Bayesian inference based modelling for gene transcriptional dynamics by integrating multiple source of knowledge

Wang, Shu-Qiang; Li, Han-Xiong

Published in:
BMC Systems Biology

Published: 16/07/2012

Published version (DOI):
10.1186/1752-0509-6-S1-S3

Publication details:
Wang, S.-Q., & Li, H.-X. (2012). Bayesian inference based modelling for gene transcriptional dynamics by integrating multiple source of knowledge. *BMC Systems Biology*, *6*(SUPPL.1), Article S3. https://doi.org/10.1186/1752-0509-6-S1-S3

Wang and Li *BMC Systems Biology* 2012, **6**(Suppl 1):S3
http://www.biomedcentral.com/1752-0509/6/S1/S3

BMC Systems Biology

RESEARCH Open Access

Bayesian inference based modelling for gene transcriptional dynamics by integrating multiple source of knowledge

Shu-Qiang Wang[1,2], Han-Xiong Li[1,2*]

From The 5th IEEE International Conference on Computational Systems Biology (ISB 2011)
Zhuhai, China. 02-04 September 2011

Abstract

Background: A key challenge in the post genome era is to identify genome-wide transcriptional regulatory networks, which specify the interactions between transcription factors and their target genes. Numerous methods have been developed for reconstructing gene regulatory networks from expression data. However, most of them are based on coarse grained qualitative models, and cannot provide a quantitative view of regulatory systems.

Results: A binding affinity based regulatory model is proposed to quantify the transcriptional regulatory network. Multiple quantities, including binding affinity and the activity level of transcription factor (TF) are incorporated into a general learning model. The sequence features of the promoter and the possible occupancy of nucleosomes are exploited to estimate the binding probability of regulators. Comparing with the previous models that only employ microarray data, the proposed model can bridge the gap between the relative background frequency of the observed nucleotide and the gene's transcription rate.

Conclusions: We testify the proposed approach on two real-world microarray datasets. Experimental results show that the proposed model can effectively identify the parameters and the activity level of TF. Moreover, the kinetic parameters introduced in the proposed model can reveal more biological sense than previous models can do.

Background

A challenge facing molecular biology is to develop quantitative, predictive models of gene regulation. The advance of high-throughput microarray technique makes it possible to measure the expression profiles of thousands of genes, and genome-wide microarray datasets are collected, providing a way to reveal the complex regulatory mechanism among cells. There are two broad classes of gene regulatory interactions: one based on the 'physical interaction' that aim at identifying relationships among transcription factors and their target genes (gene-to-sequence interaction) and another based on the 'influence interaction' that try to relate the expression of a gene to the expression of the other genes in the cell (gene-to-gene interaction).

In recent years, researchers have proposed many different computational approaches to reconstruct gene regulatory networks from high-throughput data, e.g. see reviews by Bansal et al. and Markowetz and Spang [1,2]. These approaches fall roughly into two categories: qualitative and quantitative aspects. Inferring qualitative regulatory networks from microarray data has been well studied, and a number of effective approaches have been developed [3-10]. However, these methods are based on coarse grained qualitative models [11,12], and cannot provide a realistic and quantitative view of regulatory systems. On the other hand, quantitative modelling for gene regulatory network is in its infancy. Research on quantitative models for genetic regulation has arisen only in recent years, and most of them are based on classical statistical techniques. Liebermeister et al. [13] proposed a linear model for cell cycle-related gene expression in yeast based on independent component

[1]Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong
Full list of author information is available at the end of the article

analysis. Holter et al. [14] employ singular value decomposition to uncover the fundamental patterns underlying gene expression profiles. Pournara et al. [15] and Yu et al. [16] proposed the Factor analysis model to describe a larger number of observed variables. However, these approaches are based on linear regression, and are not always being consistent with the observations in biochemical experiments which are nonlinear. Imoto et al. [17] proposed a nonlinear model with heterogeneous error variances. This model matches the microarray data well but it is not satisfying enough in revealing more biological sense. Segal et al. [18] proposed a transcription control network based model and apply their model to the segmentation gene network of Drosophila melanogaster. They reveal that positional information is encoded in the regulatory sequence and input factor distribution. However, there is still a little bit of dilemma in the model: the activity level of transcription factors is difficult to be measured or to be identified. Actually, assaying the transcription factors' activity state in a dynamic fashion is a major obstacle to the wider application of the kinetic modeling. TFs' activity levels are difficult to measure mainly due to two technical limitations: TFs are often present at low intercellular concentrations and the changes in their activity state can occur rapidly due to post-translational modifications.

Based on the above description, this paper aims to describe the transcriptional regulatory network quantitatively. In this work, a Bayesian inference based regulatory model is proposed to quantify the transcriptional dynamics. Multiple quantities, including binding energy, binding affinity and the activity level of transcription factor are incorporated into a general learning model. The sequence features of the promoter and the occupancy of nucleosomes are exploited to derive the binding energy. Compared with the previous models, the proposed model can reveal more biological sense.

Results

Case I: Circadian patterns in rat liver

Circadian rhythm is a daily time-keeping mechanism fundamental to a wide range of species. The basic molecular mechanism of circadian rhythm has been studied extensively. As a real example to test our approach, we considered the dynamics of the circadian patterns in rat liver. We employ the datasets from Almon et al [19]. This experiment was designed to examine fluctuations in gene expression in liver within the 24 hour circadian cycle in normal animals. Fifty-four male normal Wistar rats were housed in a strictly controlled stress free environment with light: dark cycles of 12 hr: 12 hr. Three animals were sacrificed at each of 18 selected time points within the 24 hour cycle. RNA was prepared from livers for gene arrays. Time point designations reflect time after lights on in hours. For details, please refer to Table S1 in additional file 1.

Analysis of the predicted activity levels of transcription factors

To test the proposed model on the above dataset, we employ two important transcriptional regulators of which activity levels indicate the variation of heat signals in a subset of gene circadian network, hsf1 and ppara. In total, we selected 7 genes to perform posterior inference of TF activities. To ensure identifiability, we included three genes that are regulated solely by hsf1 (HSP110, HSPA8 and COL4A1), and two genes that are regulated solely by ppara (ACSL1 and HMGCS1). The remaining two genes are jointly regulated by hsf1 and ppara. These genes were chosen since they exhibit the largest variance in the microarray time course, and hence are likely to provide a cleaner representation of the output of the system.

The inferred TFs' activity levels are shown in Figure 1 (a) and 1(b). Both inferred TF profiles show a noisy periodic behaviour [20]. Figure 1(c) gives the values of the parameters k_i for the four selected circadian genes (HSPA8, ACSL1, HSP90AA1 and HSPA1B). The green column represents the response k_1 to hsf1 alone, the red column is the response k_2 to ppara alone and the black column represents the joint response k_{12}. It can be seen that, for gene, HSPA8, the model predicts a clear activation by hsfl alone, which is consistent with the experimental conclusion from Yan et al [20]. The black columns of HSP90AA1 and HSPA1B demonstrate that the model predicts a significant combinatorial activation which can be verified by mutagenetic techniques, i.e. by knocking out one of the TFs.

The biological sense of kinetic parameters

Table 1 shows the relationship between scaling parameter k and the corresponding binding affinity ϕ. In table 1, 'H' indicates 'high' and 'L' indicates 'low'. We define the scaling parameter ki as 'High' if it is bigger than the mean value, as 'low', otherwise, and the same to binding affinity ϕ. From Table 1, we can find that, for most case, the scaling parameter is in accordance with the binding affinity: High scaling parameter coupling with high binding affinity, vice versa. However, gene COL4A1 and HSP110 are 2 exceptions: they have high scaling parameter but low binding affinity. Our view is that low binding affinity but high value for k_i might represent a TF which rarely binds to promoter but can strongly regulate gene expression when bound.

Figure 2 shows the results of inference on the values of the parameters c_j and ω_j. The columns on the left, shaded red, show results from our model and the white columns are the estimates obtained by the method of

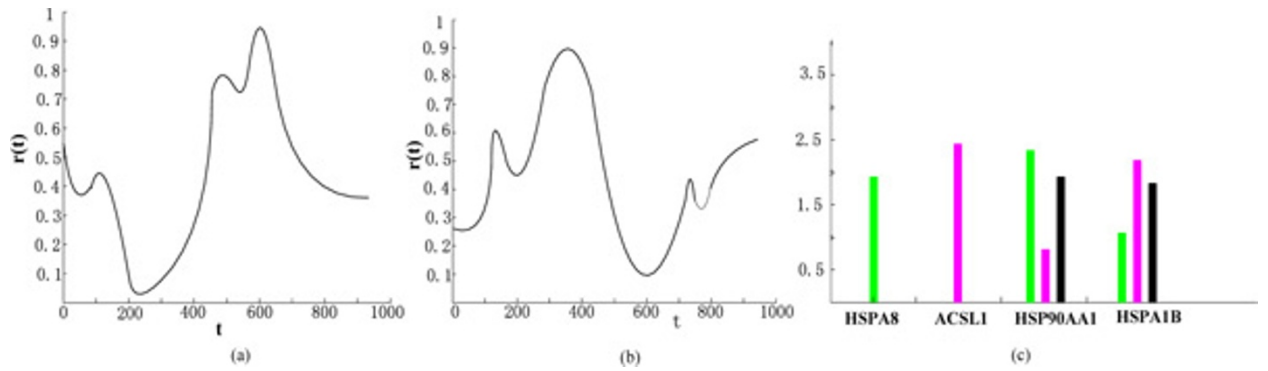

Figure 1 Results on circadian patterns data. (a) mean activity profile for hsf1, (b) mean activity profile for ppara, (c) bar-chart representation of the parameters ki, giving the logical structure of the interaction between two TFs.

Barenco et al. [21]. The parameters were assigned a vague gamma prior distribution (a = b = 0.2, corresponding to a mean of 1 and a variance of 5). The results are in good accordance with the results obtained by Barenco et. a l [21]. We can find that the parameters c_j and ω_j obtained by our method have smaller variance than that of Barenco et al. Figure 3 shows the fit of the model to the observed data at each time-point.

Case II: A yeast synthetic network for in vivo assessment

Validation of gene regulation network (GRN) inference methods has traditionally been done using in silico networks. However, depending on how realistic the choice of an in silico model is, this kind of validation approach has obvious limitations. To our knowledge, rarely the underlying model from which artificial/simulated data is generated is realistic enough. Real biological networks are fairly complex chemical reaction network models. In simulation setting one typically adds noise on top of a hypothetical simulation model, but the noise characteristics may not be realistic enough. Also, simulation models tend to be overly simplistic when compared to e.g. real gene regulatory networks. Data measured from a real biological system is, real. To overcome these problems, we use the IRMA network to evaluate out model. The IRMA network is a synthetically constructed GRN in the Saccharomyces cerevisiae genome [22], which is designed to be maximally independent in such a way that genes in the network are not regulated by genes outside of the network (i.e. by other yeast genes). However, genes in the IRMA network may regulate other genes in the genome. The network consists of five genes and there are positive and negative feedback loops and one protein to protein interaction, shown as Figure 4. Although the IRMA network contains only five genes, there are studies suggesting that the performance on smaller networks typically generalize to larger networks [1,23]. The data samples were collected every 20 min up to 5 hr in five independent experiments for the switch-on state, and every 10 min up to 3 hr in four independent experiments for the switch-off state. For details on the construction of the network and experimental procedures, we refer to [22]. One of the purposes of the IRMA network is to provide a realistic benchmark set for computational studies by providing mRNA-level measurements from a known GRN. To our knowledge, the IRMA network and dataset are the first of a kind that are meant for validation purposes. Besides, we assume that mRNA decay rates may be gene-specific, but remain constant in time [36]. The sequences of all promoters are retrieved from SCPD and SGD database. The scanning region ranges from -800 to +50 bp of each target gene.

Analysis of the predicted activity levels of transcription factors

To evaluate whether the proposed model can effectively learn the TFs' activity level and the regulation type, we first evaluate the model using the switch-on time series data. The inferred TFs' activity levels are shown in Figure 5(a) and 5(b). Both inferred TF profiles show a noisy switch-on behavior. Figure 5(c) gives the values of the parameters ki for the five target genes. The green column represents the response to

Table 1 Relationship between scaling parameter k and the corresponding binding affinity ϕ.

Gene	HSP110	HSPA8	COL4A1	ACSL1	HMGCS1	HSP90AA1-hsf1	HSPA1B- hsf1	HSPA1B- ppara
k	H	H	H	H	L	H	L	H
ϕ	L	H	L	H	L	H	L	H

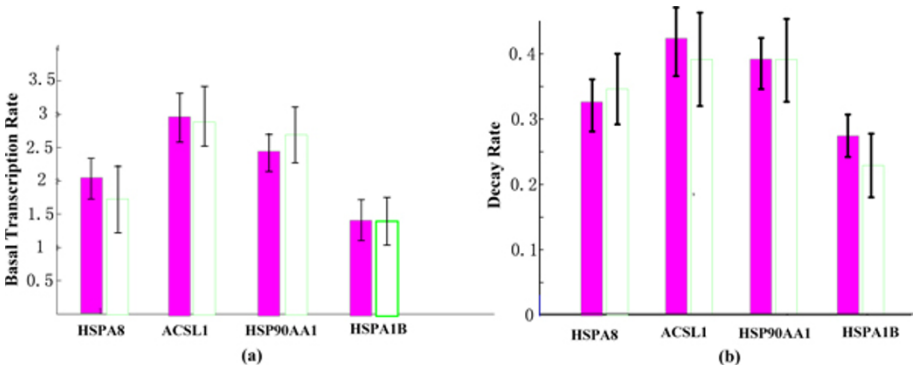

Figure 2 The bar charts for basal transcription rates and decay rates. (a) Basal transcription rates from our model and that of Barenco et al. Red are estimates obtained with our model, white are the estimates obtained by Barenco et al [21]. (b) Similar for decay rates.

the first regulator alone, the red column is the response to the second regulator alone and the black column represents the joint response, k_{12}. It can be seen that, for gene, GAL80, the model predicts a clear activation by swi5 which is consistent with the experimental conclusion [22]. For gene CBF1, the red downward column indicates that ash1 behaves as a repressor, which is verified in [22]. The black column of CBF1 demonstrates that the model predicts a significant combinatorial regulation [22].

Analysis of the kinetic parameters

Table 2 shows the relationship between scaling parameter k and the corresponding binding affinity ϕ. In table 2, the definition of 'High' and 'Low' are same as in Table 1, and the same abbreviations are employed. It can be found that gene GAL80 has the TF that rarely binds to promoter but can strongly up-regulate its expression when bound.

Figure 6 shows the results of inference on the values of the parameters c_j and ω_j. The columns on the left, shaded

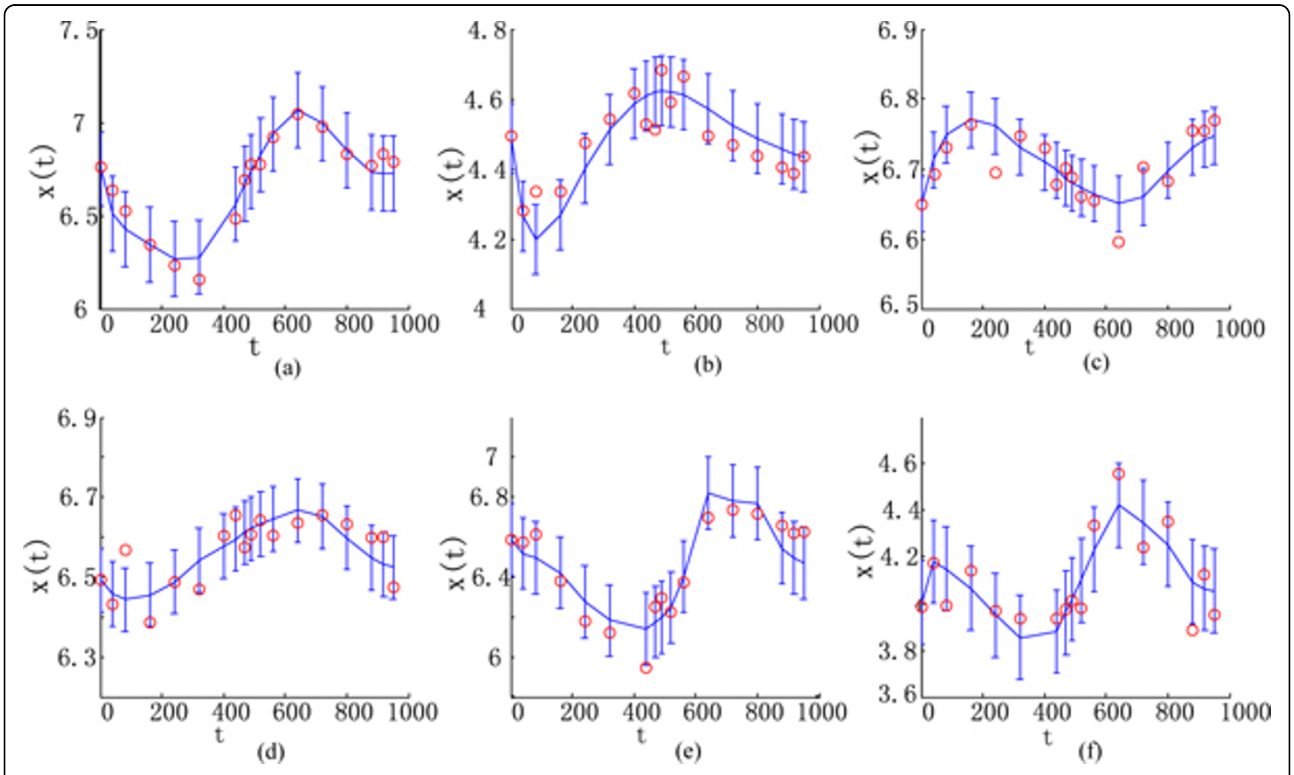

Figure 3 The predicted mean expression profiles. (a) HSPA8, (b) COL4A1, (c) ACSL1, (d) HMGCS1, (e) HSP90AA1 and (f) HSPA1B. The red circle indicates the observed value at each time-points.

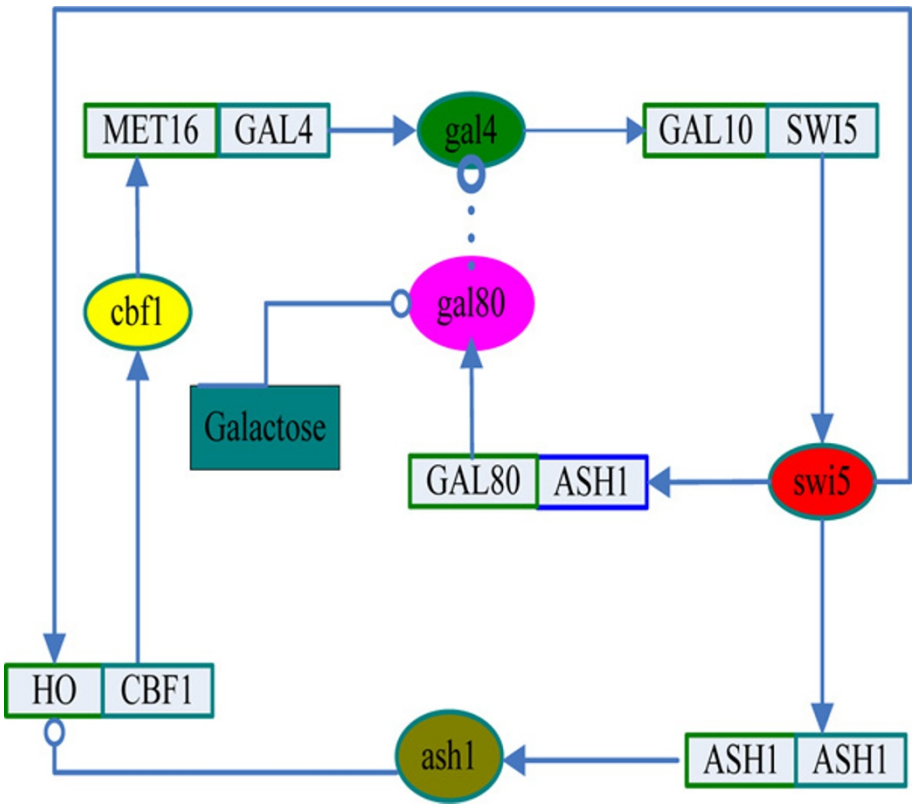

Figure 4 IRMA network. The rectangle indicates the gene while the oval represents the protein.

red, show results from our model and the white columns are the estimates obtained by Opper et al. [24]. It can be found that the results are in good accordance with the results obtained by the method of Opper et Al. [24]. It can be found that the parameters c_j and ω_j obtained by our method have smaller variance than that of Opper et al. [24].

For comparison, we also evaluate the model using the switch-off data. Figure 7 shows the fit of the model to the observed data at each time-point for both the switch-on case and switch-off case.

Discussion

In this study, two real-world microarray datasets were exploited two evaluate the efficiency of the proposed model. Comparison shows that the kinetic parameters obtained by our method have smaller variance than that

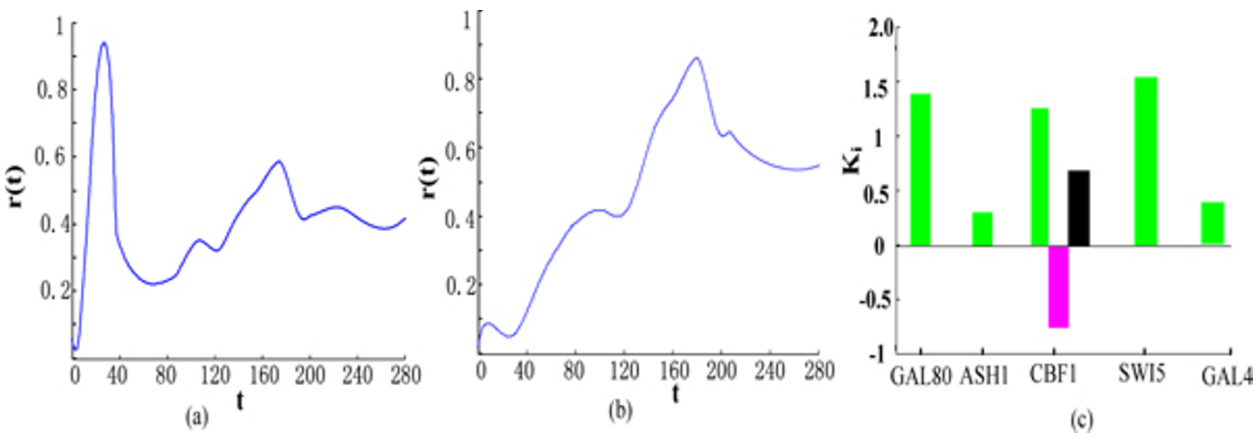

Figure 5 Results on IRMA network data. (a) mean activity profile for regulator swi5, (b) mean activity profile for regulator ash1, (c) bar-chart representation of the parameters ki.

Table 2 Relationship between k and ϕ for IRMA network data.

Gene	GAL80	GAL4	SWI5	ASH1	CBF1-swi5	CBF1-ash1
k	H	L	H	L	H	H
ϕ	L	L	H	L	H	H

of Barenco et al. [21]. One reason is that the proposed model provides a principled method for the incorporation of prior biological knowledge. This may be in the form of suitable ranges for kinetic parameters, known kinetic parameter values and suitable distributions on measurement noise. Besides, it is possible for the proposed model to circumvent the need for expensive sampling-based inference and a TFA profile can be inferred over all time, rather than just at the discrete time-points at which expression was measured.

The Bayesian inference based model of transcription rates and regulator activity levels allows us to handle these biologically relevant quantities despite the indirect measurement of the former and the lack of measurements of the latter. It also allows us to handle the inherently noisy measurement in a principled way. However, the proposed model still abstracts away some of the explicit processes that generate the actual observed expression data. A more explicit modelling of these will provide a more principled treatment of different sources of noise in the data. Furthermore, this model does not handle directly the upstream events that affect regulator activity. In fact, the current model is an open loop system, such that the regulation of regulator activity is itself viewed as exogenous to the system. By developing a richer modeling language we may capture more complex reaction models, model the upstream regulation of activity levels, and identify systems that involve feedback mechanisms and signalling networks.

Post-Transcriptional Modification Model (PTMM) have been previously used to model TF activities [25]; 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.

Conclusions

In this work, we have proposed a computational model to reverse engineer simultaneously both the activity of TFs and the logical structure of promoters by integrating multiple sources of knowledge, including time-series gene expression data, TFs' binding information and sequence features of promoters. The approach relies on a detailed model of transcription, which is an approximation to the Michaelis-Menten model from classical enzyme kinetics, and therefore should be able to capture accurately the effects that changes in TF activity have on gene expression dynamics. We testify the proposed approach on two real-world microarray datasets. Experimental results show that the proposed model can effectively identify the parameters and the activity level of TF. Moreover, the kinetic parameters introduced in the proposed model can reveal more biological sense than previous models can do.

Methods

Problem statement

A microarray experiment only measures the "observed" quantities, as shown in Figure 8, whereas the other quantities are not observed ("hidden"). The dashed oval encloses the closest quantities on the path between the TF and the target gene. Consider a transcriptional network of n genes that are regulated by m regulators, as well as a time-dependent external signal. Given the structure G and a set X of transcription rates of these genes in T time points, is it possible to reconstruct the

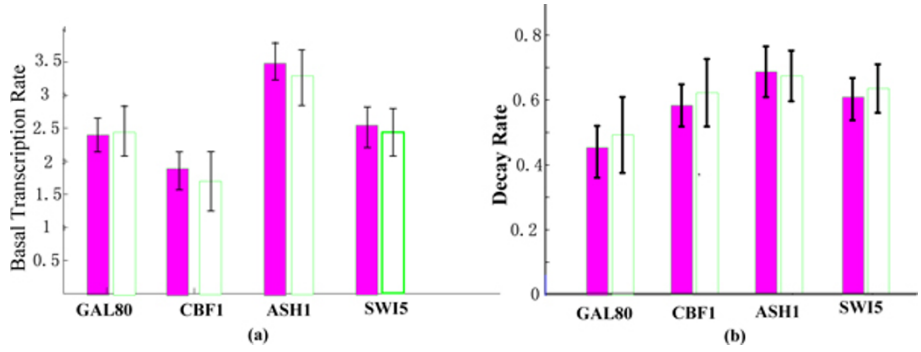

Figure 6 The bar charts for basal transcription rates and decay rates. (a) Basal transcription rates from our model and that of Opper et al. [24]. Red are estimates obtained with our model, white are the estimates obtained by Opper et al. (b) Similar for decay rates.

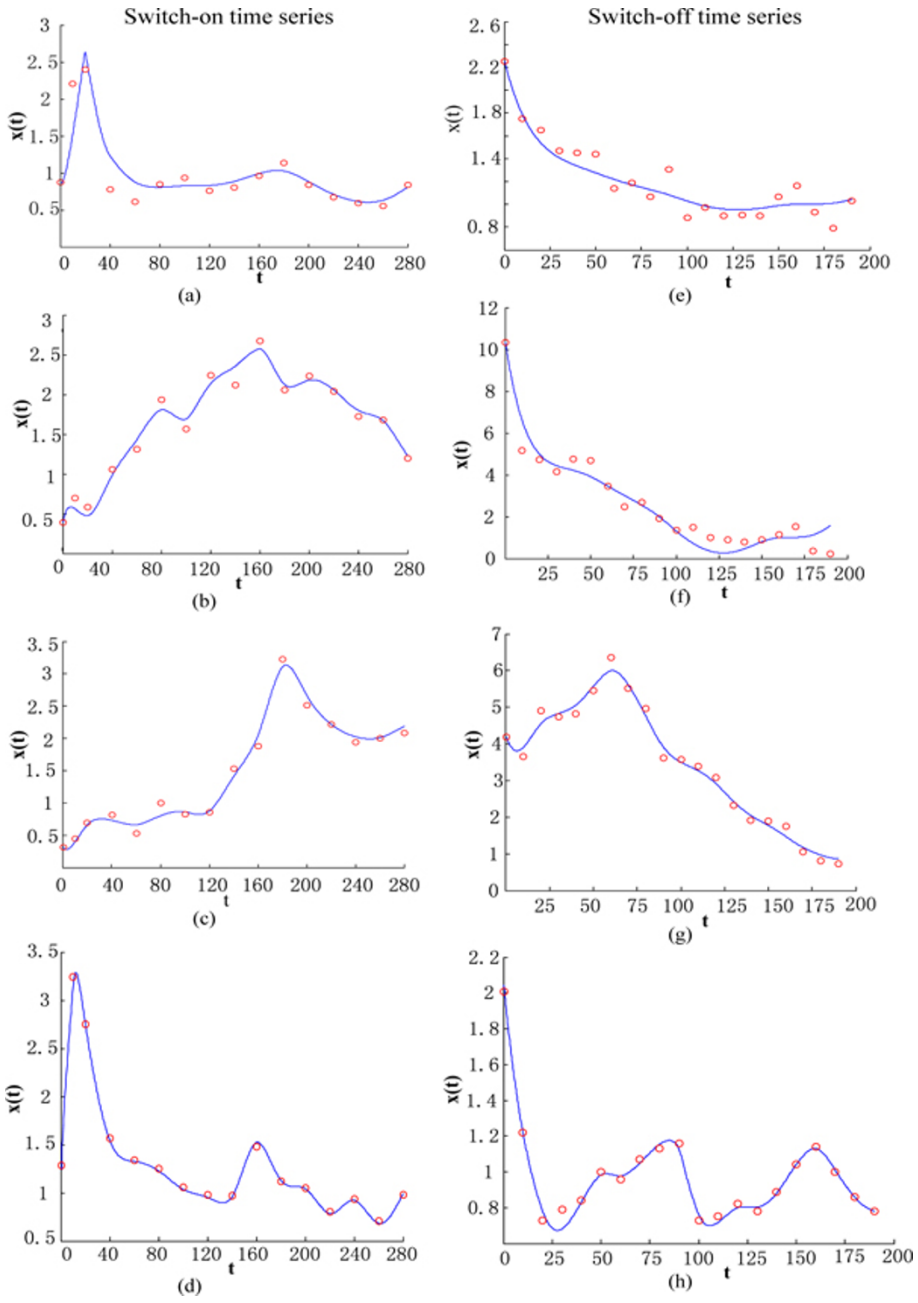

Figure 7 The predicated mean expression profiles. Expression profile and mean reconstruction of target genes. Switch-on time series: (a) GAL80, (b) ASH1, (c) CBF1, (d) GAL4, (e)-(h) The same genes in switch-off time series. The red circle indicates the observed value at each time-point.

time-varying activity level of m regulators, R, at all time points and the corresponding model parameters? This is an infinite dimensional problem that we tackle by placing a stochastic process prior over the activities of regulators.

Our approach relies on a continuous time, differential equation description of transcriptional dynamics where TFs are treated as latent on/off variables and are modelled using a switching stochastic process. The framework of the proposed method is shown in the Figure 9.

Kinematic model of regulation

Compared with the gene expression level, the gene transcription rate can capture more dynamic characteristics of transcription regulation. We here employ the

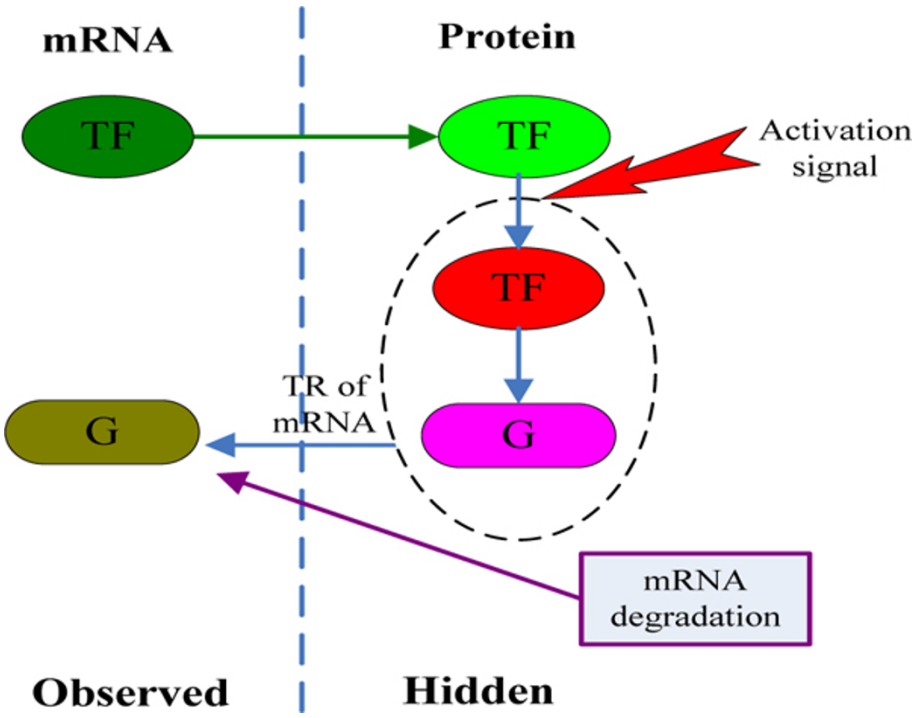

Figure 8 A qualitative molecular model of transcriptional regulation. mRNA encoding a transcription factor (TF) is translated to protein. The protein is activated and induces the transcription of a target gene at a certain rate (G). The final accumulation of G mRNA levels is determined by this transcription rate and by the rate of G's mRNA degradation.

transcription rate to model the regulation function. We first assume:

- The derived transcription rates are average rates over a cell population.
- The speed of a TF's binding to or dissociation from its target sites is assumed to be much more rapid than the transcription process, thus rapid-equilibrium approximation can be used.

Based on the above assumptions, the transcription rate of a gene is proportional to the amount of the gene bound by its regulators in all genes of the measured cell population. We first consider the case that a gene is regulated by a single activator. The corresponding regulation function can be properly described by Michaelis-Menten equation:

$$\frac{dx}{dt} = \beta \frac{r(t)}{d + r(t)} + c - \omega x, \tag{1}$$

here x represents the mRNA concentration for a particular gene, *r(t)* the concentration of active TF, β and d are kinetic constants, c a baseline expression rate and ω the mRNA decay rate.

To incorporate the sequence feature and the TF binding preference into the model, we set the binding affinity ϕ = k/d, and (1) can be reformulated as

$$\frac{dx}{dt} = \beta \frac{k\varphi r(t)}{1 + k\varphi r(t)} + c - \omega x, \tag{2}$$

here k is a scaling parameter.

We now take the regulation involving two regulators into account. Denote by $r_1(t)$ and $r_2(t)$ the concentration of two regulators, ϕ_1 and ϕ_2 the binding affinity of two regulators from their own target sites, the regulation function can be written as below:

$$\frac{dx}{dt} = \beta \frac{k_1\varphi_1 r(t)_1 + k_2\varphi_2 r_2(t) + k_3\varphi_1\varphi_2 r_1(t) r_2(t)}{(1 + \varphi_1 r_1(t))(1 + \varphi_2 r_2(t))} + c - \omega x, \tag{3}$$

Considering the general case, a gene is regulated by n regulators. There are 2^n different binding states in total. The n-dimension binary vector is employed to indicate a certain binding state, e.g., a 4-dimension vector (0 1 0 1) indicates that the second and the fourth regulators are bound to their own target sites while the first and the third are not bound. The regulation function can be written as:

$$\frac{dx_j}{dt} = \beta_j \frac{\sum_{s \in S_j} k_s \prod_{s_i=1; i=1,n} \varphi_{ij} r_i(t)}{\prod_{i=1}^{n} (1 + \varphi_{ij} r_i(t))} + c_j - \omega_j x_j \tag{4}$$

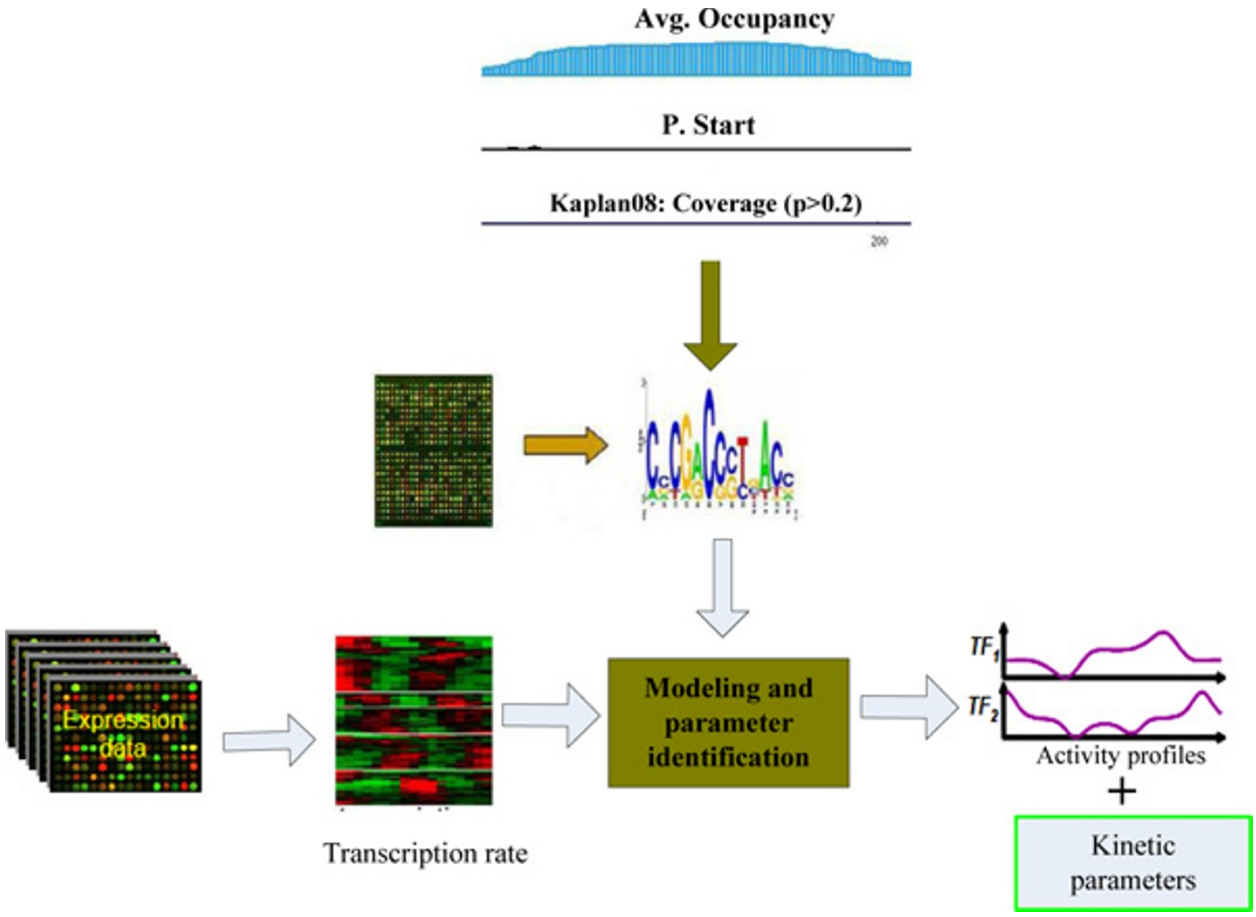

Figure 9 The framework of proposed method. The expression data and sequence features of promoters are incorporated into a general learning model. The outputs of the model are kinetic parameters and the activity levels of transcription parameters.

where S_j denotes the set of all 2^n possible state vectors, and s_i is the i_{th} element of the state vector *s*.

Modelling for binding affinity

Measuring affinities of molecular interactions in high-throughput format remains problematic, especially for transient and low-affinity interactions. We here try to describe the landscape of binding affinity in the perspective of binding energy between the various DNA-binding molecules and their target genes. Binding affinity landscapes describe how each molecule translates an input DNA sequence into a binding potential that is specific to that molecule. The presented framework models several important aspects of the binding process.

By allowing molecules to bind anywhere along the input sequence, the entire range of affinities is considered, thereby allowing contributions from both strong and weak binding sites [26,27].

- Conventional cooperative binding interactions can be explicitly modelled by assigning higher statistical weights to configurations in which two molecules are bound in close proximity.
- The cooperativity that arises between factors when both nucleosomes and transcription factors are integrated is captured automatically [28].

We first consider the simplest case that there is only one target site S_{ij} for TF *i* in the promoter of gene *j*:

$$TF_i + S_{ij} \underset{k_d}{\overset{k_b}{\longleftrightarrow}} [TF_i \cdot S_{ij}]$$

The site-specific binding affinity is given by

$$\varphi = C_i e^{-\frac{E_{ij}}{kT}} \tag{5}$$

where C_i is a constant, E_{ij} the binding free energy between TF_i and the promoter of gene *j*, *k* and *T* are the Boltzmann constant and temperature, respectively.

The above case can be expanded to the general case that binding may happen in anywhere along the input

sequence and the accessibility of target sites varies due to the occupancy of nucleosomes. The general binding affinity is modelled as

$$\varphi_{ij} = C_i \sum_{n=1}^{N} p_{ij}^{(n)} e^{\frac{-E_{ij}^{(n)}}{kT}} \quad (6)$$

where $p^{(n)}_{ij}$ is the probability of transcription factor i binding to the nth binding site in the promoter of gene j. Note that the derivation of $p^{(n)}_{ij}$ involves the information of the possible occupancy of nucleosomes. The nucleosomes positions can be predicted based on the nucleosome-DNA interaction model proposed by Kaplan et al [29]. Figure 10(b) shows the occupancy of nucleosomes for the genomic sequence shown in the Figure 10(a).

Since the positional weight matrices (PWM) are often used to represent the sequence motif [30,31], we estimate the binding energy in perspective of PWM. As the background information has been taken into the PWM, the binding free energy can be approximately calculated as below:

$$E_{ij}^{(q)} = K^{(q)} \sum_{l=1}^{L} \sum_{n=\{A,C,G,T\}} J^n{}_l \left(M_L{}^* - M_{nl}\right)$$

where $J^n{}_l = \begin{cases} 1 \text{ if } n = s(l) \\ 0 \text{ otherwise} \end{cases}$

here $K^{(q)}$ is the scaling factor, $M^*{}_L$ indicates the maximum background frequency in the motif, *s(l)* is the nucleotide in position *l*.

Regulatory network modelling using dynamic Bayesian inference

In many biological processes, the transcription factor transit from inactive to active state quickly as a consequence of fast post-translational modifications, (the time scale is micro second), so it is reasonable that we model the TF activity as a binary variable r(t) ∈ {0,1}[24,32].

For the regulation involving a single regulator, the TF activity can be seen as a two states Markov Jump Process. Based on Ref [24,33], the probability of the system being in a particular state at a given time is given by the following Master equation:

$$\frac{dp_1(t)}{dt} = n_+ p_0(t) - n_- p_1(t) \quad (7)$$

$$\frac{dp_0(t)}{dt} = n_+ p_1(t) - n_- p_0(t) \quad (8)$$

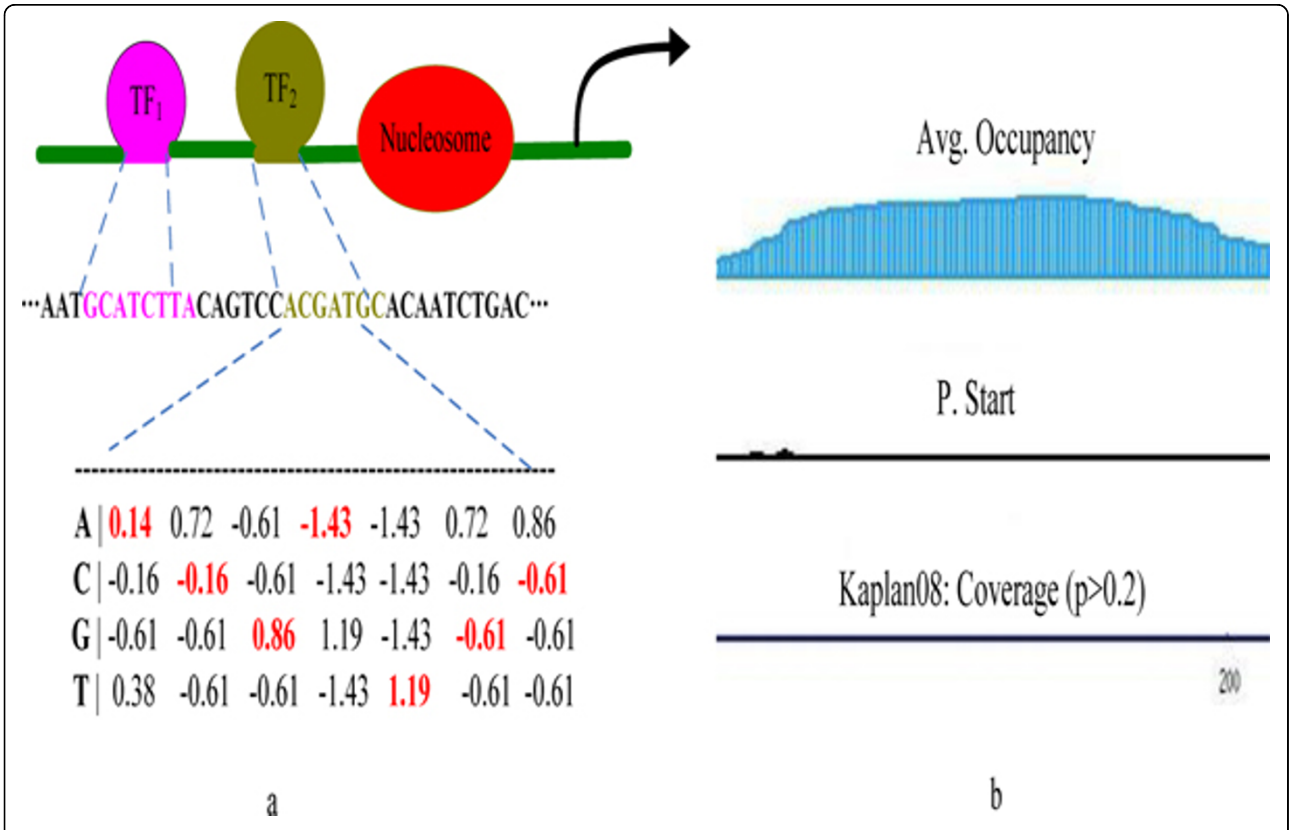

Figure 10 Employing sequence features and the occupancy of nucleosomes to estimate the binding affinity. The positional weight matrices are used to represent the sequence motif. The binding may occur anywhere along the input sequence, the entire range of affinities is considered.

here $p_1(t) = p(r(t) = 1)$, $p_0(t) = p(r(t) = 0)$ and $n_\pm$ indicates the transition rate.

The observed expression data is often assumed to be normally distributed [24]. We now define a noise model that relates the predicted mRNA concentration to the observed expression data, as shown in Figure 11.

Setting $y_j(t)$ as the observations of mRNA species j at time t, $x_j(t)$ the predicted expression and σ_j the variation, the noise model can be described as

$$y_j(t) \,|r(t) \sim N(x_j(t),\ \sigma^2{}_j)$$

Based on Refs [24,33], we define the TF's switching stochastic process as the prior distribution, then we combine the prior distribution and the noise model (likelihood) into Bayes' theorem to obtain the posterior over the process:

$$p(r \,|y,\ \Omega) = \frac{1}{S} p(y \,|r,\ \Omega) p(r)$$

where y denotes collectively the observations, Ω are the parameters involved in the regulation function and S a normalization constant.

Variational inference and model optimization

We will use a variational formulation of the inference problem [34]. Variational inference is a powerful inference method and it has been well applied for optimization by Opper and Sanguinetti [24,33]. Our model optimization is based on Ref [22]. Variational inference is used as an approximation technique: given an intractable probability distribution p, the variational approach finds an optimal approximation q within a certain family of distributions. This is usually done by minimizing the Kullback-Leibler (KL) divergence between the two distributions

$$KL[q \,\|p\,] = E_q[\log \frac{q}{p}] = \int \log \frac{q(r)}{p(r)} q(r) dr \tag{9}$$

By selecting a suitable family of approximating distributions, the inference problem is then turned into an optimization problem. It can be shown that the KL divergence is a convex functional of q and is equal to zero iff q = p [24,35]. In this case, we will choose the approximating process q to be again a Markov Jump Process, so that the required KL is given by

$$KL[q \| p_{\,post}] = KL[q \| p_{\,prior}] + \log S - E_q[\log p(y \,|r,\ \Omega)] \tag{10}$$

here S is a normalization constant, $E_q[\log\, p(y|r, \Omega)]$ the expectation of the likelihood of the observations under the approximating process.

According to Ref [24], minimization of the KL functional (11) can be represented as the saddle point problem

$$J = \max_{\tau} \min_{q} \{KL[q \,\|p_{prior}\,] + \sum_{j=1}^{n} [\tau_j(y_j - \bar{x}(t_j)) - \frac{\sigma^2}{2} \tau_j{}^2]\} \tag{11}$$

here τ is auxiliary variables. It can be found that this functional is concave in τ and convex in q. Hence we can exchange min and max. Performing the max first yields the result. This also shows that there is only a unique saddle point solution.

The optimization procedure is based on a forward-backward procedure, leading to ordinary differential equations which can iteratively be solved. Taking the regulation involving two regulators for example, the free

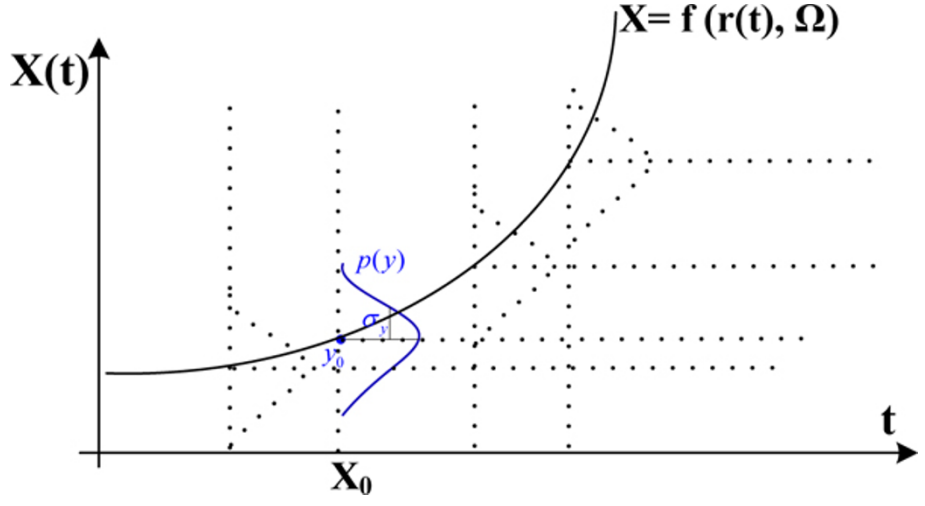

Figure 11 Normally distributed observational data. The solid line indicates the mean predicted expression while the dotted line represents the normally distributed observations.

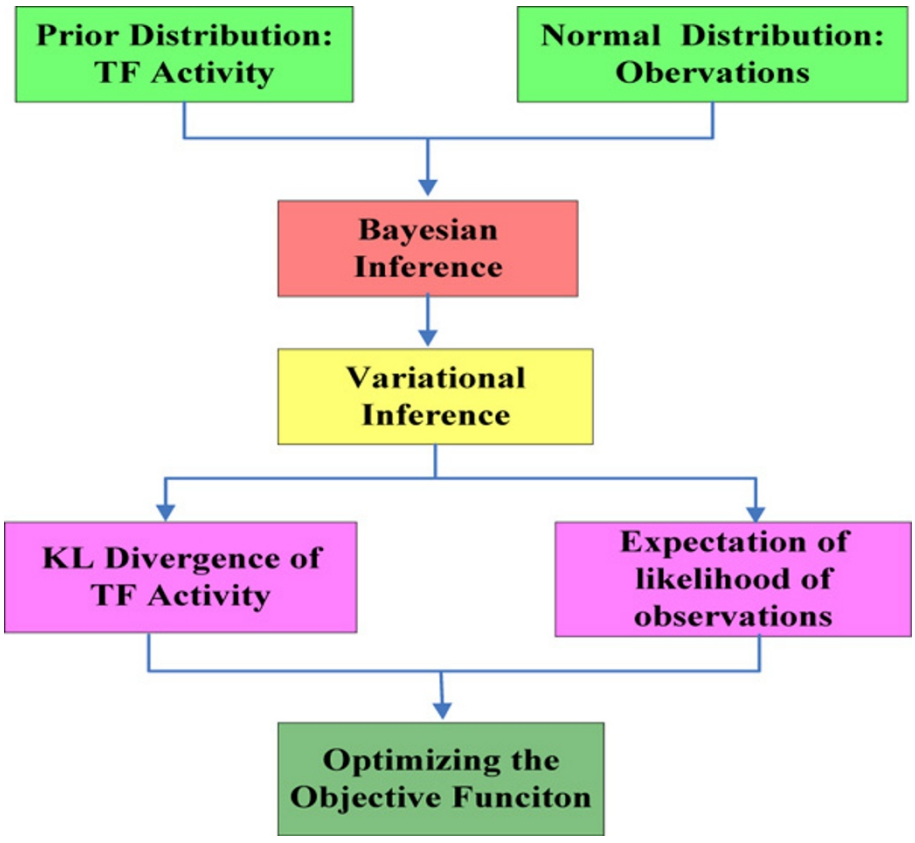

Figure 12 The framework of the inference.

energy is a functional of both the approximating processes q^1, q^2 and their transition rates n_1, n_2. However, these are not independent, but are related by the Master equation. To incorporate this constraint, we add Lagrange multipliers as

$$L(q^1,q^2,g_1,g_2) = J[q^1,q^2,g_1,g_2] + \int_0^T [\frac{dq_1{}^1(t)}{dt} + (n_{1-}+n_{1+})q_1{}^1(t) - n_{1+}]\lambda_1(t)dt + \int_0^T [\frac{dq_2{}^1(t)}{dt} + (n_{2-}+n_{2+})q_2{}^1(t) - n_{2+}]\lambda_2(t)dt \quad (12)$$

where g_1 and g_2 are the rates of jumps from the 0 to the 1 state for process q^1 and q^2, respectively.

The Lagrange multiplier functions obey the final condition $\lambda(T) = 0$. Estimation of the parameters A and b can be done directly by maximizing the approximate marginal likelihood $E_q[logp(y|r,\Omega)]$. The framework of the inference is shown in the Figure 12.

Additional material

Additional file 1: Table S1. The time series gene expression data for circadian patterns in rat liver.

Acknowledgements
This work was supported by a GRF project from Hong Kong SAR (CityU 117310) and the grant from NNSF China (51175519).

This article has been published as part of *BMC Systems Biology* Volume 6 Supplement 1, 2012: Selected articles from The 5th IEEE International Conference on Systems Biology (ISB 2011). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/6/S1.

Author details
[1]Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong. [2]School of Mechanical and Electrical Engineering, Central South University, Changsha 410083, China.

Authors' contributions
SQW proposed the method, performed the analysis; HXL supervised the work and revised the paper critically for important intellectual content.

Competing interests
The authors declare that they have no competing interests.

Published: 16 July 2012

References

1. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D: **How to infer gene networks from expression profiles.** *Mol Syst Biol* 2007, **3**:78.
2. Markowetz F, Spang R: **Inferring cellular networks - a review.** *BMC Bioinformatics* 2007, **8(Sul 6)**:S5.
3. Eisen MB, Spellman PT, Brown PO, Bostein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95**:14863-14868.
4. Davidich M, Bornholdt S: **Boolean network model predicts cell cycle sequence of fission yeast.** *PLoS One* 2008, **3**:e1672.

5. Lahdesmaki H, Shmulevich I, Yli-Harja O: **On learning gene regulatory networks under the Boolean network model.** *Mach Learn* 2004, **52**:147-167.
6. Imoto S, Goto T, Miyano S: **Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression.** *Pac Symp Biocomput* 2002, 175-186.
7. Kirimasthong K, Manorat A: **Inference of gene regulatory network by Bayesian network using Metropolis-Hastings Algorithm.** In *Proceedings of 3rd International Conference on Advanced Data Mining and Alications: 06-08 Aug, 2007; Harbin* Alhajj R, Gao H 2007, 276-286.
8. Segal E, Taskar B, Gasch A, Friedman N, Koller D: **Rich probabilistic models for gene expression.** *Bioinformatics* 2001, **17(Suppl 1)**:S243-S252.
9. Hu Z, Killion P, Iyer V: **Genetic reconstruction of a functional transcriptional regulatory network.** *Nat Genet* 2007, **39**:683-687.
10. Luscombe N, Babu M, Yu H: **Genomic analysis of regulatory network dynamics reveals large topological changes.** *Nature* 2004, **431**:308-312.
11. Werhli AV, Husmeier D: **Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge.** *Stat Appl Genet Mol Biol* 2007, **6**:Article15.
12. Segal E, Barash Y, Simon I, Friedman N: **From promoter sequence to expression: a probabilistic framework.** In *Proceedings of the Sixth Annual International Conference on Computational Biology: 18-21 April 2002; Washington* Myers G, Hannenhalli S, Istrail S 2002, 263-272.
13. Liebermeister W: **Linear modes of gene expression determined by independent component analysis.** *Bioinformatics* 2002, **18**:51-60.
14. Holter N, Mitra M, Maritan A: **Fundamental patterns underlying gene expression profiles: simplicity from complexity.** *Proc Natl Acad Sci USA* 2000, **97**:8409-8414.
15. Pournara I, Wernisch L: **Factor analysis for gene regulatory networks and transcription factor activity profiles.** *BMC Bioinformatics* 2007, **8**:61.
16. Yu T, Li K: **Inference of transcriptional regulatory network by two-stage constrained space factor analysis.** *Bioinformatics* 2005, **21**:4033-4038.
17. Imoto S, Kim S, Goto T: **Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network.** *J Bioinform Comput Biol* 2003, **1(2)**:231-252.
18. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U: **Predicting expression patterns from regulatory sequence in Drosophila segmentation.** *Nature* 2008, **451**:535-540.
19. Almon RR, Yang E, Lai W, Androulakis IP: **Circadian variations in rat liver gene expression: relationships to drug actions.** *J Pharmacol Exp Ther* 2008, **326**:700-716.
20. Yan J, Wang HF, Liu YT, Shao CX: **Analysis of Gene Regulatory Networks in the Mammalian Circadian Rhythm.** *Plos Comput Biol* 2008, **4(10)**: e1000193.
21. Barenco M, Tomescu D, Brewer D, Callard R, Stark J, Hubank M: **Ranked prediction of p53 targets using hidden variable dynamic modelling.** *Genome Biol* 2006, **7(3)**:R25.
22. Cantone I, Marucci L, Iorio F, *et al*: **A yeast synthetic network for in vivo assessment of reverse engineering and modelling aroaches.** *Cell* 2009, **137**:172-181.
23. Stolovitzky G, Monroe D, Califano A: **Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference.** *Ann N Y Acad Sci* 2007, **1115**:1-22.
24. Opper M, Sanguinetti G: **Learning combinatorial transcriptional dynamics from gene expression data.** *Bioinformatics* 2010, **26(13)**:1623-1629.
25. Shi Y, Klutstein M, Simon I, Mitchell T, Bar-Joseph Z: **A combined expression-interaction model for inferring the temporal activity of transcription factors.** *J Comput Biol* 2009, **16**:1035-1049.
26. Gertz J, Siggia E, Cohen B: **Analysis of combinatorial cis-regulation in synthetic and genomic promoters.** *Nature* 2002, **457**:215-218.
27. Tanay A: **Extensive low-affinity transcriptional interactions in the yeast genome.** *Genome Res* 2006, **16**:962-972.
28. Johnson SM, Tan FJ, McCullough HL, Riordan DP, Fire AZ: **Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin.** *Genome Res* 2006, **16**:1505-1516.
29. Kaplan N, Moore IK, Fondufe-Mittendorf Y, *et al*: **The DNA-Encoded Nucleosome Organization of a Eukaryotic Genome.** *Nature* 2008, **458**:362-366.
30. Bulyk ML: **Computational prediction of transcription-factor binding site locations.** *Genome Biol* 2003, **5(1)**:201.
31. Naum I, Gary D, Ilya P: **Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites.** *Nucleic Acids Res* 2005, **33**:2290-2301.
32. Cai L, Friedman N, Sunney X: **Stochastic protein expression in individual cells at the single molecule level.** *Nature* 2006, **440**:580-586.
33. Ocone A, Sanguinetti G: **Reconstructing transcription factor activities in hierarchical transcription network motifs.** *Bioinformatics* 2011, **27(20)**:2873-2879.
34. Guido S, Andreas R, Manfred O, Cedric A: **Switching regulatory models of cellular stress response.** *Bioinformatics* 2009, **25**:1280-1286.
35. Cover T, Thomas J: *Elements of information theory* Wiley, New York; 2006.
36. Wang YL, Liu CL, Storey JD, *et al*: **Precision and functional specificity in mRNA decay.** *Proc Natl Acad Sci USA* 2002, **99**:5860-5865.
37. Shuman S: **Structure, mechanism, and evolution of the mRNA capping apparatus.** *Prog Nucleic Acid Res Mol Biol* 2001, **66**:1-40.

doi:10.1186/1752-0509-6-S1-S3
Cite this article as: Wang and Li: Bayesian inference based modelling for gene transcriptional dynamics by integrating multiple source of knowledge. *BMC Systems Biology* 2012 **6**(Suppl 1):S3.