Mining gene expression data using a novel approach based on hidden Markov models

Xinglai Ji[a], Jesse Li-Ling[b], Zhirong Sun[a,*]

[a] *Institute of Bioinformatics, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing 100084, PR China*
[b] *Department of Medical Genetics, China Medical University, Shenyang 110001, PR China*

Received 15 January 2003; revised 28 March 2003; accepted 31 March 2003

First published online 14 April 2003

Edited by Robert B. Russell

Abstract In this work we have developed a new framework for microarray gene expression data analysis. This framework is based on hidden Markov models. We have benchmarked the performance of this probability model-based clustering algorithm on several gene expression datasets for which external evaluation criteria were available. The results showed that this approach could produce clusters of quality comparable to two prevalent clustering algorithms, but with the major advantage of determining the number of clusters. We have also applied this algorithm to analyze published data of yeast cell cycle gene expression and found it able to successfully dig out biologically meaningful gene groups. In addition, this algorithm can also find correlation between different functional groups and distinguish between function genes and regulation genes, which is helpful to construct a network describing particular biological associations. Currently, this method is limited to time series data. Supplementary materials are available at http://www.bioinfo.tsinghua.edu.cn/~rich/hmmgep_supp/.

Key words: Hidden Markov model; Gene expression data; Cluster analysis; Yeast Cell cycle

1. Introduction

Advances in microarray technology have enabled us to simultaneously measure the expression of thousands of genes under multiple experimental conditions [1–4]. This has led to an explosion of gene expression data and a great need for development of methodology to analyze and exploit such information. A major step in the analysis is the detection of gene groups with similar expression patterns that may be suggestive of associated biological functions [2]. Because of the large number of genes and complexity of biological networks, clustering has been a useful exploratory technique for the analysis of gene expression data.

A wide range of clustering algorithms have been proposed for the analysis of gene expression data. Examples include hierarchical clustering [2], self-organizing maps (SOM) [5,6], k-means [7], graph-theoretic approaches [8,9] and support vector machines [10]. Although success has been reported for application of many such clustering approaches, most of the proposed algorithms are largely heuristically motivated, and the issues of determining the 'correct' number of clusters and choosing a 'good' clustering algorithm have not yet been rigorously addressed. Moreover, such clustering solutions often partition the studied genes into disjoint sets, which implies that each gene has been associated with a single biological function or process which, however, may have oversimplified problems of biological systems. Furthermore, depending on the distances (such as Euclidean distance and Pearson correlation coefficient) obtained from pair-wise comparison of gene expression patterns, some methods could confuse the clusters because average distances may conceal the specificity of genes with multiple functions [11,12].

In this report, we propose a different strategy based on the hidden Markov models (HMM) [13] for the analysis of gene expression data. Particularly, this model-based approach assumes that each gene expression profile has been generated by a Markov chain with certain probability. Therefore, this method is currently limited to time series data. During the last several years, statistical methods of Markov source or hidden Markov modeling have become increasingly popular, partly because HMMs are very rich in mathematical structure, and hence can form the theoretical basis for a wide range of bioinformatics applications [14–19]. HMMs have been successfully applied to partition protein subfamilies [16]. In this paper, we clustered the gene expression data using a similar algorithm. To assess the clustering power of our algorithm, we compared it with two prevalent clustering algorithms (SOM and k-means) on three gene expression datasets [1–3] with internal criterion analysis [20] and on two datasets [2,3] with external criterion analysis [20] (see Section 2.3). The results show that our algorithm is comparable to, if not better than, those two algorithms. In addition, our methodology can determine the 'correct' number of clusters. Furthermore, we have successfully applied this approach to analyze yeast cell cycle gene expression data by Spellman et al. [4]. The results were compared with previous work [4] and show that our algorithm could discover the correlation between different functional groups, as well as biologically meaningful gene groups. These correlations could be used for finding out genes with multiple functions and distinguishing between function genes and regulation genes, which is helpful for reconstructing genetic networks later.

*Corresponding author. Fax: (86)-10-62772237.
E-mail address: sunzhr@mail.tsinghua.edu.cn (Z. Sun).

Abbreviations: HMM, hidden Markov model; EM, expectation maximization; SOM, self-organizing maps; FOM, figure of merit

0014-5793/03/$22.00 © 2003 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)00363-6

2. Materials and methods

2.1. Gene expression datasets

We used three gene expression datasets to compare the performance of different clustering algorithms, for two of which external evaluation criteria (see Section 2.3) were available. In this paper, we use the term 'class' to refer to a group in the external criterion, and the term 'cluster' to refer to a group of genes obtained by a clustering algorithm. Our first application was on a set of gene expression data measuring the response of human fibroblasts to serum [1]. We used a subset of 517 genes (18 time points) whose expression changed substantially in response to serum. No external criteria were available for this dataset, but according to Xu et al.'s work [9], the optimal number of clusters for this dataset is five. The second dataset was for the budding yeast *Saccharomyces cerevisiae* [2], with each gene having 18 time points. We selected four classes (74 genes in total) determined in previous work [2]. Genes in each of these four classes shared similar expression patterns and were annotated to be in the same biological pathway. The goal of this application was to compare our clustering results with known class information. The third dataset showed the fluctuation of expression levels of approximately 6000 yeast genes over two cell cycles (17 time points) [3]. We used the subset (the five-phase criterion) consisting of 384 genes whose expression levels peaked at different time points corresponding to the five phases of the cell cycle [3]. We expected clustering results to approximate this five-class partition.

To demonstrate the effectiveness of our approach, we also analyzed the yeast cell cycle dataset by Spellman et al. [4]. This dataset recorded the fluctuation of expression levels of approximately 6000 genes at 18 time points. The MATa strain was grown to logarithmic (asynchronous) growth and then arrested at the G1 phase by addition of α factor. After 2 h, cells were shifted into medium containing no α factor, and samples were taken every 7 min for 119 min. The control sample was prepared from asynchronously growing cells of the same strain in the same medium but without α factor treatment. Relative transcript abundances were determined at each time point, by hybridization to a cDNA microarray. Eight hundred genes have been identified that meet an objective minimum criterion for cell cycle regulation by periodicity and correlation algorithms [4], from which we selected 613 genes without missing values.

2.2. HMMs for clustering gene expression data

Each gene expression profile was firstly normalized to have mean 0 and standard deviation 1 (data with any missing values were filtered), and then transformed into a sequence of expression fluctuation following Eq. 1, where N is the number of time points, E is the gene expression level at each time point, S is the transformed value of the sequence and a is defined as a tolerance factor (set as 0.05 here). Thus, an N-time-point gene expression profile produced by a cDNA microarray experiment was transformed into a $(N-1)$-time-point sequence of expression fluctuation consisting of a character set {0, 1, 2}. Each character in the sequence describes how the expression level has changed, or remained unchanged at the next time point, whilst the whole sequence represents the fluctuation of gene expression. Although some experimental information was lost after this transformation, it was convenient for the following HMM clustering analysis and had little influence on the identification of co-regulated genes.

$$S_i = \begin{cases} 0 & \text{if } |E_i - E_{i+1}| < a \\ 1 & \text{if } E_{i+1} - E_i \geq a \\ 2 & \text{if } E_i - E_{i+1} \geq a \end{cases} \quad 1 \leq i \leq N-1 \tag{1}$$

A simple HMM was constructed for the expression fluctuation sequences. The main line of the HMM comprised a sequence of N states, with each corresponding to an actual cell state during the cell cycle. Each of these states could generate a character (0, 1 or 2, see Eq. 1) according to a distribution representing the regulation trend at this cell state. For convenience, we added a dummy 'BEGIN' state and a dummy 'END' state, which did not produce any observation. Thus, a gene expression fluctuation sequence could be generated by a 'random walk' through the model as follows: commencing at state 'BEGIN', choose a transition to another cell state and generate the character (0, 1 or 2) based on the distribution at this state. Then choose a transition to the next state and generate the next character. Continuing in this manner all the way to the 'END' state, following the series of time points through the model could generate a sequence of expression fluctuation. The model was then trained with the Baum–Welch method [13] (see Fig. 1) and the probability of a sequence given the HMM was calculated with a forward–backward algorithm [13].

In order to automatically partition the large sequence data set into w clusters of similar sequences, we made w copies of the HMM, one for each cluster. We called these components HMMs. At present, the number of clusters w is determined empirically. The initial lengths of the models were equal to the dimension of gene expression data and the parameters of the models were randomly initialized. The components HMM were separately trained with the expectation maximization (EM) algorithm on the gene expression data. The EM re-estimation of a component HMM was the same as the re-estimation of a single model, except that the weight that a sequence had in the re-estimation of a component was proportional to the probability of the sequence given by its component model (see Fig. 1). Thus, sequences that had a higher probability for a particular component HMM had a greater influence on re-estimating the parameters of that component, and this caused the parameters of that component to change in such a way that the component further 'specialized' in modeling those sequences. In this manner, the individual components evolved through training to represent clusters in the trained sequences. This way of using EM is called mixture modeling in statistics [22,23], and is known as '(soft) competitive learning' in the neural network literature [24]. When the models were trained, the probabilities of a sequence given by any of the component HMMs could be calculated, and these probabilities were then used for determining which cluster the sequence belonged to.

2.3. Validating clustering

According to Jain and Dubes [20], cluster validation refers to procedures that evaluate the results of cluster analysis in a quantitative and objective fashion. In the statistics literature, cluster validation procedures are divided into two main categories: external and internal criterion analysis [20].

The external criterion analysis validates a clustering result by comparing the clustering result to a given 'gold standard' that is another partition of the object and can be obtained by an independent process based on information other than the given dataset. Here, the idea was to compare clustering results by different algorithms to a known functional categorization of the genes. We used the Rand index [25]:

$$\text{Rand index} = \frac{a+d}{a+b+c+d} \tag{2}$$

where a is the number of pairs of objects in the same class in both partitions, b and c are the numbers of pairs of objects in only U or only V (suppose that U is our external criterion and V is a clustering result), and d is the number of pairs of objects in different classes in both partitions. The Rand index lies between 0 and 1. When the two partitions agree perfectly, the Rand index is 1.

The internal criterion analysis uses information from within the given dataset to represent the goodness of fit between the input dataset and the clustering results. Intuitively, genes within the same clusters are expected to have similar expression levels. Moreover, disjoint clusters are expected to be relatively far apart from each other. Therefore, we can define the ratio figure of merit (FOM) [26] to be the ratio of the within-cluster dispersion to the between-cluster separation. The ratio FOM can be written as:

$$\text{FOM}_{\text{ratio}}(e,k) = \frac{\frac{1}{n}\sum_{i=1}^{k}\sum_{x\in C_i} |R(x,e) - \mu C_i(e)|}{\frac{1}{k-1}\left(\mu_{C_i}^{\max}(e) - \mu_{C_i}^{\min}(e)\right)} \tag{3}$$

where n is the number of genes, k is the number of clusters, $R(x,e)$ is the expression level of gene x under condition e and $\mu C_i(e)$ is the average expression level in condition e of genes in cluster C_i.

3. Results and discussion

We compared the performance of various clustering algorithms (k-means, SOM and HMM) on the datasets [1–3] described in Section 2.1. We also analyzed the yeast cell cycle

$$\overline{\pi}_i = \sum_{l=1}^{L} \frac{w^l \alpha_1^l(i)\beta_1^l(i)}{P(O^l|\lambda)} \qquad 1 \le i \le N$$

$$1 = \sum_{l=1}^{L} w^l, w^l \propto P(O^l|\lambda)$$

$$\overline{a}_{ij} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T_l-1} w^l \alpha_t^l(i) a_{ij} b_j(O_{t+1}^l)\beta_{t+1}^l(j)/P(O^l|\lambda)}{\sum_{l=1}^{L}\sum_{t=1}^{T_l-1} w^l \alpha_t^l(i)\beta_t^l(i)/P(O^l|\lambda)} \qquad 1 \le i, j \le N$$

$$\overline{b}_{jk} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T_l} w^l \alpha_t^l(j)\beta_t^l(j)/P(O^l|\lambda)}{\sum_{l=1}^{L}\sum_{t=1}^{T_l} w^l \alpha_t^l(j)\beta_t^l(j)/P(O^l|\lambda)} \qquad O_t = V_k \qquad 1 \le j \le N, 1 \le k \le M$$

Fig. 1. Re-estimation of HMMs. The elements of an HMM (λ) include: N, the number of cell states; M, the number of distinct observation symbols per state (see Eq. 1, the individual symbols denoted as V = {0, 1, 2}); $\{a_{ij}\}$, the state transition probability distribution; $\{b_{jk}\}$, the observation symbol probability distribution in state j; and $\{\pi_i\}$, the initial state distribution. Supposing that there are L gene expression profiles $\{O^l\}$, the re-estimation formula for automatically partitioning these L genes to w clusters is as shown above, where $P(O^l|\lambda)$ is the probability of gene l given by HMM (λ), w^l is the weight gene l has in the re-estimation of HMM (w), $\alpha_t(i)$ is the forward variable and $\beta_t(i)$ is the backward variable.

dataset by Spellman et al. [4] using our algorithm. In our experiments, the k-means and SOM algorithms with random initialization were run 20 times to obtain reliable results, and the HMM algorithm was run 100 times. The number of clusters was varied from two to 50. The standard deviations of the measurements (ratio FOM and Rand index) were under 10% of the average values.

3.1. Independent assessment of clusters

Fig. 2 shows the ratio FOM of the three algorithms over a range of different numbers of clusters on the three gene expression datasets and the Rand index on two of the three expression datasets for which external criteria were available. We can see from the figure that similar clustering results were achieved. This indicates that our methodology could partition the gene expression data with similar quality as k-means and SOM algorithms.

All three methods produced biologically meaningful gene groups consistent with previous knowledge. A higher Rand index means a stronger correspondence to the gold standard. Fig. 2d,e shows a significant improvement of the clustering quality when the cluster number increased to a certain value, and after that, the rate of improvement dropped. This number should be the 'optimal' or 'correct' or 'meaningful' number of clusters for the gene expression data. For the gene expression

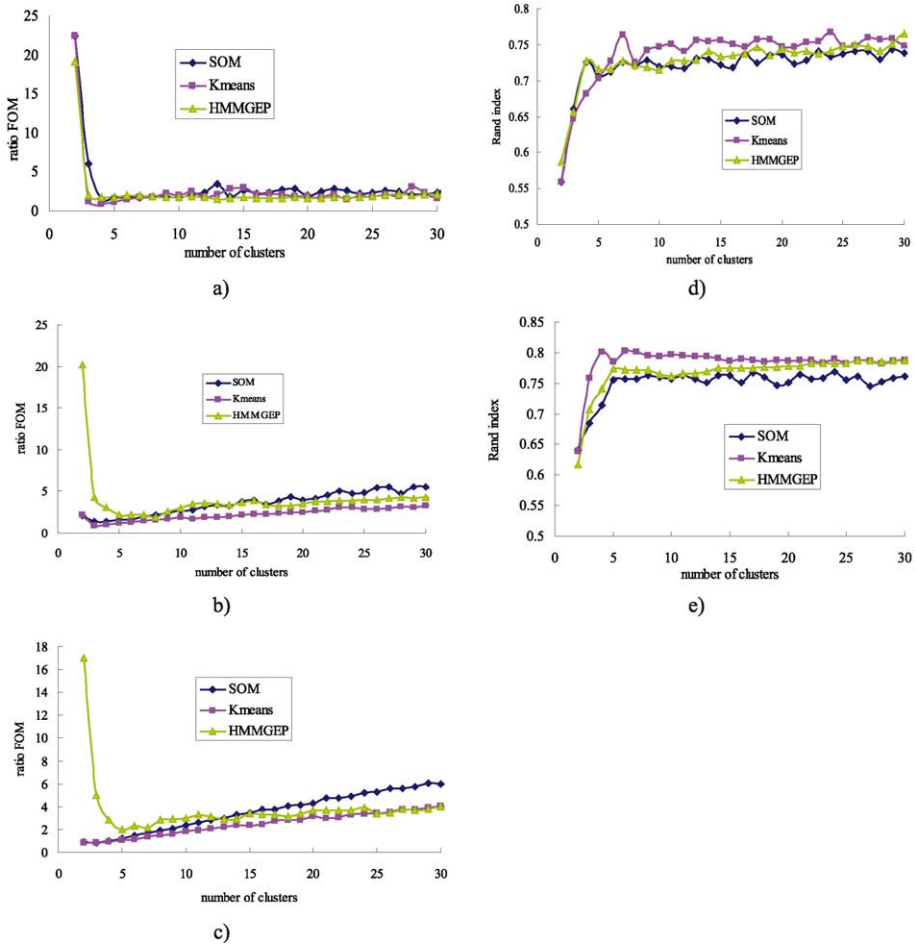

Fig. 2. Independent assessment of clusters. a: Ratio FOM for budding yeast data (four classes). b: Ratio FOM for five-phase-criterion yeast cell cycle data. c: Ratio FOM for gene expression data measuring the response of human fibroblasts to serum. d: Rand index for budding yeast data. e: Rand index for five-phase-criterion yeast cell cycle data.

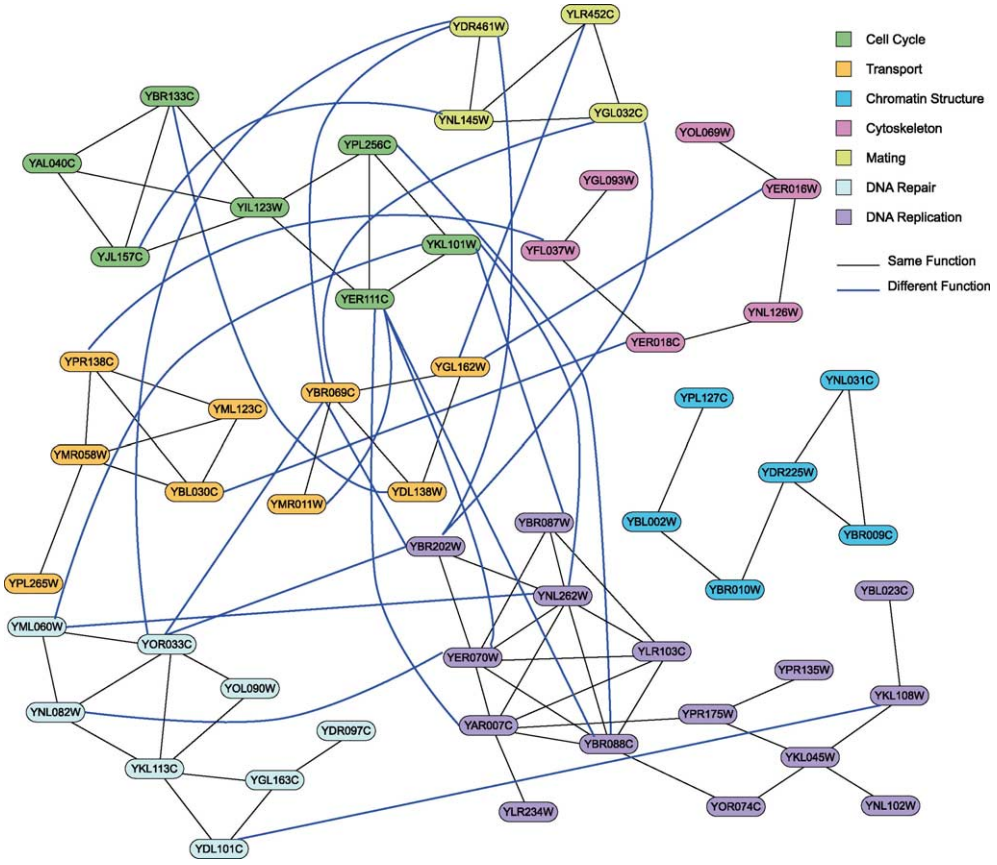

Fig. 3. The functional gene groups visualized by a Java applet [21]. Each gene is labeled in a rounded rectangle. Black lines between each pair of genes denote that the pair clustered more than 40 times over 100 calculations, and is thus probably involved in the same biological process. The blue lines between each pair of genes indicate involvement in different processes for more than 60 times. The legend in the top right describes the cellular processes in which the genes were involved (see Table 1).

data in the budding yeast *S. cerevisiae*, Fig. 2d suggests that the most 'natural' number is around 4, which is in agreement with the annotated results in previous work [2]. For the five-phase-criterion yeast cell cycle data, Fig. 2e suggests that the number is around 5, corresponding to the five phases of the cell cycle [3]. In addition, the Rand index at the 'correct' number of clusters was about 0.75, which was significant enough to demonstrate the effectiveness of our approach to successfully cluster gene expression data.

3.2. Determining the 'optimal' number of clusters

Although the clustering results were highly consistent with the external criterion, we usually had no external information about the gene expression data for clustering, and had to determine the 'optimal' number of clusters from the results. To choose from various numbers of clusters, we had to use information from within the given dataset. The ratio FOM was such a measurement.

A lower ratio FOM means a higher goodness of fit between the input dataset and the clustering results. When the number of clusters increases, the ratio FOM should significantly decrease to a valley and then fluctuate around the value. The valley should indicate the most 'natural' number of clusters. We can see from Fig. 2a–c the significant improvement of the clustering quality when the cluster number increases to a certain point, and after that, the rate of improvement drops. Fig. 2a suggests the most 'meaningful' number is around 4 and Fig. 2b suggests the number is around 5; both numbers are in agreement with the gold standard. Although no external criterion was available for the gene expression data for the response of human fibroblasts to serum, Fig. 2c suggests five-clustering results, which is consistent with the work by Xu et al. [9].

The agreement between external and internal criterion analyses suggests that the ratio FOM could be used as a reference to determine the optimal number of clusters without external or previous biological knowledge. From Fig. 2, we could also discover that the k-means and SOM methods are not as sensitive as our algorithm to the number of clusters. These results suggest that our methodology is comparable to, if not better than, the two prevalent clustering algorithms, but with the key advantage of determining the number of clusters.

3.3. Digging out co-regulated gene groups

We analyzed the yeast cell cycle dataset by Spellman et al. [4] using our HMM algorithm. In this study, we performed 100 calculations for between two and 50 clusters and found significance with a cluster number of 25 using adjusted ratio FOM measurement (the clustering results are available at the supplementary web site http://www.bioinfo.tsinghua.edu.cn/~rich/hmmgep_supp/). Since the component HMMs were initiated by randomly assigning values to the elements of the model, each calculation could potentially generate very different clusters. For statistical significance, we had to reconstruct reliable functional groups. We set 40 as the cutoff, which means that every two genes that cluster more than

Table 1
Functional gene groups illustrated in Fig. 3

No.	Open reading frame	Process	Function
1	YAL040C	cell cycle	G1/S cyclin
2	YJL157C		Cdc28p kinase inhibitor
3	YBR133C		Swe1p (kinase) regulator
4	YIL123W		cyclin
5	YER111C		transcription factor
6	YPL256C		G1/S cyclin
7	YKL101W		negative regulator of swe1 kinase
8	YBL030C	transport	mitochondrial ADP/ATP translocator
9	YPR138C		ammonia permease
10	YML123C		inorganic phosphate permease
11	YMR058W		cell surface ferroxidase
12	YPL265W		dicarboxylic amino acid permease
13	YBR009C	chromatin structure	histone H4
14	YNL031C		histone H3
15	YDR225W		histone H2A
16	YBR010W		histone H3
17	YBL002W		histone H2B
18	YPL127C		histone H1
19	YOL069W	cytoskeleton	spindle pole body component
20	YER016W		microtubule binding protein
21	YNL126W		spindle pole body component
22	YER018C		spindle pole body component
23	YFL037W		β-tubulin
24	YGL093W		spindle pole body component
25	YBR069C	transport	amino acid permease
26	YGL162W		hypoxic gene family (sterol uptake)
27	YMR011W		hexose permease
28	YDL138W		glucose permease
29	YDR461W	mating	a-factor precursor
30	YGL032C		a-agglutinin binding subunit
31	YNL145W		a-factor precursor
32	YLR452C		negative regulator of Gpa1
33	YOR033C	DNA repair	exonuclease; also recombination
34	YML060W		8-oxoguanine DNA glycosylase
35	YNL082W		MutL homolog; mismatch repair
36	YKL113C		ssDNA endonuclease
37	YOL090W		MutS homolog; mismatch repair
38	YGL163C		DNA-dependent ATPase
39	YDL101C		DNA damage-responsive protein kinase
40	YDR097C		MutS homolog; mismatch repair
41	YBR087W	DNA replication	DNA polymerase processivity factor
42	YLR103C		pre-replicative complex subunit (putative)
43	YER070W		ribonucleotide reductase
44	YNL262W		polymerase ε catalytic subunit
45	YBR088C		DNA polymerase processivity factor
46	YBR202W		MCM initiator complex
47	YAR007C		replication factor A, 69-kDa subunit
48	YLR234W		DNA topoisomerase III
49	YOR074C		thymidylate synthase
50	YKL045W		polymerase α 58-kDa subunit (DNA primase)
51	YNL102W		polymerase α 180-kDa subunit
52	YKL108W		unknown; interacts with Dpb11p
53	YBL023C		MCM initiator complex
54	YPR175W		polymerase ε 80-kDa subunit
55	YPR135W		polymerase α binding protein

40 times over the 100 clustering calculations are reserved and those that cluster for fewer than 40 times are excluded (the reason why 40 is reliable and chosen as the threshold is available at the supplementary web site). Indeed, this path is not completely unexplored. There has been some research on deducing classification and even network from correlations between genes [27,28] (the comparison is available at the supplementary web site). For convenience of visualization and analysis (see Fig. 3), the gene groups were visualized with a Java applet developed by Mrowka [21], which was originally used for visualizing protein–protein interactions. For convenience and simplicity, we present only the partial genes involved in the same cellular processes, and the pairs of genes with different cellular processes or with unknown processes are excluded. Fig. 3 shows the selected genes after filtration and simplification. In Fig. 3, the genes involved in the same cellular process are connected with black lines. Table 1 lists the functions of these genes and the biological processes they participate in.

We can see from Fig. 3 that our algorithm has the ability to dig out biologically meaningful gene groups. An examination of the clusters in the previous work [4] provides some interesting results (the details are available at the supplementary web site). The group for cell cycle contains seven genes that are involved in cell cycle regulation. Most of these genes reach peak expression in the G1 or M phase. Although YJL157C is marked for mating type and YBR133C reaches peak expression in the G2 phase, they are both related to Cdc28p, which may be the reason why they are clustered here. The precise function of YIL123W is unknown, but according to its position in the group, it may be involved in cell cycle regulation. According to their functions, this group may be placed in the CLN2 cluster by Spellman et al.'s result [4]. The group for transport contains nine genes, which can be further divided into two groups. They reach peak expression in the M phase, which reflects a surge of molecular transport during cell division (before and after the M phase). YBL030C may be involved in transport for mitochondria. The group for chromatin structure contains six genes, which are all histones. They reach peak expression in the S phase, a most important stage in the cell cycle when DNA replication and histone synthesis synchronize to form the nucleosome and the genome replicates. All six genes were grouped into the histone cluster by Spellman et al. The group for cytoskeleton contains six genes. The exact functions of YER018C and YGL093W are unknown, but we may suppose that they are involved in spindle pole body formation and are required for mitosis and karyog-

amy. The group for mating contains four genes, most of which reach peak expression in the M/G1 phase. Except YGL032C, the others were grouped into the MAT cluster by Spellman et al. Nevertheless, YNR044W, which is very similar to YGL032C, was in the MAT group. The group for DNA repair contains eight genes, which reach peak expression in the G1 phase. Many genes in this cluster were grouped into the CLN2 cluster by Spellman et al. The group for DNA replication contains 15 genes, most of which reach peak expression in the G1 phase. The function of YKL108W is unknown, but it interacts with DNA replication polymerase. Most of these genes were grouped into the CLN2 cluster by Spellman et al.

3.4. Correlation between gene groups and the network

As shown in Fig. 3, many meaningful gene groups were dug out, which shows that our algorithm could successfully discover biological information from gene expression data. We also found that the genes in different functional groups had a probability to cluster together. In Fig. 3, blue lines are appended and used for connecting the genes involved in different cellular process and clustering together more than 60 times over 100 calculations.

The correlations between different gene groups were very interesting and revealed the complexity of organisms. Genes can have more than one function and be involved in multiple biological processes, therefore partitioning genes into disjoint sets may be an oversimplification of the biological system. For example, the genes involved in cell cycle are correlated with the genes involved in many other processes, such as transport, mating, DNA repair and replication. These correlations construct a network (see Fig. 3) and the genes involved in cell cycle are located in the center.

The correlations between different gene groups would provide helpful information for the reconstruction of the genetic networks. We know that the genetic network is composed of regulation genes and function genes. It is very important to distinguish between these two types of genes while reconstructing the genetic network. We can see from Fig. 3 that there are fewer correlations between function gene groups and other groups than between regulation gene groups and other groups. It is obvious that the more complex process the genes are involved in, the more correlations there will be between these genes and other functional groups. The results show that our approach can discover some kind of information about this. For example, there were many correlations between different gene groups and regulation genes, such as transcription factors and cyclins in the group for cell cycle (see Table 1 and Fig. 3). In contrast, there was no correlation between other gene groups and the group for chromatin structure, which consisted of typical function genes, i.e. histones. This is very similar to Friedman et al.'s work [27]. In Friedman et al.'s Bayesian network, the most striking feature of the high confidence order relations is the existence of dominant genes. These dominant genes are directly involved in initiation of the cell cycle and its control. For example, in our results, YLR103C (CDC45) and YBR088C (POL30) are genes having many correlations with others and both are in the top 10 dominant genes by Friedman et al.

3.5. Conclusion

In this study, we have shown that the HMM algorithm can produce clusters with a quality as good as two other prevalent clustering algorithms (k-means and SOM methods). This is partly because our algorithm is based on a statistical model that is very rich in mathematical structure. Another reason may be that the HMM method also performs well in time series or linear sequences and the gene expression data used in this study are all time-related expression data. Accompanying the ratio FOM, our methodology has been proven to have the ability to determine the most 'natural' number of clusters without any external information, and the results were in agreement with external biological knowledge. Genes with different annotated functions or involved in different cellular processes might have similar expression patterns, which suggests that it is very important to distinguish between the function genes and regulation genes; moreover, many genes have more than one function. The biological system is very complex, and it is not enough to partition genes into disjoint clusters. Using the HMM algorithm, we could find out the correlations among genes with different functions or involved in different processes.

Not only genes with similar functions or involved in the same process, but also genes involved in different processes may have similar expression patterns and cluster together. Our algorithm could discover these complex correlations, because in our approach, each gene expression profile is classified into the component HMMs with certain probabilities, and thus the genes with multiple functions can be partitioned into various functional groups with different probabilities. Rather than clusters, it is the whole regulated network that constitutes the natural structure of the biological system. The results also show that our algorithm could provide helpful information in reconstruction of the genetic network.

We plan to carry out further work to improve and extend the application of this method. Since the HMM is Markovian, that is, the observations depend on the previous and current states only, this method is currently limited to time series data. We hope that a bootstrapping technique will compensate for this limitation. This method is not very fast, because it has to train the parameters of the model with the gene expression dataset. Introducing more biological knowledge into the model, that is, initializing the model with external information, will help to improve it. In this paper, we have used subsets of data without any missing values. With the underlying probability framework, we expect the ability to model outliers and missing values explicitly to be another potential advantage of this model-based approach.

Acknowledgements: This work was supported by NSFC Grants 39980007863. We thank Yingwu Huang for critical discussions.

References

[1] Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson, J., Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D. and Brown, P.O. (1999) Science 283, 83–87.
[2] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Proc. Natl. Acad. Sci. USA 95, 14863–14868.
[3] Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J. and Davis, R.W. (1998) Mol. Cell 2, 65–73.
[4] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders,

K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Mol. Biol. Cell 9, 3273–3297.
[5] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999) Proc. Natl. Acad. Sci. USA 96, 2907–2912.
[6] Toronen, P., Kolehmainen, M., Wong, G. and Castren, E. (1999) FEBS Lett. 451, 142–146.
[7] Tavazoie, S., Hughes, D., Campbell, M.J., Cho, R.J. and Church, G.M. (1999) Nat. Genet. 22, 281–285.
[8] del Rio, G., Bartley, T.F., del Rio, H., Rao, R., Jin, K.L., Greenberg, D.A., Eshoo, M. and Bredesen, D.E. (2001) FEBS Lett. 509, 230–234.
[9] Xu, Y., Olman, V. and Xu, D. (2002) Bioinformatics 18, 536–545.
[10] Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C., Furey, t.S., Ares, M. and Haussler, D. (2002) Proc. Natl. Acad. Sci. USA 97, 262–267.
[11] Tanay, A., Sharan, R. and Shamir, R. (2002) Bioinformatics 18 (Suppl. 1), S136–144.
[12] Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. and Barkai, N. (2002) Nat. Genet. 31, 370–377.
[13] Rabiner, L.R. (1989) Proc. IEEE 77, 257–286.
[14] Krogh, A., Larsson, B., Heijne, G.V. and Sonnhammer, E.L.L. (2001) J. Mol. Biol. 305, 567–580.
[15] Sonnhammer, E.L.L., Eddy, S.R., Birney, E., Bateman, A. and Durbin, R. (1998) Nucleic Acids Res. 26, 320–322.
[16] Krogh, A., Brown, M., Mian, I.S., Sjolander, K. and Haussler, D. (1994) J. Mol. Biol. 235, 1501–1531.
[17] Francesco, V.D., Munson, P.J. and Garnier, J. (1999) Bioinformatics 15, 131–140.
[18] Yuan, Z. (1999) FEBS Lett. 451, 23–26.
[19] Yada, T., Nakao, M., Totoki, Y. and Nakai, K. (1999) Bioinformatics 15, 987–993.
[20] Jain, A.K. and Dubes, R.C. (1988) Algorithms for Clustering Data, Prentice Hall, Englewood Cliffs, NJ.
[21] Mrowka, R. (2001) Bioinformatics 17, 669–671.
[22] Duda, R.O. and Hart, P.E. (1973) Pattern Classification and Scene Analysis, Wiley, New York.
[23] Everitt, B.S. and Hand, D.J. (1981) Finite Mixture Distributions, Chapman and Hall, London.
[24] Nowlan, S. (1990) in: Advances in Neural Information Processing System (Touretsky, D., Ed.), vol. 2, pp. 574–582, Morgan Kaufmann, San Mateo, CA.
[25] Rand, W.M. (1971) J. Am. Stat. Assoc. 66, 846–850.
[26] Yeung, K.Y., Haynor, D.R. and Ruzzo, W.L. (2001) Bioinformatics 17, 309–318.
[27] Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000) J. Comput. Biol. 7, 601–620.
[28] Dewey, T.G. and Galas, D.J. (2001) Funct. Integr. Genomics 1, 269–278.