Bioinformatics, 33(23), 2017, 3776–3783
doi: 10.1093/bioinformatics/btx508
Advance Access Publication Date: 26 August 2017
Original Paper

OXFORD

Systems biology

Efficient inference for sparse latent variable models of transcriptional regulation

Zhenwen Dai[1,2,†], Mudassar Iqbal[3,*,†], Neil D. Lawrence[1,2] and Magnus Rattray[3]

[1]Department of Computer Science, University of Sheffield, Sheffield, UK, [2]Amazon Research, Cambridge, UK and [3]Division of Informatics, Imaging & Data Sciences, Faculty of Biology, Medicine, and Health Sciences, University of Manchester, Manchester, UK

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Jonathan Wren

Received on March 27, 2017; revised on July 24, 2017; editorial decision on August 6, 2017; accepted on August 25, 2017

Abstract

Motivation: Regulation of gene expression in prokaryotes involves complex co-regulatory mechanisms involving large numbers of transcriptional regulatory proteins and their target genes. Uncovering these genome-scale interactions constitutes a major bottleneck in systems biology. Sparse latent factor models, assuming activity of transcription factors (TFs) as unobserved, provide a biologically interpretable modelling framework, integrating gene expression and genome-wide binding data, but at the same time pose a hard computational inference problem. Existing probabilistic inference methods for such models rely on subjective filtering and suffer from scalability issues, thus are not well-suited for realistic genome-scale applications.

Results: We present a fast Bayesian sparse factor model, which takes input gene expression and binding sites data, either from ChIP-seq experiments or motif predictions, and outputs active TF-gene links as well as latent TF activities. Our method employs an efficient variational Bayes scheme for model inference enabling its application to large datasets which was not feasible with existing MCMC-based inference methods for such models. We validate our method on synthetic data against a similar model in the literature, employing MCMC for inference, and obtain comparable results with a small fraction of the computational time. We also apply our method to large-scale data from *Mycobacterium tuberculosis* involving ChIP-seq data on 113 TFs and matched gene expression data for 3863 putative target genes. We evaluate our predictions using an independent transcriptomics experiment involving over-expression of TFs.

Availability and implementation: An easy-to-use Jupyter notebook demo of our method with data is available at https://github.com/zhenwendai/SITAR.

Contact: mudassar.iqbal@manchester.ac.uk

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

1 Introduction

A typical biological study of cellular response to external stress/stimuli or certain knock-outs leads to the measurement of gene expression patterns of thousands of differentially expressed genes (Galagan *et al.*, 2013; Nieselt *et al.*, 2010). Furthermore, transcription factor binding sites data from literature as well as *de novo* computational motif predictions (Gama-Castro *et al.*, 2016; Sierro *et al.*, 2008; Studholme *et al.*, 2004), in the case of small prokaryotic genomes, are accessible for many well-studied organisms. Large-scale ChIP-seq assays, e.g. (Galagan *et al.*, 2013; Minch *et al.*, 2015) are also available, detailing the genome-wide binding patterns of specific transcription factor proteins (TFs). A subsequent

computational and statistical challenge is to integrate these data in order to obtain a quantitative picture of the underlying regulatory interactions between TF proteins and target genes. In the last decade, many statistical methods have been proposed (see Marbach *et al.*, 2010 for a review) which infer gene regulatory networks by exploiting correlation patterns in the gene expression data. However, mRNA expression data alone cannot disentangle the complex wiring of regulatory interactions (Marbach *et al.*, 2012). Other experimental techniques for elucidation of regulatory interactions also have limitations, e.g. ChIP-seq experiments do not determine the effect of TF binding events on target genes and it is difficult to distinguish direct versus indirect regulatory effects in TF perturbation experiments (Siahpirani and Roy, 2017). It is therefore necessary to integrate different genomic datasets in order to infer context-specific regulatory networks.

TF proteins may be regulated at the post-transcriptional level and therefore an important consideration in modelling transcriptional regulation is that measured RNA levels often do not provide a good proxy for the concentration of active TFs. Bayesian statistical methods, especially sparse latent factor models (Carvalho *et al.*, 2008; Iqbal *et al.*, 2012; Pournara and Wernisch, 2007; Sabatti and James, 2006; Sanguinetti *et al.*, 2006) which are the main focus of this study, offer a flexible framework for data integration. These methods treat the regulator activities as latent (unobserved) variables which can be inferred from the RNA expression levels of their target genes. Sparse latent factor models also have other biological applications, e.g. modelling cellular heterogeneity in single-cell RNA-seq data (Buettner *et al.*, 2015). The core underlying hypothesis in the context of transcriptional regulation is that a large number of observed gene expression profiles can be explained by the unobserved activities of a small number of regulatory proteins. Biologically meaningful prior information on the underlying transcriptional regulatory network (between TFs and genes) can be obtained from computational motif predictions (Li *et al.*, 2002; Studholme *et al.*, 2004) or large-scale ChIP-seq experiments (Galagan *et al.*, 2013). Sparse factor models combine the prior network with relevant gene expression data in order to infer the true underlying regulatory connections driving gene expression in the experiment under study, as well as the activities of the regulators and the strength of regulatory effects.

Despite the appeal of sparse factor models for biological applications, inference in these models presents a computational challenge. Markov Chain Monte Carlo (MCMC) can be used to carry out model inference (Iqbal *et al.*, 2012; Sabatti and James, 2006) and has the advantage of quantifying uncertainty in all the inferred parameters. However, MCMC suffers from convergence issues, and becomes computationally prohibitive even for a moderate number of regulators. This lack of scalability hinders the application of sparse factor models, since a typical biological experiment involves many dozens of TFs and thousands of genes. As more ChIP-seq and gene expression data becomes available, efficient methods are therefore needed to extract biological information from these data.

Here, we present a novel method in the family of sparse factor models, named SITAR (Sparse latent varIable model of TrAnscritional Regulation), in which a spike-and-slab prior is used to induce sparsity in network connections. We propose an efficient variational inference method by deriving a closed-form variational lower-bound for our model. This adaptation of the inference scheme enables us to scale up the inference over much larger datasets than current methods based on MCMC can cope with. We test our method on synthetic data against a similar published method which uses MCMC-based inference. We then apply our method to a large-scale dataset (Galagan *et al.*, 2013; Minch *et al.*, 2015) from *Mycobacterium tuberculosis* (MTB) with ChIP-seq data for 113 TFs and matched gene expression data for 3863 genes, which include multiple time series covering hypoxia and over-expression experiments for some TFs. This is one of the largest application of its kind and the running time for our method for this dataset was about 7 h on a laptop.

The paper is organized as follows. In Section 2, we describe our model for integrating binding sites and gene expression data. We describe the choices of the prior on model parameters and present the variational inference algorithm and method for recovery of latent activities. In Section 3 we describe validation results on synthetic data and results on an application to a large-scale real dataset from MTB. We report biological validation of our predictions on the MTB dataset by comparing our inference results to results from an independent TF over-expression study which was not used for learning the model.

2 Materials and methods

We model gene expression as a weighted sum of TF activities: $e_{it} = \sum_{j=1}^{L} a_{ij} p_{jt} + \epsilon_{it}$, where e_{it} represents the expression of gene i in experiment t, a_{ij} is the control strength of TF j on gene i, p_{jt} is a proxy for the concentration of active form of TF j in experiment t and ϵ_{it} accounts for measurement errors and biological variation. In matrix notation the model is formulated as

$$\mathbf{E} = \mathbf{AP} + \epsilon, \tag{1}$$

where $\mathbf{E} \in \mathbb{R}^{N \times M}$, $\mathbf{A} \in \mathbb{R}^{N \times L}$, $\mathbf{P} \in \mathbb{R}^{L \times M}$, N is the number of genes, M is the number of experiments and L is the number of TFs. Both the control strength of TFs, $\mathbf{A}$, and the concentration of active TFs, $\mathbf{P}$, are unknown. By assuming that the noise ϵ follows an *i.i.d.* Gaussian distribution, we can define the distribution of the expression data $\mathbf{E}$ as

$$p(\mathbf{E}|\mathbf{A}, \mathbf{P}) = \prod_{t=1}^{M} \mathcal{N}(\mathbf{E}_t|\mathbf{AP}_t, \sigma^2\mathbf{I}), \tag{2}$$

where $\mathbf{E}_t$ and $\mathbf{P}_t$ indicates the tth column of $\mathbf{E}$ and $\mathbf{P}$, and σ^2 is the variance of Gaussian noise.

We define a unit variance Gaussian prior on the elements of $\mathbf{P}$, i.e. $p(\mathbf{P}_t) = \mathcal{N}(\mathbf{P}_t|0, \mathbf{I})$, and marginalize out $\mathbf{P}$ from Equation (2):

$$p(\mathbf{E}|\mathbf{A}) = \prod_{t=1}^{M} \mathcal{N}(\mathbf{E}_t|0, \mathbf{AA}^\top + \sigma^2\mathbf{I}). \tag{3}$$

Only a small subset of genes are controlled by individual TFs due to biological constraints and therefore $\mathbf{A}$ is known to be sparse. To keep inference tractable we introduce some hard constraints on the allowed connections through a binary connectivity matrix $\mathbf{X} \in \mathbb{R}^{N \times L}$ which is obtained from motif analysis or ChIP-seq data (as explained in Section 3). Entry $x_{ij} = 0$ indicates that TF j cannot control gene i, i.e. $a_{ij} = 0$. However, even if a connection is allowed by the connectivity matrix it may not be active, e.g. when $x_{ij} = 1$ then TF j does not necessarily control the corresponding gene i. To model this we introduce a latent binary variable for each pair of TF and gene, $\mathbf{S} \in \mathbb{R}^{N \times L}$, which control the connections between TFs and genes. The probability distribution of the expression matrix is modified to be:

$$p(\mathbf{E}|\mathbf{A}, \mathbf{S}) = \prod_{t=1}^{M} \mathcal{N}(\mathbf{E}_t|0, (\mathbf{A} \circ \mathbf{S})(\mathbf{A} \circ \mathbf{S})^\top + \sigma^2\mathbf{I}), \tag{4}$$

where $\mathbf{A} \circ \mathbf{S}$ indicates the element-wise multiplication between $\mathbf{A}$ and $\mathbf{S}$. We incorporate the information of the connectivity matrix

into the prior distribution of these binary variables. For the entries of S with $x_{ij} = 0$, we set $p(s_{ij}) = 1 - s_{ij}$. For $x_{ij} = 1$, we assume s_{ij} has a prior probability π_j, $p(s_{ij}|\pi_j) = \pi_j^{s_{ij}}(1-\pi_j)^{1-s_{ij}}$, and π_j follows a beta prior $p(\pi_j) = \text{Beta}(2,2)$. Finally, with a unit Gaussian prior distribution for $\mathbf{A}$, $p(a_{ij}) = \mathcal{N}(a_{ij}|0,1)$, the marginal likelihood distribution for our model is

$$p(\mathbf{E}) = \int p(\mathbf{E}|\mathbf{A},\mathbf{S})p(\mathbf{A})p(\mathbf{S}|\pi)p(\pi)\mathrm{d}\mathbf{A}\mathrm{d}\mathbf{S}\mathrm{d}\pi. \tag{5}$$

Given expression data $\mathbf{E}$, we can then write the posterior distribution of the regulatory interactions using Bayes rule:

$$p(\mathbf{A},\mathbf{S}|\mathbf{E}) = \frac{\int p(\mathbf{E}|\mathbf{A},\mathbf{S})p(\mathbf{A})p(\mathbf{S}|\pi)p(\pi)\mathrm{d}\pi}{p(\mathbf{E})}. \tag{6}$$

From this, we can estimate $\mathbf{A}$ and $\mathbf{S}$ by computing their expections $\langle\mathbf{A}\rangle_{p(\mathbf{A},\mathbf{S}|\mathbf{E})}$ and $\langle\mathbf{S}\rangle_{p(\mathbf{A},\mathbf{S}|\mathbf{E})}$, and the posterior also provides credible regions for these estimates.

2.1 Variational inference

As mentioned above, our aim is to infer the posterior distribution of the regulatory network $\mathbf{S}$, the control strength $\mathbf{A}$ and the activity profiles of transcription factors $\mathbf{P}$ by observing gene expression data $\mathbf{E}$. Unfortunately, exact inference of the posterior is infeasible due to the intractable integral in Equation (5). Sampling-based approaches such as Markov Chain Monte Carlo (MCMC) have been developed (Iqbal *et al.*, 2012) but are very time-consuming and prohibitively slow for large-scale datasets, e.g. thousands of genes and hundreds of TFs. In this work, we propose an efficient inference algorithm based on a variational approximation which reduces the computational run-time for large datasets from weeks to hours.

Variational inference avoids the evaluation of the intractable marginal likelihood by optimizing parametric posterior distributions with respect to a lower bound of the log marginal likelihood. We assume a variational posterior distribution $q(\mathbf{A},\mathbf{S},\pi)$ and derive a lower bound such as

$$\log p(\mathbf{E}) \geq \int q(\mathbf{A},\mathbf{S},\pi)\log\frac{p(\mathbf{E}|\mathbf{A},\mathbf{S})p(\mathbf{A})p(\mathbf{S}|\pi)p(\pi)}{q(\mathbf{A},\mathbf{S},\pi)}\mathrm{d}\mathbf{A}\mathrm{d}\mathbf{S}\mathrm{d}\pi. \tag{7}$$

For our model, the standard mean-field approximation $q(\mathbf{A},\mathbf{S},\pi) = q(\mathbf{A})q(\mathbf{S})q(\pi)$ is still intractable due to the covariance matrix inversion in Equation (4). In this paper, we exploit the fact that our model can be viewed as a Gaussian Process latent variable model (Lawrence, 2005) with a linear kernel and a spike-and-slab prior. This enables us to adopt the sparse Gaussian Process formulation (Titsias, 2009; Titsias and Lawrence, 2010) for our model. We first rewrite our likelihood expression (4) in the form of a Gaussian Process (GP):

$$p(\mathbf{E}|\mathbf{F}) = \prod_{t=1}^{M}\mathcal{N}(\mathbf{e}_t|\mathbf{f}_t,\sigma^2\mathbf{I}), \tag{8}$$

$$p(\mathbf{f}_t|\mathbf{A},\mathbf{S}) = \mathcal{N}(\mathbf{f}_t|0,\mathbf{K}_{ff}), \tag{9}$$

where $\mathbf{F}$ is the noise-free observation of the gene expression data $\mathbf{E}$ and $\mathbf{K}_{ff}$ is the covariance matrix of $\mathbf{F}$ computed according to our model, i.e. $\mathbf{K}_{ff} = (\mathbf{A}\circ\mathbf{S})(\mathbf{A}\circ\mathbf{S})^\top$. The sparse GP approximation introduces an auxiliary latent variable $\mathbf{U}\in\mathbb{R}^{L\times L}$ with a corresponding inducing input $\mathbf{I}\in\mathbb{R}^{L\times L}$ ($\mathbf{I}$ is an identity matrix.) This allows us to reformulate the prior distribution of $\mathbf{F}$ in terms of the auxiliary variable:

$$p(\mathbf{f}_t|\mathbf{u}_t,\mathbf{A},\mathbf{S}) = \mathcal{N}(\mathbf{f}_t|\mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}_t, \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{fu}^\top), \tag{10}$$

$$p(\mathbf{u}_t) = \mathcal{N}(\mathbf{u}_t|0,\mathbf{K}_{uu}), \tag{11}$$

where the conditional distribution (10) is derived through GP inference and $\mathbf{K}_{uu}$ and $\mathbf{K}_{fu}$ are the covariance matrices, i.e. $\mathbf{K}_{uu} = \mathbf{X}\mathbf{X}^\top$ and $\mathbf{K}_{fu} = (\mathbf{A}\circ\mathbf{S})\mathbf{X}^\top$. Note that marginalizing out the auxiliary variable $\mathbf{U}$ in Equations (10) and (11) returns the original distribution of $\mathbf{F}$ in Equation (9). Following the sparse GP formulation, we define the variational posterior distribution as $q(\mathbf{F},\mathbf{U},\mathbf{A},\mathbf{S},\pi) = p(\mathbf{F}|\mathbf{U},\mathbf{A},\mathbf{S})q(\mathbf{U})q(\mathbf{A},\mathbf{S})q(\pi)$ and obtain a lower bound of the marginal likelihood:

$$\mathcal{L} = \mathcal{F} - \mathrm{KL}(q(\mathbf{U})\,||\,p(\mathbf{U})) - \mathrm{KL}(q(\mathbf{A},\mathbf{S})q(\pi)\,||\,p(\mathbf{A})p(\mathbf{S}|\pi)p(\pi)), \tag{12}$$

where $\mathcal{F} = \langle\log p(\mathbf{E}|\mathbf{F},\mathbf{U},\mathbf{A},\mathbf{S})\rangle_{p(\mathbf{F}|\mathbf{U},\mathbf{A},\mathbf{S})q(\mathbf{A},\mathbf{S})q(\mathbf{U})}$. Since $\mathbf{A}$ and $\mathbf{S}$ are often strongly correlated in the posterior distribution, their variational posterior is defined as a conditional distribution,

$$q(\mathbf{S}) = \prod_{i=1}^{N}\prod_{j=1}^{L}\gamma_{ij}^{s_{ij}}(1-\gamma_{ij})^{(1-s_{ij})},$$

$$q(a_{ij}|s_{ij}=1) = \mathcal{N}(a_{ij}|\mu_{ij},c_{ij}), \tag{13}$$

where γ_{ij} is the posterior probability of TF j controlling the gene i and μ_{ij} and c_{ij} are the posterior mean and variance of the control strength. Note that the distribution $q(a_{ij}|s_{ij}=0)$ is not defined explicitly, because, as the switch variable is zero, the control strength does not influence the likelihood anymore, so that $q(a_{ij}|s_{ij}=0)$ will only appear inside the KL divergence, which makes it always equal to the prior distribution $p(\mathbf{A})$. With the above posterior distribution of $q(\mathbf{A},\mathbf{S})$, the first expectation in Equation (12) can be solved analytically.

$$\begin{aligned}\langle\log p(\mathbf{E}|\mathbf{F},\mathbf{U},\mathbf{A},\mathbf{S})\rangle_{p(\mathbf{F}|\mathbf{U},\mathbf{A},\mathbf{S})q(\mathbf{U})q(\mathbf{A},\mathbf{S})} &= -\frac{NM}{2}\log 2\pi\sigma^2 \\ &\quad -\frac{1}{2\sigma^2}\Big\langle\sum_{t=1}^{M}\mathbf{u}_t^\top\mathbf{K}_{uu}^{-1}\Psi_2\mathbf{K}_{uu}^{-1}\mathbf{u}_t\Big\rangle_{q(\mathbf{U})} \\ &\quad +\sum_{t=1}^{M}\frac{1}{\sigma^2}\mathbf{e}_t^T\Psi_1\mathbf{K}_{uu}^{-1}\langle\mathbf{u}_t\rangle_{q(\mathbf{U})} - \frac{1}{2\sigma^2}\sum_{t=1}^{M}\mathbf{e}_t^\top\mathbf{e}_t \\ &\quad -\frac{M}{2\sigma^2}\psi_0 + \frac{M}{2\sigma^2}\mathrm{Tr}(\mathbf{K}_{uu}^{-1}\Psi_2)\end{aligned} \tag{14}$$

where ψ_0, Ψ_1 and Ψ_2 denote the expectation of the covariance matrices w.r.t. $q(\mathbf{A},\mathbf{S})$, i.e. $\psi_0 = \mathrm{Tr}(\langle\mathbf{K}_{ff}\rangle_{q(\mathbf{A},\mathbf{S})})$, $\Psi_1 = \langle\mathbf{K}_{fu}\rangle_{q(\mathbf{A},\mathbf{S})}$, $\Psi_2 = \langle\mathbf{K}_{fu}^\top\mathbf{K}_{fu}\rangle_{q(\mathbf{A},\mathbf{S})}$. For the linear kernel used in this paper, ψ_0, Ψ_1 and Ψ_2 can be derived analytically (we call them psi-statistics) as:

$$\psi_0 = \sum_{i=1}^{N}\sum_{j=1}^{L}\gamma_{ij}(\mu_{ij}^2 + c_{ij}), \tag{15}$$

$$(\Psi_1)_{id} = \sum_{j=1}^{L}\gamma_{ij}\mathbf{x}_{dj}\mu_{ij}, \tag{16}$$

$$(\Psi_2)_{dd'} = \sum_{i=1}^{N}\left(\sum_{j=1}^{L}\gamma_{ij}\mathbf{z}_{dj}\mathbf{z}_{d'j}(\mu_{ij}^2 + c_{ij}) + \sum_{j=1}^{L}\sum_{j'\neq j}\gamma_{ij}\gamma_{ij'}\mathbf{z}_{dj}\mathbf{z}_{d'j'}\mu_{ij}\mu_{ij'}\right). \tag{17}$$

The optimal distribution of $q(\mathbf{U})$ can be derived analytically from Equation (12) by setting its derivative to be zero:

$$q(\mathbf{u}_t) = \mathcal{N}(\mathbf{u}_t|\mathbf{K}_{uu}(\mathbf{K}_{uu}+\Psi_2)^{-1}\Psi_1^T\mathbf{e}_t, \mathbf{K}_{uu}(\mathbf{K}_{uu}+\Psi_2)^{-1}\mathbf{K}_{uu}). \tag{18}$$

By substituting the optimal variational distribution of $q(\mathbf{U})$, the variational lower bound can be formulated in closed form which enables us to perform inference efficiently by optimizing the model

parameters and variational parameters with respect to the closed-form lower bound.

2.2 Recovering the activities of TFs

Using the lower bound of the log-marginal likelihood derived in the previous subsection, we can efficiently infer the posterior distribution of the connectivity $q(\mathbf{S})$ and the control strength if the link is connected $q(a_{ij}|s_{ij}=1)$. Besides these posterior distributions, we are also interested in the posterior distribution of the latent activity profiles of TFs $p(\mathbf{P}|\mathbf{E})$. As they are marginalized out in our model, their posterior distribution can be estimated as:

$$\begin{aligned} p(\mathbf{P}|\mathbf{E}) &= \int p(\mathbf{P}|\mathbf{E},\mathbf{A},\mathbf{S})p(\mathbf{A},\mathbf{S}|\mathbf{E})\mathrm{d}\mathbf{A}\mathrm{d}\mathbf{S} \\ &\approx \int p(\mathbf{P}|\mathbf{E},\mathbf{A},\mathbf{S})q(\mathbf{A},\mathbf{S})\mathrm{d}\mathbf{A}\mathrm{d}\mathbf{S} \end{aligned} \tag{19}$$

where we approximate the true posterior $p(\mathbf{A},\mathbf{S}|\mathbf{E})$ by the estimated variational posterior $q(\mathbf{A},\mathbf{S})$. According to the model definition, we can derive $p(\mathbf{P}|\mathbf{E},\mathbf{A},\mathbf{S})$ as:

$$\begin{aligned} p(\mathbf{P}|\mathbf{E},\mathbf{A},\mathbf{S}) &= \frac{p(\mathbf{E}|\mathbf{P},\mathbf{A},\mathbf{S})p(\mathbf{P})}{p(\mathbf{E}|\mathbf{A},\mathbf{S})} \\ &= \prod_{t=1}^{M}\mathcal{N}\left(\mathbf{p}_t|\sigma^2\Sigma_p(\mathbf{A}\circ\mathbf{S})^\top\mathbf{e}_t,\Sigma_p\right) \end{aligned} \tag{20}$$

where $\Sigma_p=(\sigma^{-2}(\mathbf{A}\circ\mathbf{S})^\top(\mathbf{A}\circ\mathbf{S})+\mathbf{I})^{-1}$. Due to the matrix inversion in Σ_p, the posterior in Equation (19) is not analytically tractable. However, since we only need to infer it once after optimizing the variational posterior $q(\mathbf{A},\mathbf{S})$, we numerically estimate the posterior mean and variance of the activity profiles through Monte Carlo integration.

3 Results and discussion

3.1 Simulation study

To assess the proposed model, we first generate synthetic data where the true regulatory network, control strengths and activities of TFs are known. To mimic a real network, we take a subset of the connectivity matrix from the real dataset from Iqbal *et al.* (2012) which was obtained from motif analysis. The resulting connectivity matrix contains 353 genes and 20 TFs. The control strength of each interaction and TF activities are sampled from a unit Gaussian. Then 94 gene expression measurements are generated according to the model with noise variance $\sigma^2=0.1$. For this validation experiment, we relied on this relatively smaller size network, to be able to run MCMC method long enough to ascertain the convergence and compare the efficiency and accuracy of SITAR against MCMC given the ground truth. As shown in Supplementary Figure S1, not all parameters are converged even after one week's run-time.

We apply the proposed model (SITAR) and the existing MCMC method from Iqbal *et al.* (2012) to the synthetic data. Both methods recover the underlying regulatory network with similar accuracy (93% for SITAR and 92% for MCMC) which is defined as proportion of correctly predicted positive and negative regulatory interactions. TF-gene links were called positive if the corresponding posterior probability was greater than 0.5, negative othewise. The latent motif activities are correctly recovered by both methods, as shown in Figure 2 for MCMC and Figure 1 for SITAR. The control strength and activity profiles are recovered up to an ambiguity of their sign. In order to compare with the ground truth, we correct the sign of the predicted control strength and TF latent activity according to the ground truth control strength. The underlying motivation for this simulation study was to show that given the ground truth network, SITAR is at least as accurate as the MCMC method. At the same time, we want to emphasize on the computational efficiency of our method where one single run of the method took about an hour on a laptop, achieving mean squared error MSE = 0.007 between the input gene expression data and model prediction, which was better than the MCMC even after a week-long run (Fig. 2). Performance of SITAR for synthetic data generated using different noise variances and varying the number of independent gene expression datasets was also studied as shown in Supplementary Tables S1 and S2, respectively.

3.2 Application: MTB hypoxia regulatory network

Next, we apply our method to much larger real data from *Mycobacterium tuberculosis* (MTB), involving a large-scale ChIP-seq assay as well as gene expression data measuring response to hypoxia treatment. MTB is known to have a robust hypoxia response network, involving a large number of TFs, facilitating its clinical latency within the host. Earlier analysis of the ChIP-seq data shows that MTB has a highly complex regulatory network, more diversified binding patterns of TFs as compared to previous models of promoter proximal binding, and context-specific occupancy of TF binding sites (Galagan, 2014). For this study, we downloaded 78 samples of gene expression data from GEO (GSE43466, samples GSM1084307 to GSM1084384), which included 10 hypoxia-relevant TF over-expression data, and 68 samples comprising three overlapping time-series covering hypoxia. We obtained pre-processed ChIP-seq data for 113 MTB TFs (through personal communication with James Galagan, the same data are now publicly available at http://genome.tbdb.org/annotation/genome/tbdb/Resources.html). This constituted a prior topology matrix with 113 columns representing TFs and 3863 rows representing genes with at least one TF connection and for which corresponding expression data was available. The binary entries of this matrix represent the binding of corresponding TF (column) and gene (row). We combined expression data from 78 individial samples in a matrix for all genes present in prior topology matrix, thus obtaining an expression data matrix with 3863 rows representing genes and 78 columns representing samples. This matrix was standardized to zero mean and unit variance.

With these many regulators in the model, which is still much less than the total TFs in MTB, a latent factor model with MCMC-

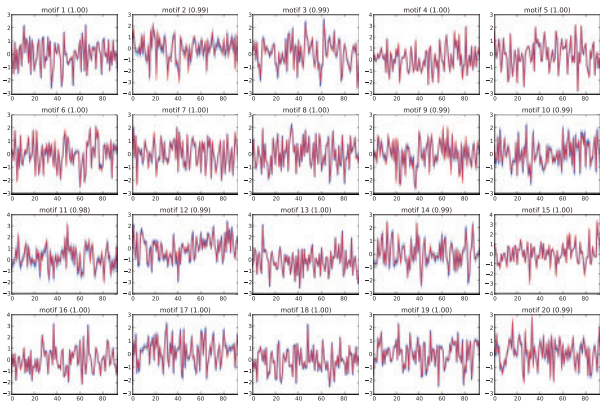

Fig. 1. The recovered motif activities (blue) by SITAR are compared with the ground truth activities (red). Their Pearson correlations are shown in the parentheses. For all subplots, x-axis shows the time and y-axis shows the normalized activities of corresponding motifs (Color version of this figure is available at *Bioinformatics* online.)

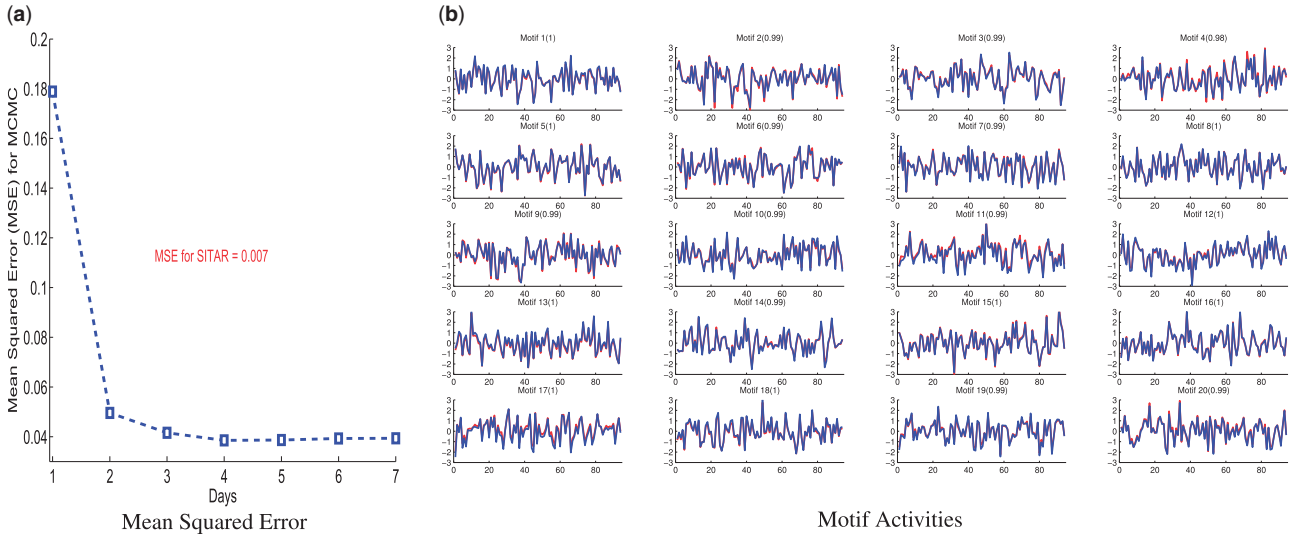

Fig. 2. Application of MCMC method (Iqbal *et al.*, 2012) to synthetic data, (**a**) Mean Squared Error (MSE) against the running time for MCMC(single MSE value for SITAR is shown in red), (**b**) Recovered motif activities by MCMC (red) are compared with the ground truth activities (blue), Pearson correlations are shown in the parentheses. For all subplots, *x*-axis shows the time and *y*-axis shows the normalized activities of corresponding motifs (Color version of this figure is available at *Bioinformatics* online.)

based inference would not be feasible while one run of SITAR was completed in just 7 h on a laptop with an 8-core intel processor (E5-2650v2). The results of our methods are shown in Figure 3 where we show that although the prior TF-gene links were obtained from high quality ChIP-seq data (unlike computational motif predictions), there are still large numbers of links for most of the TFs which were switched off by the model after integration with gene expression data (from 21 501 prior interactions obtained form ChIP-seq data, 8645 were switched off by the model). Among the original 113 TFs, there are 14 TFs which were switched off altogether since they have no significant targets. This confirms the quality of prior data on one hand, but also the ability of the model to discard the non-functional links conditional on the gene expression data under use. We also infer the latent activity of all TFs and in Figure 3 we show clusters of activities of TFs showing dynamic patterns in response to hypoxia. Here, we only show the clusters of latent profiles of 99 TFs plotted for three time-series (68 out of the 78 samples used in the model). For all TFs, we also show plots comparing recovered latent activity with gene expression data of corresponding TFs (see Supplementary Figs S3–S9).

In order to validate our network predictions we used an independent large-scale TF over-expression (TFOE) study in (Rustad *et al.*, 2014; Turkarslan et al., 2015) which was not used to learn the model. This data resulted from a systematic experiment of over-expressing 206 MTB TFs in order to quantify regulatory effects of each transcription factor. We downloaded and extracted the data (http://networks.systemsbiology.net/) for the TFs and genes in our prior network. As shown in Figure 4, we report the enrichment of predicted targets (of given TFs individually) which are showing differential expression in corresponding over-expression experiment ($|\log_2(\mathrm{FC})| \geq 0.5$), for example, for a known regulator involved in hypoxia response (Galagan *et al.*, 2013), RV3133c (DosR), 38% of its 200 significant targets satisfy the over-expression criteria, against 14% among the 3663 non-targets. Although over-expression criteria cannot be considered a completely reliable indicator of TF-gene connection due to complex regulatory control, there are still many regulators, e.g. RV0576, RV1846c, RV2557c among others, whose targets are highly enriched among the differentially expressed genes.

The right-hand panel in Figure 4 shows the correlation of the connection strength of TF-gene links predicted by the model against the over-expression indicator, i.e. log2FC. Here again, we have a number of regulators with very strong correlation which is plotted alongside the random background (calculated by randomly reassigning TF-gene connections). Among these, we have RV3133c (DosR) again, with very high correlation, which along with RV0081 are well-known primary hypoxia response regulators, identified in a number of studies in the literature (Galagan *et al.*, 2013; Park *et al.*, 2003). Other known regulators which score highly in this validation analysis include RV3574, RV0324, RV0757 (PhoP), RV1255c and RV2034 among others. Besides these known regulators, we also have few predictions which might be the novel regulators with significant role in hypoxia response and which might be worth analyzing further. These include RV0576 (ArsR family transcriptional regulator), RV1846c (Blal family transcriptional repressor), RV0818 (PhoB) and RV2359 (Zur, zinc uptake regulatory protein) and others as shown in Figure 4.

Furthermore, in order to compare the enrichments shown in Figure 4, we downloaded a relatively smaller hypoxia regulatory network (a Cytoscape session file from Galagan *et al.*, 2013) and performed similar validation analysis using same over-expression data as with our predicted network. After filtering for matching TFs and genes in our network and over-expression data, we have a final network including 39 TFs and 2763 target genes. In Figure 5(a), we show the enrichment of differential targets of TFs in this published network, again plotted alongside non-targets (out of total 2763 genes in that network). Overall, the matching TFs in both networks have similar level of enrichment despite the difference in the number of targets (since our network has a larger number of targets). Also, there are more TFs in our data with highest proportion of differential targets. Lastly, in order to ascertain if our predictions of links were more enriched in differential targets compared to ChIP-seq data (prior), in Figure 5(b), we plot the enrichment for 54 TFs (analyzed earlier in Fig. 4). The majority of predicted TFs are significantly enriched compared to the prior network with only a few exceptions. Overall, this analysis leads us to believe that the model is making biologically meaningful predictions.

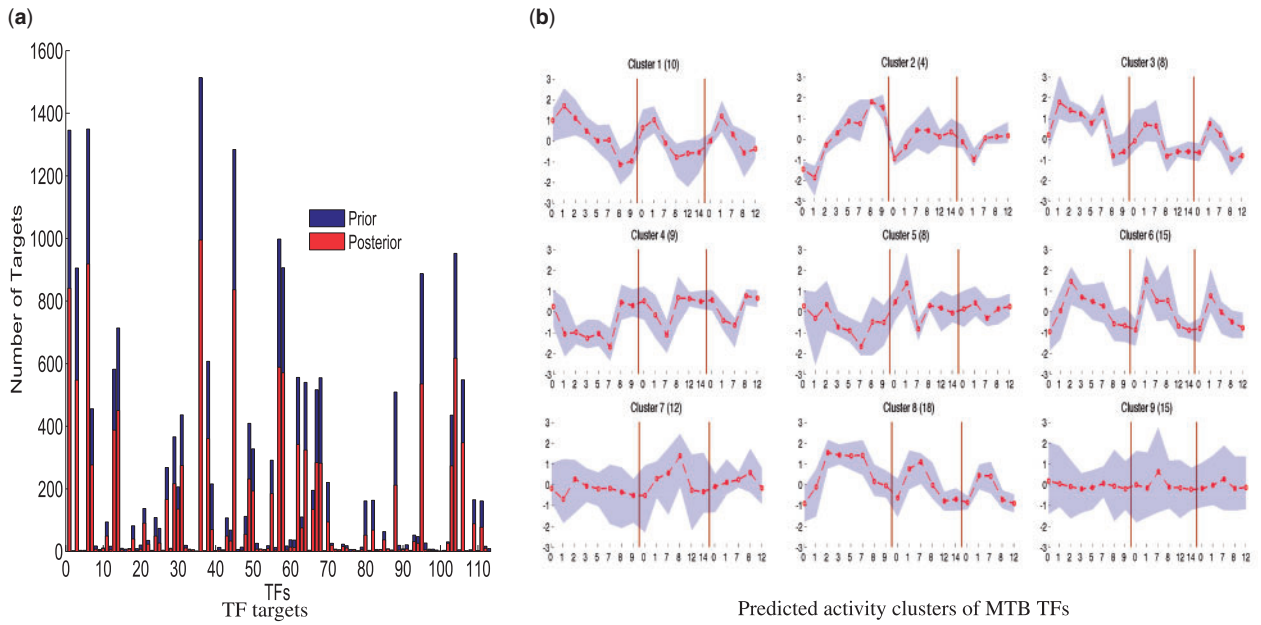

Fig. 3. Results of SITAR for MTB data. (**a**) Number of links (targets) for each TF in the prior network based on ChIP-seq data (blue) and posterior links, as predicted significant by SITAR. We used >0.5 cut-off on the posterior probability of links to decide if the link is supported or not. (**b**) Predicted TF activity profiles clustered into 9 clusters (using K-means method). The shaded area represents the activities of cluster members and red line shows the mean profile of the cluster, while the cluster number and number of its members are given in the suplot title. Vertical lines in each plot separate three separate hypoxia experiments (time series SG2, SG6 and SG7 respectively). Out of total 78 samples used in the model, in these plots, we only use 68 samples corresponding to three time-series which are partly replicated, overall cover day0 to day14 of hypoxia, each of them covering a subset of the days, with some overlap with other time series (for detail of the experimental design, see GSE43466). The x-axis shows the time (in Days) for experiments for individual time-series, while y-axis shows the normalized, replicate-averaged, activity of the corresponding TFs

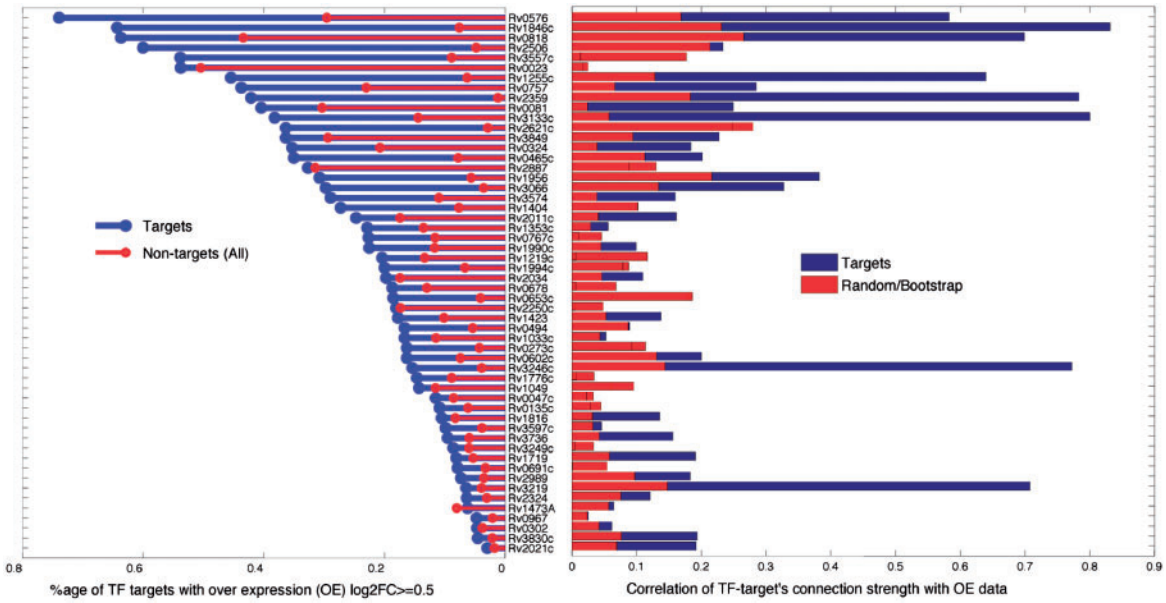

Fig. 4. Validation of our predictions against TF over-expression data from Rustad *et al.* (2014). The *y*-axis, showing 54 TFs which have at least 10 significant targets predicted by our method, is shared among two subplots. For each TF, the left panel shows the proportion of targets with log2 fold change (log2FC) ≥0.5 (blue for targets, red for non-targets), while in this data the average ratio of enrichments for targets and non-targets is 3.5. The right panel shows the absolute correlation of over-expression log2FC against the connection strength predicted by the model, blue for targets while red is the background calculated by randomly permuting the connection strength for the given TF targets, averaged over 100 iterations (Color version of this figure is available at *Bioinformatics* online.)

4 Conclusions

In conclusion, we present a Bayesian sparse factor analysis model coupled with a highly efficient inference scheme to make quantitative inferences of regulatory networks, including binary TF-gene interactions as well as latent activities of TFs, using binding sites and gene expression data in prokaryotic systems. As MCMC is commonly used for inference in these specific type of models, we validate our method against a study with similar modelling scheme for

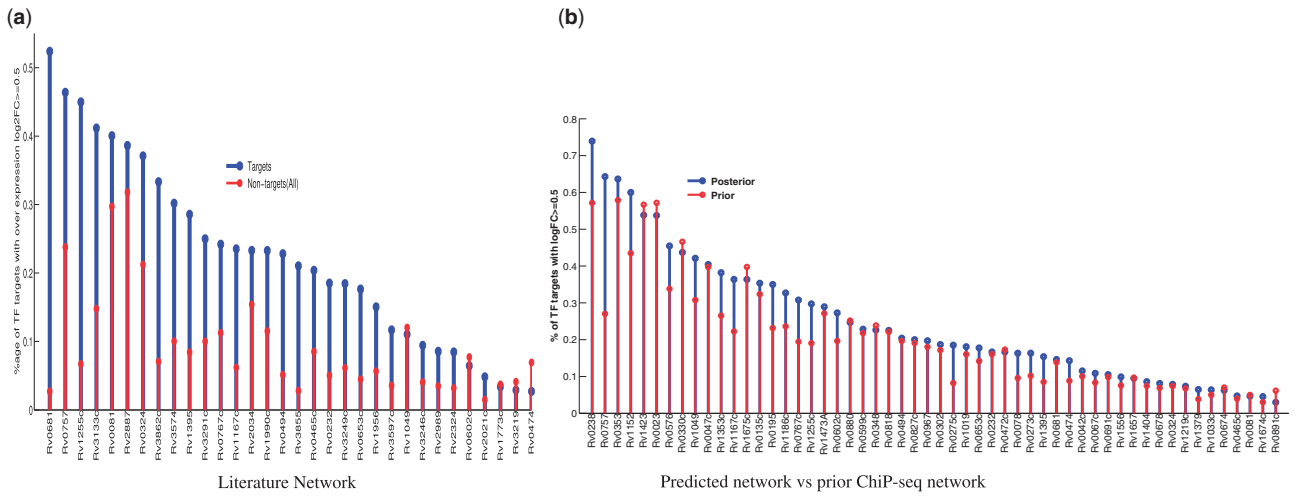

Fig. 5. (**a**) Validation of a smaller MTB network published in Galagan *et al.* (2013) against the TF over-expression data. We only used TFs which are in our input network and for which there is OE data. For each TF, we plot the proportion of targets with absolute log2 fold change (log2FC) $\geq$0.5 (blue for targets, red for non-targets in that network). These analysis was further restricted to TF with at least 10 targets. (**b**) Based on 54 TFs reported in Figure 4, further comparison of our predictions (blue) with the prior network (red) obtained from ChIP-seq data only. The average ratio of enrichments for targets over non-targets for (a) is 3.2 while 1.25 for (b) (Color version of this figure is available at *Bioinformatics* online.)

synthetic data where ground truth is known. Our method reproduced the network underlying synthetic data with high accuracy but with much higher efficiency than MCMC-based method. This led us to apply it to much larger real data on *M. tuberculosis* which constitute one of the largest application of regulatory network inference. There is one recent study (Arrieta-Ortiz *et al.*, 2015) where genome-wide network inference was performed on *B. subtilis* data of similar scale, but there are significant differences in the methodology. Inference was done in an iterative two-step procedure, first TF activities were estimated directly from known regulatory interactions using NCA (Liao *et al.*, 2003), which were then used in the prediction of regulatory direction and strength. On the other hand, our approach is significantly different in the sense that we employ a probabilistic model providing simultaneous inference of control strength and latent activities and providing a degree of uncertainty in our estimates.

We perform further validation of our predictions using independent transcriptomics data, compare our predictions against existing network from literature and make novel predictions about the role of certain regulators in MTB's response to hypoxia treatment. As more and more ChIP-seq and gene expression data becomes available, we believe our method will be a useful tool to make practical inference of large-scale networks regulating gene expression in prokaryotes. Also, since methodology is generic, we can imagine adaptation of our method for other problems in biology and beyond, especially in single-cell RNA-seq applications (see Buettner *et al.*, 2016). Another future direction for our work would be to use a non-linear kernel in our GP formulation or take into account interactions among hidden factors (as in Asif and Sanguinetti, 2011).

Acknowledgements

We thank James Galagan (Biomedical Engineering and Microbiology, Boston University) for his assistance with *M. tuberculosis* ChIP-seq data. Data used was generated in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institute of Health, Department of Health and Human Services, under contract no. HHSN272200800059C.

Funding

Medical Research Council (MRC), UK [MR/M012174/1 to M.I. and M.R.].

Conflict of Interest: none declared.

References

Arrieta-Ortiz,M. *et al.* (2015) An experimentally supported model of the *Bacillus subtilis* global transcriptional regulatory network. *Mol. Syst. Biol.*, **11**, 839.

Asif,H. and Sanguinetti,G. (2011) Large-scale learning of combinatorial transcriptional dynamics from gene expression. *Bioinformatics*, **27**, 1277–1283.

Buettner,F. *et al.* (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, **33**, 155–160.

Buettner,F. *et al.* (2016) Scalable latent-factor models applied to single-cell rna-seq data separate biological drivers from confounding effects. *bioRxiv*, 087775.

Carvalho,C. *et al.* (2008) High-dimensional sparse factor modeling: applications in gene expression genomics. *J. Am. Stat. Assoc.*, **103**, 1438–1456.

Galagan,J. (2014) Genomic insights into tuberculosis. *Nat. Rev. Genet.*, **15**, 307–320.

Galagan,J. *et al.* (2013) The *Mycobacterium tuberculosis* regulatory network and hypoxia. *Nature*, **499**, 178–183.

Gama-Castro,S. *et al.* (2016) Regulondb version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.*, **44**, D133–D143.

Iqbal,M. *et al.* (2012) Extracting regulator activity profiles by integration of *de novo* motifs and expression data: characterizing key regulators of nutrient depletion responses in *Streptomyces coelicolor*. *Nucleic Acids Res.*, **40**, 5227–5239.

Lawrence,N. (2005) Probabilistic non-linear principal component analysis with gaussian process latent variable models. *J. Mach. Learn. Res.*, **6**, 1783–1816.

Li,H. *et al.* (2002) Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl. Acad. Sci. USA*, **99**, 11772–11777.

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

Marbach,D. *et al.* (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. USA*, **107**, 6286–6291.

Marbach,D. *et al*. (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**, 796–804.

Minch,K. *et al*. (2015) The dna-binding network of *Mycobacterium tuberculosis*. *Nat. Commun*., **6**, 5829.

Nieselt,K. *et al*. (2010) The dynamic architecture of the metabolic switch in *Streptomyces coelicolor*. *BMC Genomics*, **11**, 10.

Park,H.-D. *et al*. (2003) Rv3133c/dosr is a transcription factor that mediates the hypoxic response of mycobacterium tuberculosis. *Mol. Microbiol*., **48**, 833–843.

Pournara,I. and Wernisch,L. (2007) Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, **8**, 61.

Rustad,T. *et al*. (2014) Mapping and manipulating the mycobacterium tuberculosis transcriptome using a transcription factor overexpression-derived regulatory network. *Genome Biol*., **15**, 502.

Sabatti,C. and James,G. (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.

Siahpirani,A. and Roy,S. (2017) A prior-based integrative framework for functional transcriptional regulatory network inference. *Nucleic Acids Res*., **45**

Sierro,N. *et al*. (2008) Dbtbs: a database of transcriptional regulation in bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res*., **36**, D93–D96.

Studholme,D. *et al*. (2004) Bioinformatic identification of novel regulatory DNA sequence motifs in streptomyces coelicolor. *BMC Microbiology*, **4**, 14.

Titsias,M. (2009). Variational learning of inducing variables in sparse Gaussian processes. In: *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS-09)*, vol. 5, pp. 567–574.

Titsias,M.K. and Lawrence,N. (2010). Bayesian Gaussian process latent variable model. In: *Proceedings of International Workshop on Artificial Intelligence and Statistics*, pp. 844–851.

Turkarslan,S. *et al*. (2015) A comprehensive map of genome-wide gene regulation in *Mycobacterium tuberculosis*. *Sci. Data*, **2**, 150010.