# Probabilistic Inference of Transcription Factor Concentrations and Gene-specific Regulatory Activities for Time-independent Data

Hafiz Muhammad Shahzad Asif and Guido Sanguinetti

Department of Computer Science,
University of Sheffield, Sheffield, United Kingdom
Shahzad.Asif@sheffield.ac.uk
G.Sanguinetti@sheffield.ac.uk

**Abstract.** Estimation of the quantitative relationship between transcription factor proteins and genes within the gene regulatory network is a major task in modeling biological systems. High-throughput experimental techniques provide architectural information about the regulatory network but concentration of transcription factor proteins and their role in transcription regulation is not well known. Probabilistic inference of concentrations of transcription factor proteins and gene-specific activities for time-independent gene expression data is discussed in this paper. Model proposed in [1] is employed to incorporate time-independent nature of gene expression data to measure the transcription factor profiles. Results are similar to ones obtained using the original model but with larger credibility intervals due to the non-temporal relationship between concentrations of transcription factor proteins in different experimental conditions.

## 1 Introduction

Biological cells are complex systems made of several thousand proteins that interact with each other and produce different proteins while sensing different environmental conditions. This information processing task is carried out by the dynamical system composed of genes and transcription factors, transcription network, that determines the rate of production of different proteins [2]. In this dynamical system, input signal causes the changes in the transcription factor activities resulting in the change of production rate of other genes. Quantitative estimation of transcription factor activities and gene regulation in regulatory networks is essential for modeling cellular processes. Owing to recent advances in high-throughput techniques [3, 4], some connectivity information is available and there is a need to analyze this qualitative connectivity information to generate quantitative network structures.

Many different mathematical modeling techniques are available for gene transcription analysis. In network component analysis (NCA), a dimensionality reduction approach is used to generate network of genes and transcription factor

activities (TFAs)[5]. In this work, network structure based on hidden regulatory signals is extracted using constrained maximum-likelihood procedure using Gaussian and i.i.d. assumptions. Another work focusing on the same task using singular value decomposition (SVD) is discussed in [6]. Main limitation of these methods is the difficulty in associating confidence interval with results of these methods due to the non-probabilistic nature of these methods. A more reliable approach is discussed in [7] using dynamical Bayesian networks (DBN) but computationally expensive.

It is important to associate credibility intervals with the results obtained using gene transcription analysis. To achieve this objective, a different class of probabilistic modeling techniques are available in literature. State space models (SSM) are used in [8] to extract the transcription networks from gene expression data. SSMs are very suitable for gene expression data analysis as SSMs also incorporate the hidden state variables of the system. In [8], task of gene transcription analysis is discussed using classical and Bayesian approach. Variational approximations are used for estimating the distribution over model parameters and can also be used to infer the structure of the true generating model. Priors over all the model parameter are taken from conjugate distribution for the sake of efficiency of the model. They have tested their model on data which is discussed in [9]. However, prior connectivity information was not incorporated in inferring the TFAs. Use of this information can help in reducing the search space and genome-wide applications become feasible [1].

A fully probabilistic framework extending the linear regression model of Liao [5] is given in [1]. SSM is used to model the concentration of proteins and regulatory strengths are given separate Gaussian priors . Due to this characteristic, this model can effectively be used to reconstruct genome-wide transcriptional regulation. There are such situations where gene expression data is not in time series so analysis of that experimental data do not involve HMM based approach. Estimation of the concentrations of transcription factor proteins using such data is necessary as experimental measurement of these quantities is a difficult task. Inference of transcription factor profiles under time course settings was done in [1] using SSM giving results which are largely confirmed in biological literature. Here, we employ the model proposed in [1] using gene expression data that do not have temporal dependency to estimate the concentrations of transcription factor proteins.

## 2 Methods

We first briefly review the model proposed in [1] which uses the time course data. Later, we will use this model for inference in time-independent gene expression data.

### 2.1 Model for Time-series Gene Expression Data

Log gene expression data from a time-series microarray experiment is collected in a matrix form $\mathbf{Y} \in \Re^{N \times T}$, where $N$ is number of genes and $T$ is the total

number of time points in the data. We assume gene expression to be driven by $M$ transcription factors. The model we use is a log-linear approximation to the non-linear relationship between changes in transcription factor activity and gene expression. In practice, we use a discrete-time model where gene expression for gene $n$ is modelled as a linear combination of the the activity of its regulators

$$y_n(t) = \sum_{m=1}^{q} X_{nm} b_{nm} c_m(t) + \mu_n + \epsilon_{nt} \tag{1}$$

$X$ here is a binary matrix whose $nm$ entry is one if and only if gene $n$ is regulated by transcription factor $m$. This matrix is assumed to be known from literature or other experimental techniques such as ChIP-on chip. The activity matrix $B$ encodes the regulatory strength with which transcription factor $m$ effects the gene $n$. To incorporate the baseline expression for each gene, a constant vector $\mu = [\mu_n]$ is used. To model the dynamics of the transcription factor concentrations, 1st order Markov chain are used as shown in equation 2. This matrix $C$ represents the relative (log)-concentration of the transcription factor $m$ at specific time instant $t$

$$c_m(t) = \gamma_m c_m(t-1) + \eta_{mt} \tag{2}$$

It is important to state that the only observed variables are the gene expression levels $y_n(t)$; the transcription factor activity profiles $c_m(t)$, the regulatory strengths $b_{nm}$ and the baseline expression levels $\mu_n$ are all treated as latent random vectors. The prior distribution for $c_m(t)$ is given by the state-space model 2, while $b_{nm}$ and $\mu_n$ are given zero mean spherical Gaussian priors.

The joint distribution for the observed and latent variables is

$$p(\mathbf{Y}, \mathbf{B}, \mathbf{C}, \mu) = \left[ \prod_{n=1}^{N} \prod_{t=1}^{T} \mathcal{N}\left( y_n(t) | \mu_n + \sum_{m=1}^{q} \mathbf{X}_{nm} b_{nm} c_m(t), \sigma^2 \right) \right] \times \left[ \prod_{n=1}^{N} \prod_{m=1}^{q} \mathcal{N}\left( b_{nm} | 0, \alpha^2 \right) \right] \mathcal{N}\left( \kappa | \mathbf{0}, \mathbf{K} \right) \mathcal{N}\left( \mu | \mathbf{0}, \mathbf{I} \right) \tag{3}$$

where $\kappa$ is a vector obtained by concatenating the transcription factor concentrations at various time points. Notice that the state-space model prior implies that the prior covariance matrix $K$ is banded.

Exact marginalisation in this model is impossible due to the multiplicative nature of the model. Variational approximation is employed to obtain posterior distribution over the model parameters. Approximating posterior distribution factorizes over the hidden variables as

$$q(\mathbf{B}, \mathbf{C}, \mu) = q_1(\mathbf{B}) q_2(\mathbf{C}) q_3(\mu) \tag{4}$$

Using variational EM algorithm, approximating distribution is constructed iteratively by starting with any distribution $q_2(\mathbf{C})$ and $q_3(\mu)$ and average the

joint likelihood under them. Then $q_1(\mathbf{B})$ that maximizes the variational lower bound can be computed. Using the updated distribution $q_1(\mathbf{B})$, we can compute the updates for updated distributions $q_2$ and $q_3$ and repeat this until the convergence is achieved.

### 2.2 Model for Time-Independent Gene Expression Data

To incorporate the time-independent nature of the gene expression data, temporal dependency in the form of Markov process shown in equation 2 is no longer required. Experimental data containing independent point or different experimental conditions can be used to estimate the concentration of transcription factor proteins as shown above for time-series data. Using $\gamma = 0$ also give the required solution but computation expensive as the size of the matrix $K'$ becomes very large. Revisiting the equation 1, now concentration matrix $C$ represent the concentration of transcription factor proteins at different experimental conditions. This also simplify the process of estimating the matrix $K'$ for posterior distribution of transcription factor profiles as shown below. In this case, the row vector of concentrations is formalized as

$$\mathbf{c}(1) \dots \mathbf{c}(T) \sim \mathcal{N}(0, \mathbf{K}) \tag{5}$$

The covariance matrix for posterior distribution of transcription factor profiles $K$ has simpler form now. For time-series data case, $K$ was a banded matrix of size $Tq \times Tq$ [1]. For genome-wide applications, size of this matrix becomes very large while increasing the time and space complexity for inversion so an optimized inversion algorithm for banded matrix was used for the sake of efficiency. But in case of time-independent gene expression data, matrix $K$ is an identity matrix as there is no need to incorporate the Markov process of equation 2 hence simplifying the calculation for posterior estimation of $C$ and increasing the performance. Using the distribution given in equation 5 in the joint likelihood and estimating the posterior for $C$, one obtains that

$$q_2(\mathbf{C}) = \mathcal{N}(\mathbf{c}(1) \dots \mathbf{c}(T) | \nu, \mathbf{K}') \tag{6}$$

with

$$\mathbf{K}' = \left( \mathbf{K^{-1}} + \mathbf{I_T} \otimes \frac{1}{\sigma^2} \sum_{n=1}^{N} \chi_n \langle \mathbf{b_n} \mathbf{b_n^T} \rangle_{q_1} \chi_n \right)^{-1}$$

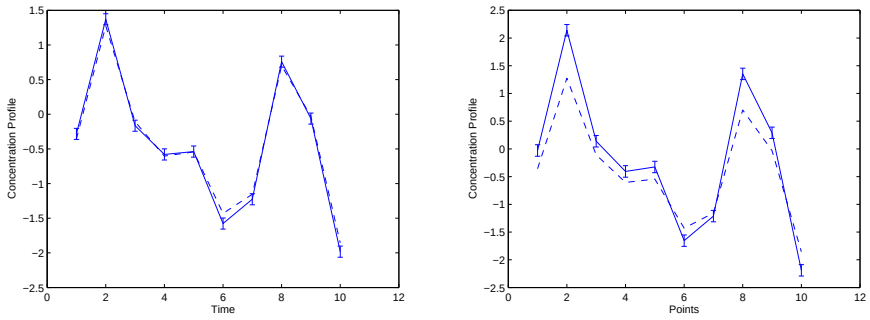
$$\nu = \mathbf{K}' \left( \frac{y_n - \langle \mu_\mathbf{n} \rangle_{q_3}}{\sigma^2} \chi_n \langle \mathbf{b}_n \rangle_{q_1} \right)$$

Calculating $K$ is much more efficient now that can further be improved if the posterior estimation is done in the following way.

$$\langle \mathbf{c}(t) \rangle = \left( \mathbf{I_q} + \frac{1}{\sigma^2} \sum_{n=1}^{N} \chi_n \langle \mathbf{b_n} \mathbf{b_n^T} \rangle_{q_1} \chi_n \right)^{-1} \left( \frac{y_n - \langle \mu_\mathbf{n} \rangle_{q_3}}{\sigma^2} \chi_n \langle \mathbf{b}_n \rangle_{q_1} \right)$$

## 3 Results

Here, we present some preliminary results comparing the time-dependent model with the time-independent model. We test on a very simple synthetic data set generated using the time-dependent model. We used the time-independent model to the simulated gene expression data to infer the transcription factor protein concentration and gene-specific regulatory activities from microarray data. Figure 1 shows the comparison of the results for both time-series and time-independent cases using artificial data. From the results, it can be seen that in both cases results are similar with slight differences in confidence intervals associated with the estimated concentration profiles of transcription factor proteins. Another measure would be to compare the ratios of variance of the expected values of a particular transcription factor protein concentration and the associated average error for both times-series and time-independent data. This come out to be 11.9185 for the time series case and 17.8964 for time-independent data. Here, figure 1a shows better result as the data used here is taken from a time-series experiment.




(a) (b)

**Fig. 1.** (a) Estimated concentration profile using time-series data (b) Estimated concentration profile using time-independent data. Dashed line shows the original concentration profile while solid line is the estimated concentration profile.

## 4 Conclusion

In this paper, inference of transcription regulation for gene-specific activities is modeled for gene expression data containing different experimental conditions. State space model in variational framework is used to provide the basis for inference in transcription networks. Computational complexity is a prominent feature of this model which is better in case of time-independent data. Also,

using specific structure of the regulatory network, genome-wide application are possible using time-series and time-independent gene expression data.

## References


1. Sanguinetti, G., Lawrence, N., Rattray, M.: Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. Bioinformatics **22**(22) (2006) 2775
2. Alon, U.: An introduction to systems biology: design principles of biological circuits. Chapman & Hall/CRC (2007)
3. Boyer, L., Lee, T., Cole, M., Johnstone, S., Levine, S., Zucker, J., Guenther, M., Kumar, R., Murray, H., Jenner, R., et al.: Core transcriptional regulatory circuitry in human embryonic stem cells. Cell **122**(6) (2005) 947–956
4. Harbison, C., Gordon, D., Lee, T., Rinaldi, N., Macisaac, K., Danford, T., Hannett, N., Tagne, J., Reynolds, D., Yoo, J., et al.: Transcriptional regulatory code of a eukaryotic genome. Nature **431** (2004) 99–104
5. Liao, J., Boscolo, R., Yang, Y., Tran, L., Sabatti, C., Roychowdhury, V.: Network component analysis: reconstruction of regulatory signals in biological systems. Proceedings of the National Academy of Sciences **100**(26) (2003) 15522–15527
6. Alter, O., Golub, G.: Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. Proceedings of the National Academy of Sciences **101**(47) (2004) 16577–16582
7. Nachman, I., Regev, A., Friedman, N.: Inferring quantitative models of regulatory networks from expression data. Bioinformatics **20 Suppl 1** (August 2004)
8. Beal, M., Falciani, F., Ghahramani, Z., Rangel, C., Wild, D.: A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. Bioinformatics **21**(3) (2005) 349–356
9. Rangel, C., Angus, J., Ghahramani, Z., Lioumi, M., Sotheran, E., Gaiba, A., Wild, D.L., Falciani, F.: Modeling T-cell activation using gene expression profiling and state-space models. Bioinformatics **20**(9) (2004) 1361–1372