BIOINFORMATICS

Vol. 00 no. 00 2010
Pages 1–9

NETGEM: Network Embedded analysis of Temporal Gene Expression Model

Vinay Jethava[1], 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 Division, 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 Temporal analysis of gene expression data has been limited to identifying genes whose expression varies with time and/or correlation between genes what have similar temporal profiles. Often, the methods do not consider the underlying network constraints that connect the genes. In addition to identifying changes in the genes, it is becoming increasingly evident that interactions change substantially. Thus far, there is no systematic method to relate the temporal changes in gene expression to the dynamics of interactions between them in the context of a regulatory network. The availability of this data opens up possibilities for discovering new mechanisms of regulation and provides valuable insight into identifying time-sensitive interactions. Furthermore, such a framework would also allow for studies on the effect of a genetic perturbation on the dynamics of the interactions.

Results We present NETGEM, a tractable model rooted in Markov dynamics, for analyzing temporal profiles of genetic expressions arising out of known protein interaction networks evolving with unknown dynamics. The model treats the interaction strengths as random variables which are modulated by suitable priors. This approach is necessitated by the extremely small sample size of the available observations. The model is amenable to a linear time algorithm for efficient inference. When applied to real data NETGEM was successful in identifying (i) temporal interactions and determining their strength, (ii) functional categories of the actively interacting partners and (iii) dynamics of interactions in perturbed networks.

Availability: The source code for NETGEM is available from http://www.sysbio.se/BioMet

Contact: goutham@chalmers.se

1 INTRODUCTION

Gene expression microarrays are increasingly being used to determine transcriptional regulation in response to a genetic or environmental perturbation. This has generated a vast amount of quantitative data which have compelled the use of statistical methods to identify differentially expressed genes and thereby, deduce transcriptional regulation. The general theme of these methods is to cluster genes, assuming that co-expressed genes are co-regulated. Often the inference is presented as a static network of genes that are activated or repressed by relevant transcription factors. This representation is similar to a wiring diagram of electrical circuits (Stigler *et al.*, 2007). Important parameters of regulation such as amplitude of the signal, time lag, etc in such networks can only be studied by explicitly modelling the dynamics of such a system. This has spurred interest in analyzing time series gene expression data.

Conventional methods of time series analysis may not apply to this problem as the data has many interesting properties e.g. observations close together in time are more closely related (Glass & Kaplan., 1993). The analysis is further complicated by the fact that extremely small number of observations from different time points are available relative to variables (genes). There is the inherent risk of many genes having similar expression profile, just by random chance. Recognizing these problems, it is only recently that dedicated methods are being developed to infer temporal regulation of transcription (Leek *et al.*, 2006; Ernst & Joseph, 2006; Ramoni *et al.*, 2002), although temporal gene expression data are available much earlier. These, and other methods reviewed recently (Androulakis *et al.*, 2007) do not consider any dependency of observations between time points and hence are not suitable for the problem at hand.

Previous methods which have focused on identifying temporal changes in the genes and/or identifying correlations in their temporal expression profiles have ignored the dynamics of interactions between them. In recent work (Song *et al.*, 2009), time-sensitive interactions were identified based on local neighbourhood selection with $L1$-regularization to obtain sparse networks. The analysis learns the topology of the network from the data, and assumes a smooth variation in the network interactions strengths to overcome the unreliability of results due to the small number of observations. This approach is extremely insightful in cases where the topology of the underlying network is unknown.

However, when network structure is known, as in the case of regulatory networks, there is an obvious benefit in incorporating this information into analyzing the dynamics of the interactions. Furthermore, it is of fundamental biological interest in determining

*to whom correspondence should be addressed time-sensitive components of the networks. Currently there are no models which can be used for this purpose.

The interaction networks of baker's yeast, *Saccharomyces cerevisiae* are arguably the most well-constructed with a high level of confidence (Petranovic & Vemuri, 2009). Therefore, we used this network to study and validate our models. The genes and proteins in yeast are classified according to their biological function (Mewes *et al.*, 2007) to a high degree of resolution. This allows the possibility to relate functional classification of the network components with the temporal interactions between them.

This line of argumentation leads to two very fundamental questions: (i) can we distill observations about temporal characteristics of a group of functionally similar genes? (ii) would it be possible to model the effect of a genetic perturbation (gene deletion or addition) while comparing temporal interactions between the reference strain and its perturbed mutant?

In this paper, we introduce NETGEM, which stands for "Network Embedded analysis of Temporal Gene Expression Model", a generative model for temporal expression data which is capable of capturing the interaction dynamics in the network. Our approach incorporates network effects into a basic underlying Markovian dynamics, and also handles variation across closely related species. A fundamental premise of the model is that the evolution of the interaction strengths can be modeled in terms of the functional categories of the interacting genes. To the best of our knowledge, this is the first time such a model has been investigated. Our case studies with synthetic and real data show promising results.

One could of course consider a simple Hidden Markov Model (HMM) for modeling the interaction dynamics. Unfortunately, such a model becomes intractable as the number of hidden states is exponential in the number of interactions in the network. The problem of learning such a model is further complicated by the small number of available observation samples. NETGEM addresses this problem by assuming that the interaction strengths evolve *independently* of each other. This assumption leads to a model where one can derive efficient inference procedures which have linear time complexity in number of temporal observations. To handle the problem of low sample size, we adopt a Bayesian approach by introducing appropriate priors over the parameters governing the evolution of the interactions. Information from multiple strains which are slight perturbations of a reference strain are incorporated by effects determined by the underlying network. The basic assumption is that interactions near the point of perturbation (gene deletions) are affected more strongly than those far away from the point of intervention.

The remainder of this manuscript is organized as follows: Section 2.1.2 describes the construction of the high confidence network. Section 2.3 presents the independently evolving weights assumption for tractable inference. Section 2.4 presents the variant for handling multiple strains. Section 2.6 presents the main contribution of the paper, i.e., the generative model NETGEM. We present the experiments on synthetic and real datasets in Section 3 and conclude in Section 4.

2 METHODS

2.1 Datasets and Interaction network

2.1.1 Gene expression Data Temporal gene expression datasets from *Saccharomyces cerevisiae* were downloaded from Gene Expression Omnibus using accession numbers XXXXX and GSE9644 . The two datasets were obtained using Affymetrix platform. The first dataset contained the expression profiles of the genes in *S. cerevisiae* during the gradual increment in the availability of glucose. Therefore, the cells experience a nutrient transition from glucose starvation to nitrogen starvation. 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). The data was measured at eight time points. Beyond a certain concentration of glucose, nitrogen became the limiting nutrient. The cells underwent changes related to growth rate as well as metabolism. This dataset is referred to as EXP1 in this manuscript. 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 (Cipollina *et al.*, 2008). The reference strain is referred to as REF, the strain in which *sfp1* was deleted is referred to as MUT. The data was measured at six time points after the pulse. This dataset is referred to as EXP2 in this manuscript. All raw data was normalized and preprocessed in R programming environment using BioConductor suite of tools (Gentleman *et al.*, 2004).

2.1.2 Construction of the interaction network The yeast interaction network was constructed using previously published data (Gavin *et al.*, 2002; Ho *et al.*, 2002; Gavin *et al.*, 2006; Krogan *et al.*, 2006; Ito *et al.*, 2001; Uetz *et al.*, 2000) as well as data 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). We used only those interactions that were backed by at least two independent sources, resulting in a high-confidence protein interaction network. We excluded protein-DNA interactions since the result of this interaction is the gene expression and including these interactions would result in a cyclic relationship between the interactions and gene expression. We overlaid the gene expression data onto the protein interaction network. Therefore, inherent in this is the assumption that gene expression is translated into protein abundance uniformly among all proteins.

2.2 Model description

This section presents the basic observation model which relates the observed gene expression data to the high confidence interactions network. We begin by defining the following notation,

2.2.1 Notation 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, $\{t_a, (t_a + 1), \ldots, t_b\}$.

In the following sections, we will describe the building blocks of our generative model, NETGEM. We begin by describing the overall process dynamics as follows.

2.2.2 Observation model 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):

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

where $Z(t)$ is the normalization constant.

2.2.3 Evolution model We assume that the weights evolve according to the Markov chain, i.e.,

$$P(\mathbf{w}(t+1) = \mathbf{w}_{t+1}|\mathbf{w}(t) = \mathbf{w}_t) = \mathbf{Q}(\mathbf{w}_t, \mathbf{w}_{t+1}) \quad (2)$$

where $\mathbf{Q}(\mathbf{w}_t, \mathbf{w}_{t+1})$ is the probability of the transition from state $\mathbf{w}_t$ at time t to state $\mathbf{w}_{t+1}$ at time $(t+1)$. In general, if there are S strains present, then each will have corresponding transition probability matrix $\mathbf{Q}_s$.

The problem, then, is to estimate the transition probability matrix for strains, $\mathbf{Q}_s$, based on the observed gene expression values over multiple strains, $\mathbf{x}^{1:S}(t)$, are the observed variables and the strengths of the interaction network, $\mathbf{w}^s(t)$, is the hidden variable at time t. A direct HMM based approach requires $O(\mathcal{W}^{2N_e}T)$ 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 for most practical problems.

2.3 Independent weights dynamics

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 allows the characterization of the overall probability distribution in terms of the interaction strengths,

$$\hat{P}(\mathbf{w}^t) = \prod_{e\in E} P_e(w^t_e) \quad (3)$$

$$\hat{P}(w^{t+1}_e = w_l | w^t_e = w_m) = q_e(l, m) \quad (4)$$

This 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.

2.4 Analysis of Perturbed Networks

We consider the problem of multiple strains which are just slightly altered versions of the 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. 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

$$w^s_e(t) = w_e(t)\Gamma^s_e \quad (5)$$

The perturbing parameters Γ^s_e are determined deterministically from the underlying network G by

$$\Gamma^s(i,j) = (1-\gamma^s_i)(1-\gamma^s_j) \quad (6)$$

where $\gamma^s_i \in [0,1]$ 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.

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

Intuitively, 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, $\Gamma_e \approx 1$.

2.5 Incorporating functional categories via mixtures

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

$$Q_e = \sum_{h=1}^{H} \alpha_{e,h} Q_h \quad (7)$$

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 $e \in E$.

2.5.1 Prior on transition probabilities, (Θ) The learning of the model parameter is significantly difficult when there are few time points. To alleviate this, we propose a prior, Θ, on the transition probabilities, Q_h. The individual rows, $\vec{q}_l$, of the transition probability , Q_h, can be thought about as drawn from a multinomial distribution. The Dirichlet distribution is the conjugate distribution (Gelman *et al.*, 2003) of the multinomial distribution and hence naturally suited as a prior distribution. We model the transition probabilities matrices as Dirichlet distributions, such that the prior on the transition probabilities matrix, Q, given the parameter, Θ, is

$$P(\vec{q}_l|\vec{\theta}_l) \sim Dir(q_{l1}, \ldots, q_{l\mathcal{W}}; \theta_{l1}, \ldots, \theta_{l\mathcal{W}}) \quad (8)$$

$$= \frac{1}{B(\vec{\theta}_l)} \prod_{m=1}^{\mathcal{W}} q_{lm}^{\theta_{lm}-1} \quad (9)$$

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).

2.5.2 Choice of prior parameters (Λ) We incorporate the effect of the functional classification of genes on the mixture components, $\vec{\alpha}_e$, for an edge, e, by using a Dirichlet prior of the form:

$$P(\vec{\alpha}_e) \sim Dir(\alpha_{e,1}, \ldots, \alpha_{e,H}; \lambda_{e,1}, \ldots, \lambda_{e,H}) \quad (10)$$

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

$$\lambda_{e,h} = \begin{cases} \lambda_p & \text{if genes } i \text{ or } j \text{ in class } h \\ \lambda_o & \text{otherwise} \end{cases} \quad (11)$$

2.5.3 Functional Category (Y) We define the random variable $Y_e(t)$ which denotes the functional category chosen at time t for edge e; such that the event $Y^t_e = h$ implies that $P(w^{t+1}_e = w_m|w^t_e = w_l, y^t_e = h) = q_h(l,m)$.

2.6 NETGEM: a generative model

We now present a unifying view of the NETGEM model as a generative probabilistic model for gene expression data, $\mathbf{X}^{1:S}(1:T)$, for multiple strains $\{1, \ldots, S\}$ and observation time points $\{1, \ldots, T\}$. The quantities known apriori are the number of classes, H; the set of edges, E; the number of strains, S, and the number of observation points T.

Figure 1 shows the graphical model corresponding to the NETGEM model. The inference is done over the hidden variables $\Omega = \{y_e(t), w_e(t)\}$; and the parameters to be learnt are $\Psi = \{Q_h, \alpha_{e,h}\}$.

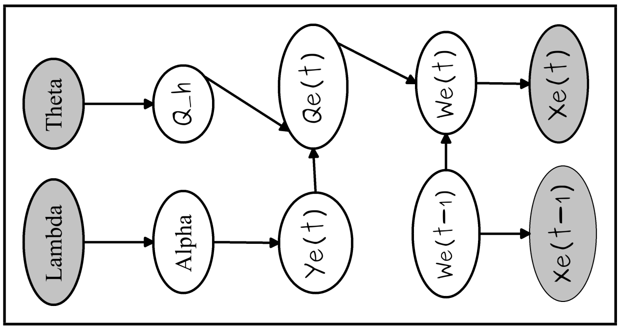

Fig. 1. Graphical model corresponding to the NETGEM generative model. The shaded nodes are known/observed. The inference is done over the hidden variables $\Omega = \{y_e(t), w_e(t)\}$; and the parameters to be learnt are $\Psi = \{Q_h, \alpha_{e,h}\}$

In addition, the hyper-parameters, Θ and Λ need to be specified which are described in Sections 2.5.1 and 2.5.2 respectively. Then, NETGEM models the following generative process:

1. Initialization

   a. For each functional category $h \in H$; Choose $Q_h|\Theta$, s.t. each row, $\vec{q}_i \sim Dir(\vec{\theta}_i)$

   b. For each edge $e \in E$; Choose mixture components, $\alpha|\Lambda$ s.t. $\vec{\alpha}_e \sim Dir(\vec{\lambda}_e)$

2. For each time, $t \in \{1, \ldots, T\}$

   a. For each edge $e \in E$

      (1) Choose $Y_e(t) \sim P(Y_e(t) = h) = \alpha_{e,h}$, The choice of $Y_e(t)$ fixes $Q_e(t)$ as $Q_e(t)|\{Y_e(t) = h\} = Q_h$

      (2) Choose $w_e(t) \sim P(w_e(t)|w_e(t-1), Q_e(t-1))$ as in (4)

      (3) Compute $w_e^s(t) = \Gamma_e^s w_e(t)$

   b. Choose $X^s(t)|W^s(t)$ as in (1)

The key principle underlying the model is that the interaction dynamics are governed by the functional categories. In particular, for each edge e and time t, we generate a functional category, $Y_e(t)$. This functional category is *solely* responsible for the change of interaction strength from $w_e(t)$ to $w_e(t+1)$. The interaction strengths affect the observed gene expression data, through the strain damping model and the Ising probability distribution (1). On the whole, this is equivalent to a mixture of the corresponding functional categories with appropriate mixing proportions.

We now outline the expectation maximization procedure (Dempster *et al.*, 1977) 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,

$$\begin{array}{lllll} \text{E-step:} & \mathcal{L}(\Psi; \Psi^{(n)}) & = & E_{\Omega^|}[\ln P(\mathbf{x}^{1:S}(1:T), \Omega(1:T)|\Psi(1:T))] \\ \text{M-step:} & \hat{\Psi}^{(n+1)} & = & \arg\max_\Psi (\ln P(\Psi) + \mathcal{L}(\Psi; \Psi^{(n)})) \end{array} \tag{12}$$

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

The maximization step in (12) can be done separately for $q_h(l, m)$ and $\alpha_{e,h'}$ independently. We use the priors in (8) and (10)-(11), and the constraints

$$\sum_m q_h(l, m) \quad = \quad 1 \quad \forall h \tag{13}$$

$$\sum_h \alpha_{e,h} \quad = \quad 1 \quad \forall e \tag{14}$$

to obtain the following update equations:

$$\alpha_{e,h'}^{(n+1)} = \frac{(\lambda_{e,h'}-1)+\sum_{t=1}^{T-1}\sum_{l,m,h}\xi_e^t(l,m,h,h')}{\sum_{h'}(\lambda_{e,h'}-1)+\sum_{t=1}^{T-1}\sum_{l,m,h,h'}\xi_e^t(l,m,h,h')} \tag{15}$$

$$q_h^{(n+1)}(l, m) = \frac{(\theta_{lm}-1)+\sum_e\sum_{t=1}^{T-1}\sum_{h'}\xi_e^t(l,m,h,h')}{\sum_m(\theta_{lm}-1)+\sum_e\sum_{t=1}^{T-1}\sum_{m,h'}\xi_e^t(l,m,h,h')} \tag{16}$$

where $\xi_e^t(l, m, h, h')$ is defined in appendix 1.

2.6.1 Simplification: Independent weights dynamics model We now consider the simplification that each edge has its own evolution characterization. In other words, suppose there are $H = E$ functional classes with corresponding priors Λ defined such that each row $\vec{\lambda}_e$ corresponding to edge e has only one zero entry. The inference in such a model is considerably simpler since there is no learning of mixture components, $\alpha_{e,h}$.

We call this model the independent weights dynamic, (IWD), model and present the derivation for the inference and parameter learning in appendix 2.

2.6.2 Relation to the multiple strains The multiple strains in section 2.4 effect the probability $P(X^s|W^0)$. If the strains are all identical (Section 2.3), then, the observations $x_i^s(t)$ can be considered as i.i.d. samples, i.e., the damping factor $\Gamma^s(i, j) = 1$ for all i, j.

3 IMPLEMENTATION

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. We present the results on two genetic expressions datasets in 3.2.

3.1 Validation of the models on synthetic datasets

We consider a model, M, which provides estimates $\{\hat{\mathbf{Q}}^M\}$ and $\{\hat{\mathbf{w}}_e^M(t)\}$ for the true transition probabilities $\mathbf{Q}$ and the actual interaction strengths $\{w_e(t)\}$ for the edges $e \in E$ for the time period $t \in T$. Then, a standard hidden Markov model attempts to estimate $\hat{\mathbf{Q}}^{HMM}$ directly. Our methods, respectively, the independent weights dynamics (IWE) model and the generative model (NETGEM), estimate the transition probabilities, $\hat{Q}_e$ defined over the edges $e \in E$ and then construct a rank-1 approximation, $\hat{\mathbf{Q}}^{IWE}$ and $\hat{\mathbf{Q}}^{NETGEM1}$.

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, $\hat{\mathbf{Q}}^{HMM}$, of size $\mathcal{W}^N \times \mathcal{W}^N$, using an standard implementation of HMM. 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 our the quality of the results obtained by our models against the results obtained by the HMM for $N = 20$ random trials.

We use two metrics to compare our results with respect to the standard hidden Markov model implementation over the large state space, namely,

[1] Supplemental material

1. *F1-score:* F1-score is the harmonic mean of the precision, P, and recall, R , i.e., $F1 = \frac{2PR}{(P+R)}$. In the context of our problem, the precision, P_M, and recall, R_M, for a model, M, are defined as

$$P_M = \frac{\sum_{e\in E}\sum_{t=1}^{T}\mathbf{1}_{\hat{w}_e^M(t)=1}\mathbf{1}_{w_e(t)=1}}{\sum_{e\in E}\sum_{t=1}^{T}\mathbf{1}_{w_e(t)=1}} \tag{17}$$

$$R_M = \frac{\sum_{e\in E}\sum_{t=1}^{T}\mathbf{1}_{\hat{w}_e^M(t)=1}\mathbf{1}_{w_e(t)=1}}{\sum_{e\in E}\sum_{t=1}^{T}\mathbf{1}_{\hat{w}_e^M(t)=1}} \tag{18}$$

   where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. Thus, F-score measures the accuracy of the predictions about strengths of interactions made by the model.

2. *Frobenius norm:* We measure the Frobenius norm between the estimated (by model M) transition probability, $\hat{\mathbf{Q}}^M$ and the true transition probability, $\mathbf{Q}$ used to generate the data, as

$$||\hat{\mathbf{Q}}^M - \mathbf{Q}||_F = \sum_{i,j}|q_{ij}^M - q_{ij}|^2 \tag{19}$$

   where q_{ij} denotes the $(i,j)^{th}$ element of the matrix $\mathbf{Q}$. Thus, this measures the accuracy with which we can estimate the transition probability, Q.

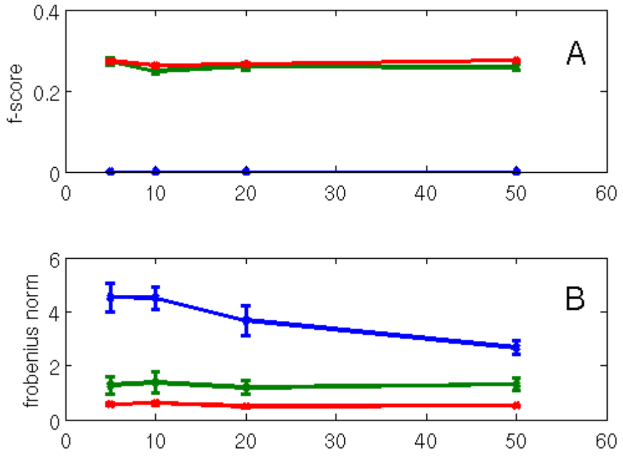

Fig. 2. This figure compare the results obtained using the Independent weights dynamics model (IWE) and the generative model (NETGEM) with result of an standard HMM implementation with increasing number of observations (time points), T, available for 20 random trials. The colors denote HMM (blue), IWE (green), NETGEM (red) (a) F-score for the estimated weight evolution *versus* the true weight evolution sequence (b) Frobenius norm of the difference between the true evolution characteristics for the system and the estimated characteristics.

Figure 2 shows the comparison between different methods with increasing length of the sequence, T. The results indicate that the standard HMM requires longer sequences to get comparable results with the IWD and NETGEM approach.

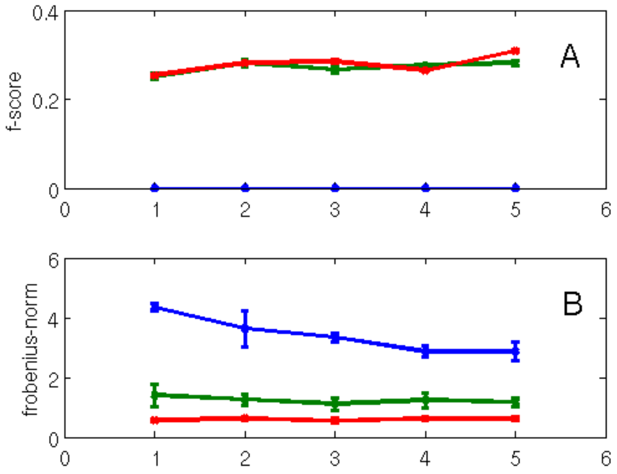

Fig. 3. This figure compare the results obtained using the Independent weights dynamics model (IWE) and the Mixture Model (NETGEM) with result of an standard HMM implementation with increasing number of strains, S, available for 20 random trials. The colors denote HMM (blue), IWD (green), NETGEM (red). (a) F-score for the estimated weight evolution vs the true weight evolution sequence (b) Frobenius norm of the difference between the true evolution characteristics for the system and the estimated characteristics.

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 models improves slightly with additional number of strains.

3.2 Inferring interaction dynamics from gene expression data

This section presents the experiments on real genetic expression datasets. We begin by describing a procedure for identifying the interactions which show a high degree of temporal variation in their strengths.

3.2.1 Identifying time-sensitive interactions We apply our algorithm to infer the interaction strengths for two genetic network scenarios. We are interested in the edges which show considerable change during the measured time points. Towards this end, we compute the change score, $s(e)$, of an interaction strength $w_e(1:T)$ as follows

$$s(e) = \frac{1}{T}\sum_{t=1}^{T-1}(w_e(t+1) - w_e(t))^2 \tag{20}$$

We fit an exponential distribution to it and consider the weights falling in the top-5% ($p = 0.05$) tail of the distribution. The interaction between the nodes was visualized as a graph in the Cytoscape environment (Shannon *et al.*, 2003).

3.2.2 Interaction dynamics in response to nutrient availability (Experiment 1): The data in this experiment captures the changes in gene expression during the gradual transition from glucose to ammonia as the growth-limiting nutrient. Genes that have already been grouped into eight clusters (Farzadfard *et al.*, 2010) were connected according to the protein interaction network and the interaction strength between them determined from gene expression using the model (Figure 4). The data in this experiment captures the changes in gene expression during the gradual transition from glucose to ammonia as the growth-limiting nutrient. Genes that have already been grouped into eight clusters (Farzadfard *et al.*, 2010) were connected according to the protein interaction network and the interaction strength between them determined from gene expression using the model. At the beginning of the experiment (t = 0), the cells were starved for glucose and were progressively exposed to increased glucose availability. After time t = 9.6, ammonia became the limiting nutrient. Subsequent time points capture the changes that occur in the presence of excess glucose. In response to this transition, we observed corresponding changes in glucose and ammonia metabolism, which are reflected in the interactions between the genes responsible for the synthesis of proteins and lipids. For example, we observed a strong positive (inductive) interaction between the genes of carbohydrate metabolism and protein synthesis (top left corner of the network) during glucose starvation. The interaction strength gradually decreased until t = 9.6 hours and subsequently turned negative (repressive). Since the genes in this cluster predominantly belong to amino acid synthesis, our results indicate a gradual repression in amino acid and protein synthesis upon the onset of ammonia starvation.

The transition in the growth-limiting nutrient also brought about an abrupt change in the interactions between genes responsible for ribosome biosynthesis (cluster of genes on the extreme right of each network). Our model identified a momentary repression in the synthesis of ribosomes at t = 9.6 hours, when the growth limitation was exerted by ammonia. These genes were constitutively active during all other time points, as is expected because ribosome biosynthesis is an essential cellular process. The temporary arrest in ribosome biosynthesis was attributed to the control exerted by Sfp1 transcription factor (Farzadfard *et al.*, 2010). Our model also identified positive interactions between amino acid metabolism and cell cycle and negative interactions between genes of storage carbohydrates and lipid metabolism (Figure 4). These interactions are available as supplementary files.

3.2.3 Interaction dynamics in response to network perturbation (Experiment 2): In order to test the utility of incorporating the damping our model in capturing the impact of perturbations in the network, we used the dataset in which the key transcription factor Sfp1 was deleted (Cipollina *et al.*, 2008). We chose this dataset since Sfp1 was previously identified to be one of the most important transcription factors that governed the response to nutrient availability in yeast (Farzadfard *et al.*, 2010). We first identified the interactions that are sensitive to time and perturbation, as described in 3.2.1. The histogram of the number of edges was fit to an exponential distribution (Figure 5) and only those interactions that passed the threshold cutoff of $p <= 0.05$ were considered for subsequent analysis. In this manner, we identified 171 interactions among the genes that were already identified to have differentially expressed between REF and MUT (see section 2.1.1 for strain nomenclature). An important aspect of novelty in our model is the incorporation of the damping effect in the model. This model ensures that interactions further from the point of perturbation in the network are affected to a lesser degree than those closer to it. The effect of damping is very sensitive to the network and in the network we considered in this study, a majority of the edges appear to be relatively unaffected by the perturbation (Figure 6).

After assessing the effect of perturbation for our network, we identified temporal changes in the interactions in the REF strain as well as the MUT strain independently (Figure 6). We observed some overlap in the actively interacting genes between REF and MUT. Many of these genes were hexose transporters and those responsible for pH homeostasis. The genes (and the functional categories) that are common to the strain indicate that they are not responsive to the mutations. The utility of NETGEM in identifying the temporal interactions that are different between REF and MUT is indicated by the identification of many genes that are responsible for ribosome biosynthesis and amino acid metabolism (Figure 6). The results concur with the known role of Sfp1 in coordinating metabolism with ribosome biosynthesis. These interactions were identified by considering gene expression profiles in REF and MUT, using the damping model. Indeed the functional classification of the genes between which interactions change significantly indicate that Sfp1 transcription factor has widespread control over coordinating ribosome biosynthesis, pH homeostasis, transport of proteins and drugs, etc (Table 2).

In Table 2, the list of 20 topmost functional categories whose transitional probabilities exhibit the maximum degree of change in interaction strengths out of 260 possible functional categories are shown. This is obtained by considering the total probability of change in the transition probability matrix, Q_h, i.e., $\sum_{i \neq j} Q_h(i, j)$. The fact that many of these functional categories have already been identified (Cipollina *et al.*, 2008) to be sensitive to the perturbation gives substantial credibility to our findings.

4 CONCLUSION

There is a trade off 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. NETGEM is a systematic model that relates temporal changes in gene expression data to the dynamics of interactions in the context of a regulatory network. We believe that NETGEM achieves an optimal balance between model complexity and the data requirement, while allowing ample flexibility to adjust the parameters. The framework of the model will also inherently facilitate analyzing the effect of a perturbation in the network. For a given regulatory network and a gene expression data, NETGEM was able to identify time-sensitive interactions in the network and determine their strength. It was able to deduce most active functional categories that interacted. In addition to these, the NETGEM uses a damping feature that models the effect of a network perturbation by localizing more activity around the point of perturbation. These three novel features that NETGEM offers reflect its advantage over many other time-series models that have been developed recently. Of particular interest is its ability to capture abrupt changes in the interaction patterns. For example, NETGEM identified momentary arrest in ribosome biosynthesis during the transition in the nutrient that limits growth from glucose to ammonia (Experiment 1). We identified many actively interacting genes that were implicated to play an important role in the biological conditions from which we obtained

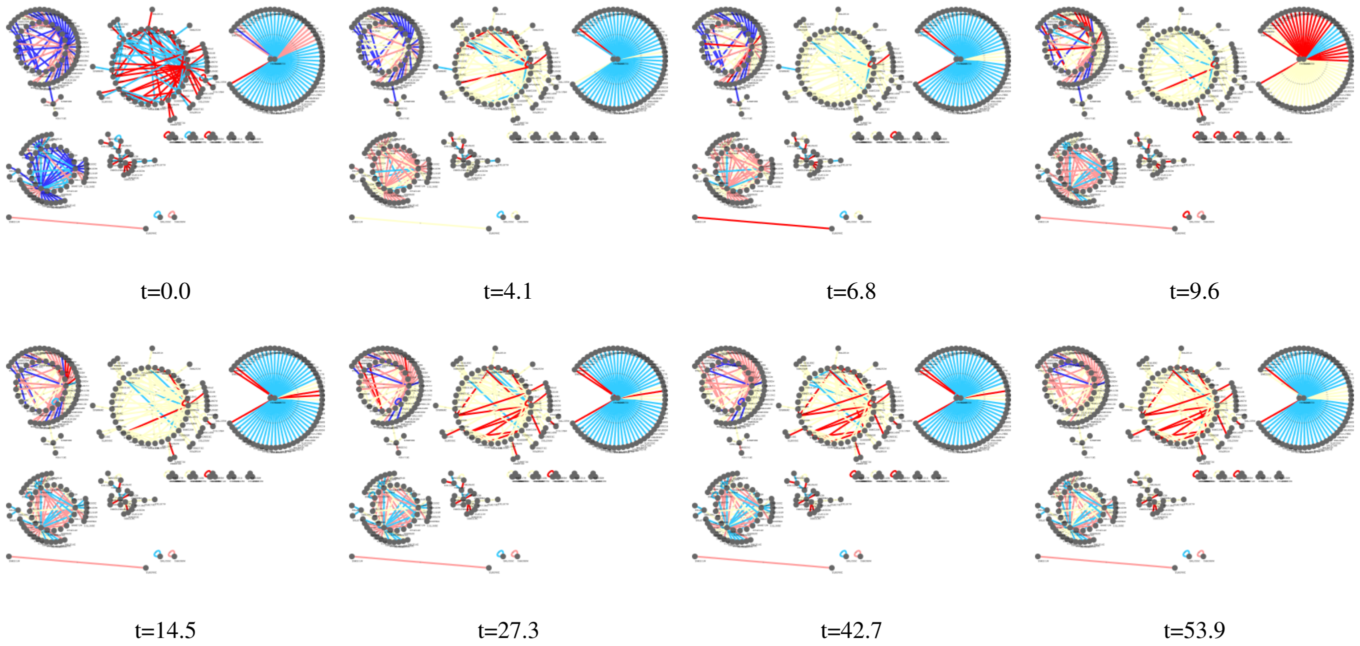

Fig. 4. Time varying interaction strengths between the genes from Experiment 1. Each network is composed of all the genes from the eight clusters previously identified (Farzadfard *et al.*, 2010), and is shown for the eight time points for which gene expression was measured. The time stamps (in hours) are indicated below each network. The edge colors denote their interaction strength, which was classified as strong repressing (red), low repressing (pink), no effect (yellow), low inducing (light blue) and strong inducing (dark blue).

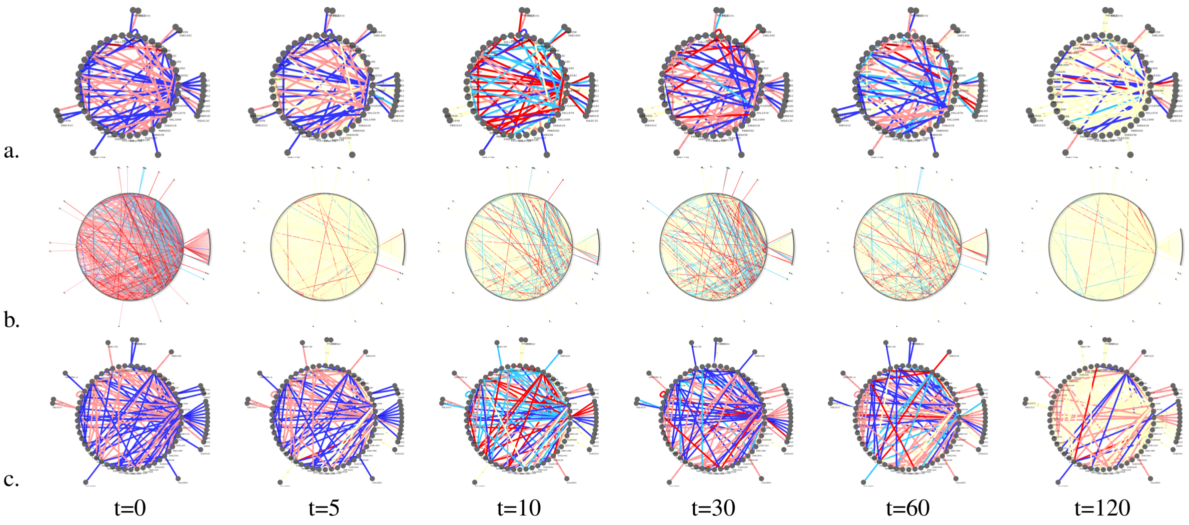

Fig. 6. Dynamics of temporal interactions between genes in REF (a), MUT (b) and in both strains combined, using the damping model (c) for Experiment 2. All the genes identified to be significantly changed (Cipollina *et al.*, 2008) were combined into one network. The color of the edges in the network indicates the interaction strength, which was classified as strong repressing (red), low repressing (pink), no effect (yellow), low inducing (light blue) and strong inducing (dark blue). The time stamps (in minutes) are indicated below each network.

the data. This lends the promise that new insights obtained from using NETGEM are also physiologically relevant. Given that the inputs to NETGEM are the topology of the network and temporal variation of the nodes, it is evident that this methodology has widespread applications in analyzing network dynamics, beyond biological systems.

MIPS ID	Description of functional categories
01.03.01	purin nucleotide/nucleoside/nucleobase metabolism
10	cell cycle and DNA processing
10.01.09	DNA restriction or modification
10.03.01	mitotic cell cycle and cell cycle control
11.04.02	tRNA processing
12	protein synthesis
12.01.01	ribosomal proteins
14.13.01	cytoplasmic and nuclear protein degradation
16	protein with binding function or cofactor requirement (structural or catalytic)
16.02	peptide binding
16.03.03	RNA binding
16.21	complex cofactor/cosubstrate/vitamine binding
20.01.10	protein transport
20.01.15	electron transport
20.01.27	drug/toxin transport
20.09.03	peroxisomal transport
20.09.07	vesicular transport (Golgi network, etc.)
30	cellular communication/signal transduction mechanism
32.01.04	pH stress response
32.05.05	virulence, disease factors
32.07	detoxification
32.07.07	oxygen and radical detoxification
34.07.02	cell-matrix adhesion
40.01.03	directional cell growth (morphogenesis)
40.01.05	growth regulators / regulation of cell size
42.04.03	actin cytoskeleton

Table 1. Description of the functional categories corresponding to the MIPS IDs identified in Table 2.

REF	MUT	JOINT
32.01.04	32.01.04	32.01.04
20.09.03	20.09.03	20.09.03
10	16.21	16.21
16.21	32.05.05	11.04.02
11.04.02	10	10
32.05.05	11.04.02	12
20.01.27	40.01.05	32.05.05
30	30	30
16.03.03	10.01.09	16.03.03
20.09.07	10.03.01	32.07
12	14.13.01	20.09.07
32.07	20.01.10	40.01.03
20.01.15	16.03.03	20.01.27
42.04.05	20.09.07	10.01.09
34.07.02	42.04.03	34.07.02
10.03.01	16	32.07.07
40.01.03	34.07.02	16.02
10.01.09	12.01.01	42.04.05
42.29	01.03.01	42.29

Table 2. This table shows the list of functional categories which show the maximum amount of variation in time for the strains: (a) REF (b) MUT and (c) in both strains combined, using the damping model. The description for the categories can be found in Table 1. The categories indicated in blue are those which are known to have been enriched in the original dataset (Cipollina *et al.*, 2008).

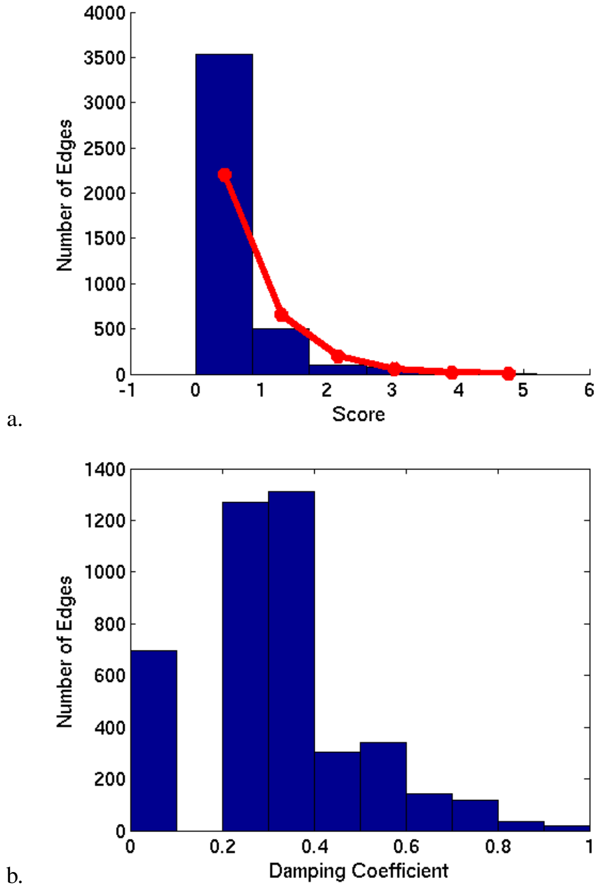

Fig. 5. This figure presents histograms characterizing the change in the interaction strengths over all edges when the inference is done over JOINT (both strain) in Experiment 2 (a). We also show the histogram of the damping coefficients for the edges in the perturbed strain in Experiment 2 (b). It is important to note that the damping coefficients are dependent on the network topology.

5 ACKNOWLEDGEMENTS

GNV acknowledges funding from Chalmers Research Foundation and salary from EC-funded project, UNICELLSYS. The authors also acknowledge Torbjörn Karfunkel for assistance with constructing the protein interaction network and discussion on the method.

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.

Cipollina, C. *et al.* (2008) Saccharomyces cerevisiae SFP1: at the crossroads of central metabolism and ribosome biogenesis *Microbiology,* **154** , 1686–1699.

James M. Coughlan and Huiying Shen. Shape matching with belief propagation: Using dynamic quantization to accommodate occlusion and clutter. In *Generative-Model Based Vision workshop at CVPR*, 2004.

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.

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

Farzadfard, F. *et al.* (2010) Metabolic and transcriptional dynamics during the transition from carbon limitation to nitrogen limitation in Saccharomyces cerevisiae. *Genome Biology (in review)*.

Gavin, A. C., *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature,* **440**, 631–636.

Gavin, A. C., *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature,* **415**, 141–147.

Gelman, A. *et al.* (2003) *Bayesian Data Analysis, Second Edition (Texts in Statistical Science).* 2 edition,, Chapman & Hall/CRC.

Gentleman, R. C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology,* **5** (10), R80.

Glass, L. & Kaplan., D. (1993) Time series analysis of complex dynamics in physiology and medicine. *Med Prog Technol,* **19**, 115–128.

Ho, Y. *et al.* (2002) Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature,* **415**, 180–3.

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

Ito, T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *PNAS,* **98**, 4569–74.

Krogan, N. J. *et al.* (2006) Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature,* **440**, 637–43.

Leek, J. T. *et al.* (2006) EDGE: extraction and analysis of differential gene expression. *Bioinformatics,* **22** (4), 507–508.

Mewes, H. W. *et al.* (2007) MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Res,* **36**.

Mewes, H. W. *et al.* (2002) Mips: a database for genomes and protein sequences. *Nucleic acids research,* **30** (1), 31–34.

Petranovic, D. & Vemuri, G. N. (2009) Impact of yeast systems biology on industrial biotechnology. *J Biotechnol,* **144** (3), 204–11.

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. *et al.* (2002) Cluster analysis of gene expression dynamics. *Proc Natl Acad Sci U S A,* **99** (14), 9121–9126.

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

Shannon, P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks In *Genome Research* **13** (11), pp. 2498 –2504.

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

Stark, C. *et al.* (2006) Biogrid: a general repository for interaction datasets. *Nucleic Acids Res,* **34** (Database issue).

Stigler, B. *et al.* (2007) Reverse engineering of dynamic networks. *Ann N Y Acad Sci,* **1115**, 168–77.

Uetz, P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. :. *Nature,* **403**, 623–7.

Xenarios, I. *et al.* (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.