BIOINFORMATICS

Vol. 20 Suppl. 1 2003, pages i283–i289
DOI: 10.1093/bioinformatics/bth937

Robust inference of groups in gene expression time-courses using mixtures of HMMs

Alexander Schliep[1,], Christine Steinhoff[1] and Alexander Schönhuth[2]*

[1]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany and [2]Center for Applied Computer Science, University of Cologne, Weyertal 80, 50937 Cologne, Germany

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: Genetic regulation of cellular processes is frequently investigated using large-scale gene expression experiments to observe changes in expression over time. This temporal data poses a challenge to classical distance-based clustering methods due to its horizontal dependencies along the time-axis. We propose to use hidden Markov models (HMMs) to explicitly model these time-dependencies. The HMMs are used in a mixture approach that we show to be superior over clustering. Furthermore, mixtures are a more realistic model of the biological reality, as an unambiguous partitioning of genes into clusters of unique functional assignment is impossible. Use of the mixture increases robustness with respect to noise and allows an inference of groups at varying level of assignment ambiguity. A simple approach, partially supervised learning, allows to benefit from prior biological knowledge during the training. Our method allows simultaneous analysis of cyclic and non-cyclic genes and copes well with noise and missing values.

Results: We demonstrate biological relevance by detection of phase-specific groupings in HeLa time-course data. A benchmark using simulated data, derived using assumptions independent of those in our method, shows very favorable results compared to the baseline supplied by *k*-means and two prior approaches implementing model-based clustering. The results stress the benefits of incorporating prior knowledge, whenever available.

Availability: A software package implementing our method is freely available under the GNU general public license (GPL) at http://ghmm.org/gql

Contact: schliep@molgen.mpg.de

Supplementary information: Supplemental material can be found at http://algorithmics.molgen.mpg.de/ExpMix

1 INTRODUCTION

Life is a dynamic process over time. An understanding of the cellular mechanisms, which govern its peculiarities on the level of genes and their regulation can only be gained from experiments that reflect this time dependence. The final goal in the analysis of large-scale time-course gene expression datasets available is the inference of regulatory networks. A number of methods have attempted to solve this problem directly (Friedman *et al*., 2000; Chen *et al*., 1999). Because of the large number of genes, their complex relationships and the very large amount of noise in microarray measurements, it has been a standard procedure for a first analysis to identify groups of genes with similar temporal regulatory patterns or time-courses. The relevant approaches can be broadly categorized by the following criteria: their applicability to either cyclic (Spellman *et al*., 1998; Whitfield *et al*., 2002), non-cyclic (Bar-Joseph *et al*., 2002; Ramoni *et al*., 2002) or both types of time-courses (Schliep *et al*., 2003), whether they take dependencies along the time-axis into account like the former methods or not (Eisen *et al*., 1998; Gasch *et al*., 2000; Tavazoie *et al*., 1999; Rifkin and Kim, 2002), and, finally, if the clustering is based on statistical models or some sort of distance function. All these approaches compute a partition of the time-courses, requiring assignment of each gene to a single group. However, the biological reality does not agree with this. Consider the MAP-kinase pathway as an example. A stimulus such as serum induction after starvation activates a number of processes. On one hand, it leads to cell division and proliferation but on the other hand space limiting factors or high levels of serum may also induce stress genes. One would expect MEKK and MAP kinase phosphatases to be activated as well during the whole course of the experiment as a consequence of serum induction. G-proteins and early response genes are only activated in the very beginning of the serum induction while cyclins, like cyclin A and cyclin B1 are induced quite late in the experiment. Hence, one would expect the time-course of MEKK to be equally similar to those G-proteins and cyclins A and B.

To cope with this reality, we model a set of gene expression time-courses as a mixture model. Compared with clustering, the use of mixtures increases robustness of the estimation process in the presence of noise. The individual components

*To whom correspondence should be addressed.

are hidden Markov models (HMMs), which have been successfully used in a wide range of applications (MacDonald and Zucchini, 1997), mainly for their flexibility in encoding 'grammatical' constraints of time-courses. In addition, we found that their graphical structure benefits the analysis process, as it affords a high degree of interactivity and accessibility.

The estimation or learning algorithms used are only proven to arrive at local maxima. Owing to the complexity of the problem a high degree of dependency on initial conditions is to be expected. Typically, only unlabeled (no information about the correct group assignment is known) data are used in clustering. We propose to additionally use labeled data. That the combination can be beneficial has been discovered first in the context of learning classifiers. In fact, the decrease in classification error is exponential in the amount of unlabeled data (Castelli and Cover, 1994). Since then a number of approaches were developed following the same general idea. Szummer and Jaakkola (2002) propose two estimation procedures for classifying text documents by constructing weighted graphs, Blum and Chawla (2001) partition graphs by mincuts controlled by labeled examples. Belkin and Niyogi (Belkin, 2003) infer the (minimal) sub-manifold that contains the data from the complete dataset and use the labeled samples for classification on it. Nigam *et al*. (2000) present a method based on statistical models for text classification. Here, analogous to our approach, the EM algorithm is extended to gain from labeled examples when inferring groups in data. We show that there is a large improvement in convergence to good local optima on typical data, even if only small amounts of labeled data are supplied.

The potential benefit is particularly relevant in biological applications as, generally speaking, the understanding of complex biological systems is still too limited to formulate very detailed mathematical models. Methods which have the ability to include prior knowledge and thus integrate more biological facts should outperform those which do not. This integration is feasible, as typically small amounts of high-quality annotation regarding regulation or function of genes are available.

Our method allows simultaneous analysis of cyclic and non-cyclic time-courses in a mixture modeling framework using flexible graphical models based on HMMs, which explicitly model horizontal dependencies, as mixture components using a partially supervised learning approach to obtain parameter estimates robustly and reproducibly. Background noise is accounted for in a dedicated noise component, missing data are flexibly and consistently handled.

2 METHODS

Our method is based on the well-established framework of mixture modeling (McLachlan and Basford, 1988; McLachlan and Peel, 2000) in which we employ HMMs as simple, robust and flexible models for time-course data. The combination of the two is novel—a clustering approach based on the same class of HMMs has been described earlier (Schliep *et al*., 2003). Also a novelty is the application of a simple, nevertheless efficient, extension to the classical Expectation-Maximization (EM) algorithm for estimating mixture parameters from data, when high-quality annotation for some genes is known a priori. Lastly, our method includes a robust decoding procedure, which allows to infer groups of genes in time-course datasets, such that the assignment to groups is unambiguous.

2.1 Mixtures of HMMs

We use HMMs [see Rabiner (1989) for an excellent introduction] with continuous emissions governed by a normal distribution in each state. The HMM topology—the number of states, the set of possible transitions—is essentially a linear chain (following Schliep *et al*. 2003), neglecting a possible transition from the last to the first state to accommodate cyclic behavior. Note that we do not expect states to have a semantic in terms of regulation. They simply reflect regions of a time-course with similar levels of expression. There are usually fewer states than time-points, as several similar successive measurements will be accounted for by the same state by making use of its self-transition. It is important to point out that our approach is not limited to such models but rather accommodates arbitrary HMM topologies. As many of the successful applications in time-course modeling (MacDonald and Zucchini, 1997; Knab *et al*., 2003) show, more complex models, capturing more of the 'grammar' observable in the time-courses, which in the case of gene expression is imposed by the regulatory mechanisms at work, should improve the quality of the results greatly.

We combine K of such HMMs $\lambda_1, \ldots, \lambda_K$ to a probability density function (pdf) for a gene expression time-course by use of a convex combination of the K component probability density functions induced by the HMMs, denoted $p_j(\cdot, \lambda_j)$. The mixture pdf is parameterized by $\Theta = [\lambda_1, \ldots, \lambda_K, (\alpha_1, \ldots, \alpha_K)]$ and defined as

$$p(\cdot|\Theta) := \sum_{j=1}^{K} \alpha_j \, p_j(\cdot, \lambda_j).$$

As the former is just a usual mixture (McLachlan and Basford, 1988; McLachlan and Peel, 2000), the well-known theory applies. The resulting likelihood function can be optimized with the EM algorithm (Dempster *et al*., 1977; Wu, 1983; Boyles, 1983; Bilmes, 1998) to compute maximum-likelihood estimates for Θ, or learning the mixture.

2.1.1 Partially supervised learning Analogous to Nigam *et al*. (2000), we propose partially supervised learning as an extension to the usual EM-based mixture estimation. This allows the training to benefit from prior knowledge about genes, e.g. when it is known that they are regulated by the same regulatory pathway. The benefits of even very small

quantities, e.g. <1% of all time-courses, of labels are large. They improve the robustness of the estimation process with respect to noise and the quality of the local optimum to which the mixture likelihood converges to during the learning. In the following, we will argue why the modified EM algorithm still converges in the case of partially supervised learning.

To apply the EM algorithm, one assumes the existence of unobservable (or hidden) data $Y = \{y_i\}$ that indicates which component has produced each O^i in the set of time-courses O. Thus, we can formulate a complete-data log-likelihood function $\log L(\Theta|O, Y)$.

If we are given labeled time-courses, we do not have to guess the corresponding y_i. While the labels do not reveal the parameters of the mixture component, they however indicate whether two labeled time-courses have been created by the same or by distinct components. We denote the set of labeled time-courses with O_L and the set of unlabeled ones with O_U. For a time-course O^i from O_L, we set the value of y_i to its component label l_i and maintain this assignment throughout the running time by setting $\mathbf{P}[\lambda_j|O^i] = 1$ for $j = l_i$ and zero else. This can be thought of as conditioning the relevant distributions and the likelihood on the known labels, yielding a Q-function [cf. Bilmes (1998); the Θ^t are the estimates for the maximum likelihood in the t-th iteration], which splits into two sums,

$$Q(\Theta, \Theta^t) := \sum_{O^i \in O_L} \log\left[\alpha_{l_i} p_{l_i}(O^i|\lambda_{l_i})\right] + \sum_{O^i \in O_U} \sum_{j=1}^{K} \log[\alpha_j p_j(O^i|\lambda_j)]\mathbf{P}[j|\Theta^t, O^i],$$

and for which the usual local convergence result holds.

2.1.2 Decoding mixtures Even neglecting the high experimental error rates, which obfuscate the assignment, biology gives us no reason to believe that genes can be assigned unambiguously to clusters due to the high complexity of interacting networks. If we subscribe to the view that, ultimatively, clusters are formed based on functional categories, the use of genes in multiple regulatory pathways negates the possibility of a unique assignment for all genes. Mixture estimation—a non-statistical analogon is fuzzy clustering (Pedrycz, 1990)—circumvents this dilemma, which can easily lead most clustering methods astray in the learning phase.

The simplest way of decoding a mixture, that is inferring groups in the data, is to interpret the mixture components as clusters and assign each time-course to the cluster which maximizes the probability of the cluster given the time-course O, $\mathbf{P}[\lambda_j|O]$. However, a mixture encodes much more information. Inspection of the discrete distribution $d(O) := \{\mathbf{P}[\lambda_i|O]\}_{1\le i\le K}$ reveals the level of ambiguity in making the assignment, which can be quantified easily and sensibly by computing the entropy $H[d(O)]$. Choosing a threshold on the entropy yields a grouping of the data into $K+1$ groups; K corresponding to the K mixture components and one collecting all time-courses, which exhibit an ambiguity level exceeding the threshold and which remain unassigned for that reason.

It is not easily possible to automate the choice of threshold. However, an interactive graphical user interface displaying the time-courses and their assignment to clusters when the user changes the threshold provides an effective way to settle on a value.

We deal with missing values in the following manner. Each state of an HMM can either emit a real-valued variate according to its Gaussian state emission pdf or with a low probability equal to the proportion of missing values in all the time-courses, a special missing symbol. This circumvents the need for replacing missing values with estimates, for example through interpolation. The models we use afford a high degree of variability as far as the time-points of changes allowed by the model are concerned. Hence, groups will typically contain time-courses having the same qualitative behavior. The time at which, e.g. an up-regulation occurs will often vary. Synchronous subgroups of such clusters are found with the Viterbi-decomposition introduced in Schliep *et al.* (2003).

The Graphical Query Language (GQL) software is based on the freely available GHMM-library (GHMM, 2003, http://ghmm.org) and implements the method described using a portable graphical user interface.

2.2 Data

It is still open to debate what constitutes an appropriate evaluation of analysis methods for gene-expression time-courses. Regulation is not well understood and the annotation available is too sparse and too inconsistent to allow a large-scale automated evaluation as it is routinely done in machine learning. Real biological data usually provides anecdotal evidence; none of the prior approaches provides a reasonably sized biological dataset for benchmarking.

We tested our approach on HeLa cell-cycle data and resorted to artificial data for benchmarking. The assumptions made in creating the artificial data were disjoint from those made by any of the methods compared.

Whitfield We used published data from a time-course experiment (Whitfield *et al.*, 2002), in which the authors measured genome-wide gene expression of synchronized HeLa cells. We used the raw data from doubly thymidine experiment three as provided by the authors in the Supplementary information. In this dataset, HeLa cells, which have been arrested in S phase by a double thymidine block, were measured every hour from 0 to 46 h. For reasons of comparison, we excluded clones showing missing values from further analysis. Furthermore, the data were pre-processed by extracting all these genes with an absolute fold change of at least two in at least one time point. This resulted in a dataset containing 2272 expression time courses. Additionally, we used a list of

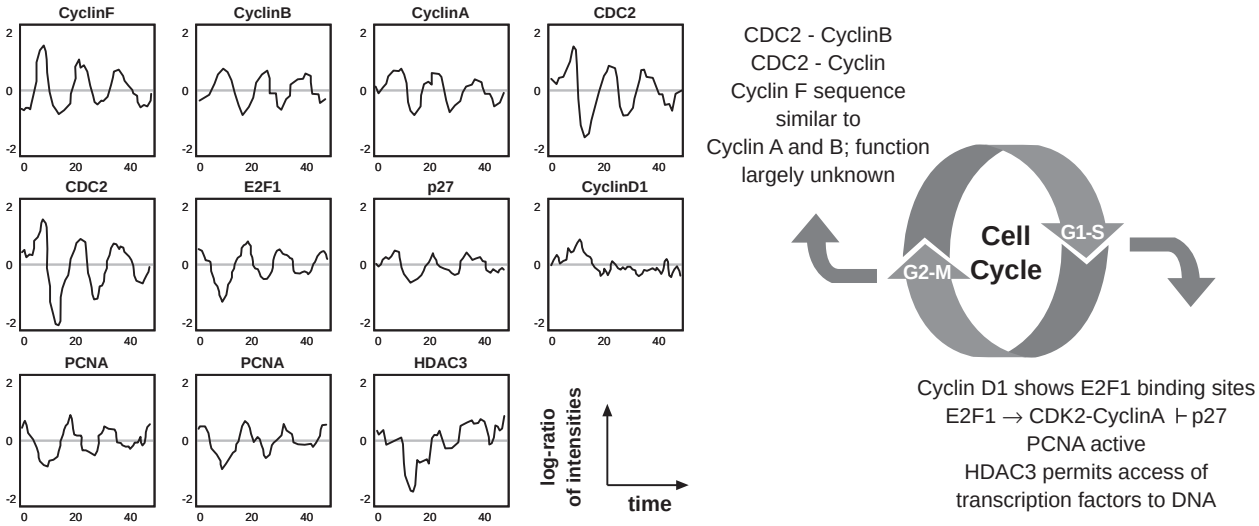

Fig. 1. Time courses of selected cell cycle regulators obtained from the biological literature.

Table 1. The simulated dataset `SIM` consists of a total of 3500 time-courses of length 30 (equal step-width in $[0, 2\pi]$) in six classes

Class	Description	Size	Function
C1	Up-regulation	500	$0.15 \cdot x - 0.7 + N(1, 0.3)$
C2	Noise	1000	$0 + N(1, 0.6)$
C3	Down-regulation	500	$-0.3 \cdot x - 0.3 + N(1, 0.3)$
C4	Cyclic 1	500	$N(1, 0.1) \cdot \sin[1.2 \cdot N(1, 0.05) \cdot x + 0.8 \cdot 2\pi] + N(0, 0.4)$
C5	Cyclic 1	100	$N(1, 0.0075) \cdot \sin[1.2 \cdot N(1, 0.1) \cdot x + 0.6 \cdot 2\pi] + N(0, 0.5)$
C6	Cyclic 1	900	$N(1, 0.9) \cdot \sin[1.5 \cdot N(1, 0.025) \cdot x + 0.5 \cdot 2\pi] + N(0, 0.5)$

The time-courses were obtained by sampling from the respective class models. The normal distribution is denoted as $N(\mu, \sigma)$.

Table 2. The results on `SIM` for k-means clustering, CAGED (Ramoni *et al*., 2002), Splines (Bar-Joseph *et al*., 2002) and HMM mixtures with no, 0.9% (5 per class) and 1.7% (10 per class) labeled time-courses per class

Method	Description	Specificity (%)	Sensitivity (%)
M1	k-means, Euclidean distance	85.55	71.87
M2	CAGED	41.00	99.70
M3	Splines	47.29	39.38
M4	HMM mixtures	93.00	79.14
M5	HMM mix., 0.9% labeled	96.40	96.90
M6	HMM Mix., 1.7% labeled	96.60	96.99

By comparing the known classes in `SIM` with the computed clustering for all pairs of time-courses, we computed true and false positives as well as true and false negatives, abbreviated TP, FP, TN and FN, respectively. True positive is defined as a pair of time-courses with equal class that are assigned to the same cluster. To quantify the performance, we computed the standard sensitivity, #TP/(#TP + #FN), and specificity, #TP/(#TP + #FP).

genes which have been described in literature to be regulated dependent on the cell cycle [Table 2, Whitfield *et al*. (2002)].

Simulated data To facilitate benchmarking and evaluation we tried to design a method for creating simulated datasets, which makes very mild assumptions about the nature of the data but reflects the realities of microarray experiments. Our proposed approach is independent from the underlying assumptions and peculiarities of the statistical model in our method, as it is independent from the assumptions in other methods. We assume three broad categories of genes, cell-cycle regulated, non-cell-cycle regulated and unregulated genes. We choose the sine function as a 'true' model for the first, linear functions for the second and const = 0 for the third category (Table 1). Randomization is performed by modifying the argument to a function, changing phase and frequency, and the resulting function value, modifying amplitude, shifting all values and, finally, adding noise.

3 RESULTS

The method requires as input a collection of initial models. Given that the training procedure will only arrive at local optima, one would expect a large degree of dependence on input. Surprisingly, random choices of linear component models performed well. We used k models of size within a prescribed range and default emission parameters $\mu = 0$, $\sigma = 0.1$. Subsequently, we computed a random k-partition of the dataset and estimated each of the models with one group

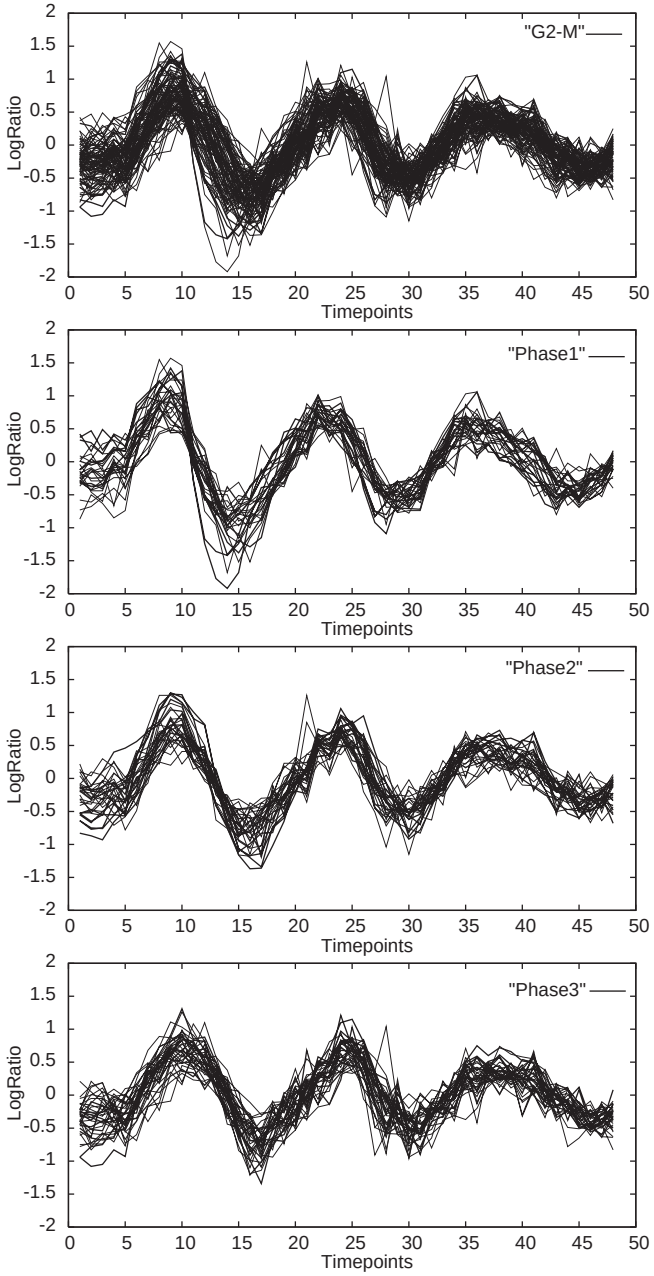

Fig. 2. A group obtained by computing a mixture model using 9 labeled and 2263 unlabeled time-courses from the Whitfield dataset (top). It contains five of the labeled time-courses. The group was decomposed, using the Viterbi decomposition, into three subgroups, corresponding to synchronous genes, resulting in a first subgroup containing mainly G2 genes (phase 1), the second having G_2 as well as G_2/M genes (phase 2) and the third having mostly G_2/M genes (phase 3, bottom).

of the partition using the Baum–Welch algorithm (Rabiner, 1989). We avoided overfitting by limiting the number of steps. If labeled data were given, only the labeled data were used to estimate models, which had labeled data assigned; the

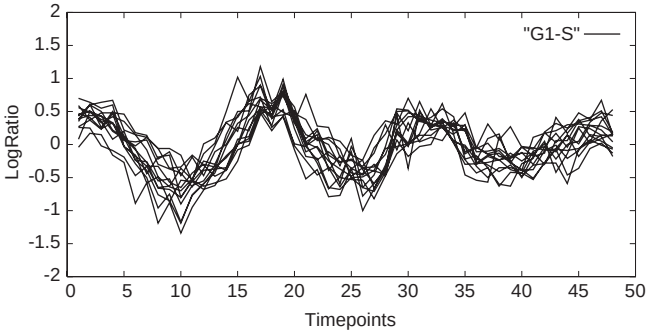

Fig. 3. Another group containing cell cycle related genes obtained by computing a mixture model using 9 labeled and 2263 unlabeled time-courses from the Whitfield dataset. This group contains only genes belonging to phases G_1/S and S, four of which were labeled input.

unlabeled data were used as described above for the remaining models. We added a designated noise-component, which is simply a one-state model with $N(0, \sigma)$-emissions for a large value of σ, exempted from training. This mixture component that accounts for time-courses, which do not fit any of the other mixture components well avoids unnecessary 'broadening' of the other components. As we use a default mixture model, we can apply standard criteria for model selection such as the Bayesian information criterion (BIC) (Hastie *et al.*, 2001) to decide between different numbers of components. We repeated the experiments 20 times for varying numbers of k and used BIC to choose a best k. Best performance is shown. To allow comparisons, a cluster assignment was obtained from the mixture we estimated using an entropy cutoff value of $\log(k)$.

Whitfield Cell cycle regulators as, for example, different cyclins, E2F, PCNA and HDAC3 are known to be active in different stages in the cell cycle. Furthermore, they have a regulatory impact on each other, either directly (as E2F1 acting on Cyclin D) or indirectly (as E2F1 has impact on p27 and vice versa via CDK2-Cyclin A). Thus, one expects to find these patterns of regulatory activity in the underlying gene expression dataset. Cyclin B and Cyclin A both act while being bound to CDC2 during the transition from G_2 to M phase. In fact, as shown in Figure 1 they are coordinately regulated and there is a clear phase shift compared to E2F1, e.g. which is active in the transition of the G_1 phase to S phase. Cyclin F is known to have a similar sequence as Cyclins A and B but the function is largely unknown (Kraus *et al.*, 1994). In Figure 1 one can see that it is apparently regulated temporarily equally as Cyclins A and B. CDC2 is present twice on the array, and as shown here, the expression profile are almost equal. On the other hand, E2F1 regulates p27, which is demonstrated by a clear phase shift in the time course. PCNA is needed for the initiation of S phase as well and is regulated clearly

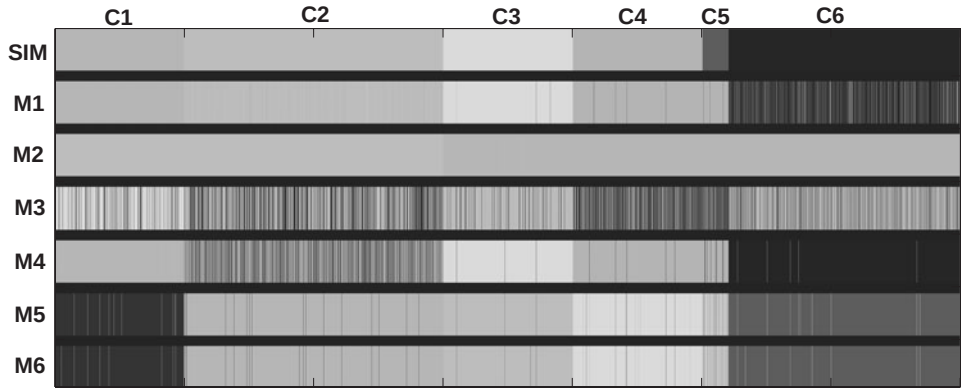

Fig. 4. Cluster assignment of time-courses in `SIM`: the first column shows clusters C1–C6 from `SIM` as blocks of different gray levels, the second to the seventh column shows the cluster assignments obtained by methods M1–M6 (see Table 2 for a definition). This allows a qualitative assessment of errors made. A good clustering should display solid blocks (see the Web supplement for a color version of this figure).

in G_1/S phase. HDAC3 which is needed to permit the access to the DNA and thus allowing transcriptional regulation is expressed during G_1/S transition as well.

We used a collection of 35 random linear 24-state models. We used the five G_2/M phase genes described above as a seed for one cluster and four genes of the G_1/S phase for a second one. Decoding the mixture resulted in two groups containing the labeled time-courses of size 91 and 14, respectively (Figs 2 and 3). We computed a Viterbi-decomposition of the larger group. The first subgroup contained 26 genes known to be G_2 and one G_2/M, the second 11 G_2 and 19 G_2/M, the third 31 G_2/M, 2 M/G_1 and 1 G_1/S. The second group a shown in Figure 3, contained 12 G_1/S and 2 S-phase genes. Both CDC2 representatives (Fig. 1) are found in the same component (phase 1, Fig. 2). Furthermore, cyclin A (phase 2), cyclin B (phase 3) are assigned to different, slightly phase shifted components compared with the one in which CDC2 is captured. Moreover, all time-courses that are assigned to the different phases of our G_2, G_2/M phase cluster are known to be cell cycle regulated in their respective phase (Whitfield *et al*., 2002). The same holds for the G_1/S, S phase cluster. This can be followed up in detail in our Web supplement. Thus, the modest amount of prior information used resulted in highly specific (sub-)groups of synchronously expressed genes.

Simulated data Most remarkable is the very good performance of the simplest method, *k*-means clustering using Euclidean distance, which is not tailored to time-course data on the `SIM` dataset. As shown in Table 2, two of the more involved methods, Caged (Ramoni *et al*., 2002) and the Spline-based clustering by Bar-Joseph *et al*. (2002) only reach a specificity of $<$50%. The main error made by Caged in deciding on too few clusters (this cannot be controlled by the user), which leads to merging of several classes (C1 and C2, respectively C3–C6, cf. Fig. 4) into one cluster. The HMM mixture perform quite well, achieving a high degree of over 90% specificity and over 75% sensitivity. The tests also show very clearly the impressive effect of partially supervised learning. It suffices to have labels for 30 or $<$1% of all time-courses (cf. M5 in Table 2), to obtain a specificity and sensitivity exceeding 95%. More labels do not yield further significant improvements.

4 DISCUSSION

We present a robust, simple and efficient approach to analyze gene expression time-course data using a mixture of HMMs. The method can easily make use of prior knowledge about time-courses due to a partially supervised training procedure, which greatly increases robustness (see Supplementary information) and the quality of the local optima found. Simultaneous analysis of cyclic and non-cyclic time-courses is possible and neither missing values nor realistic levels of noise pose a serious problem. Besides their computational advantages (for experiments demonstrating a higher robustness to noise compared with clustering see the Supplementary information) and their better fit to biological reality, mixtures allow a quantification of the assignment uncertainty. Moreover, an interactive exploration of assignments at various levels of uncertainty is supported.

We demonstrate biological relevance by analysis of a HeLa time-course dataset for which we infer synchronous groups specific to cell cycle phases. A comparison on simulated data created using mild assumptions distinct from ours and those implicit to other methods, yielded favorable results. Our flexible framework combined with an effective graphical user interface implemented in GQL supports interactive, exploratory knowledge discovery making full use of biological expert knowledge.

ACKNOWLEDGEMENTS

Thanks to Terry Speed for helpful discussions. Thanks to Wasinee Rungsarityotin and Benjamin Georgi for work on GHMM, the GHMM Python interface and GQL. The third author would like to acknowledge funding from the BMBF through the Cologne University Bioinformatics Center (CUBIC).

REFERENCES

Bar-Joseph,Z., Gerber,G., Gifford,D.K., Jaakkola,T.S. and Simon,I. (2002) A new approach to analyzing gene expression time series data. *Proceedings of the Sixth Annual International Conference on Research in Computational Molecular Biology (RECOMB'02)*, Washington, DC, 18–21 April, ACM Press, pp. 39–48.

Belkin,M. (2003) Problems of learning on manifolds. Ph.D. thesis, University of Chicago.

Bilmes,J.A. (1998) A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *Technical Report TR-97-021*, International Computer Science Institute, Berkeley, CA.

Blum,A. and Chawla,S. (2001) Learning from labeled and unlabeled data using graph mincuts. In C.E.Brodley and A.P.Danylok, (eds), *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*, Williamstown, MA, 28 June–1 July. Morgan Kaufmann, pp. 19–26.

Boyles,R. (1983) On the convergence of the EM algorithm. *J. R. Stat. Soc.*, **45**, 47–50.

Castelli,V. and Cover,T.M. (1994) On the exponential value of labeled samples. *Pattern Recognit. Lett.*, **16**, 105–111.

Chen,T., Filkov,V. and Skiena,S.S. (1999) Identifying gene regulatory networks from experimental data. *Proceedings of Third Annual International Conference on Computational Molecular Biology* (*RECOMB'99*), ACM Press, pp. 94–103.

Dempster,A., Laird,N.M. and Rubin,D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, **39**, 1–38.

Eisen,M., Spellman,P., Brown,P. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci., USA*, **95**, 14863–14868.

Friedman,N. Linial,M., Nachman,I. and Pe'er,D. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

GHMM (2003) The General Hidden Markov Model library. URL: http://ghmm.org

Hastie,T., Tibshirani,R. and Friedman,J.H. (2001) *The Elements of Statistical Learning*. Springer.

Knab,B., *et al*. (2003) Model-based clustering with hidden Markov models and its application to financial time-series data. In Schader,M., Gaul,W. and Vichi,M. (eds), *Between Data Science and Applied Data Analysis*. Springer, pp. 561–569.

Kraus,B., Pohlschmidt,M., Leung,A.L., Germino,G.G., Snarey,A., Schneider,M.C., Reeders,S.T. and Frischauf,A.M. (1994) A novel cyclin gene (CCNF) in the region of the polycystic kidney disease gene (PKD1). *Genomics*, **24**, 27–33.

MacDonald,I.L. and Zucchini,W. (1997) *Hidden Markov and Other Models for Discrete-Valued Time Series*. Chapman & Hall, London.

McLachlan,G. and Basford,K. (1988) *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, Inc., New York, Basel.

McLachlan,G. and Peel,D. (2000) *Finite Mixture Models*. *Wiley Series in Probability and Statistics*. Wiley, New York.

Nigam,K., McCallum,A., Thrun,S. and Mitchell,T. (2000) Text classification from labeled and unlabeled documents using EM. *Mach. Learn.*, **39**, 103–104.

Pedrycz,W. (1990) Fuzzy sets in pattern recognition: methodology and methods. *Pattern Recognit.* **23**, 121–146.

Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–285.

Ramoni,M.F., Sebastiani,P. and Kohane,I.S. (2002) Cluster analysis of gene expression dynamics. *Proc. Natl Acad. Sci., USA*, **99**, 9121–9126.

Rifkin,S.A. and Kim,J. (2002) Geometry of gene expression dynamics. *Bioinformatics*, **18**, 1176–1183.

Schliep,A., Schonhuth,A. and Steinhoff,C. (2003) Using hidden Markov models to analyze gene expression time course data. *Bioinformatics*, **19** (Suppl. 1), I255–I263.

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) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Szummer,M. and Jaakkola,T. (2002) Partially labeled classification with Markov random walks. *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, Vancouver, BC, Canada, 3–8 December. MIT Press.

Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.

Whitfield,M.L., Sherlock,G., Saldanha,A.J., Murray,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. and Botstein,D. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.

Wu,C. (1983) On the convergence of the EM algorithm. *Ann. Stat.*, **11**, 95–103.