. However, how to adaptively determine the parameters K and σ_0 is still an ongoing research.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 4. The simulated data with overlapping biclusters. (a) The heat map of expression data. (b) The state labels output by CCC biclustering. (c) The state labels output by HMM. (d)(e)(f)(g)(h) Five biclusters identified after post-processing. (f) and (h) are small fake biclusters due to noise. Two overlapping biclusters are merged in (d) since they have similar expression values in the overlapping time points.

The proposed HMM method is only for biclustering time-course gene expression data instead of more general gene expression data in time independent conditions [6]. Another limitation is that the current model assumes all genes in the same cluster have similar gene expression values at the same time point. The assumption that all genes in the same cluster have similar gene expression trend at the same time point will be more reasonable. One simple way to do this is using the differences between the expression values at two consecutive time points. The assumption of time-homogeneous Markov chain is also not very reasonable. But simply removing this assumption may lead overfitting and fragmentation of biclusters. This will be one of the major challenges of further research.

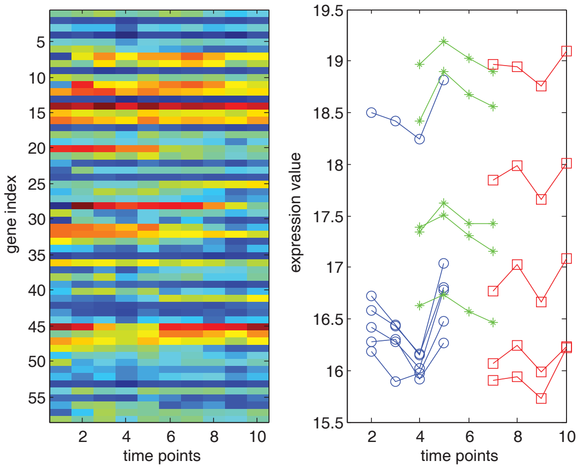


Fig. 5. Real data. (a) The T cell expression data with 58 genes and 10 time points. (b) The profiles of three biclusters with different expression values. The gene expression values before normalization are used.

Acknowledges

Funding: This work is supported by the National Natural Science Foundation of China (Grant No. 60970091 and 60873205), and Knowledge Innovation Program of the Chinese Academy of Sciences (Grant No. kjcx-yw-s7).

Conflict of interest statement. None declared.

References

[1] T. Hughes, M. Marton, A. Jones, C. Roberts, R. Stoughton, C. Armour, H. Bennett, E. Coffey, H. Dai, Y. He *et al.*, "Functional discovery via a compendium of expression profiles," *Cell*, vol. 102, no. 1, pp. 109–126, 2000.

[2] E. Segal, M. Shapira, A. Regev, D. Peer, D. Botstein, D. Koller, and N. Friedman, "Module networks: Discovering regulatory modules and their condition specific regulators from gene expression data," *Nature genetics*, vol. 34, no. 2, pp. 166–176, 2003.

[3] G. R. Howell, D. G. Macalinao, G. L. Sousa, M. Walden, I. Soto, S. C. Kneeland, J. M. Barbay, B. L. King, J. K. Marchant, M. Hibbs, B. Stevens, B. A. Barres, A. F. Clark, R. T. Libby, and S. W. M. John, "Molecular clustering identifies complement and endothelin induction as early events in a mouse model of glaucoma." *J Clin Invest*, vol. 121, no. 4, pp. 1429–1444, Apr 2011. [Online]. Available: http://dx.doi.org/10.1172/JCI44646

[4] L. E. M. Hopcroft, M. W. McBride, K. J. Harris, A. K. Sampson, J. D. McClure, D. Graham, G. Young, T. L. Holyoake, M. A. Girolami, and A. F. Dominiczak, "Predictive response-relevant clustering of expression data provides insights into disease processes." *Nucleic Acids Res*, vol. 38, no. 20, pp. 6831–6840, Nov 2010. [Online]. Available: http://dx.doi.org/10.1093/nar/gkq550

[5] J. DeRisi, V. Iyer, and P. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, no. 5338, pp. 680–686, 1997.

[6] Y. Kluger, R. Basri, J. Chang, and M. Gerstein, "Spectral biclustering of microarray data: coclustering genes and conditions," *Genome Research*, vol. 13, no. 4, pp. 703–716, 2003.

[7] Y. Cheng and G. Church, "Biclustering of expression data." in *Proceedings of International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology*, vol. 8, 2000, pp. 93–103.

[8] S. Bergmann, J. Ihmels, and N. Barkai, "Iterative signature algorithm for the analysis of large-scale gene expression data," *Physical review E*, vol. 67, no. 3, p. 031902, 2003.

[9] X. Gan, A. Liew, and H. Yan, "Discovering biclusters in gene expression data based on high-dimensional linear geometries," *BMC bioinformatics*, vol. 9, no. 1, p. 209, 2008.

[10] P. Carmona-Saez, R. Pascual-Marqui, F. Tirado, J. Carazo, and A. Pascual-Montano, “Biclustering of gene expression data by non-smooth non-negative matrix factorization,” *BMC bioinformatics*, vol. 7, no. 1, p. 78, 2006.

[11] E. Meeds and S. Roweis, “Nonparametric bayesian biclustering,” Citeseer, Tech. Rep., 2007.

[12] C. Li and G. Biswas, “A Bayesian approach to temporal data clustering using hidden Markov models,” in *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 2000, pp. 543–550.

[13] X. Ji, J. Li-Ling, and Z. Sun, “Mining gene expression data using a novel approach based on hidden Markov models,” *FEBS letters*, vol. 542, no. 1-3, pp. 125–131, 2003.

[14] A. Schliep, A. Schönhuth, and C. Steinhoff, “Using hidden Markov models to analyze gene expression time course data,” *Bioinformatics*, vol. 19, no. suppl 1, pp. i255–i263, 2003.

[15] A. Schliep, C. Steinhoff, and A. Schönhuth, “Robust inference of groups in gene expression time-courses using mixtures of HMMs,” *Bioinformatics*, vol. 20, no. suppl 1, p. i283, 2004.

[16] S. Madeira and A. Oliveira, “Biclustering algorithms for biological data analysis: a survey,” *IEEE Transactions on computational Biology and Bioinformatics*, pp. 24–45, 2004.

[17] L. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[18] J. Bilmes, “A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models,” *International Computer Science Institute*, vol. 4, p. 126, 1998.

[19] S. Barkow, S. Bleuler, A. Prelić, P. Zimmermann, and E. Zitzler, “Bicat: a biclustering analysis toolbox,” *Bioinformatics*, vol. 22, no. 10, pp. 1282–1283, 2006.

[20] S. C. Madeira, M. C. Teixeira, I. Sá-Correia, and A. L. Oliveira, “Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm.” *IEEE/ACM Trans Comput Biol Bioinform*, vol. 7, no. 1, pp. 153–165, 2008. [Online]. Available: http://dx.doi.org/10.1109/TCBB.2008.34

[21] C. Rangel, J. Angus, Z. Ghahramani, M. Lioumi, E. Sotheran, A. Gaiba, D. Wild, and F. Falciani, “Modeling t-cell activation using gene expression profiling and state-space models,” *Bioinformatics*, vol. 20, no. 9, pp. 1361–1372, 2004.

[22] J. Reimand, M. Kull, H. Peterson, J. Hansen, and J. Vilo, “g:profiler — a web-based toolset for functional profiling of gene lists from large-scale experiments,” *Nucleic acids research*, vol. 35, no. suppl 2, pp. W193–W200, 2007.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE