BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 9 2011, pages 1277–1283
doi:10.1093/bioinformatics/btr113

Gene expression Advance Access publication March 2, 2011

Large-scale learning of combinatorial transcriptional dynamics from gene expression

H. M. Shahzad Asif[1,2] and Guido Sanguinetti[1,*]

[1]School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK and [2]Department of Computer Science and Engineering, University of Engineering and Technology, Lahore, Pakistan

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Knowledge of the activation patterns of transcription factors (TFs) is fundamental to elucidate the dynamics of gene regulation in response to environmental conditions. Direct experimental measurement of TFs' activities is, however, challenging, resulting in a need to develop statistical tools to infer TF activities from mRNA expression levels of target genes. Current models, however, neglect important features of transcriptional regulation; in particular, the combinatorial nature of regulation, which is fundamental for signal integration, is not accounted for.

Results: We present a novel method to infer combinatorial regulation of gene expression by multiple transcription factors in large-scale transcriptional regulatory networks. The method implements a factorial hidden Markov model with a non-linear likelihood to represent the interactions between the hidden transcription factors. We explore our model's performance on artificial datasets and demonstrate the applicability of our method on genome-wide scale for three expression datasets. The results obtained using our model are biologically coherent and provide a tool to explore the concealed nature of combinatorial transcriptional regulation.

Availability: http://homepages.inf.ed.ac.uk/gsanguin/software.html.

Contact: g.sanguinetti@ed.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 6, 2010; revised on February 15, 2011; accepted on February 27, 2011

1 INTRODUCTION

Understanding the control of gene expression is one of the major goals of systems biology. While gene expression is a complex process with multiple control points, perhaps the most fundamental is the control of mRNA transcription by DNA-binding proteins, transcription factors (TFs). A fundamental difficulty in elucidating this process from the experimental point of view is measuring active TF concentrations: TFs are often expressed at low levels, and their activity state is frequently determined by fast post-translational modifications which are difficult to measure directly.

A possible solution to this impasse has arisen due to the availability of experimental tools to determine the *connectivity* of the transcriptional regulatory network, i.e. which TFs bind specific target genes. In particular, the large-scale take-up of chromatin immunoprecipitation techniques (ChIP-on-chip) has meant that, for model organisms such as yeast and *Escherichia coli*, this connectivity is now available on a high-throughput scale (Lee *et al.*, 2002). As a result, several authors have recently proposed to integrate connectivity and gene expression data in an inference-based approach to modelling transcription, whereby TF activity is treated as a latent variable to be reconstructed from observations of target gene's expression. Broadly speaking, inferential approaches to TF activity reconstruction have used one of two strategies: one approach is to use a very simplistic, typically log-linear model of transcription to infer the activity of a very large number of TFs (Liao *et al.*, 2003; Sabatti and James, 2006; Sanguinetti *et al.*, 2006; Asif *et al.*, 2010). This approach is relatively well established and has already led to several novel insights in biological studies in a range of situations (Davidge *et al.*, 2009; Partridge *et al.*, 2007); however, the simplicity of the models, imposed by the computational constraints of working with large datasets, has meant that important features of transcriptional regulation have been neglected. More recently, other authors have focused on inferring TF activities in small subnetworks but employing more realistic models of transcription based on differential equations (Barenco *et al.*, 2006; Lawrence *et al.*, 2006). These approaches are computationally more expensive but allow to model biologically more plausible effects such as saturation (Rogers *et al.*, 2007), rapid transitions (Sanguinetti *et al.*, 2009) and non-linear interactions between TFs (Opper and Sanguinetti, 2010).

In this article, we aim at retaining some of the desirable features of small-scale inference approaches in a model capable of learning TF activity on a genome-wide scale. We focus on the problem of modelling interactions between multiple TFs; this is a crucial mechanism that allows cells to integrate signals (Ptashne and Gann, 2002). We present what, to our knowledge, is the first statistical method for reconstructing combinatorial interactions between TFs from target genes' expression levels. We achieve this by modelling TF activity as binary switches (which naturally allow for saturation) within a factorial Hidden Markov Model (FHMM) with a non-linear emission model which models combinatorial interactions between multiple TFs at a promoter.

We propose a fast structured variational approximation for inference in large-scale systems. As our model includes non-linear interaction, it is relatively more highly parametrized than simpler models. We, therefore, extensively tested our model on simulated data to check its identifiability. We then applied it to three real time course datasets in *Saccharomyces cerevisiae* and *E.coli*, using network architectures derived from ChIP-on-chip experiments or curated databases of biological interactions. The key purpose of

*To whom correspondence should be addressed.

our analysis of real data is to investigate the extent to which non-linear combinatorial effects are evident from the expression data. Perhaps not surprisingly, we find that the length of the time series is a critical factor in reducing the uncertainty of the model's predictions, and thus enabling the recovery of non-linear interactions. Despite this, specific examples of biologically meaningful combinatorial effects are recovered, showing that computational prediction of combinatorial interactions is indeed possible from the analysis of mRNA time series.

2 MODEL

Suppose that N genes are regulated by M TFs over T conditions/time point. Throughout this article, we will assume TFs to be binary variables who can either be on or off (Sanguinetti *et al.*, 2009). This modelling assumption corresponds to two biological assumptions: TFs switch fast from active to inactive form and vice versa, and the number of TF molecules per cell is sufficient to saturate the downstream transcriptional machinery. Let g_i^t be the mRNA expression level of gene i in condition t, and let $\{T_j\}_i \quad j \in \mathcal{J}_i \in \{1,\ldots,M\}$ be the set of TFs binding gene i. Our model for (log) gene expression is given by

$$g_i^t = \boldsymbol{e}_t^T \boldsymbol{\theta}_i + \epsilon \tag{1}$$

where $\boldsymbol{\theta_i}$ is a set of expression parameters specific for gene i, $\boldsymbol{e_t}$ is composed of the states of the transcription factors and their two-point interactions and ϵ is measurement noise. In the simple case of two TFs, this would become

$$g_i^t = A_i^1 T_t^1 + A_i^2 T_t^2 + A_i^{12} T_t^1 T_t^2 + b + \epsilon. \tag{2}$$

Gene expression is, therefore, digitized with four expression levels corresponding to the four possible joint states of the two regulators. This can be viewed as a steady-state approximation to the combinatorial transcription model of Opper and Sanguinetti (2010). The assumption of binary states of the TFs is mainly due to the transient behaviour of these regulators that makes it harder to measure experimentally at the sampling rate used in most of the cases.

To cast the model (1) in a Bayesian framework, we need to specify prior distributions over the various components. The prior for the parameters $\boldsymbol{\theta}_i$ is assumed to be a zero mean Gaussian with variance encoded by a hyperparameter α^2,

$$\boldsymbol{\theta}_i \sim \mathcal{N}(0, \alpha^2).$$

The choice of prior over the TF activity is dictated by the experiment we are modelling. If the experimental design consists of a number of independent conditions, then a uniform prior over the TF states at each condition may be justified. While this experimental design is indeed very widely used, in this article we will focus on the time-course experimental design. The derivations for independent conditions experimental design can be easily worked out using a similar methodology. In the time-course experimental design, the natural prior distribution for the TF activity is given by a factorial HMM [FHMM, Ghahramani and Jordan (1997)]. Therefore, the prior probability defines a series of *a priori* independent Markov chains consisting of sequences of binary states, one for each TF,

$$p\left(T_1^j, \ldots, T_T^j\right) = \prod_{t=1}^{T} p\left(T_{t+1}^j | T_t^j, \tau_j\right).$$

Each of these Markov chains depends on a matrix of hyperparameters, the *transition probabilities*, encoding the prior probability of the TF switching from active to inactive form. As the TFs are assumed to be binary, by normalization there are only two independent hyperparameters in each transition matrix. Finally, the model is completely specified by the assumption that the observation error in Equation (1) is zero mean Gaussian and i.i.d., so that

$$p(G|T, \boldsymbol{\Theta}) = \prod_{i=1}^{N} \prod_{t=1}^{T} \mathcal{N}\left(g_i^t | \boldsymbol{e}_t^T \boldsymbol{\theta}_i, \sigma^2\right).$$

Here G, T and $\boldsymbol{\Theta}$ are collective names for all the observations, TF states and gene specific parameters, respectively.

Before discussing how inference can be performed in this model, it is important to observe that, as the parameters $\boldsymbol{\Theta}$ and the TF states T only appear in the model (1) through their product, a basic identifiability problem exists for this model. To clarify the issue, if we take the simple case of a gene regulated by two TFs, we see that Equation (2) is left invariant by the transformation

$$\begin{aligned} &T_t^1 \to 1 - T_t^1 \forall t \in \{1,\ldots,T\} \\ &b \to -A_1 + b, \quad A_1 \to -A_1 \\ &A_2 \to A_2 - A_{12}, \quad A_{12} \to -A_{12}. \end{aligned} \tag{3}$$

This ambiguity, which is common to all statistical models involving multiplication of latent variables, cannot be resolved without prior knowledge. This is occasionally available: for example, it may be known that a given TF activates/ represses a specific target, or that the TF is on/ off in a specific condition. Notice that knowledge about the sign of regulation for a *single* target gene or for a *single* condition/ time point is sufficient to remove the ambiguity for all other conditions/ targets of the same TF. Another important observation is that the presence or absence of a combinatorial interaction is not affected by the identifiability problem. Only the sign of the combinatorial term A_{12} changes under the transformation (3).

3 INFERENCE

Our goal is to infer from observations of gene expression both the state of TFs and the gene-specific expression parameters θ. Bayesian inference in model (1) is analytically intractable. Stochastic inference approaches such as Gibbs sampling are often employed in these cases; unfortunately, we found that the computational costs of such an approach were too high (see Supplementary Material). We, therefore, develop a fast structured mean-field approximation which is capable of performing inference in very large-scale problems.

Variational Bayesian inference is an optimization-based approximate inference technique originally developed in statistical physics. The basic idea is to approximate the posterior distribution over the latent variables and parameters with a simpler distribution. Variational techniques convert a complex problem into a simpler problem by decoupling the degrees of freedom in the original problem (Jordan *et al.*, 1999). This decoupling is obtained by expanding the problem to include additional parameters also know as variational parameters that are optimized according to the problem under consideration. Compared with stochastic approximations like Gibbs sampling, this optimization process is usually very efficient computationally, and has the advantage of allowing an unambiguous monitoring of convergence.

Variational inference relies on the following general lower bound on the log likelihood:

$$log[p(\boldsymbol{G}|\boldsymbol{\phi})] \geq \langle log\, p(\boldsymbol{G}, \boldsymbol{\Theta}, \boldsymbol{T}|\boldsymbol{\phi}) \rangle_{q(\boldsymbol{\Theta}, \boldsymbol{T})} + H(q) \tag{4}$$

which follows from Jensen's inequality. Here $\langle\rangle$ shows the expectation of the joint likelihood under the approximating distribution q, H denotes the entropy of the distribution and ϕ collectively denote the hyperparameter α and σ. It can be shown that the lower bound (4) is saturated if and only if the approximating distribution q is equal to the posterior distribution $p(\boldsymbol{\Theta}, \boldsymbol{T}|\boldsymbol{G}, \boldsymbol{\phi})$. In our case, the approximating distribution q is assumed to be a structured mean-field approximation

$$q(\boldsymbol{\Theta}, \boldsymbol{T}) = q(\boldsymbol{\Theta}) \prod_i q(\boldsymbol{T}^i). \tag{5}$$

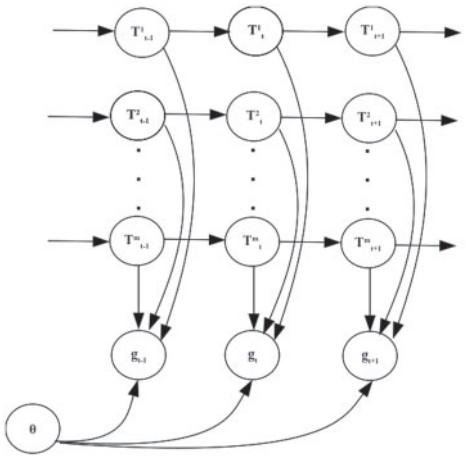

Fig. 1. Graphical representation of the model.

Therefore, we assume the approximating distribution to factor across parameters and transcription factors, but *not* across time points. The joint likelihood of the model is given by

$$p(\boldsymbol{G},\boldsymbol{\Theta},\boldsymbol{T})=p(\boldsymbol{G}|\boldsymbol{T},\boldsymbol{\Theta})\,p(\boldsymbol{\Theta}|\alpha^2)\,p(\boldsymbol{T}). \quad (6)$$

We will use a variational EM algorithm to optimize iteratively the lower bound w.r.t. $\boldsymbol{\Theta}$ and each of the TFs $\boldsymbol{T}^i$; the reader is referred to Beal (2003) for a more thorough discussion of variational EM algorithms in HMMs. The lower bound (4) is guaranteed to increase after each step of this iterative process, and the convergence of the algorithm can be monitored through evaluation of the lower bound.

3.1 E-step

In the E-step, the approximate posterior distribution over the TF states is calculated. Averaging out the parameters $\boldsymbol{\Theta}$ in Equation (6), we readily recognize the result as the joint likelihood for the observations and the hidden states in a standard (i.e. fixed parameters) FHMM. The posterior distribution over each TF can be easily obtained using the standard *forward backward* (FB) algorithm (Bishop, 2006) that provides the probabilities for both states (i.e. on or off) of TFs over all the time point of the gene expression measurements. Further using the factorization across TFs given in Equation (5), we use the FB algorithm independently for each hidden layer of FHMM (Fig. 1) to provide the single time state marginals of the approximate posterior distribution $q(\boldsymbol{T})$. Further details, including pseudocode, are given in the Supplementary Material.

3.2 M-step

Taking expectations of the log of the joint likelihood under $\boldsymbol{T}$, one can see that the approximate posterior distribution over the parameters of $\boldsymbol{\Theta}_i$ is given by a multivariate normal

$$q(\boldsymbol{\Theta})=\prod_{i=1}^{N}\mathcal{N}(\theta_i|\boldsymbol{m_i},\boldsymbol{\Sigma_i}). \quad (7)$$

The mean and covariance of this multivariate normal distribution are given by

$$\boldsymbol{\Sigma_i}^{-1}=\frac{1}{\sigma^2}\sum_{t=1}^{T}\boldsymbol{X}_i\langle\boldsymbol{e}_t\boldsymbol{e}_t^T\rangle_{q(\boldsymbol{T})}\boldsymbol{X}_i+\alpha^{-2}\boldsymbol{I}$$

$$\boldsymbol{m_i}=\frac{1}{\sigma^2}\left[\sum_{t=1}^{T}g_i^t\langle\boldsymbol{e}_t^T\rangle_{q(\boldsymbol{T})}\boldsymbol{X}_i\right]\boldsymbol{\Sigma_i}^{-1}$$

Here $\langle\rangle_{q(\boldsymbol{T})}$ denotes the expectation under $q(\boldsymbol{T})$, and $\boldsymbol{X}_i$ denotes a diagonal matrix with the i-th row of the connectivity matrix $\boldsymbol{X}$ along the diagonal. For more details about the method and implementation, refer to Supplementary Material.

As the length of the time series is usually very limited, we will not attempt to infer hyperparameters of the model such as the transition matrices and observation noise variance (even if point estimation of hyperparameters by type II maximum likelihood is in principle straightforward). Rather, these hyperparameters will be fixed heuristically: transition matrices will be set to give a prior expectation of few transitions within the time under consideration; and noise variance will be fixed after preliminary inspection of the data. Experiments on synthetic data showed that the model predictions to be fairly insensitive to the specific values of the transition matrices.

4 RESULTS

While our model is still relatively simple, the addition of non-linear interaction terms means that more parameters need to be estimated. On top of that, asymptotically exact inference is computationally unfeasible in large-scale examples. Therefore, as a first analysis we perform a thorough test of the proposed model using artificial data to verify its identifiability in a realistic simulated situation. We then use three real datasets; in all cases, the main purpose is to probe the extent to which combinatorial regulations can be learned from the expression data. These datasets are the classic and much studied yeast cell cycle dataset (Spellman *et al.*, 1998), the yeast metabolic cycle dataset (Tu *et al.*, 2005) and the *E.coli* micro-aerobic shift dataset (Partridge *et al.*, 2007). Finally, we compare our results with those obtained with two different methods: a simplified version of the method by Shi *et al.* (2009) and the TFInfer method by Sanguinetti *et al.* (2006); Asif *et al.* (2010).

4.1 Synthetic results

As a first analysis, we performed a series of experiments on artificial data generated with known parameters to benchmark and check the consistency of the model. Specifically, two aspects of the inferential problem need to be investigated:

(1) Is the model identifiable given realistic data, i.e. in a large-scale example with relatively few time points?

(2) Does the efficient variational approximation developed in Section 3 give an accurate representation of the posterior uncertainty over the random variables?

We discuss here Issue 1; a discussion of Issue 2 is given in the Supplementary Material where a comparison with a Gibbs sampling approach is given. We generated an artificial dataset with 1000 genes, 50 transcription factors and 20 time points. We used the connectivity information from yeast cell regulatory network (Lee *et al.*, 2002) with random initialization for the gene-specific parameters. We then ran the variational EM algorithm to infer the posterior probabilities over TF states and gene-specific parameters, and compared with the true parameter values/ TF states. The results for parameter estimation are given in Figure 2, displaying true parameter values with posterior mean estimates. In most cases, it is clear that the parameters inferred using the variational EM algorithm match closely with the true values. In a few cases, the inferred parameters are anticorrelated with the true parameter values; these correspond to TFs whose activity was inferred to be the opposite of the true activity. As we noted earlier, this ambiguity is unavoidable and cannot be resolved without further knowledge.

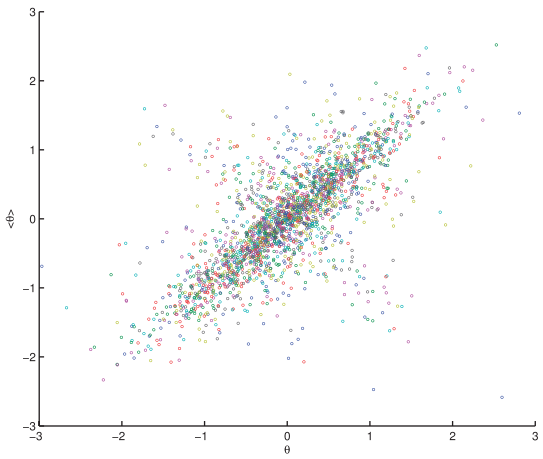

Fig. 2. Comparison of inferred parameters with true values of $\boldsymbol{\Theta}$.

While Figure 2 gives support to the identifiability of the *mean* predictions of our model, the Bayesian nature of the model means that estimates of the uncertainty of the predictions are also available. These estimates can be precious to assess the statistical significance of predicted interactions: for example, we could say that two TFs regulate combinatorially a certain gene at 5% significance level if the absolute value of the posterior mean of the predicted combinatorial term in Equation (2) is greater than twice the predicted SD. We are interested in quantifying what fraction of combinatorial interactions can be recovered at a certain significance level as a function of the length of the time series and the experimental noise. To do this, we generated multiple artificial datasets with different numbers of time points (Table 1, column 1) and varying corrupting noise levels ($\sigma^2 = 0.1, 0.5, 1.0$). In all cases, the number of genes and transcription factors, as well as the network architecture and true parameter values, was kept fixed ($N = 200, M = 50$). Table 1 reports the fraction of combinatorial regulatory interactions which were recovered at 5% significance level for specific lengths of the time series and different values of the Gaussian noise in gene expression. Not surprisingly, this percentage increases monotonically with the length of the time series and decreases when the additive observation noise is increased. Also, it appears that the level of noise somehow determines the proportion of combinatorial interactions that can be recovered *even for long* time series. Empirically, it appears that, with this network structure, more than 40 time points do not lead to a significant change in the proportion of combinatorial interactions recovered.

4.2 Micro-aerobic shift in *E.coli*

Partridge *et al.* (2007) studied the transcriptomic response of *E.coli* to the withdrawal of oxygen in a chemostat culture under controlled growth conditions. *Escherichia coli* is a metabolically versatile bacterium and responds to changes from aerobic to micro-aerobic conditions by activating TF proteins that act as oxygen sensors. The probabilistic approach described in Sanguinetti *et al.* (2006) was used to infer the states of six crucial regulators of oxygen sensing and metabolism (FNR, MetE, MetJ,ArcA, CpxR and SigE) from the mRNA expression of 302 target genes. The analysis revealed insights in the dynamics of the key regulators upon oxygen withdrawal, as well as biologically interesting predictions about the timing of TF activity. The dataset consists of four time points taken at 5, 10, 15 and 60 min and measured relative to a sample taken immediately before the perturbation. Connectivity information about the regulatory network was obtained from the ecocyc database (http://ecocyc.org/) and is available for 6 TFs and 302 genes in the Supplementary Material of Partridge *et al.* (2007). In this dataset, no combinatorial interactions were predicted at a significance level of 5%. In the light of the analysis on synthetic data, this is probably due to the very short time series.

Table 1. Combinatorial interactions found using synthetic data with different number of time points

T	$\sigma^2 = 0.1$		$\sigma^2 = 0.5$		$\sigma^2 = 1$	
	$A_{ij}(\%)$	Average posterior SD	$A_{ij}(\%)$	Average posterior SD	$A_{ij}(\%)$	Average posterior SD
10	18	0.2273	5	0.4009	3	0.5027
20	28	0.1655	10	0.3016	6	0.3953
30	40	0.1342	25	0.2550	8	0.3364
40	54	0.1088	33	0.2248	18	0.2993
50	54	0.0996	33	0.2003	18	0.2710

A_{ij} is the percentage of combinatorial interactions recovered from the data. σ^2 stands for the noise level in the synthetics data.

4.3 Yeast cell cycle data

Spellman *et al.* (1998) used microarray hybridization to measure the expression profiles of most of the yeast genes over a complete cell cycle. Three time-series experiments were conducted on three different strains of yeast and these experiments were synchronized by three independent methods; α factor-based synchronization, size-based synchronization and cdc15-based synchronization. We use the cdc15 synchronized data, consisting of 6181 gene expression profiles over 24 time points. The connectivity information for the yeast regulatory network was obtained in Lee *et al.* (2002) using ChIP-on-chip for 113 TFs measuring their binding to 6270 genes. These two datasets are relatively old but well studied and serve as the standard benchmark for validating the model described here. We preprocessed these two datasets such that all the genes are bound by at least one TF and each TF is regulating at least one gene; that gave us a network of 1975 genes and 104 TFs and expression profiles of 1975 genes. The data were analysed using the variational approximation, since the large size of this network rules out the application of the Gibbs sampling algorithm.

Once again, the predictions in terms of TF activities matched well the predictions of previous models [such as Liao *et al.* (2003); Sanguinetti *et al.* (2006)], in particular recovering the periodic pattern of key cell cycle regulators such as SWI5 and ACE2. An analysis of the predicted interaction terms reveals that about 5% of the combinatorial interactions [A_{12} in (2)] are significant at 5% level as shown in Figure 3. This accounts for 186 combinatorial interactions out of a total of 3886 possible pairwise interactions allowed by the structure of the regulatory network.

A more detailed analysis of the results obtained (across transcription factor profiles) using the Model 2 reveals that some of the TFs in the yeast regulatory network have a much

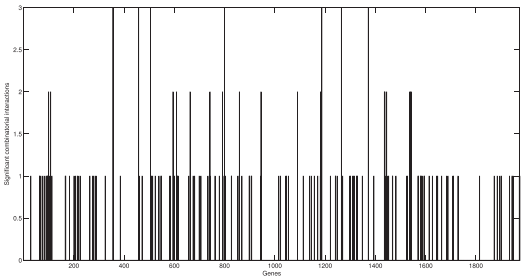

Fig. 3. Number of $A_{ij} \geq 2$ SD for 1975 genes of (Spellman *et al.*, 1998).

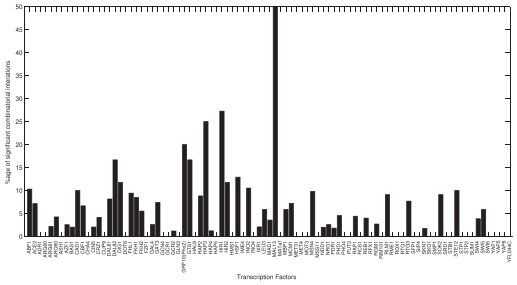

Fig. 4. Percentage of combinatorial interactions for 104 TFs of yeast dataset (Spellman *et al.*, 1998).

higher proportion of significant combinatorial interactions than the average. Figure 4 shows the percentage of significant combinatorial regulation for all the TFs in this dataset. It can be seen from this plot that a group of TFs (DAL82, Pho2, GTS1, HAP3, HIR1, MAL13) have 15% or more significant combinatorial interactions compared with the overall average of 5% significant combinatorial interactions. Looking at the biological function of these highly interacting proteins, we found that our results are often plausible in terms of the underlying biology. The transcription factor Pho2 found to be actively involved in combinatorial regulation by our model is known to behave in a combinatorial manner (Bhoite *et al.*, 2002). Pho2 is functionally active in many biological processes such as histidine biosynthesis and phosphate utilization (Daignan-Fornier and Fink, 1992). Similarly, HAP3 is a global regulator of respiratory gene expression and contains sequence contributions to both complex assembly and DNA binding (Xing *et al.*, 1993), (Hahn *et al.*, 1988). The contributions of these transcription factors to multiple biological processes indicates that plausibly these TFs will need cofactors to achieve specificity in gene regulation.

Our model predicted that DAL82 regulatory activities contains a higher percentage of significant combinatorial regulation. DAL82 is a positive regulator of allophanate-inducible genes and is one of four transcription factors that are required for this process (Scott *et al.*, 2000). Experimental evidence in this case suggests that DAL81 protein is required for DAL82-dependent transcription activation. As shown in Figure 4, our model also predicted the higher percentage of combinatorial activity for DAL81 ($\sim$10%). GTS1 is a transcriptional coactivator for the genes that exhibits the metabolism of carbohydrates, requiring interactions with other regulators to induce gene expression (Xu and Tsurugi, 2007).

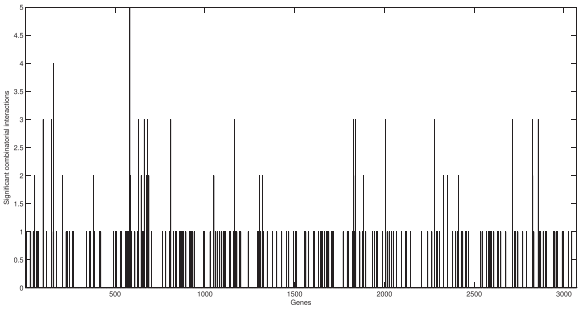

Fig. 5. Number of $A_{ij} \geq 2$ SD for 3070 genes for yeast dataset (Tu *et al.*, 2005).

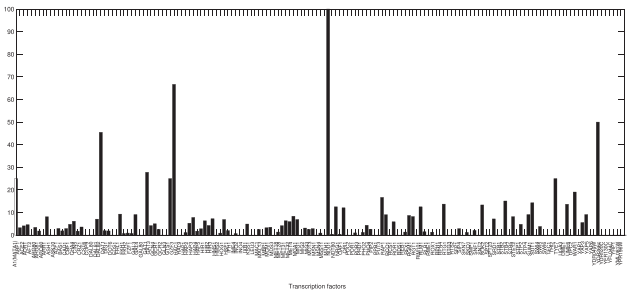

Fig. 6. Percentage of combinatorial interactions for 177 TFs for yeast dataset (Tu *et al.*, 2005).

4.4 Metabolic cycle data

Tu *et al.* (2005) studied the yeast metabolic cycle (YMC) that governs the genome-wide transcription of genes in a periodic manner. Budding yeast under nutrient-limited conditions goes through robust cycles of respiratory bursts that in turn causes almost half of the yeast genome to express periodically. In this experiment, total RNA was prepared after every 25 min over a period of three consecutive metabolic cycles. In order to use this dataset with our model, we fused the network connectivity available from two ChIP-chip experiments (Harbison *et al.*, 2004; Lee *et al.*, 2002) and removed the genes that were not regulated by any TFs in the connectivity information. The TFs not involved in regulating any genes were also eliminated leaving a network of 3070 genes and 177 TFs. Our probabilistic approach can handle the false positive that could arise from this dataset by assigning higher uncertainty to the regulatory interactions that are not eminent from the data.

Once again, the predicted activity profiles of most regulators showed a good agreement with previously reported results Sanguinetti *et al.* (2006) using different inference models (data not shown). In particular, our model confidently predicted a periodic behaviour for many of the regulators, which is in agreement with the experimental design. The details about the extent of the combinatorial regulation in this dataset are shown in Figure 5 where $\sim$3% of the possible combinatorial interactions are found to be statistically significant. Out of a total of 10 876 possible combinatorial interactions in this dataset, only 322 were predicted to have posterior mean greater than 2 SD.

Further analysis across the transcription factor profiles showed that a small proportion of the TFs in this dataset have significantly higher combinatorial interactions as shown in Figure 6. The most

Table 2. Comparison of different techniques for inference of the states of transcription factors. The states inferred with different methods are compared using the Hamming distance (HD) between the vectors of states

Method	Dataset		
	Partridge *et al.* (2007)	Spellman *et al.* (1998)	Tu *et al.* (2005)
FHMM (Shi *et al.*, 2009)	Run time: 6 s MSE: 0.0189 HD with cFHMM = 0.0667 HD with TFInfer = 0.0667	Run time: 4.7 h MSE: 0.1381 HD with cFHMM = 0.2688 HD with TFInfer = 0.2015	Run time: 5.5 h MSE: 0.4332 HD with cFHMM = 0.2502 HD with TFInfer = 0.2280
cFHMM (combinatorial factorial HMM)	Run time: 22 s MSE: 0.0423 HD with FHMM = 0.0677 HD with TFInfer = 0.1333	Run time: 42 h MSE: 0.1391 HD with FHMM=0.2688 HD with TFInfer=0.2708	Run time: 335 h MSE: 0.4125 HD with FHMM = 0.2502 HD with TFInfer = 0.3021
TFInfer (Asif *et al.*, 2010)	Run time: 45 seconds MSE: 0.0399 HD with FHMM = 0.0667 HD with cFHMM = 0.1333	Run time: 10 h MSE: 0.1156 HD with FHMM = 0.2015 HD with cFHMM = 0.2708	Run time: 115 h MSE: 0.3811 HD with FHMM = 0.2280 HD with cFHMM = 0.3021

prominent of these highly interacting TFs are as follows: DAL82, GAT1, GTS1, GZF3, MTH1, PUT3, STB2, THI2, UPC2, VMS1. Some of these TFs appear to have consistently combinatorial behaviour between the cell cycle and the metabolic cycle; e.g. DAL82 and GTS1 could be interpreted as 'housekeeping' combinatorial TFs. GAT1, a positive regulator of nitrogen catabolite repression (NCR), is an essential regulator of the NCR-sensitive genes along with another transcription factor GLN3. The model for regulatory circuit of GAT1–GLN3 combination is discussed in Coffman *et al.* (1996). The majority of the other TFs predicted to have high combinatorial behaviour are clearly associated with the metabolic processes: GZF3 is a catabolite repressor, MTH1 regulates glucose sensing, THI2 regulates thiamine biosynthesis, and UPC2 regulates sterol biosynthesis. This is perhaps not surprising, as metabolic genes have higher expression changes within the metabolic cycle, and hence presumably a lower level of noise. However, this highlights an important feature of our model: even if the absolute fraction of combinatorial interactions recovered is rather low, predictions have higher confidence for the specific biological processes investigated in the given experiment.

4.5 Comparison with other methods

To assess the relative merits of our method (which we denote as combinatorial FHMM, cFHMM), we performed an extensive comparative study with two recently published methods for reconstructing TF profiles. Shi *et al.* (2009) used FHMMs with inputs to simultaneously infer TF activities and post-transcriptional regulation in TFs; in our case, we are interested only in the TF inference part of the model, so that their model reduces to a simplified form of our model without the non-linear interactions of TFs. This method is denoted as FHMM. The other method we compare to is the TFInfer model (Sanguinetti *et al.*, 2006; Asif *et al.*, 2010). This is a log-linear models using a discrete time state space model for the TF activities. To compare the binary TF states obtained with the other two methods with the TFInfer results, we binarize the inferred TF activities using the average of the inferred temporal profile of each TF in the network (activity 0 if below average, 1 if above). We use three criteria to evaluate the performance of our method with these methods; run-time, mean squared error (MSE) in reconstructing gene expression profiles and the Hamming distance between the inferred states of the TFs.

It should be stressed that the method proposed here models the non-linear interactions of the transcription factors at the promoters, something that neither of the competitor methods can do. The flipside of this extra flexibility is that more time is required to execute the algorithm.

As an initial benchmark, we conducted experiments on simulated data (40 time points) with two different connectivities, the *E.coli* connectivity data (302 genes, 6 TFs, average 60 targets per TF) and the yeast connectivity with varying network sizes (25, 50 or 75 TFs). The data were generated from the cFHMM model; however, we noted that both cFHMM and FHMM managed to give good reconstructions of the TF profiles (obviously FHMM could not capture the coefficients of the non-linear effects). This is essentially due to the sparsity of the connectivity; in particular, the connectivity matrix in the yeast data is sparser, so that FHMM is a very good model for most genes. For the denser *E.coli* network, the performance of cFHMM was significantly better, particularly in terms of MSE (results shown in the Supplementary Material).

Table 2 presents the comparison of the results obtained using our method with two other methods on the real datasets considered in this study. In the *E.coli* dataset, the results of FHMM and cFHMM are similar in terms of TF reconstruction (average Hamming distance 0.067); this is probably due to fact that we did not find any combinatorial interactions at 5% significance level. In the other datasets, we obtained a relatively larger Hamming distances between FHMM and both cFHMM and TFInfer (0.2688 and 0.2502, respectively). These datasets contained many more time points, which allowed the recovery of a small but non-negligible number of combinatorial interactions, leading to the predictions of cFHMM (which does take these interactions into account) to be significantly different from the two linear methods.

5 CONCLUSION

We present a novel method to infer combinatorial interactions between transcriptional regulators from expression data and network

connectivity data. To our knowledge, this is the first statistical method which simultaneously infers TF activities and their combinatorial interactions in large-scale networks. We model TF activities as latent binary variables with Markovian dynamics; gene expression is determined by the latent TF activities through a non-linear likelihood which allows for pairwise interactions between TFs. According to our model, gene expression is digitized; digitized levels of gene expression have recently been shown to yield computational savings and more robust predictions (Tuna and Niranjan, 2010). The principal novelty of our work in this perspective is to connect the level of discretization with the state of underlying regulators.

FHMMs have been previously used to model TF activities (Shi *et al.*, 2009); in that work, further dependencies were included between TF mRNA expression levels and their predicted activities, which enabled to predict possible post-transcriptional modifications in TFs. Naturally, it should be possible to combine both our approach and their approach to give a model capable of simultaneously inferring TF activities, combinatorial interactions and post-transcriptional regulations. This would also allow to remove the assumption, hard-wired into our model as well as many other related models, that TF activity is independent of their mRNA expression levels. While in many cases this assumption is justified by the fact that measurement of TF gene expression are often poor proxies for their activity state, it is plausible that, at least in some situations, mRNA expression levels of TF genes will bear some influence on their activity.

Funding: H.M.S.A. acknowledges funding from University of Engineering & Technology Lahore under the Faculty development Program. G.S. is supported by the Scottish Government through the Scottish Informatics and Computer Science Alliance (SICSA) initiative.

Conflict of Interest: none declared.

REFERENCES

Barenco,M. *et al.* (2006) Ranked prediction of p53 targets using hidden variable dynamical modelling. *Genome Biol.*, **7**, R25.

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

Bhoite,L.T. *et al.* (2002) Mutations in the Pho2 (Bas2) transcription factor that differentially affect activation with its partner proteins Bas1, Pho4, and Swi5. *J. Biol. Chem.*, **277**, 37612–37618.

Bishop,C. (2006) *Pattern Recognition and Machine Learning*. Springer, New York.

Coffman,J. *et al.* (1996) Gat1p, a GATA family protein whose production is sensitive to nitrogen catabolite repression, participates in transcriptional activation of nitrogen-catabolic genes in Saccharomyces cerevisiae. *Mol. Cell. Biol.*, **16**, 847.

Daignan-Fornier,B. and Fink,G. (1992) Coregulation of purine and histidine biosynthesis by the transcriptional activators BAS1 and BAS2. *Proc. Natl Acad. Sci. USA*, **89**, 6746.

Davidge,K. *et al.* (2009) Carbon monoxide-releasing antibacterial molecules target respiration and global transcriptional regulators. *J. Biol. Chem.*, **284**, 4516.

Ghahramani,Z. and Jordan,M. (1997) Factorial hidden Markov models. *Mach. Learn.*, **29**, 245–273.

Hahn,S. *et al.* (1988) The HAP3 regulatory locus of Saccharomyces cerevisiae encodes divergent overlapping transcripts. *Mol. Cell. Biol.*, **8**, 655.

Harbison,C. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Jordan,M. *et al.* (1999) An introduction to variational methods for graphical models. *Mach. Learn.*, **37**, 183–233.

Lawrence,N.D. *et al.* (2006) Modelling transcriptional regulation using Gaussian processes. In *Advances in Neural Information Processing Systems 19*. MIT Press, p. 785.

Lee,T.I. *et al.* (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.

Liao,J. *et al*. (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl Acad. Sci. USA*, **100**, 15522–15527.

Opper,M. and Sanguinetti,G. (2010) Learning combinatorial transcriptional dynamics from gene expression data. *Bioinformatics*, **26**, 1623.

Partridge,J. *et al.* (2007) Transition of Escherichia coli from aerobic to micro-aerobic conditions involves fast and slow reacting regulatory components. *J. Biol. Chem.*, **282**, 11230.

Ptashne,M. and Gann,A. (2002) *Genes & Signals*. Cold Spring Harbor, New York.

Rogers,S. *et al.* (2007) Bayesian model-based inference of transcription factor activity. *BMC Bioinformatics*, **8**, S2.

Sabatti,C. and James,G.M. (2006) Bayesian sparse hidden components analysis for transcription regulation networks . *Bioinformatics*, **22**, 739–746.

Sanguinetti,G. *et al.* (2006) Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, **22**, 2775–2781.

Sanguinetti,G. *et al.* (2009) Switching regulatory models of cellular stress response. *Bioinformatics*, **25**, 1280–1286.

Scott,S. *et al.* (2000) Roles of the Dal82p domains in allophanate/oxalurate-dependent gene expression in Saccharomyces cerevisiae. *J. Biol. Chem.*, **275**, 30886.

Asif,H.M.S. *et al.* (2010) TFInfer: a tool for probabilistic inference of transcription factor activities. *Bioinformatics*, **26**, 2635.

Shi,Y. *et al.* (2009) A combined expression-interaction model for inferring the temporal activity of transcription factors. *J. Comput. Biol.*, **16**, 1035–1049.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Tu,B. *et al.* (2005) Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science*, **310**, 1152.

Tuna,S. and Niranjan,M. (2010) Reducing the algorithmic variability in transcriptome-based inference. *Bioinformatics*, **26**, 1185.

Xing,Y. *et al.* (1993) Mutations in yeast HAP2/HAP3 define a hybrid CCAAT box binding domain. *EMBO J.*, **12**, 4647.

Xu,Z. and Tsurugi,K. (2007) Role of Gts1p in regulation of energy-metabolism oscillation in continuous cultures of the yeast Saccharomyces cerevisiae. *Yeast*, **24**, 161–170.