BIOINFORMATICS

Vol. 00 no. 00 2010
Pages 1–6

NETGEM: Network Embedded analysis of Temporal Gene Expression using Mixtures

Vinay Jethava [1], Torbjorn Karfunkel[2], Chiranjib Bhattacharyya [1], Devdatt Dubhashi[2],Goutham N. Vemuri[3]*

[1]Computer Science and Automation Department, Indian Institute of Science, Bangalore, INDIA
[2]Department of Computer Science, Chalmers University of Technology, Göteborg, SWEDEN
[3]Systems Biology, Department of Chemical and Biological Engineering, Chalmers University of Technology, Göteborg, SWEDEN

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation

Results

Availability: The source code for NETGEM is available from http://129.16.106.142/

Contact: goutham@chalmers.se

1 INTRODUCTION

Microarrays have become a routine tool in biological enquiry, geared to measure global gene expression in response to genetic or environmental perturbations. Gene expression microarrays present a snapshot of the transcriptional profile of all the genes at the time of measurement. The outcome is a vast amount of data, which has been analyzed using several statistical methods including hierarchical clustering (Eisen *et al.*, 1998), k-means clustering (Tavazoie *et al.*, 1999), self organizing maps (Tamayo *et al.*, 1999), singular value decomposition (Rifkin & Kim, 2002). The key focus of the methods has been clustering of genes that have similar expression profile, based on the assumption that co-expressed genes are likely to be regulated. An inherent drawback of the clustering approaches is their unsuitability in the analysis of temporal expression data.

This has led to growing interest towards development of dedicated algorithms to handle the temporal data. One of the key challenges is the small number of observations (time points), owing to cost and/or biological limitations. Several methods have been investigated including significance analysis (Tusher *et al.*, 2001; Leek *et al.*, 2006), autoregressive curves based model (Ramoni *et al.*, 2002), hidden markov models (HMM) (Schliep *et al.*, 2003; Yoneya & Mamitsuka, 2007), mixture models (Schliep *et al.*, 2004; Costa *et al.*, 2005), clustering methods (Ernst & Joseph, 2006), association rules (Nam *et al.*, 2009). A review of the methods is available in Androulakis *et al.* (2007). However, the previous methods assume an time-invariant network topology, such as the protein-protein interaction network or the genetic network inferred from microarray data.

Song *et al.* (2009) first investigated the problem of learning the temporally-varying interaction networks based on local neighbourhood selection with $l1$-regularization to obtain sparse networks. The analysis assumes a smooth variation in the network interactions strengths to overcome the unreliability of results due to the small number of observations available in most biological experiments.

This paper investigates a markovian model for analyzing the rewiring problem when the underlying interactions network is known with a certain confidence. In other words, the observed expression levels are controlled by a known network but the interaction strengths are varying with time. The dynamics of temporal evolution in a rewiring network is a matter of study, and is hypothesized to be stochastic in nature. In this work, we postulate that the evolution of the interaction strengths are markovian in nature. This means that the temporal evolution of the interactions strengths can be characterized in terms of a transition probability matrix. Thus, the problem is one of learning the transition probability matrix characterizing for the temporal evolution of interaction strengths defined on edges in the network based on the observed expression data. Hidden markov models (HMM) (Rabiner, 1989; Cappé *et al.*, 2007) provide a natural method for the study of such systems. However, a naive HMM implementation for this inference problem would have an exponentially large state space and is NP-hard.

One of the contributions of this paper is a principled approach that performs approximate inference to estimate the transition probability matrix that best characterizates of the temporal evolution of the hidden interaction strengths. The main assumption is that the interaction strengths evolve *independently* of each other. The analysis is closely related to the Factorial HMM (Ghahramani & Jordan, 1997) and allows us to model the evolution characteristics of each interaction (edge) in the known network independently. Then, the problem is the learning of the transition probability matrix for each edge based on the observed gene expression levels. We employ a bayesian approach (Gelman *et al.*, 2003; Beal, 2003), which have often been used to solve inference problem where there are few observation samples available, to solve the problem of learning the transition probability matrices for each interaction (edge) in the known network.

There has been considerable effort in establishing a hierarchy of genes based on their functionality (Bader *et al.*, 2003; Mewes

*to whom correspondence should be addressed

et al., 2002; Stark *et al.*, 2006; Xenarios *et al.*, 2000; Zanzoni, 2002). This poses the natural question of the relationship between the functional classification of the genes and the temporal evolution of the interaction. Further, can we distill some observations about evolution characteristics of a group of functionally similar genes. The second contribution of this paper is a novel method for incorporating the functional classifications in the analysis of temporal expression data. We model the evolution of interaction strength for a gene pair as a mixture of evolution characteristics of the functional categories. The problem then becomes the learning of the evolution characteristics for the functional categories as well as the mixing proportions for each interaction edge in the known network.

BIO-REF NEEDED Often, gene expression level measurements are available for multiple strains which are slightly different perturbed versions of the original version. For example, a couple of genes might be knocked out of the network. Traditional methods have treated each of the slightly perturbed strains separately. However, it might be expected that in the case that the perturbed strain only slightly varies from the original strain (i.e. only a few genes are knocked out), the interactions (edges) near the knocked out genes will show a significant change in their evolution characteristics while interactions (edges) far from the knocked out genes would have the same evolution characteristics as in the reference strain. The third contribution of this paper is a simple damping model which allows systematic incorporation of the microarray data available across several slightly perturbed strains in the inference algorithm.

This leads to the approximate inference algorithm, NETGEM, which models the gene interactions for a known network in terms of the functional hierarchy of the genes using the expression data over multiple strains. We applied the algorithm to publicly available time-series gene expression data in Saccharomyces cerevisiae. The available of a highly curated interaction network for this organism makes it an ideal platform for testing the method. We selected two time-series datasets in which the nutritional environment changed with time, one without any genetic perturbations and one with a deletion in the $Sfp1$ transcription factor. The first dataset consists of expression of genes during the gradual transition from carbon starvation to nitrogen starvation in a D-stat under aerobic or anaerobic conditions (Farzadfard et al., 2010). Almost a fourth of the genome underwent transcriptional changes in response to the transition. The dominant transcription factor that brought about these changes was $Sfp1$, which is known to assimilate signals from the environment and coordinates growth with metabolism (Marion et al., 2004). The second dataset measures the temporal changes in gene expression upon sudden exposure of a strain of $S.cerevisiae$ in which $Sfp1$ was deleted to glucose (Cipollina et al., 2009).

The remainder of this manuscript is organized as follows: Section 2.2 discusses the construction of the high confidence network, Section 2.3 presents the factorial approximation, Section 2.5 presents the mixture model and Section 2.4 presents the strain damping model.

2 METHODS

2.1 Dataset

Temporal gene expression datasets were downloaded from Gene Expression Omnibus using accession numbers XXXXX and XXXXX. The two datasets were obtained using Affymetrix platform. The first dataset contained the expression profiles of the genes in S. cerevisiae during the transition from carbon limitation to nitrogen limitation under aerobic or anaerobic conditions. The transition was achieved by gradual increment of glucose availability in the feed to the cells, while keeping the nitrogen concentration constant in a D-stat (Farzadfard et al., 2010). Beyond a certain concentration of glucose, nitrogen became the limiting nutrient. The cells underwent changes related to growth rate as well as metabolism. Analysis of genes whose expression significantly changed indicated that Sfp1 transcription factor played a dominant role in the bringing out the response to transition. In the interest of coherence, we chose a dataset that contains the temporal gene expression profiles in sfp1 deletion mutant and its isogenic reference at different time points after pulsing steadily growing cells with glucose. The data was measured at six time points after the pulse. These data were analyzed using conventional methods, assuming that all time points are independent.

2.2 Construction of the interaction network

The yeast interaction network was constructed using data from previously published datasets. Interactions between proteins that occurred in at least two independent datasets were considered. These interactions were downloaded from BIND Bader *et al.* (2003), MIPS Mewes *et al.* (2002), MINT Zanzoni (2002), DIP Xenarios *et al.* (2000) and BioGRID Stark *et al.* (2006) and literature data (5-6 references). The construction of this high-confidence network was described in detail previously (Musigkain et al., 2010). The transcriptional regulatory network (interactions between transcription factors and genes) was downloaded directly from YEASTRACT Teixeira *et al.* (2006). The two networks were combined and the nature of interactions was not distinguished for the analysis.

2.3 Factorial model

We assume that the base underlying network of interactions is known as a graph $G = (V, E)$ as described in the previous section. Under different conditions, some of the edges are switched on or off, or, more generally set at various levels of activation, $\mathcal{W}$. Also, the same edge may be active in one strain and not in others at any given time point. Thus, we model the state of the network by activation levels, $\mathbf{w}^s(t) = \{w_e^s(t)\}_{e\in E}$, where $w_e^s(t)$ is the activation level of the edge e at time t in strain s.

We use the notation $x_e^s(t)$ to denote the expression levels for genes, i and j, consisting the edge, $e = (i, j) \in E$, for strain, s, at time t. Similarly, $x_e^{1:S}(t_a : t_b)$ denotes the observations for gene expression levels for edge, $e = (i, j)$, over the set of strains, $\{1, \ldots, S\}$; for the time interval, $\{t_a, (t_a + 1), \ldots, t_b\}$.

The observed gene expression levels, $\mathbf{x}^s(t)$, for an strain s at time t are modeled as an Ising system Song *et al.* (2009):

$$P\left(\mathbf{x}^s(t)|\mathbf{w}^s(t)\right) = \frac{1}{Z}\exp\left(-\sum_{(i,j)\in E} w_e^s(t)x_i^s(t)x_j^s(t)\right) \quad (1)$$

We assume that the weights evolve according to the markov chain, i.e., $P(\mathbf{w}^s(t+1) = \mathbf{w}_{t+1}|\mathbf{w}^s(t) = \mathbf{w}_t, \mathbf{w}^s(t-1) = \mathbf{w}_{t-1}, \ldots) = P\left(\mathbf{w}^s(t+1) = \mathbf{w}_{t+1}|\mathbf{w}^s(t) = \mathbf{w}_t\right)$ for each strain. However, the strains in the given problem are just slightly altered networks where a few genes have been knocked out of the network. Therefore, most of the network remains the same across strains with only the "close" neighbourhood of the knocked out genes being affected. Thus, if one looks at a "far" edge, e_{far}, the activation strength, $w_{e_{far}}(t)$, should be the same across strains the gene expression data for the edge strains, $x_{e_{far}}^{1:S}(t)$, should be like i.i.d. samples, generated with the same activation strength, $w_{e_{far}}(t)$. In

the following discussion, we present a heuristic method which incorporates the ideas mentioned above into the inference problem.

2.4 Strain Damping Model

We assume that the weights corresponding to the reference strain $\mathbf{w}(t)$ evolve according to a Markov law given by a matrix Q, where $Q(l,m) = P(\mathbf{w}(t+1) = \mathbf{w}_m | \mathbf{w}(t) = \mathbf{w}_l)$ with the property that $\sum_m Q(l,m) = 1$ for all the initial states $\mathbf{w}_l$. For other strains, we assume that the corresponding values are just slightly perturbed; thus

$$w_e^s(t) = w_e(t)\Gamma_e^s \tag{2}$$

The perturbing parameters Γ_e^s are determined deterministically from the underlying network G by

$$\Gamma^s(i,j) = (1-\gamma_i^s)(1-\gamma_j^s) \tag{3}$$

where $\gamma_i^s \in [0,1]$ is a label determined by how far the gene i is in the underlying network to one of the genes knocked out in strain s. We note that the deterministic nature of the damping implies that all strains evolve similarly, i.e., $Q^s = Q \ \forall \ s$. This allows us to incorporate the information for gene expression levels in the different strains while learning the temporal evolution characteristics.

There is a tradeoff between using more sophisticated conditional probability models $p(\mathbf{w}^s(t)|\mathbf{w}^0(t))$ involving more parameters to be learnt and the limited amount of experimental data.

We compute the damping factor, γ_i^s, for the genes as follows: If the gene, i, is knocked out in strain s, then we label it as $\gamma_i^s = 0$. Now, we diffuse the labels across the graph such that $\gamma_i^s = \frac{1}{d(i)}\beta \sum_{j \in N(i)} \gamma_j^s$, i.e., the damping factor at a node is the average of the damping factors at its neighbours.

Intutitively, while $\Gamma_e = 0$ for an edge directly incident to one of the knocked out genes, the perturbation gradually damps out with distance from the knocked out gene and for an edge e far away from one of the knocked out genes, $\Gamma_e \approx 1$.

Thus, the problem is a simple HMM Rabiner (1989) with $\mathbf{x}^{1:S}(t)$ and $\mathbf{w}(t)$ as the observation and the hidden variable at time t, and $Q = P(\mathbf{w}(t+1)|\mathbf{w}(t))$ the unknown parameter to be learnt. We note that an application of the standard forward-backward algorithm to compute the probability distribution over the weight states requires $O(\mathcal{W}^{N_e}T)$ computations, where, $\mathcal{W}$ is the number of possible discrete states for an edge activation strength, N_e is the total number of edges, and T is the time period for which observations are made. This is prohibitively expensive and in the following section, we outline an approximation to solve this problem.

2.4.1 *Factorial approximation*

As noted in the previous section, applying the standard forward backward algorithm is prohibitively expensive for moderate sized graphs. So, we make the simplifying assumption that the weights are evolving *independent* of each other. This leads to the factorial approximation Ghahramani & Jordan (1997); Mclachlan & Krishnan (1996) for weight distribution, i.e.,

$$\hat{P}(\mathbf{w}^t) = \prod_{e \in E} P_e(w_e^t) \tag{4}$$

$$\hat{P}(w_e^{t+1} = w_l | w_e^t = w_m) = q_e(l,m) \tag{5}$$

LEMMA 2.1. *The factorial weights assumption gives a rank-1 tensor approximation Horn & Johnson (1990) to the global weight transition probability.*

For example, if the interaction network consists of two edges having transition probabilities $A = \{a_{ij}\}$ and B respectively, the approximation to the original matrix, $\hat{Q}$ is given as the kronecker product

$$\hat{Q} = A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B \\ a_{21}B & a_{22}B \end{bmatrix}$$

In general, if there are E edges with associated transition probability matrices, $Q_1, \ldots, Q_E$, we obtain the approximation, $\hat{Q}$

$$\hat{Q} = Q_1 \otimes Q_2 \otimes \ldots \otimes Q_E$$

We discuss the quality of approximation in case the original probability matrix Q is a higher order tensor elsewhere.

We use the factorial approximation where the parameter to be learnt is the transition matrix Q_e for each edge, $e = (i,j) \in E$. We solve the Expectation Maximization (EM)Dempster *et al.* (1977) for MAP problem for each edge, e,[1]

$$\begin{aligned} \text{E-step:} \quad \mathcal{L}(Q_e; Q_e^{(n)}) &= E_{w_e^|}[\ln P(\mathbf{x}_e^{1:S}(1:T), \mathbf{w}_e(1:T)|Q_e)] \\ \text{M-step:} \quad \hat{Q}_e^{(n+1)} &= \arg\max_{Q_e}(\ln P(Q_e) + \mathcal{L}(Q_e; Q_e^{(n)})) \end{aligned} \tag{6}$$

where $W_e^|$ is the conditioned variable, $w_e(1:T)|\mathbf{x}_e^{1:S}(1:T), Q_e^{(n)}$ and $Q_e^{(n)}$ is the MAP estimate for the transition probability, Q_e, at the n^{th} iteration of the algorithm. We assume that the data for strain, s, is independently generated based on the Ising model in (1) with the weights that are damped versions (2) of the weights in the original strain. This leads to the observation model specified as:

$$o_e^t(l) = P(x_e^{1:S}(t)|w_e(t) = w_l) \tag{7}$$

$$= \frac{1}{Z}\prod_{s=1}^{S} P(x_e^s(t)|w_e(t) = w_l) \tag{8}$$

$$= \frac{\exp\left\{-w_l\left(\sum_{s=1}^{S} x_i^s(t)x_j^s(t)\Gamma^s(i,j)\right)\right\}}{\sum_{l=1}^{\mathcal{W}} \exp\left\{-w_l\left(\sum_{s=1}^{S} x_i^s(t)x_j^s(t)\Gamma^s(i,j)\right)\right\}} \tag{9}$$

The update equations for computing the forward and backward probability distributions are given as:

$$f_e^{t+1}(m) = P(x_e^{1:S}(1:t), w_e(t+1) = w_m|Q_e^{(n)}) \tag{10}$$

$$= o_e^{t+1}(m)\sum_{l=1}^{\mathcal{W}} f_e^t(l)q_e^{(n)}(l,m) \tag{11}$$

$$b_e^t(l) = P(x_e^{1:S}((t+1):T)|w_e(t) = w_l, Q_e^{(n)}) \tag{12}$$

$$= \sum_{m=1}^{\mathcal{W}} q_e^{(n)}(l,m)o_e^{t+1}(m)b_e^{t+1}(m) \tag{13}$$

and the joint probability as

$$\xi_e^t(l,m) = P(w_e(t,t+1) = (w_l, w_m)|x_e^{1:S}(1:T), Q_e^{(n)}) \tag{14}$$

$$\propto f_e^t(l)q_e^{(n)}(l,m)o_e^{t+1}(m)b^{t+1}(m) \tag{15}$$

Often, there is domain knowledge available which can be incorporated in the form of prior distribution. For example, one may know that most of the edges are inactive about 50% of the time.

2.4.2 *Dirichlet prior*

We model the transition probabilities matrices as dirichlet distributions, such that the prior on the transition probabilities matrix, Q, given the parameter, Θ, is

$$P(\vec{q}_l|\vec{\theta}_l) \sim Dir(q_{l1}, \ldots, q_{l\mathcal{W}}; \theta_{l1}, \ldots, \theta_{l\mathcal{W}}) \tag{16}$$

$$= \frac{1}{B(\vec{\theta}_l)}\prod_{m=1}^{\mathcal{W}} q_{lm}^{\theta_{lm}-1} \tag{17}$$

where $\vec{\theta}_l = [\theta_{l1}, \ldots, \theta_{l\mathcal{W}}]$ and $B(\vec{\theta}_l)$ is the multinomial beta function Gelman *et al.* (2003).

[1] The quality of the factorial approximation is discussed in the appendix

This leads to the update equation for the MAP estimate for transition probabilities, $q(l,m)$, obtained by the maximization step in (6) as

$$q_e^{(n+1)}(l,m) = \frac{(\theta_{lm}-1)+\sum_{t=1}^{T-1}\xi_e^t(l,m)}{\sum_m(\theta_{lm}-1)+\sum_{t=1}^{T-1}\sum_{m=1}^{\mathcal{W}}\xi_e^t(l,m)} \quad (18)$$

2.4.3 Cluster Similarity: We can often group the network interactions into categories based on domain knowledge about the functional classification of genes. For example, one might model the genes that participate in sugar metabolism as one component, while treating genes involved in DNA synthesis as another component. This allows us a simplification that we need to consider only evolution over the *components*(or *clusters*), A_k of edges, which are parameterized by the cluster transition, Q_k for cluster A_k.

Then, the update equations for the transition probability matrix, Q_k, for the cluster, A_k are as follows,

$$q_k^{(n+1)}(l,m) = \frac{(\theta_{lm}^{(k)}-1)+\sum_{e\in A_k}\sum_{t=1}^{T-1}\xi_e^t(l,m)}{(\theta_{ij}^{(k)}-1)+\sum_{e\in A_k}\sum_{t=1}^{T-1}\sum_m\xi_e^t(l,m)} \quad (19)$$

where $\theta^{(k)}$ is the dirichlet parameter matrix for cluster, A_k.

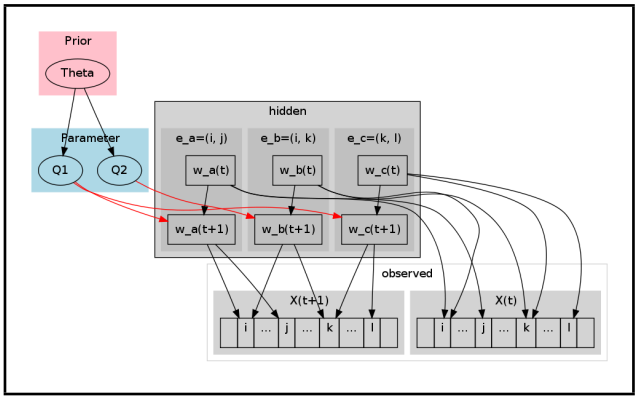

Fig. 1. Single functional classification model. Here, the edges a, c belong to evolution class 1, while edge b belongs to evolution class 2. X_t is the observed gene expression level at time t; $w_e(t)$ is the hidden variable denoting interaction strength for edge e at time t; Q_i are the evolution characteristic for class i; and θ is the prior based on domain knowledge.

Figure 1 shows the graphical model corresponding to the known network model with cluster similarity. We present an extension of the current model to handle multiple functional categories for genes.

2.5 Mixture Model

This section presents our approach to incorporating the functional categories of genes in our analysis. Since, genes belong to multiple categories, a mixture model is a naturally suited model to handle the influence of multiple functional categories in the inference procedure. This allows us to explore the relationship between functional categories and the temporal evolution characteristics of the genes which fall in the same functional category. We now define the problem concretely. There are H possible gene categories. Each gene can be a member of one or more hierarchical classes, $\mathcal{C} = \{C_1,\ldots,C_H\}$, where the hierarchical class C_h is characterized by evolution matrix, Q_h. The evolution probability matrix, Q_e, for each edge, $e \in E$, is given as

$$Q_e = \sum_{h=1}^{H}\alpha_{e,h}Q_h \quad (20)$$

where $\alpha_{e,h}$ denotes the influence of hierarchical class C_h in the edge, e, such that $\sum_h \alpha_{e,h} = 1$ for all edges $e \in E$. We define the random variable $\mathbf{y}^{1:T} = \{y_e^t\}$ for all edges, $e \in E$, and times, $t = \{1,\ldots,T\}$ where y_e^t denotes the component from which the evolution characteristics are chosen at time t for edge e such that the event $y_e^t = h$ implies that $P(w_e^{t+1} = w_m | w_e^t = w_l, y_e^t = h) = q_h(l,m)$.

We now outline the expectation maximization procedure Bilmes (1998); Mclachlan & Krishnan (1996) which iteratively learns the unknown quantity, $\Psi = \{Q_h, \alpha_{e,h}\}$, for $h \in \mathcal{H} = \{1,\ldots,H\}$ and $e \in E$ where Q_h is the class evolution probability matrix for class, C_h, and $\alpha_{e,h}$ is the mixing proportion for edge, e and class C_h; and $\Omega = \{\mathbf{y}^{1:T}, \mathbf{w}^{1:T}\}$ is the hidden variable. Let $\Psi^{(n)} = \{\alpha_{e,h}^{(n)}, Q_h^{(n)}\}$ be the estimates at the n^{th} iteration. Then,

$$\begin{array}{lll} \text{E-step:} & \mathcal{L}(\Psi;\Psi^{(n)}) & = E_{\Omega|}[\ln P(\mathbf{x}^{1:S}(1:T),\Omega(1:T)|\Psi(1:T))] \\ \text{M-step:} & \hat{\Psi}^{(n+1)} & = \arg\max_\Psi(\ln P(\Psi)+\mathcal{L}(\Psi;\Psi^{(n)})) \end{array} \quad (21)$$

where $\Omega|$ is the conditioned variable $(\mathbf{w}^{1:T}, \mathbf{y}^{1:T}|\mathbf{x}^{1:S}(1:T), \Psi^{(n)})$.

The factorial approximation in the previous section allows us to compute the probability distribution over the edges independently. [2] The observation model, $o_e^t(l) = P(x_e^{1:S}(t)|w_e^t = w_l)$, remains unchanged as in (7)-(9). The forward iterates, $f_e^t(l,h)$ and backward iterates, $b_e^t(l,h)$ can be computed as follows:

$$\begin{aligned} f_e^t(m,h) &= P(x_e^{1:S}(1:t), w_e^t = w_m, y_e^t = h|\Psi_e^{(n)}) && (22)\\ &= P(x_e^{1:S}(t)|w_m)\sum_{w_l}\sum_{h'}\Big[P(y_e^t = h|\alpha^{(n)}) \\ &\times P(w_m|w_e^{t-1} = w_l, y_e^{t-1} = h')\times f_e^{t-1}(l,h')\Big] && (23)\\ &= o_e^t(m)\sum_{l=1}^{\mathcal{W}}\sum_{h'=1}^{H} f_e^{t-1}(l,h')\alpha_h^{(n)}q_{h'}^{(n)}(l,m) && (24)\\ b_e^t(m,h) &= P(x_e^{1:S}((t+1):T)|w_e(t) = w_m, y_e^t = h, \Psi_e^{(n)}) && (25)\\ &= \sum_{w_l}\sum_{h'}\Big[P(x_e^{1:S}(t+1)|w_e^{t+1} = w_l)b_e^{t+1}(l,h') \\ &\times P(w_e^{t+1} = w_l|w_m, y_e^t = h)P(y_e^{t+1} = h'|\alpha^{(n)})\Big] && (26)\\ &= \sum_{m=1}^{\mathcal{W}}\sum_{h'=1}^{H} q_h^{(n)}(m,l)o_e^{t+1}(l)\alpha_{h'}^{(n)}b_e^{t+1}(l,h') && (27) \end{aligned}$$

The conditional probability $P(\Omega_e^t = (w_l,h), \Omega_e^{t+1} = (w_m,h')|\mathbf{x}_e^{1:S}(1:T), \Psi^{(n)})$ denoted by $\xi_e^t(l,m,h,h')$ can be computed as

$$\xi_e^t(l,m,h,h') \propto f_e^t(l,h)\alpha_{h'}^{(n)}q_h^{(n)}(l,m)o_e^{t+1}(m)b_e^{t+1}(m,h') \quad (28)$$

The likelihood term, $\mathcal{L}(\Psi;\Psi^{(n)})$, in (21) can be expressed in terms of the conditioned edge probabilities, ξ_e^t, in (28) as

$$\mathcal{L}(\Psi;\Psi^{(n)}) = \sum_{e\in E}\sum_{t=1}^{T-1}\mathbf{E}_{\xi_e^t}[\ln q_h(l,m) + \ln\alpha_{e,h'}] \quad (29)$$

subject to the constraints

$$\sum_m q_h(l,m) = 1 \quad \forall h \quad (30)$$

$$\sum_h \alpha_{e,h} = 1 \quad \forall e \quad (31)$$

2.5.1 Domain knowledge: We incorporate the effect of the functional classification of genes on the mixture components, $\vec{\alpha}_e$, for an edge, e, by

[2] Part of this section may be moved to appendix.

using a dirichlet prior of the form:

$$P(\vec{\alpha}_e) \quad \sim \quad Dir(\alpha_{e,1},\ldots,\alpha_{e,H};\gamma_{e,1},\ldots,\gamma_{e,H}) \tag{32}$$

with the prior parameter, $\gamma_{e,h}$, for the edge, $e=(i,j)$, of the form

$$\gamma_{e,h} = \begin{cases} \gamma_p & \text{if genes } i \text{ or } j \text{ in class } h \\ \gamma_o & \text{otherwise} \end{cases} \tag{33}$$

The maximization step in (21) can be done separately for $q_h(l,m)$ and $\alpha_{e,h'}$ independently. We use the priors in (16) and (32)-(33), and the constraints in (30)-(31) to obtain the following update equations:

$$\alpha_{e,h'}^{(n+1)} = \frac{(\gamma_{e,h'}-1)+\sum_{t=1}^{T-1}\sum_{l,m,h}\xi_e^t(l,m,h,h')}{\sum_{h'}(\gamma_{e,h'}-1)+\sum_{t=1}^{T-1}\sum_{l,m,h,h'}\xi_e^t(l,m,h,h')} \tag{34}$$

$$q_h^{(n+1)}(l,m) = \frac{(\theta_{lm}-1)+\sum_e\sum_{t=1}^{T-1}\sum_{h'}\xi_e^t(l,m,h,h')}{\sum_m(\theta_{lm}-1)+\sum_e\sum_{t=1}^{T-1}\sum_{m,h'}\xi_e^t(l,m,h,h')} \tag{35}$$

3 EXPERIMENTS

We present the experiments performed on synthetic and actual datasets in this section. We compare the factorial weights and the mixture model against a standard implementation of HMM which learn the evolution characteristic for the system jointly in 3.1 and present the results on a synthetic dataset in 3.2.

3.1 Validation

We choose a synthetic graph with $N=5$ nodes (genes) and $H=10$ functional classes. The small value of N allows direct estimation of the transition probability matrix, Q, of size $\mathcal{W}^N \times \mathcal{W}^N$, using an standard implementation of HMM[3].

The genes are randomly assigned classes such that each node is a member of $N_{av}=1.5$ classes on average. The weights take possible values in $\mathcal{W}=\{-1,1\}$, and the evolution characteristic, Q_e for each weight is a mixture based on the interacting genes. We compare the rank-1 approximations to the transition probability matrix of the original HMM, $\hat{Q}_{large}$, obtained by our methods to the estimated transition probability matrix obtained from standard implementation of HMM for $N=20$ random trials.

Figure 2 shows the comparison between different methods with increasing length of the sequence, T. We note that the standard HMM requires longer sequences to get comparable results with factorial and mixture model approach. Figure 3 compares the methods with increasing number of strains present (with at most 1 gene knocked out). We note that the performance of the factorial weights assumption shows a slight improvement with increasing number of strains.

3.2 Synthetic dataset

We generate a "random" graph $G=(V,E)$ with C major components, $\{A_1,\ldots,A_C\}$ with input parameters p_i and p_c, where p_i is the probability, of an edge between two vertices in the same component, and p_c is the probability of an edge between two vertices in different components.

The activation level, $w_e(t)$ defined on an edge, e, belonging to component, A_k, is a markov chain with transition probability matrix Q_k. The problem is the estimation of the unknown transition probability matrices Q_k for each component, A_k.

[3] http://people.cs.ubc.ca/ murphyk/Software/HMM/hmm.html

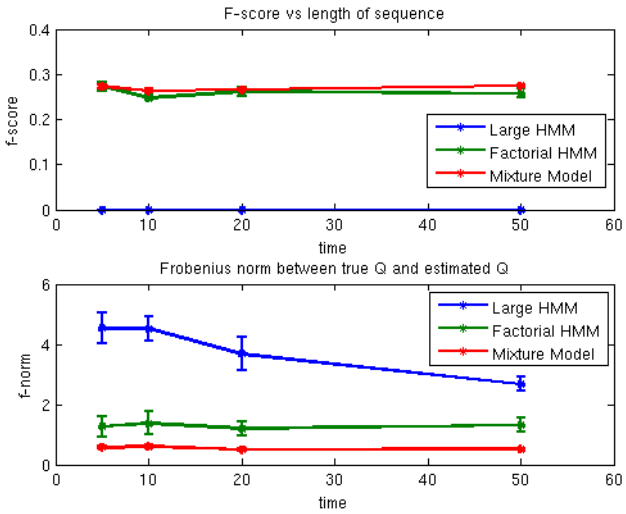

Fig. 2. (a) F-score for the estimated weight evolution vs the true weight evolution sequence with increasing length of sequence, T. (b) Frobenius norm of the difference between the true evolution characteristics for the system and the estimated characteristics.

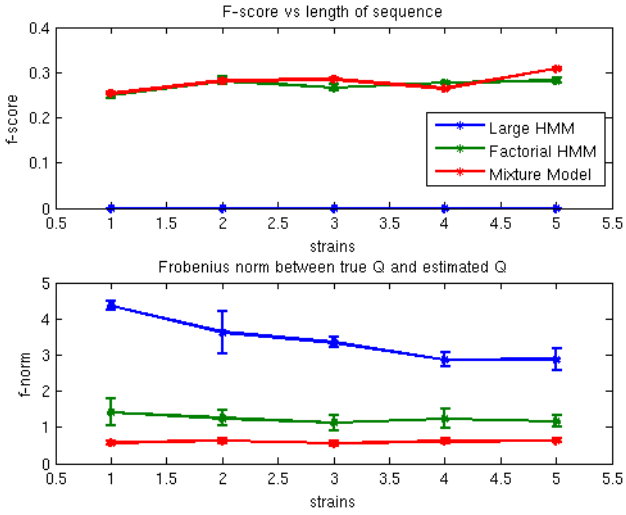

Fig. 3. (a) F-score for the estimated weight evolution vs the true weight evolution sequence with increasing number of strains. (b) Frobenius norm of the difference between the true evolution characteristics for the system and the estimated characteristics.

We use a noisy dirichlet prior for the estimation as follows:

$$\Theta^{(k)} = Q_k + \mathcal{N}(0,\sigma^2 I) \tag{36}$$

The experiment is conducted for 20 trials with a graph of size $N=50$ and number of components, C chosen randomly between 2 and 10. Figure 4 shows the F-scores for the experiments done with multiple number of strains.

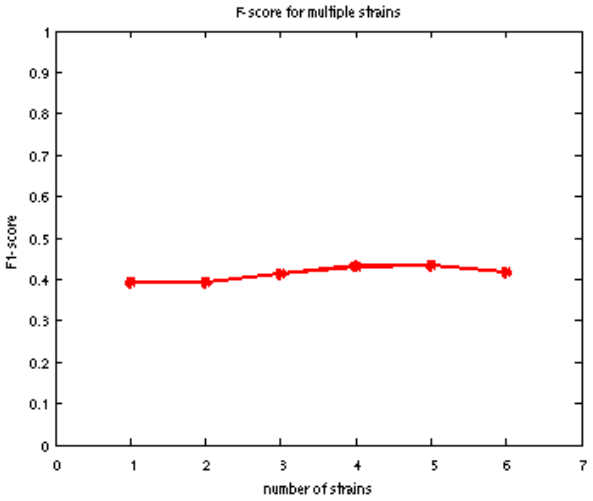

Fig. 4. F-score for the synthetic dataset for multiple strains. We observe that increase in the number of strains provides more information about the activation strengths in the original network, which is visible in the slight increase in the F1-scores. However, since genes are being knocked out in each of the strains, the expression levels over multiple strains are not equivalent to i.i.d. samples.

4 CONCLUSION

REFERENCES

Androulakis, I. P., Yang, E. & Almon, R. R. (2007) Analysis of time-series gene expression data: methods, challenges, and opportunities. *Annual Review of Biomedical Engineering,* **9** (1), 205–228.

Bader, G. D., Betel, D. & Hogue, C. W. (2003) Bind: the biomolecular interaction network database. *Nucleic acids research,* **31** (1), 248–250.

Beal, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London.

Bilmes, J. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report University of Washington.

Cappé, O., Moulines, E. & Ryden, T. (2007) *Inference in Hidden Markov Models*. Springer Series in Statistics, Springer.

Costa, I. G., Schönhuth, A. & Schliep, A. (2005) The graphical query language: a tool for analysis of gene expression time-courses. *Bioinformatics,* **21** (10), 2544–2545.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological),* **39** (1), 1–38.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America,* **95** (25), 14863–14868.

Ernst, J. & Joseph, Z. B. (2006) Stem: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics,* **7** (1).

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2003) *Bayesian Data Analysis, Second Edition (Texts in Statistical Science)*. 2 edition,, Chapman & Hall/CRC.

Ghahramani, Z. & Jordan, M. I. (1997) Factorial hidden markov models. *Machine Learning,* **29** (2-3), 245–273.

Horn, R. A. & Johnson, C. R. (1990) *Matrix Analysis*. Cambridge University Press.

Leek, J. T., Monsen, E., Dabney, A. R. & Storey, J. D. (2006) EDGE: extraction and analysis of differential gene expression. *Bioinformatics,* **22** (4), 507–508.

Mclachlan, G. J. & Krishnan, T. (1996) *The EM Algorithm and Extensions*. 1 edition,, Wiley-Interscience.

Mewes, H. W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkötter, M., Rudd, S. & Weil, B. (2002) Mips: a database for genomes and protein sequences. *Nucleic acids research,* **30** (1), 31–34.

Nam, H., Lee, K. & Lee, D. (2009) Identification of temporal association rules from time-series microarray data sets. *BMC Bioinformatics,* **10** (Suppl 3), S6+.

Rabiner, L. R. (1989) A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE,* **77** (2), 257–286.

Ramoni, M. F., Sebastiani, P. & Kohane, I. S. (2002) Cluster analysis of gene expression dynamics. *Proc Natl Acad Sci U S A,* **99** (14), 9121–9126.

Rifkin, S. A. & Kim, J. (2002) Geometry of gene expression dynamics. *Bioinformatics,* **18** (9), 1176–1183.

Schliep, A., Schönhuth, A. & Steinhoff, C. (2003) Using hidden markov models to analyze gene expression time course data. In *ISMB (Supplement of Bioinformatics)* pp. 255–263.

Schliep, A., Steinhoff, C. & Schönhuth, A. (2004) Robust inference of groups in gene expression time-courses using mixtures of hmms. In *ISMB/ECCB (Supplement of Bioinformatics)* pp. 283–289.

Song, L., Kolar, M. & Xing, E. P. (2009) Keller: estimating time-varying interactions between genes. *Bioinformatics,* **25** (12).

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A. & Tyers, M. (2006) Biogrid: a general repository for interaction datasets. *Nucleic Acids Res,* **34** (Database issue).

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Dmitrovsky, S. K. E., Lander, E. S. & Golub, T. R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America,* **96** (6), 2907–2912.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999) Systematic determination of genetic network architecture. *Nature genetics,* **22** (3), 281–285.

Teixeira, M. C., Monteiro, P., Jain, P., Tenreiro, S., Fernandes, A. R., Mira, N. P., Alenquer, M., Freitas, A. T., Oliveira, A. L. & Sá-Correia, I. (2006) The yeastract database: a tool for the analysis of transcription regulatory associations in saccharomyces cerevisiae. *Nucleic Acids Res,* **34** (Database issue).

Tusher, V. G., Tibshirani, R. & Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America,* **98** (9), 5116–5121.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M. & Eisenberg, D. (2000) Dip: the database of interacting proteins. *Nucl. Acids Res.,* **28** (1), 289–291.

Yoneya, T. & Mamitsuka, H. (2007) A hidden markov model-based approach for identifying timing differences in gene expression under different experimental factors. *Bioinformatics,* **23** (7).

Zanzoni, A. (2002) Mint: a molecular interaction database. *FEBS Letters,* **513** (1), 135–140.

1 FACTORIAL MODEL DERIVATION

2 MIXTURE MODEL DERIVATION