2004 IEEE International Workshop on Biomedical Circuits & Systems

A TWO-STAGE METHODOLOGY FOR GENE REGULATORY NETWORK EXTRACTION FROM TIME-COURSE GENE EXPRESSION DATA

Zeke S. H. Chan[1], N. Kasabov[2] and Lesley Collins[3]

[1,2]Knowledge Engineering and Discovery Research Institute (KEDRI) , Auckland University of Technology, Auckland, New Zealand and [3]Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

ABSTRACT

The discovery of gene regulatory networks (GRN) from time-course gene expression data (gene trajectory data) is useful for (1) identifying important genes in relation to a disease or a biological function; (2) gaining an understanding on the dynamic interaction between genes; (3) predicting gene expression values at future time points and accordingly, (4) predicting drug effect over time.

In this paper, we propose a two-stage methodology that is implemented in the software "Gene Network Explorer (GNetXP)" for extracting GRNs from gene trajectory data. In the first stage, we apply a hybrid Genetic Algorithm and Expectation Maximization algorithm on clustering the large number of gene trajectories using the mixture of multiple linear regression models for fitting the trajectory data. In the second stage, we apply the Kalman Filter to identify a set of first-order differential equations that describe the dynamics of the representative trajectories, and use these equations for discovering important gene interactions and predicting gene expression values at future time points. The proposed method is demonstrated on the human fibroblast response gene expression data.

1. INTRODUCTION

Discovering the Gene Regulatory Network (GRN) that governs the dynamics and interaction of genes is an important task for medical purposes. In this paper, we introduce a two-stage methodology that is implemented in a software called the "Gene Network Explorer" (GNetXP, Fig. 1) for extracting GRN from time-course gene expression data (gene trajectory data).

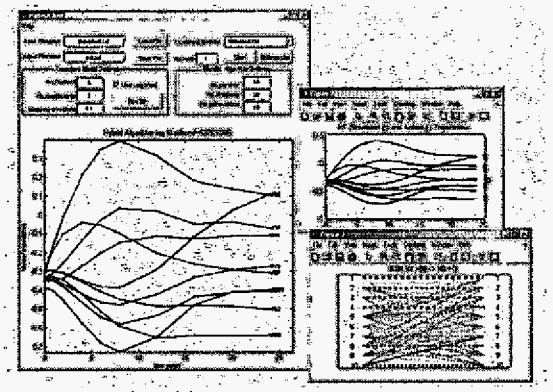

Fig. 1 A screeen shot of GNetXP

In the first stage, the original set of gene trajectory data (several hundreds to thousands) are clustered based on their trajectory similarities, primarily for reducing the number of trajectories to be processed. GNetXP uses a model-based clustering approach - the mixture of Multiple Linear Regression models (MLRs) [1] – to account for the temporal information of the data in the clustering process. Moreover, the learning algorithm is hybridized with a Genetic Algorithm (GA) [2] for improving the optimality and consistency of the clustering solution.

In the second stage, we model the representative trajectories (in our case, the centroids) of the cluster groups with a set of first-order differential equations, which enable easy elucidation of the gene dynamics and interaction. GNetXP applies the the Kalman Filter (KF) algorithm [3] for parameter estimation using the Expectation Maximization algorithm (EM) [4]. The reason for using KF is that it can handle noisy, and even missing or irregularly spaced data, which are common problems with time-course gene expression data.

As an illustration, we apply GNetXP on analysing the human fibroblasts to serum response time-course gene expression data [5] and compare some of results with the findings from the literature. In section 2, we briefly describe the implementation of the model-based clustering algorithm and the hybrid GA approach. In section 3, we discuss using the first order differentiation equations for modeling GRN and KF for parameter estimation. Experimental results on the human fibroblast data is described in section 4.

2. THE HYBRID GENE TRAJECTORY CLUSTERING ALGORITHM

2.2 The Clustering Model: Mixture of MLRs

The clustering model is a mixture of G MLRs (one for each cluster), each of which represents a single gene trajectory cluster given by

$$\mathbf{Y}_i = S(\mu_k + \gamma_i) + \varepsilon_i$$
$$\gamma_i \sim N(0,\Gamma) \quad \varepsilon_i \sim N(0,\sigma^2\mathbf{I}) \qquad (1)$$

0-7803-8665-5/04/$20.00 ©2004 IEEE

where $\mathbf{Y}_i = [y_{i,1}, y_{i,2}, \ldots y_{i,l}]^t$ is the ith gene trajectory of length l, S is the $l\times(p+1)$ regression matrix (or basis matrix) where p is the regression order, μ_k is the $(p+1)$-vector of regression coefficients, γ_i and ε_i are uncorrelated Gaussian noises for the regression coefficients and the trajectory respectively. Here we use the Vandermonde function as the regression matrix S, while the spline basis function or time-series functions are also possible. Let $\mathbf{z}_i = (z_{i1},\ldots,z_{iG})$ be the cluster membership vector for the ith trajectory where z_{ik}=1 if the ith trajectory belongs to the kth cluster and 0 otherwise. The standard method for mixture model learning is to treat z_i as missing variables and apply the Expectation Maximization algorithm (EM) [1, 4], which maximizes the complete data log likelihood.

2.3 The Hybrid Clustering Algorithm

The choice of initial conditions for the clustering model makes significant difference to the quality of the final solution. It is because the search space is highly multi-modal, and since EM is a local optimizer that performs hill-climbing from the initial solutions, the proximity between the optimized solution and the global optimum is very sensitive to the proximity between the initial solutions and the global optimum. In the standard EM clustering method, the initial centers are randomly chosen from the data[1]. The clustering solutions are therefore, often sub-optimal and inconsistent.

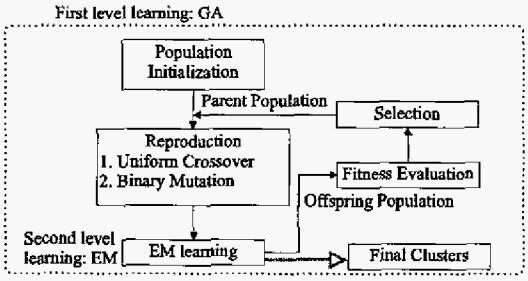

Fig. 2. The hybrid clustering algorithm.

The hybrid algorithm improves the initial estimates by using GA to select the optimal subset of data as the initial cluster centers. It combines the strengths of GA and EM to produce a global yet efficient clustering algorithm. It consists of two levels, as depicted in Fig. 2. At the higher level (the "wrapper" level), GA searches for the optimal subset of genes to be the initial cluster centers. At the lower level, the local learning method (EM) performs local clustering from these initial centers. The genetic operators include uniform crossover, mutation and the repair operator (a mutation operator for ensuring all solutions are feasible). The elitist scheme $(\mu+\lambda)$ is used instead of the Roulette Wheel to achieve faster convergence.

[1] In some implementations of EM, the initial centers are approximated using the K-means algorithm. This method is however, not applicable to the mixture of MLRs because K-means does not incorporate the temporal information into the similar measure of gene trajectories, giving poor approximation to the true objective function.

3. GRN MODELING USING THE KALMAN FILTERING METHOD

3.1 Discrete-Time Approximation of the First-Order Differential Equations

Once the gene trajectory clusters are found, we use the first-order differential equations to model the GRN of the representative trajectories of the clusters. We represent the true gene trajectory as unobserved variables $\{\mathbf{x}_t\}$ called the state variables and the observed gene trajectory as $\{\mathbf{y}_t\}$. The state space representation of the first-order differential equations is given by:

$$\mathbf{x}_{t+1} = \mathbf{\Phi}\mathbf{x}_t + \mathbf{w}_t \quad (2)$$

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\mu} + \mathbf{v}_t \quad (3)$$

$$\mathrm{cov}(\mathbf{v}_t) = \mathbf{R} \qquad \mathrm{cov}(\mathbf{w}_t) = \mathbf{Q} \quad (4)$$

where $\mathbf{\Phi}$ is the state transition matrix that relates $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$, $\mathbf{A}$=$\mathbf{I}$ is the linear connection matrix that relates $\mathbf{x}_t$ to $\mathbf{y}_t$, $\mathbf{w}_t$ and $\mathbf{v}_t$ are uncorrelated white noise sequences whose covariance matrices are $\mathbf{Q}$ and $\mathbf{R}$ respectively.

Besides being a tool widely used for modeling biological processes, there are two advantages in using the first-order differential equations.

First, gene relations can be elucidated from the transition matrix $\mathbf{\Phi}$. Significant gene interactions can be identified as those elements whose absolute value is greater than a pre-defined threshold. Such information can be expressed in an influence matrix or network diagram, as shown in Fig. 2.

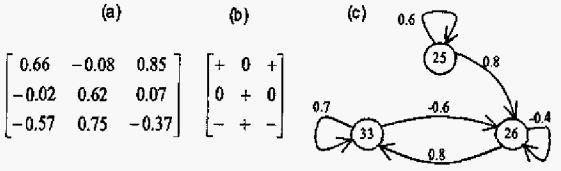

Fig. 3 (a) The transition matrix Φ for GRN of genes (33, 25, 26); (b,c) the corresponding influence matrix and network diagram, respectively, when a threshold of 0.3 is used.

Second, they can be easily manipulated with KF to handle irregularly sampled data, which allow parameter estimation, likelihood evaluation and model simulation and prediction. The main limitation in using the differential equations is that given n is the number of genes modeled, it requires estimation of $O(n^2)$ parameters – i.e. n^2 parameters for the transition matrix $\mathbf{\Phi}$, n parameters for the bias $\boldsymbol{\mu}$ and $n(n\text{-}1)/2$ parameters for the noise covariance $\mathbf{Q}$, which restricts the size of GRN we can model from the limited amount of data. For this reason, trajectory clustering is an integral part of the GRN discovery process as a tool for problem dimension reduction.

4. ANALYSING THE HUMAN FIBBROBLAST TIME SERIES DATA WITH GNETXP

The data set used for this study was reported in [5], which contains gene expression data for the response of human fibroblast cells to fetal bovine serum (FBS). The addition of serum to fibroblast induces changes in the expression of many genes, resulting in fibroblast cell growth. This response has been used in the past as a model for studying cell growth, cell cycle progression and fibroblast wound response [5-7]. The data contains expression data of 8618 recorded at 12 irregularly spaced time points during the physiological response of fibroblasts to serum using cDNA microarrays. The log-normalized data of 517 genes is shown in Fig. 4.

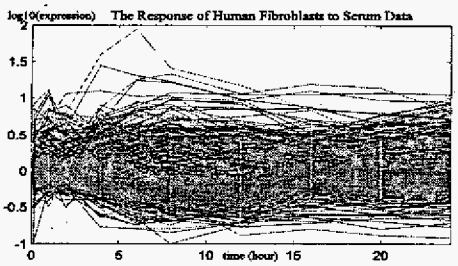

Fig. 4. The human fibroblasts to serum data (517 genes).

We use the default settings of GNetXP, which are as follows. The order for the MLRs is chosen to be P=6 since the coefficients for higher orders are negligibly small ($<10^{-4}$). The population sizes are μ=10 for the parents and λ=20 for the offspring. Uniform crossover is applied at a high crossover rate p_c=0.9, which is a common choice to facilitate transmission of optimal schema in the population. While binary mutation is often applied at the mutation rate p_m=1/N that yields an average of one inversion per string, we set p_m to relative high rate of p_m=(G/N) to yield an average of G inversions per string for increasing the diversity of the search. Each GA runs for 20 generations (obviously, the larger the number the generation, the higher the solution quality becomes. Here we use a relatively small number of generations for time saving and demonstration purpose only). For EM, the stopping criterion is when the maximum log likelihood increases by less than 1.

4.1 Clustering Performance Analyses

To verify the effectiveness of the hybrid clustering algorithm, we (i) compare the model likelihood (the log likelihood) returned by the two algorithms in clustering the data over {5, 7, 9 11, 13, 15} clusters, and (ii) compare the consistency of the grouping of genes. All results are averaged over 10 runs.

Table 2. Maximum log likelihood of the mixture of MLRs identified by the hybrid algorithm and the standard EM on different number of clusters.

No. Clusters	Hybrid Method		Standard EM	
	mean	std.	mean	std.
5	**2200.50**	**0.08**	2114.60	50.52
7	**2590.20**	**0.09**	2511.40	68.57
9	**2898.50**	**0.03**	2809.10	71.05
11	**3086.20**	**0.74**	3019.10	31.11
13	**3246.70**	**2.46**	3136.70	54.21
15	**3350.40**	**1.69**	3228.70	57.53

The likelihood comparison in Table 2 show that for all number of clusters, the hybrid algorithm scores higher model likelihood values, which shows that more globally optimal solutions are achieved. In addition, the standard deviations of the likelihood values are much smaller, which shows that it records more consistent results over different runs.

Next, we examine some examples of gene groupings returned by both algorithms. These groupings are based on the findings from previous fibroblast response gene expression studies [5, 6], which show that certain key genes can be expressed in a similar fashion during a time-course experiment and should thus be clustered together upon microarray analysis. Results tabulated in table 3 shows that the hybrid algorithm produces more consistent gene groupings. Highly reliable clustering algorithms are essential to microarray expression analysis as inconsistent clustering may cause misleading assumptions on the genes having similar expression patterns. Improvements by the hybrid algorithm in clustering reliability are therefore of practical importance.

Note that GA's superior performance incurs higher computational costs, requiring a total of (λ×max.generations)=(20×10)=200 EM evaluations. However, with the human fibroblast data that has 517 genes and 12 time points, each EM evaluation requires less than 10 seconds (running in Matlab on a Pentium IV 2.4GHz), and hence GA poses little problem with computation time. In addition, we can easily apply parallel computing techniques to achieve speed up with GA.

4.2 GRN Identification and Analysis

After limited trials, we find that the interactions between clusters are most easily elucidated using 10 clusters.

Table 1. Some examples of gene groupings obtained in 10 runs of the standard EM and the hybrid algorithm.

Genes/Transcription Factors	No.Genes	Functional Group	Freq. of Grouping	
			Std. EM	Hybrid Alg.
c-FOS, JunB, MAPKP1	3	Transcription factors	10/10	10/10
c-FOS, JunB, MAPKP1, **Dec1, A20**	5	Transcription factors	8/10	10/10
CyclinA, CyclinB1, Cdc2, Cdc28	4	Cell cycle progression	9/10	10/10
CyclinA, CyclinB1, Cdc2,Cdc28, **Madp2, CENP-f**	6	Cell cycle progression	8/10	10/10
p18, Wee1-like, DP2(E2F2)	3	Cell cycle inhibitors	6/10	10/10

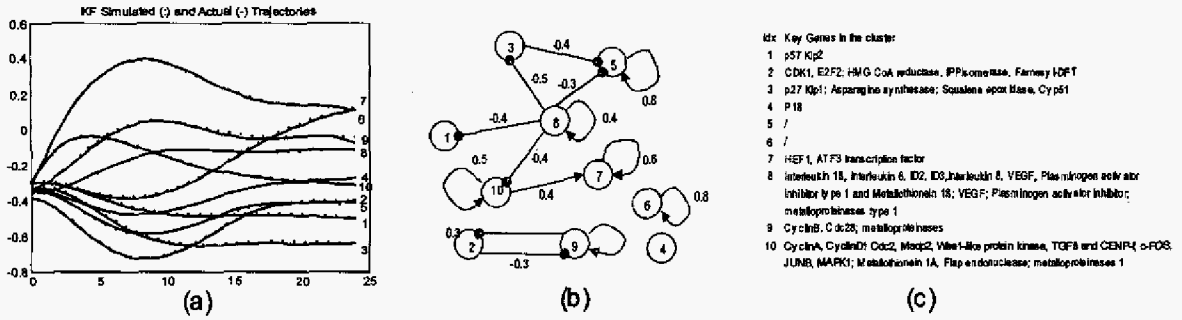

Fig. 5 (a) Actual and KF simulated trajectories; (b) network diagram of the GRN; (c) list of key genes in each of the 10 clusters.

Fig. 5 shows (a) the model simulated trajectories, (b) the network diagram of the discovered GRN and (c) the list the key genes in each of the 10 clusters.

First of all, note that the model tracks closely to the actual trajectories (Fig. 5(a)), showing that the first-order differential equations is sufficient even for the complex trajectories in this case.

Next, we examined the biological appropriateness of the discovered GRN on two examples. The cell cycle genes (CyclinA, CyclinD1 Cdc2, Madp2, Wee1-like protein kinase, TGFβ and CENP-f) are grouped together in cluster 10 and exhibit a large coefficient (0.5) of being self-regulatory. This cluster appears to also have an up-regulatory effect on cluster 7, which contains Human enhancer of filamentation-1 (HEF1) and the ATF3 transcription factor. This agrees with previous findings [8, 9] that HEF1 is involved in integrin-based signaling that affects cell growth and death, and is strongly up-regulated by TGFβ, which is a cytokine that regulates re-modeling of tissue extracellular matrix during wound healing.

Cluster 8 consists of Interleukin 1β, Interleukin 6, ID2, ID3, Interleukin 8, VEGF, Plasminogen activator inhibitor type 1 and Metallothionein 1β, regulate a number of other clusters, including self-regulation and the down-regulation of cluster 10 and 1, 3 and 5. These results indicate that cluster 8 contains some fibroblast response regulatory genes that control key responses to serum stimulation. Interleukin 6 has been shown in gene expression studies to affect 57 genes in normal human fibroblast cells [9]. Some of these genes include Metallothionein 1A and Flap endonuclease 1 (cluster 10), Asparagine synthesase (cluster 3), and the some cluster 8, of which two are of particular importance. The first is the Vascular endothelial growth factor (VEGF), an important regulatory gene that in turn is regulated by ID2 and ID3 [10]. Another key cluster 8 gene is Plasminogen activator inhibitor type 1. This gene inhibits the activation of plasminogen to plasmin [9]. Down-regulation of this gene will decrease plasmin levels, which will in turn down-regulate other key growth factors including the TGFβ (cluster 10) and the metalloproteinases (cluster 8, 10 and 9). Plasmin itself plays a central role in wound repair as it degrades fibrin, a major component of the haemostatic clot [9]. Down-regulation of the production of plasmin would have the effect therefore of stopping cell growth, the first stage of wound repair.

5. CONCLUSIONS

The proposed two-stage method for GRN discovery is described and is demonstrated on the human fibroblast data using the software GNetXP.

11. ACKNOWLEDGMENTS

This research is supported by the KEDRI postdoctoral fellow research fund and by FRST New Zealand (NERF/AUT X0201).

12. REFERENCES

[1] S. Gaffney and P. Smyth, "Curve Clustering with Random Effects Regression Mixtures," presented at Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Key West, Florida, 2003.

[2] J. H. Holland, *Adaptation in natural and artificial systems*: The University of Michigan Press, Ann Arbor, MI, 1975.

[3] R. H. Shumway, *Applied Statistical Time Series Analysis*. Englewood Cliffs, New Jersey: Prentice Hall, 1988.

[4] A. P. Dempster, N. M. Larird, and D. B. Rubin, "Maximum likelihood for incomplete data via the EM algorithm," *Journal of Statistics Society*, vol. B, pp. 1-38, 1977.

[5] V. R. Iyer, M. B. Eisen, D. T. Ross, and G. Schuler, "The transcriptional program in the response of human fibroblasts to serum," *Science*, vol. 283, pp. 83-87, 1999.

[6] S. M. Swamy, P. Tan, Y. Z. SZhu, J. Lu, H. N. Achuth, and S. Moochhala, "Role of phenytoin in wound healing: microarray analysis of early transcriptional responses in human dermal fibroblasts," *Biochem. Biophys. Res. Commun.*, vol. 314, pp. 661-666, 2004 Feb.

[7] J. A. Winkles, "Serum- and polypeptide growth factor-inducible gene expression in mouse fibroblasts," *Prog. Nucleic Acid Res Mol Biol.*, pp. 41-78, 1998.

[8] M. Zheng and P. J. McKeown-Longo, "Regulation of HEF1 expression and phosphorylation by TGF-beta 1 and cell adhesion," *J.Biol.Chem.*, vol. 277, pp. 39599-39608, 2002.

[9] M. R. Dasu, H. K. Hawkins, R. E. Barrow, H. Xue, and D. N. Herndon, "Gene expression profiles from hypertrophic scar fibroblasts before and after IL-6 stimulation," *J.Pathol.*, vol. 202, pp. 476-485, 2004.

[10] D. Sakurai, N. Tsuchiya, A. Yamaguchi, Y. Okaji, N. H. Tsuno, T. Kobata, K. Takahashi, and K. Tokunaga, "Crucial role of inhibitor of DNA binding/differentiation in the vascular endothelial growth factor-induced activation and angiogenic processes of human endothelial cells," *J.Immunol.*, vol. 173, pp. 5801-5809, 2004.