A Customized Class of Functions for Modeling and Clustering Gene Expression Profiles in Embryonic Stem Cells

Shenggang Li[1], Miguel Andrade-Navarro[2], and David Sankoff[1]

[1] Department of Mathematics and Statistics, University of Ottawa, Canada
[2] Max Delbrück Center for Molecular Medicine, Berlin-Buch, Germany

Abstract. Based on the trajectories of individual genes, we address the problem of clustering time course gene expression data for embryonic stem cells (ESC) differentiation. We propose a class of functions determined by only two parameters but flexible enough to model realistic time courses. This serves as a basis for a mixed model clustering method. This method takes into account (1) genetic function profile induced or controlled by other regulators, (2) unobservable random effects producing heterogeneity within gene clusters, and (3) autoregressive components defining the stochastic and autocorrelation structures. We employ an EM algorithm to fit the mixture model and clustering follows monitoring via Bayesian posterior probabilities. Our method is applied to a mouse ESC line during the first 24 hours of differentiation period. We assess the biological credibility of the results by detecting significantly associated FatiGO Gene Ontology terms for each cluster.

1 Introduction

Cluster analysis of gene expression profiles over a time course often treat each sampling time as one of the dimensions of a multivariate variable, reducing the problem of clustering the gene trajectories to one of clustering multivariate observations. Since the object of gene expression clustering is to detect common trajectories, however, this "static" multivariate characterization loses power and precision, since the temporal order plays no role in the analysis; indeed the times can be permuted in any order without changing the results of the cluster analysis.

Gene expression profiles during cell growth, differentation, tissue activation of various kinds, and during cyclical changes tend to follow well-defined patterns and are often highly autocorrelated. These facts lead to the incorporation of more "dynamic" considerations into clustering procedures. Model-based clustering techniques [9] incorporate dynamical features by implementing formal statistical modeling and parametric models in preference to pure data mining.

A number of model-based clustering methods for dynamical gene expression data have been proposed [1,5,14,17]. Here a group of genes following the same probability model fall into the same cluster, taking into account autoregressive and other random effects. For example, [19] discusses autoregressive models for clustering time course gene expression data. [2,11] apply cubic spline and B-spline mixture models to analyze gene expression time series data.

A.L.C. Bazzan, M. Craven, and N.F. Martins (Eds.): BSB 2008, LNBI 5167, pp. 92–103, 2008.

Purely autoregressive models, however, only consider autocorrelative structure without explicitly taking account of other properties of functional interactions among genes. Although a p-order autoregressive model may explain a p-order polynomial time trend effect, it fails to match exponential time growth effects such as $\exp(\lambda t)$ or periodic effects such as $\sin(\omega t)$. Spline techniques can represent data trends but still remain essentially black boxes [8] with respect to genetic functions. [8] suggests a novel clustering technique based on gene functional curves, making use of plausible biological models for gene expression dynamics.

1.1 The Proposed Model

Here we treat a gene profile over a time course as composed of the following components

1. impact of related genes or other regulation mechanisms (modeled as a dynamical system of differential equations),
2. autoregressive factors representing autocorrelation structure and feedback or loop effects,
3. random components allowing for heterogeneity among the individual genes in each cluster.

This method is flexible in assigning both dynamic and fixed functional components into an integrated time series model with random mixtures. We note that by appropriately incorporating randomness into the model we can avoid the tendency to generate clusters based purely on chance properties of the data.

2 The Mixture Time Series Model for Functional Curve Clustering

We denote an expression profile (or observation vector over time course) of an individual gene g_i as $Y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})$. In this notation T refers to the total number of time points sampled over a time course. Our objective is to cluster target genes into distinct clusters in terms of the similarities of gene functional behavior. Genes are grouped together on the basis of the form of their expression profiles (curve shapes) rather than similarity in Euclidean distance. The model-based clustering analysis assumes that the genes in every cluster will perfectly fit an underlying mixture time series model, whereany gene profile Y_i is considered as an observation of a functional curve following the probability model:

$$\begin{aligned} Y_{it} = {} & \text{Gene Network Interactions} + \text{Auto Effects} \\ & + \text{Time–dependent Random Effect} + \text{Noise} \end{aligned} \tag{1}$$

If all target genes $g_i, i = 1, 2, \ldots, N$ satisfy the model above, then the $Y_{it}, i = 1, 2 \ldots, N$ are generated by the following model:

$$Y_{it} = f^{(c)}(t|\Phi^{(c)}) + \sum_{j=t-p}^{t-1} \alpha_{jc} Y_{ij} + \gamma_c + \varepsilon \quad (t = 1, 2 \ldots, T \text{ and } i = 1, 2, \ldots, N), \tag{2}$$

where the "within" random effect $\gamma_c \sim \text{normal}(0,\tau_c^2)$ $(c = 1, 2, \ldots, C)$ are independent of each other, indicating that genes in the same cluster have a unified correlation structure, ε is the i.i.d. white noise $\sim \text{normal}(0,\sigma^2)$, the trend curve $f^{(c)}(t|\Phi^{(c)})$ contains the internal function effect of g_i, combined with the other network interactions (or regulatory effect on g_i), and $\Phi^{(c)}$ is the parameter set for the fixed mean curve in cluster c $(c = 1, 2, \ldots, C)$. We have included autoregressive items in the model because we wish to capture feedback or loop effects in genetic pathways.

The key problem here is to determine the form of function $f^{(c)}(t|\Phi^{(c)})$. We will estimate it with the trend deduced from a genetic transcription model. The B spline techniques previously employed to deal with this problem [11] have no direct biological functional interpretation. Instead we are inspired Chen's linear transcription model [7] to elaborate a function $f^{(c)}(t|\Phi^{(c)})$. Chen's model is a nonlinear dynamic system with the following form:

$$\frac{dr}{dt} = Cp - Vr \frac{dp}{dt} = Lr - Up \tag{3}$$

where r is the mRNA concentration, p is the protein concentration, L contains the translational constants, V is the degradation rate of mRNA and U is the degradation rate of proteins. According to Theorem 1 in [7], the solution to model (3) has form:

$$x(t) = Q(t)e^{t\lambda} \tag{4}$$

where $Q(t) = \{q_{ij}\}$ is a $2n \times 2n$ matrix whose elements are polynomial functions of t. In [7] it is stated that a gene system should be stable system and its expression should not have an exponential or a polynomial growth rate, which implies that $Q(t)$ is a constant and $x(t)$ is actually an ordinary exponential function. However, an exponential function is monotonic and cannot fully reflect the wave-like shapes that characterize many gene functional curves. Consequently we do not adopt (4) directly, though we will utilize features of (3) and (4) to develop a more effective functional curve.

To do so so, we have studied the functional gene curves of mouse embryonic stem cells (mESC) in the first 24 hours of differentiation (for experimental details and the data bank, see [10]) and found that gene expression of these genes can be well described by the following hyperbolic function:

$$f(t) = \frac{1}{\exp[b(t-a) + \sqrt{1 + (t-a)^2}]} \tag{5}$$

To see the connection between (5) on one hand and (3) and (4) on the other, we note that the gene's growth rate satisfies:

$$\frac{df(t)}{dt} = -[b + \frac{t-a}{\sqrt{1 + (t-a)^2}}]f(t), \tag{6}$$

where the expression (6) indicates that this gene system also remains stable, but is more flexible than the one in (3). Here $\frac{t-a}{\sqrt{1+(t-a)^2}}$ is restricted to the

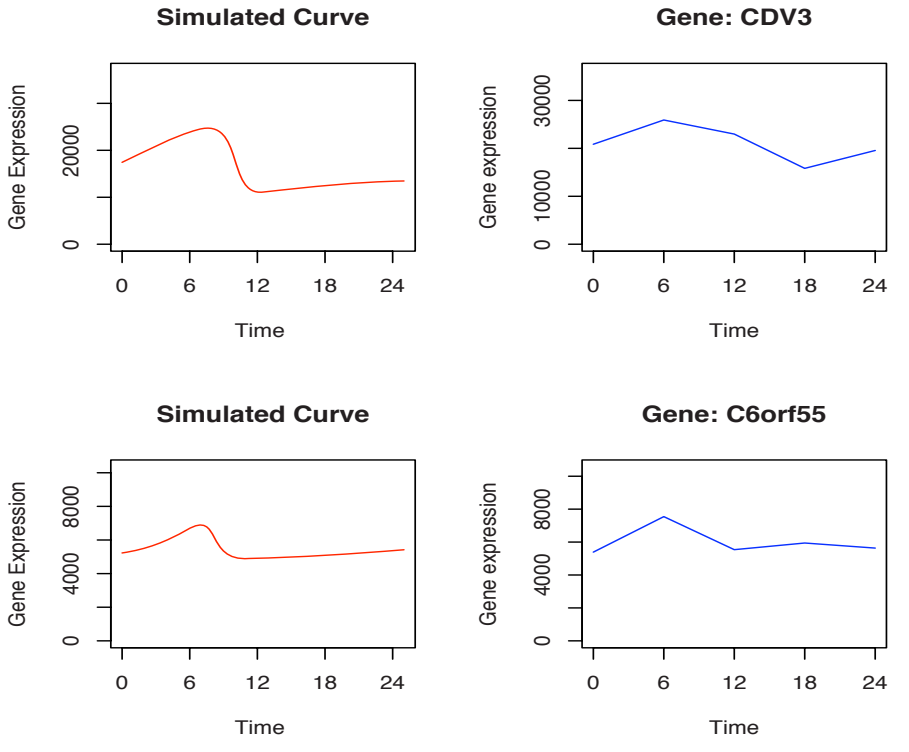

Fig. 1. Simulated curve *vs.* actual gene curve

range [-1,1], but can model behaviours more varied than a the constant growth rate. In Figure 1, the gene profiles (over the time course 0 to 24 hours) of Cdv3(CDV3 homolog (mouse)) and C6orf55 (chromosome 6 open reading frame 55) are plotted against the corresponding simulated hyperbolic functional curves. In the same way, Figure 2 presents the plots for genes Riok2 (RIO kinase 2) and Etnk1(Ethanolamine kinase 1). These genes are highly active over the time course in the early period of stem cell differentiation. These figures illustrate how parameter settings are available to produce functional curve shapes similar to those of actual gene profiles.

For curve fitting purposes, we can include linear and quadratic parts without affecting system stability. Thus, we define gene functional curves as follows:

$$f(t|\beta_l) = \beta^{(1)} + \beta^{(2)} \exp\{-\beta^{(3)}(t-\beta^{(4)}) - \sqrt{1+(t-\beta^{(4)})^2}\} + \beta^{(5)}t + \beta^{(6)}t^2 \quad (7)$$

From (2) and (7), by setting different curve parameters $\beta_c = (\beta_c^{(1)}, \beta_c^{(2)}, \beta_c^{(3)}, \beta_c^{(4)}, \beta_c^{(5)})$, random effects γ_c and auto regression coefficients α_{jc} $(c = 1, 2, \ldots, C)$ we can determine different gene clusters.

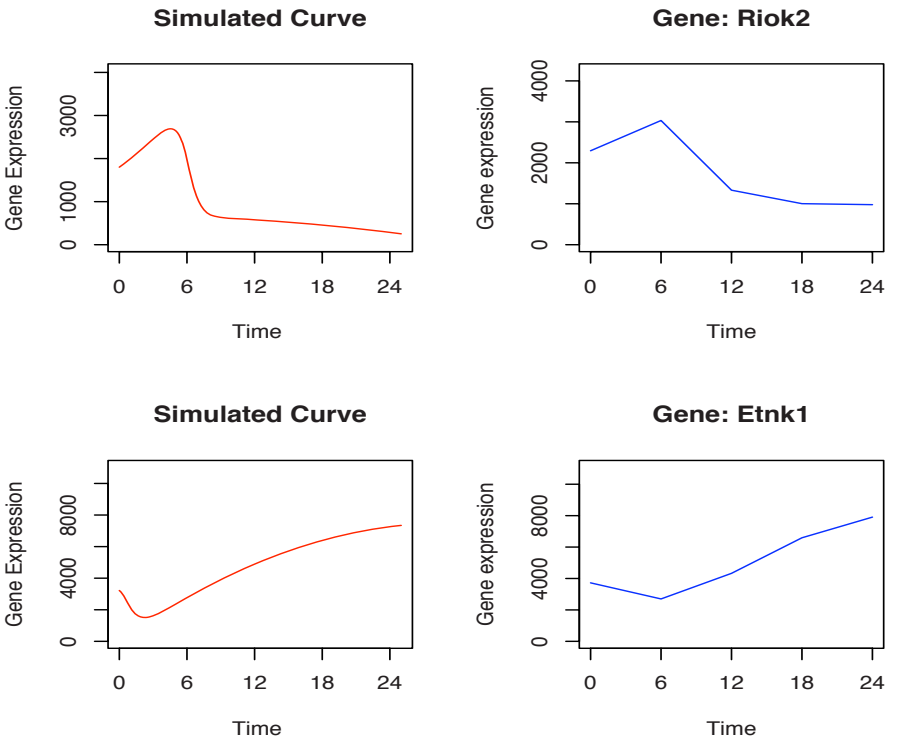

Fig. 2. Simulated curve *vs.* actual gene curve

To implement cluster recognition, we assume the following mixture-density model:

$$P(Y|\Theta,\Omega) = \sum_{c=1}^{C} \omega_c p_c(Y|\theta_c), \tag{8}$$

where the parameter set $\Theta = (\beta_1,\ldots,\beta_C,\alpha_{j1},\ldots,\alpha_{jC},\gamma_1,\ldots,\gamma_C)$ (j=1,2...,p) and $\Omega = (\omega_1,\ldots,\omega_C)$ such that $\sum_{c=1}^{C}\omega_c = 1$ and $p_c(Y|\theta_c)$ is the density function generated by the model (2) defined previously. Given observed samples $Y = (Y_1, Y_2, \ldots, Y_N)$, the log-likelihood expression for the above mixture density is:

$$\log(P(Y|\Theta,\Omega)) = \log(\sum_{c=1}^{C}\omega_c p_c(Y|\theta_c)) = \sum_{i=1}^{N}\log(\sum_{c=1}^{C}\omega_c p_c(Y_i|\theta_c)), \tag{9}$$

where $Y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})$ $(i = 1, 2, \ldots, N)$ represents the profile of gene g_i over the time course $t = 1, 2, \ldots, T$. Since it is hard to optimize (9) by ordinary analytical methods, we will apply the well known expectation maximization(EM) algorithm [13] to estimate these parameters. The details of the EM algorithm approach and the corresponding inference technique is given in the Appendix.

3 Numerical Tests

3.1 Clustering Results

In this section we illustrate the result of clustering gene profiles for V6.5 mouse embryonic stem cells over the time course [0h, 6h, 12h, 18h, 24h],which represents the early period of mESC differentiation. At the first stage, we select as target genes 419 with high differential expression as detected by the SAM method [18]. A standardization procedure is first applied to all gene profiles before running the clustering program. We arbitrarily choose a small number of clusters, namely $C = 3$, in order to facilitate the evaluation of the method, and set the initial cluster proportion to be $\frac{1}{3}$. Thus, we will optimize (9) with respect to $\Theta = (\theta_1, \theta_2, \theta_3)$, $\Omega = (\omega_1, \omega_2, \omega_3)$ and σ^2.

Figure 3 presents the results of clustering the 419 genes. The three clusters clearly have different patterns over the time course. Briefly, cluster one is basically up-regulated compared to cluster three which has a distinct down-regulated pattern. Cluster two contains the largest number of genes. Note that that these clusters have different "compactness" levels. For instance, cluster three is characterized by the fact that the gene functional curves are more closely intertwined while cluster two is more loosely arranged, indicating a larger intra-cluster variation. This effect is captured by the different random gene effect within the three clusters, a feature of the mixture model. In visually assessing the clusters, we must recall that the curves are not clustered by the overall proximity to each other, but by the similarities of their patterns.

We choose the order of autoregressive component to be 1 (i.e. $p = 1$) to avoid the increased noise effect associated with larger order components demonstrated in numerical experiments [11].

To validate the clustering, we characterized the genes in each cluster by compiling Gene Ontology (GO) terms. We also searched the FatiGO server to compare the three clusters. Table 1 displays the distribution of genes featured by different biological processes.

Table 1. Distribution (%) of genes according to biological process

Biological Functions	Cluster one	Cluster two	Cluster three
RNA metabolic	17.14	40.57	42.65
Cellular lipid metabolic	13.33	2.83	1.47
Cellular protein metabolic	45.71	29.25	35.29
Nucleoside, nucleic metabolic metabolic	25.64	50.91	52.94
Regulation of cellular process	23.29	43.64	27.42
Cellular macromolecule metabolic	43.84	28.18	35.59
Protein metabolic	43.84	30.02	38.24
Transcription	18.57	33.33	20.97
Cellular biosynthetic process	21.92	13.76	27.42
Biosynthetic process	24.66	14.16	28.97
N	94	153	85

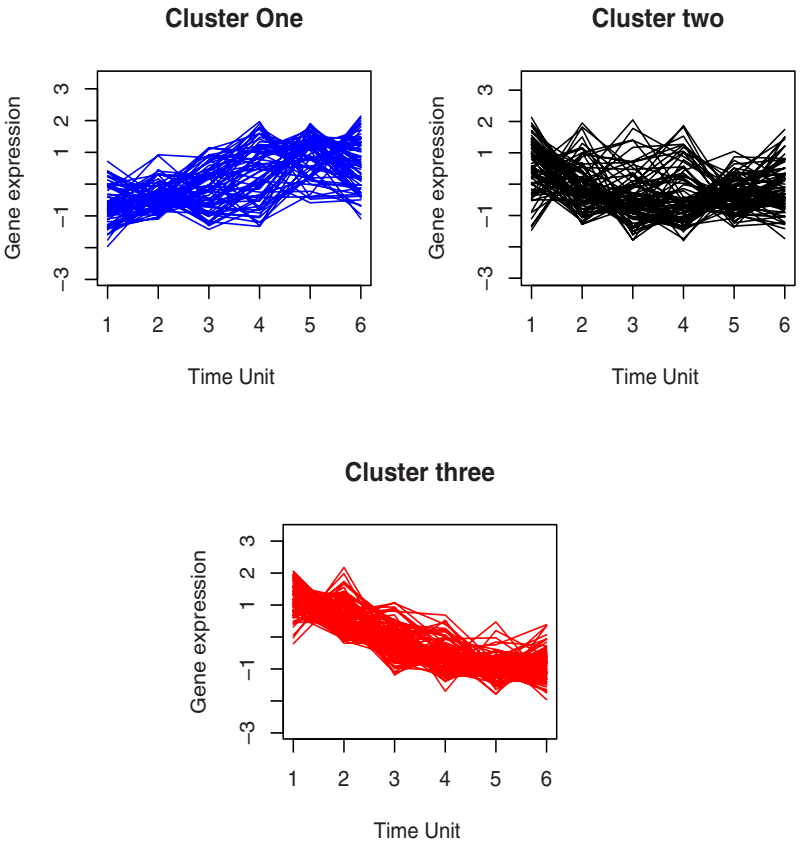

Fig. 3. Clustering result for gene profiles of V6.5 embryonic stem cells

3.2 Biological Analysis

Cluster one includes 119 genes, of which 94 had associated GO terms, largely with protein metabolism, processing of macromolecules and catalytic functions, such as Sc4mol, the Rpl and Rps family (Rpl4 Rpl13 Rpl14 Rpl23 Rpl41 Rps2 Rps6 Rps12 Rps17 Rpl22 Rps27 Rps28 Rps27l), Uba52, Csnk1e, Otx2, Igfbp2 and Gpi1. These genes follow a generally up-regulated trend during ESC differentiation and participate in ESC metabolism or protein synthesis processes. For example, Otx2 is an early stage murine ESC marker playing a cental role in gastrulation, essential for the early specification of the neuroectoderm destined to become fore midbrain [15]. Csnk1e, which undergoes a persistent up-regulation pattern over the time course, performs a catalytic function for serine/threonine kinase activity. The Rpl/Rps family is involved in ribosomal structure and hence protein biosynthesis. Note that there was some evidence, discounted by the authors as well as our results here, that Rpl4 and Rps24 are among the genes constituting a unique molecular signature in human ESC [3].

There are 198 genes in cluster two, 153 with GO annotations. These genes are involved in RNA metabolism, regulation of cellular process and DNA-dependent transcription. Many key ESC markers such as Nanog, Sox21, Zfp57, Cbx7, Vcl, Dnajb6, Msi2h and Cggbp1 fall into this cluster. They generally usually down-regulated for the first 18 or 24 hours and are up-regulated later. By checking the correlation distance, we found these gene profiles are remarkably similar to the expression patterns of undifferentiated ESC markers such as OCT4, Sox2 and Foxd3. From the point of view of stem cell differentiation, these genes are crucial for stem cell self-renewal and maintenance of the inner cell mass (ICM) of the blastocyst. With the loss of cell proliferation and pluripotency, they undergo a significant and persistent down-regulation pattern [6,16].

Many members of cluster two are also related to the cell cycle, cell proliferation or growth. For example, C2ORF29, Cdv3 and Rbbp7 are involved in cell proliferation. Socs3 and Ltbp4 function in the regulation of cell growth. Nmyc1, Cdk2ap1, Nipbl, Cdc34, Mcm5, Nipbl, BC068171 and Ccne1 are cell-cycle related genes. They mediate the progression through the cell cycle, particularly the G1/S transition of the mitotic cell. That these genes are in cluster two agrees with theories [12], whereby the cells initially derived from a population of stem cells undergo rapid cell division during early differentiation; throughout this period, expression of genes keeping control of the cell cycle or proliferation should be attenuated.

Another 102 genes are found in cluster three, 85 with GO annotations. Compared with cluster one, cluster three is enriched for genes in charge of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process, such as Ctbp2, Cdk2ap1, Apex1, Hmgb2, Mybbp1a Ran, Gars, H2afz, Sod2, Nola2, Mki67ip, Etv4, Atp5l, Bzw1, Nr0b1, Klf5, Rbm14, Polr2f, Ankrd17, Lsm3, Zfp36l1, Ddx5, Sfrs1, Rbm3, Cars and Psmc5. This cluster is also associated with microtubule-based movement, signal transduction and biosynthetic processes. These include Arpc4, Lefty1, Ptprf, Stmn1, Cfl1, Trh and Eif4ebp1. Lefty1 is an important "stemness" marker expressed in the left half of gastrulating mouse embryos and involved in the TGF-beta signaling pathway. Ptprf has an intrinsic protein tyrosine phosphatase activity (PTPase) and plays a role in cell adhesion receptor and the insulin signaling pathway. Ptprf has also been identified in human ESC studies. Stmn1 is involved in signal transducing, participates in the regulation of the microtubule (MT) filament and promotes disassembly of microtubules. Trh plays a role in cell-cell signaling and neuroactive ligand-receptor interaction. Arpc4 is related to cytoskeleton and actin filament polymerization.

Since clusters two and three have a similar down-regulated pattern, especially in the first 12 hour period, it is not surprising that some of their members have similar gene functions in signalling, ATP/GTP binding, actin binding and microtubules. For example, the genes Pfn1, Wdr1, Efna2, Lefty2, Ak7, Ptch1, Riok2, Arf6 and Actg1 are classified into cluster two. This is not surprising; maintenance of the pluripotent state of ESC requires intrinsic signalling as well as extrinsic environmental elements, involving microtubules. It is known that environmental factors such as cell-surface receptors and cytokines play an important

Table 2. Over-represented GO terms for each cluster from GOstat

Cluster	Biological Function	In cluster	In genome	P-value
I	primary metabolic process	60	5694	0.000107
	cellular protein metabolic process	34	2421	7.44E-05
	structural constituent of ribosome	14	147	4.41E-10
Total		94	14456	
II	gene expression	57	2452	4.21E-08
	RNA metabolic process	46	2082	1.59E-05
	transcription, DNA-dependent	35	1704	0.00215
	ribonucleoprotein complex biogenesis	11	174	0.000302
	translational initiation	5	40	0.00266
	glycolysis	5	42	0.00326
Total		153	14456	
III	gene expression	33	2452	2.02E-05
	translation	13	369	2.50E-05
	amino acid,derivative metabolic process	9	264	0.00111
	ribonucleoprotein complex	10	382	0.00267
Total		85	14456	

role in the maintenance of stem cell functions [12]. Transcription profiles of this type generally take a persistent down-regulated pattern during differentiation. Examples include the actin binding related gene Ptprf and receptor activity-associated gene Tagln, which decrease sharply throughout the first 24 hours in ESC differentiation.

Table 2 uses a t-test to evaluate the over-representation of various kinds of GO terms in each cluster. The analysis in Table 2 confirms and deepens that of Table 1. For example, the high value for RNA metabolism in cluster 2 possibly results from the selection of genes involved in transcriptional regulation. The higher proportion of cellular protein metabolic process in cluster 1, amino acid derivative metabolic processes and ribonucleoprotein complex, in cluster 3, are all closely related to protein biosynthesis. This similarity possibly originates from genes involved in translation process(cluster 3). We also found interesting to see five glycolitic enzymes in cluster 2, and a number of genes involved in the formation of RNA-protein complexes in clusters 2 and 3. In the case of cluster 2, this includes five genes involved in translation initiation.

References

1. Aach, J., Church, G.M.: Aligning gene expression time series with time warping algorithms. Bioinformatics 17, 495–508 (2001)
2. Bar-Joseph, Z., Gerber, G., Gifford, D., Jaakkola, T., Simon, I.: Continuous representations of time-series gene expression data. Journal of Computational Biology 10, 341–356 (2003)

3. Bhattacharya, B., Miura, T., Brandenberger, R., Mejido, J., Luo, Y., Yang, A.X., Joshi, B.H., Ginis, I., Thies, R.S., Amit, M., Lyons, I., Condie, B.G., Itskovitz-Eldor, J., Rao, M.S., Puri, R.K.: Gene expression in human embryonic stem cell lines: unique molecular signature. Blood 103, 2956–2964 (2004)
4. Bilmes, J.A.: A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report 97-021. International Computer Science Institute, Berkeley, CA (1997)
5. Brumback, B.A., Rice, J.: Smoothing spline models for the analysis of nested and crossed samples of curves. Journal of the American Statististical Association 93, 961–976 (1998)
6. Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., Smith, A.: Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell 113, 643–655 (2003)
7. Chen, T., He, H.L., Church, G.M.: Modeling gene expression with differential equations. In: Pacific Symposium on Biocomputing, pp. 29–40 (1999)
8. Chudova, D., Hart, C., Mjolsness, E., Smyth, P.: Gene expression clustering with functional mixture models. In: Advances in Neural Information Processing, vol. 16. MIT Press, Cambridge (2004)
9. Fraley, C., Raftery, A.E.: How many clusters? Which clustering method? Answers via model-based cluster analysis. Computer Journal 41, 578–588 (1998)
10. Hailesellasse Sene, K., Porter, C.J., Palidwor, G., Perez- Iratxeta, C., Muro, E.M., Campbell, P.A., Rudnicki, M.A., Andrade-Navarro, M.A.: Gene function in mouse embryonic stem cell differentiation. BMC Genomics 8, 85 (2007)
11. Luan, Y., Li, H.: Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics 19, 474–482 (2003)
12. Martinez Arias, A., Stewart, A.: Molecular Principles of Animal Development. Oxford University Press, NY (2002)
13. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions. John Wiley and Sons, New York (1997)
14. Medvedovic, M., Yeung, K.Y., Bumgarner, R.E.: Bayesian mixture model based clustering of replicated microarray data. Bioinformatics 20, 1222–1232 (2004)
15. Morsli, H., Tuorto, F., Choo, D., Postiglione, M.P., Simeone, A., Wu, D.K.: Otx1 and Otx2 activities are required for the normal development of the mouse inner ear. Development 126, 2335–2343 (1999)
16. Niwa, H., Burdon, T., Chambers, I., Smith, A.: Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. Genes and Development 12, 2048–2060 (1998)
17. Ramoni, M.F., Sebastiani, P., Kohane, I.S.: Cluster analysis of gene expression dynamics. Proceedings of the National Academy of Sciences USA 99, 9121–9126 (2002)
18. Tusher, V.G., Tibshirani, R., Chu, G.: Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences USA 98, 5116–5121 (2001)
19. Wu, F., Zhang, W.J., Kusalik, A.J.: Dynamic model-based clustering for time-course gene expression data. Journal of Bioinformatics and Computational Biology 3, 821–836 (2005)

A Appendix: EM Algorithm of Mixture Model

A.1 Inference of the Log-Likelihood Function for the Mixture Model

For convenience,we only consider the first order autoregressive item (i.e. $p = 1$ in model (2)). According to model (2), we infer that if the gene g_i belongs to cluster c then its expression profile $Y_i = (y_{i1}, \ldots, y_{iT})$ at time point t satisfies the following conditional probability model:

$$P(y_{it}|\theta_c, Y_{i(t-1)})) = \text{Normal}(f^{(c)}(t|\Phi^{(c)}) + \alpha_c y_{i(t-1)}), {\gamma_c}^2 + \varepsilon^2) \tag{10}$$

Thus, the log-likelihood for Y_i over time course $t = 1, 2, \ldots, T$ is

$$L_c(Y_i|\theta_c) = \log[P(y_{i1}|\theta_c) \prod_{j=2}^{T} P(y_{ij}|\theta_c, y_{i(j-1)})] \tag{11}$$

From expression (10), (11) can be finally expressed as

$$L_c(Y_i|\theta_c) = C_0 - T Log(\sigma_i) - \frac{1}{2\sigma_i}[(y_{i1} - f(1|\beta_c) - \mu_0)^2 + \sum_{j=2}^{T}(y_{ij} - f(j|\beta_c) - \alpha_{(c)} y_{i(j-1)})^2] \tag{12}$$

where the variance of Y_i, $\sigma_i = \sqrt{\gamma_c^2 + \varepsilon^2}$, C_0 is a constant and μ_0 represents the mean at the initial time point. The log-likelihood of the mixture model of total observed samples can be expressed as

$$L(Y|\Theta, \Omega) = \sum_{i=1}^{N} \log(\sum_{j=1}^{C} \omega_j P_j(Y_i|\theta_j)) \tag{13}$$

where $P_j(Y_i|\theta_j)$ is identified by $\exp(L_j(Y_i|\theta_j))$, and $\omega_j, (j = 1, \ldots, C)$ are the unknown cluster membership parameters such that $\sum_{j=1}^{C} \omega_j = 1$.

A.2 EM Algorithm Approach

The optimization for (12) can be carried out by using an EM algorithm. To see the specific computational steps, the reader can refer to [4].

1. Set initial parameters $\Theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_C^{(0)})$ for the given number of clusters C and cluster proportion $\Omega^{(0)} = (\omega_1^{(0)}, \ldots, \omega_C^{(0)})$.
2. Define cluster membership (a random variable) l, where $l \in \{1, 2, \ldots, C\}$ and $l = c$ if gene g_i belongs to cluster c. For $k = 0, 1, 2, \ldots$ repeat the following iterative steps until a given threshold is reached.
3. E-step: Calculate the posterior of the cluster membership l given Y_i, $\Theta^{(k)}$ and $\Omega^{(k)}$ at the k^{th} iterative step from the following procedure:

$$P(l|Y_i, \Theta^{(k)}, \Omega^{(k)}) = \frac{\omega_l^{(k)} P_l(Y_i|\theta_l^{(k)})}{\sum_{j=1}^{C} \omega_j^{(k)} P_j(Y_i|\theta_j^{(k)})} \tag{14}$$

4. M-step: Maximize the expected posterior log-likelihood with respect to cluster membership parameters Ω and model parameters Θ given observed Y and current parameter estimates:

$$\omega_l^{(k+1)} = \frac{1}{N} \sum_{i=1}^{N} P(l|Y_i, \Theta^{(k)}, \Omega^{(k)}) \tag{15}$$

and

$$\text{Maximize } F = \sum_{l=1}^{C} \sum_{i=1}^{N} [L_l(Y_i|\theta_l^{(k+1)}) P(l|Y_i, \Theta^{(k)}, \Omega^{(k)})] \tag{16}$$

where (16) can be maximized by non-linear optimization techniques such as BFGS or the conjugate gradients method. Note that solving (16) is much easier than directly optimizing (9).