Bioinformatics Advance Access published September 22, 2007

Inferring Transcriptional Regulatory Networks from High-throughput Data

Rui-Sheng Wang [a,b,*], Yong Wang [c] Xiang-Sun Zhang [c] Luonan Chen [b]

[a]School of Information, Renmin University of China, Beijing 100872, China, [b]Department of Electronics, Information and Communication Engineering, Osaka Sangyo University, Osaka 574-8530, Japan, [c]Academy of Mathematics and Systems Science, CAS, Beijing 100080, China

Associate Editor: Dr. Olga Troyanskaya

ABSTRACT

Motivation: Inferring the relationships between transcription factors (TFs) and their targets has utmost importance for understanding the complex regulatory mechanisms in cellular systems. However, the transcription factor activities (TFAs) cannot be measured directly by standard microarray experiment owing to various post-translational modifications. In particular, cooperative mechanism and combinatorial control are common in gene regulation, e.g. TFs usually recruit other proteins cooperatively to facilitate transcriptional reaction processes.

Results: In this paper, we propose a novel method for inferring transcriptional regulatory networks (TRN) from gene expression data based on protein transcription complexes and mass action law. With gene expression data and TFAs estimated from transcription complex information, the inference of TRN is formulated as a linear programming problem which has a globally optimal solution in terms of L_1 norm error. The proposed method not only can easily incorporate ChIP-Chip data as prior knowledge but also can integrate multiple gene expression datasets from different experiments simultaneously. A unique feature of our method is to take into account protein cooperation in transcription process. We tested our method by using both synthetic data and several experimental datasets in yeast. The extensive results illustrate the effectiveness of the proposed method for predicting transcription regulatory relationships between TFs with co-regulators and target genes.

Availability: The software TRNinfer is available from http://intelligent.eic.osaka-sandai.ac.jp/chenen/TRNinfer.htm.

Contact: chen@eic.osaka-sandai.ac.jp, zxs@amt.ac.cn

1 INTRODUCTION

With the rapid advance of biological science, tremendous amounts of biological data have been produced by high-throughput technologies. In particular, microarray gene expression data become an increasingly common information source that can quantitatively provide insights for understanding complex mechanisms in a cell at a system-wide level (Hughes *et al*. (2002)). In addition to clustering microarray data for functional analysis, a hot topic on gene expression data analysis is the reconstruction of gene regulatory network which aims to reveal the underlying network of gene-gene interactions from the measured dataset of gene expression (Hartemink (2005); Basso *et al*. (2005)). In the last few years, a number of methods have been developed for inferring gene regulatory networks from time-course data such as bayesian networks (Beal *et al.* (2005); Hughes *et al*. (2002); Husmeier (2003)), differential equations (Chen and Aihara (2002); de Jong (2002)), and optimization techniques (Wang *et al.* (2006)).

In contrast to reconstruction of gene-gene interactions, recently inferring direct relationships between transcription factors (TFs) and target genes attracts much attention. The transcription regulation of genes is achieved by DNA-binding proteins that attach to specific DNA promoter regions. These DNA-binding proteins are known as transcription regulators or TFs which recuit other proteins to form chromatic-modifying complexes and transcription apparatus to initiate RNA syntheis (Lee *et al.* (2002)). In recent years, many experimental and computational techniques have been used to identify TFs and their target genes in several organisms such as *S. cerevisiae*, *E. Coli* and *Drosophila* (Iyer *et al.* (2001); Steensel *et al.* (2003); Harbison *et al.* (2004)), of which one important technique is ChIP-on-chip, also known as genome-wide location analysis. This technique combines a modified chromatin immunoprecipitation (ChIP) assay with microarray technology and can identify the DNA sequences (target genes) occupied by specific DNA-binding proteins (TFs) in cells. However, ChIP-on-chip technique is not so reliable and suffers from a large proportion of false positives (Boulesteix and Strimmer (2005)).

Recently, several research groups integratively analyzed gene expression data and ChIP connectivity to infer transcription factor activities (TFAs) since TFAs cannot be measured directly by standard microarray experiment. For instance, Liao *et al.* (2003) have developed a method called network component analysis (NCA) which incorporates a prior qualitative knowledge about gene-TF interactions to infer the true regulatory activities. This method was extended as PLS-based network component analysis by Boulesteix and Strimmer (2005) which offers an efficient and sound way to infer true TFAs for any given connectivity matrix without much restriction like NCA. In addition to the methods for predicting the activities of TFs, several groups also attempted to recover the network structure between TFs and their targets using the gene expression levels of both transcription factors and target genes (Segal *et al.* (2003); Xiong *et al.* (2004)). However, most existing algorithms for inferring transcription regulatory networks from gene expression data assume that one TF without other cooperative proteins participates in the regulation process of a gene and directly use its mRNA level as activity profile. As is well known to us, this assumption is not biologically reasonable, since cooperative mechanism and combinatorial control are common

*to whom correspondence should be addressed

in gene regulation and a TF usually recruits other proteins cooperatively to facilitate transcriptional function (Banerjee and Zhang (2003); Eisbacher *et al.* (2003); Charron *et al.* (1999)).

Although ChIP-chip technique can detect target genes occupied by specific DNA-binding proteins (TFs), generally a single TF cannot mediate transcription reactions. In the transcription process, TFs and several proteins cooperatively regulate the expression of a gene by combining into a protein transcription complex (TC) (Remenyi *et al.* (2004)). Those proteins in a TC generally cannot be detected by ChIP-chip technique because they are non-DNA-binding proteins, despite their similar roles in regulating a target gene as TFs. A transcription complex (TC) is a protein complex which is formed through a series of elementary biochemical reactions. This reaction mechanism obeys the law of mass action which means that the rate of any given elementary reaction is proportional to the product of the concentrations of the reactants. This law can be used to analyze the behavior of biochemical systems through dynamical equations and has been applied in reconstruction of biochemical reaction pathways from time course data (Crampina *et al*. (2004); Srividhya *et al.* (2007)). In this work, mass action law in transcription process and the expression levels of genes in TCs are used to approximate TFAs. This idea is motivated by the fact that a transcription process is often achieved by a TF with other cooperative proteins in a TC which is formed by a series of biochemical reactions. According to these biochemical reactions, the activity level of a TF can be viewed as a nonlinear function with respect to the gene expression levels of the TF and its cooperative proteins in the TC instead of the mRNA level of a single constituent gene, which is consistent with biological experiments (Banerjee and Zhang (2003); Eisbacher *et al.* (2003); Charron *et al.* (1999)).

In this paper, we propose a novel method for inferring transcriptional regulatory networks based on transcriptional factor activities (TFAs). We firstly approximate TFAs by respective transcription complexes (TCs) in transcription process based on mass action law. Then, with gene expression data and TFAs, inferring the direct interactions between TFs and target genes is formulated as a linear programming problem which is computationally fast and has a globally optimal solution in terms of L_1 error. The proposed method (TRNinfer) not only exploits the activities of transcriptional complexes which contain protein-protein interaction information but also can easily incorporate ChIP-Chip data as prior knowledge. In addition, the proposed approach can integrate multiple gene expression datasets from different experiments which can significantly alleviate the scarcity of data. A unique feature of TRNinfer is to take into account protein cooperation in transcription process. We tested our method by using both simulated data and several experimental data in yeast. Extensive results demonstrate the effectiveness and efficiency of the proposed method for predicting transcription regulatory relationships between TFs with co-regulators and target genes.

2 METHODS

2.1 Transcriptional regulatory network

Transcription regulation has been responsible for organismal complexity and diversity in the course of biological evolution and adaptation. It can be represented by a set of differential equations with gene expression levels and TFAs as variables, i.e. the dynamical model of TRN

$$\dot{x}(t) = f(a(t)) - Kx(t) \tag{1}$$

where $x(t) = (x_1(t), \cdots, x_m(t))^T$ denotes the expression level of genes with m being the number of genes, and $\dot{x}(t) = (\dot{x}_1(t), \cdots, \dot{x}_m(t))^T$ with time points $t = t_1, \cdots, t_n$ denotes the difference of gene expression levels. $a(t) = (a_1(t), \cdots, a_c(t))$ denotes the activity level of TFs (TFAs) which are generally functions of x. In above equation, the first term denotes the synthesis rates of mRNAs and the second one denotes the degradation rates where $K = \text{diag}(k_1, ..., k_m)$.

Generally, transcriptional regulations are nonlinear, i.e. $f(\cdot) = (f_1(\cdot), ..., f_m(\cdot))^T$ is a nonlinear function vector. However, due to the complex and unclear structures of biological systems, linear or additive models are often adopted. If the first order approximation of f is adopted, the linear form of (1) is

$$\dot{x}(t) = Ja(t) + b(t) \tag{2}$$

where $J = [J_{ij}]_{m\times c} = \partial f(a)/\partial a$ is an $m \times c$ Jacobian matrix or connectivity matrix. In the model, clearly if $J_{ij} > 0$, then we interpret TF j as an activator of gene i. On the other hand, if $J_{ij} < 0$, then TF j is a repressor to gene i. $b(t) = (b_1(t) - k_1x_1(t), \cdots, b_n(t) - k_nx_n(t))$ is a vector representing the external stimuli or environment conditions with self degradations. b_i is set to zero when there is no external input. Rewriting (2), we have the following matrix form which is a linear TRN.

$$\dot{X} = JA + B \tag{3}$$

where $\dot{X} = (\dot{x}(t_1), \cdots, \dot{x}(t_n))$ and $B = (b(t_1), \cdots, b(t_n))$ are $m \times n$ matrices, and $A = (a(t_1), \cdots, a(t_n))$ is a $c \times n$ matrix.

Biological systems are inherently nonlinear and show multi-stability and nonlinear oscillation which may not be so appropriately expressed by a linear model. Assume that a transcription regulatory network can be expressed by a set of nonlinear differential equations with each gene expression levels and TFAs as variables, i.e. the nonlinear TRN

$$\dot{x}(t) = f(u(t)) - Kx(t) \tag{4}$$

where $u(t) = (u_1(t), \cdots, u_n(t))^T$, $u_i(t) = \sum_{j=1}^{n} J_{ij}a_j(t) + b_i(t)$ for $i = 1, \cdots, m$. $f(u(t)) = (f(u_1(t)), \cdots, f(u_m(t)))^T$ is nonlinear functions. $f(u_i)$ specifies the production efficiency of gene i, as a function of the weighted accumulated effects of all a_j modified by the gene's activation threshold or external stimulus $b_i(t)$. $f(u_i)$ is generally expressed as a sigmoidal transfer function:

$$f(u_i(t)) = \frac{1}{1 + e^{-u_i(t)}}$$

which is a value-constrained nonlinear function. It has desirable properties such that it can closely represent actual systems, act as a molecular switch to control gene expression, and handle saturation and repression (Veitia (2003)).

In contrast to the dynamical model of (1), when the system in (1) is at an equilibrium, $\dot{x}(t) = 0$ or $x(t) = K^{-1}f(a(t))$. Liao *et al.* (2003) suggested $f(\cdot)$ can be approximated by a log-linear model, i.e. the expression level of a gene is such a function of the connection strength of each regulatory pair and regulators' activities:

$$\frac{x_i(t)}{x_i(0)} = \prod_{j=1}^{c} \left(\frac{a_j(t)}{a_j(0)}\right)^{J_{ij}}. \tag{5}$$

This equation can be written in the following matrix form after taking the logarithm

$$\log X = J \log A \tag{6}$$

where X and A are the time courses of relative gene expression levels and TFAs. For this case, we will solve J under steady state (model (12)).

We will see in the subsequent sections that for linear model (3), nonlinear model (4) or log-linear model (6), whichever TRN model is adopted, the inference process of transcription regulatory networks can all be formulated into a linear programming (LP) framework.

2.2 Approximating transcription factor activity

The TFA levels cannot be measured directly by standard microarray experiment due to various post-translational modifications and biochemical reactions (Boulesteix and Strimmer (2005)). Most existing algorithms for inferring transcription regulatory networks directly use the mRNA level of a TF as its activity (Segal *et al.* (2003); Xiong *et al.* (2004)). This assumption is not biologically reasonable since transcription process is usually achieved by one or more TFs with several cooperative proteins. These TFs and proteins form a transcriptional complex (TC) through a series of biochemical reactions (Banerjee and Zhang (2003); Eisbacher *et al.* (2003)). Therefore, the TF activity level cannot be simply determined by its mRNA but depends on the expression levels of all genes in the TC. In this work, from the biochemical reactions which obey law of mass action, we estimate the TFA levels from the protein composition of TCs and the expression levels of individual genes. Mass action law means that the rate of any given elementary reaction is proportional to the product of the concentrations of the reactants (Crampina *et al.* (2004)). Let us consider the following general chemical reaction:

$$\mathrm{A}_0 + \mathrm{A}_1 + \mathrm{A}_2 \rightleftharpoons_{k_{-1}}^{k_1} \mathrm{A}, \tag{7}$$

where k_1 and k_{-1} are the rate constants for the forward and backward reactions. $\mathrm{A}_0, \mathrm{A}_1, \mathrm{A}_2$, and A are the molecules of reactants (here, proteins, mRNAs or genes). According to the law of mass action, the velocities of the above reactions can be given by

$$\nu_1 = k_1 a_0 a_1 a_2, \quad \nu_{-1} = k_{-1} a,$$

where a_0, a_1, a_2, and a are the concentrations of reactants $\mathrm{A}_0, \mathrm{A}_1, \mathrm{A}_2$, and A respectively. Most reactions involve a number of simultaneous elementary steps. The rate of change of the concentration of any given reactant is then a sum of the rates of change due to the elementary reactions in which that reactant participates (Crampina *et al.* (2004)). Applying the law of mass action, the governing equations of the above reactions are given by

$$\begin{aligned} \frac{da_i}{dt} &= -k_1 a_0 a_1 a_2 + k_{-1} a \quad \text{for } i = 0, 1, 2, \\ \frac{da}{dt} &= k_1 a_0 a_1 a_2 - k_{-1} a \,. \end{aligned} \tag{8}$$

The above framework can be used to approximate the activities of TFs according to the transcriptional elementary reactions and the gene members of the TC. For example, assume that a TF A_0 recruits two proteins A_1 and A_2 to form a transcription complex A for regulating a transcription process as shown in (7) and Fig. 1 , i.e. the TC is composed of

$$\mathrm{A} = \{\mathrm{A}_0, \mathrm{A}_1, \mathrm{A}_2\}.$$

Let x_0 and x_1, x_2 represent the mRNA expression levels of the corresponding genes. After the translation process, the concentrations a_0, a_1, a_2 of individual proteins A_0, $\mathrm{A}_1, \mathrm{A}_2$ before various modifications and reactions are proportional to their mRNA expression levels: $a_i \propto x_i$. Therefore, assuming that the reactions generating a in (8) is much faster that the reactions synthesizing x in (1) or (2), according to the law of mass action and the biochemical reactions (7), the activity level a of A (represented by the activity of TF A_0) can be given approximately by

$$a = k_0 a_0 a_1 a_2 \approx k x_0 x_1 x_2 \tag{9}$$

which is estimated by the expression levels of individual genes in the TC, and acts as the overall activity of TF A_0 after biochemical reaction modifications. Obviously, the overall activity a of TF A_0 after recruiting proteins A_1 and A_2 is not simply proportional to its gene expression level x_0 but dependent on all x_i in the TC. In the following section, we will use gene expression data with such estimated activity levels of TFs to reconstruct transcription regulatory networks.

It is worth noting that we also use the mRNA level of a constituent gene as its protein concentration in estimating the activity level of a transcription complex. This assumption may not be always true since some post-translational modification events such as phosphorylation, degradation can affect transcriptional activities (Tootle and Rebay (2005)). However, as mentioned by Tootle and Rebay (2005), studies on post-translational modifications of transcription factors are only in the initial stages, and it is a still difficult task to quantify the effects of these biological events. Although using the mRNA levels of all constituent genes in a TC is one step forward, more elaborated formulations are needed in the future.

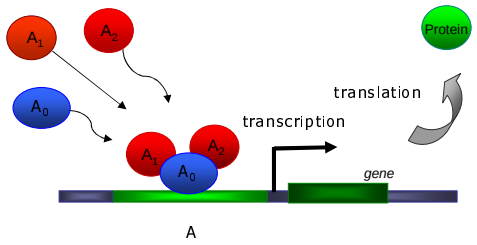

Fig. 1. An illustration for a transcription complex participating in transcription process.

2.3 Inferring transcription regulatory network

Assume that there are multiple datasets for one organism which may be measured under various environments by different labs. Let X^k denote the kth gene expression data $(m \times n_k)$ and A^k denote the kth activity data of TF complexes $(t \times c)$. Then, according to the differential equation (3) by ignoring the last term, we have

$$\dot{X}^k = JA^k$$

where J is an $m \times c$ matrix representing the direct regulation relationships between TFs and genes. For one dataset X, we intend to find a connection matrix J to minimize total errors:

$$\min_J |\dot{X} - JA| + \lambda |J|.$$

For all L datasets, J should be as consistent as possible with all datasets, which can be achieved by

$$\min_J \sum_{k=1}^{L} |\dot{X}^k - JA^k| + \lambda |J| \tag{10}$$

where the first term is to minimize the error between real data and the reconstructed model, whereas the second term is the sparsity term which forces J sparse by using L_1 norm. Since a biological gene network is expected to be sparse, this term is intended to overcome the defects that conventional approaches often have densely connected regulatory relationships among nodes. λ is a positive parameter which balances the error and sparsity term in the objective function. (10) is an optimization problem with positive combination of L_1 norm of variables J_{ij} which can be transformed into a linear programming (LP) problem through a well-known procedure (see Supporting Materials). Owing to L_1 norm, generally the optimal solution of (10) has as many zeros for $|\dot{X}^k - JA^k|$ and $|J|$ as possible, which exactly serves our purpose, i.e. consistent and sparse structure. Also the proposed framework can handle multiple datasets even if they are obtained from different experiments or conditions. Although the solution may depend on λ, it is a single parameter which can be tuned in a relatively easy manner or be simply tested for a range of its value.

For the sigmoidal nonlinear model of the TRN (4), we have

$$f^{-1}(\dot{x}_i(t) + k_i x_i(t)) = u_i(t) = \sum_{j=1}^{n} J_{ij} a_j(t) + b_i(t)$$

which is a typical linear form similar to (2). Such a fact suggests that we can efficiently infer the nonlinear transcriptional regulatory network by a similar

LP model:

$$\min_J \sum_{k=1}^{K} |f^{-1}(\dot{X}^k + KX^k) - JA^k| + \lambda|J|. \quad (11)$$

In contrast to model (10) or (11) corresponding to the dynamical TRNs (3) and (4), for the log-linear static TRNs, (6), a similar LP problem in the same framework can be obtained:

$$\min_J \sum_{k=1}^{L} |\log X^k - J \log A^k| + \lambda|J|. \quad (12)$$

Notice that it is a major task to derive $m \times c$ elements of J in computational biology because J is the strength matrix of the regulatory interactions between TFs and target genes. In this work, we estimate $\dot{x}(t)$ by using linear difference scheme $\dot{x}(t_j) = [x(t_{j+1}) - x(t_j)])/[t_{j+1} - t_j]$ (Yeung *et al.* (2002); Wang *et al.* (2006)). Above models can employ multiple microarray data sets from different conditions or experiments simultaneously which will alleviate significantly the scarcity of time points in a single data set. In order to make the inference more accurate when only a single data set is available, cubic spline interpolation methods may be explored to calculate the derivatives of $x(t)$ (de Boor (1978)).

3 EXPERIMENT RESULTS

In this section, we conducted several numerical experiments to evaluate the method by using multiple synthetic datasets and yeast microarray gene expression data. The algorithm is implemented in the Fortran programming language on a 1.66 GHz Pentium 4 PC and the software (TRNinfer) is available from http://intelligent.eic.osaka-sandai.ac.jp/chenen/TRNinfer.htm.

3.1 Synthetic data

3.1.1 Synthetic time-course data The first example is a synthetic transcriptional regulatory network with three genes and three transcriptional factors (see Figure 2(a)). In this network, TF1 with the cooperation of protein P1 and protein P2 forms a TC1 to regulate $tf2$ (activation) and gene $p3$ (repression), while TF3 and protein 3 form a TC3 to regulate gene $p2$ and $tf1$. TF2 as a special TC2 regulates $tf3$ (repression) and gene $p1$ (activation). These relationships are governed by the following nonlinear differential equations:

$$\begin{aligned}
\dot{x_1}(t) &= 2a_{\mathrm{TF2}}(t) - x_1(t) + \xi_1(t) \\
\dot{x_2}(t) &= 1.5a_{\mathrm{TF3}}(t) - x_2(t) + \xi_2(t) \\
\dot{x_3}(t) &= -3a_{\mathrm{TF1}}(t) - x_3(t) + \xi_3(t) \\
\dot{x}_{tf1}(t) &= 0.5a_{\mathrm{TF3}}(t) - x_{tf1}(t) + \xi_4(t) \\
\dot{x}_{tf2}(t) &= 2.2a_{\mathrm{TF1}}(t) - x_{tf2}(t) + \xi_5(t) \\
\dot{x}_{tf3}(t) &= -2a_{\mathrm{TF2}}(t) - x_{tf3}(t) + \xi_6(t)
\end{aligned} \quad (13)$$

where $x_i(t)$ denotes the expression level of the gene i, x_{tfi} is the expression level of the transcription factor i, and $a_{\mathrm{TF1}}(t) = x_{tf1}(t)x_1(t)x_2(t), a_{\mathrm{TF2}}(t) = x_{tf2}(t), a_{\mathrm{TF3}}(t) = x_{tf3}(t)x_3(t)$. ξ_i represents all of noises for microarray data.

We randomly chose the initial condition of the system and took several points of x as a measured time-course dataset. With five different initial conditions, we obtained 5 different datasets with 5 time points respectively. According to the method for computing transcription factor activities (9), $x_{tf1}(t)x_1(t)x_2(t)$, $x_{tf2}(t)$ and $x_{tf3}(t)x_3(t)$ respectively represent the activities of TF1, TF2 and TF3 (or TC1, TC2 and TC3 in Figure 2(a)). With the time-course data and the activity levels of TFs, we applied LP model (10) to reconstruct the connectivity matrix of TFs and target genes (i.e. the Jacobian matrix J). The true regulatory structure and coefficients can be recovered accurately by TRNinfer. Specifically, when a single dataset is used and the sparse parameter $\lambda = 0$, we obtained a connection matrix shown in the first subtable of Table 1, where the inferred connection matrix is dense and the corresponding entries are not accurate, compared with the original coefficients in (13) or Figure 2(b). By increasing the sparse parameter to $\lambda = 0.05$, the inferred connection matrix is shown in the second subtable of Table 1. We can see that it is sparser than the one obtained at $\lambda = 0$ and the corresponding entries are also more accurate. Such results justify the benefit brought by the sparse parameter. Moreover, when multiple datasets are used, we obtained a more accurate connection matrix which illustrates the advantage of employing more datasets (see the third and forth subtables of Table 1).

Table 1. The connection matrix inferred by TRNfinfer on simulated data, where L denotes the number of datasets employed and λ is the sparse parameter.

$L = 1,\ \lambda = 0.0$	$p1$	$p2$	$p3$	$tf1$	$tf2$	$tf3$
TC1 (TF1)	0.98	-2.80	-1.57	-1.90	-3.20	-2.62
TC2 (TF2)	1.73	0.85	-0.49	0.58	1.71	-0.20
TC3 (TF3)	0.70	-0.77	1.13	-1.02	-4.34	-2.07
$L = 1,\ \lambda = 0.05$	$p1$	$p2$	$p3$	$tf1$	$tf2$	$tf3$
TC1 (TF1)	0.11	0.0	-3.06	0.0	2.26	0.0
TC2 (TF2)	2.0	0.0	0.0	0.0	0.5	-1.0
TC3 (TF3)	0.0	1.45	0.0	0.47	0.0	0.0
$L = 5,\ \lambda = 0.0$	$p1$	$p2$	$p3$	$tf1$	$tf2$	$tf3$
TC1 (TF1)	0.12	0.0	-3.17	0.0	2.31	0.0
TC2 (TF2)	2.0	0.0	0.0	0.0	0.5	-1.0
TC3 (TF3)	0.0	1.49	0.0	0.50	0.0	0.0
$L = 5,\ \lambda = 0.5$	$p1$	$p2$	$p3$	$tf1$	$tf2$	$tf3$
TC1 (TF1)	0.0	0.0	-3.01	0.0	2.20	0.0
TC2 (TF2)	2.0	0.0	0.0	0.0	0.0	-1.0
TC3 (TF3)	0.0	1.49	0.0	0.50	0.0	0.0

The results for a simulated example with noises $\xi_i(t)$ (see Table 4 in Supporting Materials) also illustrate the effectiveness of the LP method on suppressing noises and the advantages of the sparse parameter and multiple datasets.

3.1.2 The hemoglobin data Now we use a network of seven hemoglobin solutions (denoted by $M_1, \cdots, M_7$) and their absorbance spectra which were measured in Liao *et al.* (2003) to evaluate our method. Each solution contains a combination of three components: oxyhemoglobin, methemoglobin, and cyanomethemoglobin. According to Beer-Lambert law, the absorbance spectra of the mixture can be described as a linear combination of the composition proportions of three components and the absorbance spectra of each pure solution (Liao *et al.* (2003)). The mixing diagram representing the compositions of pure components is shown in Fig. 3. The mixture composition is known and we test if or not TRNinfer can correctly identify the compositions of each mixture and the corresponding concentration. Here LP model (12) is adopted since the data set is not time-course and

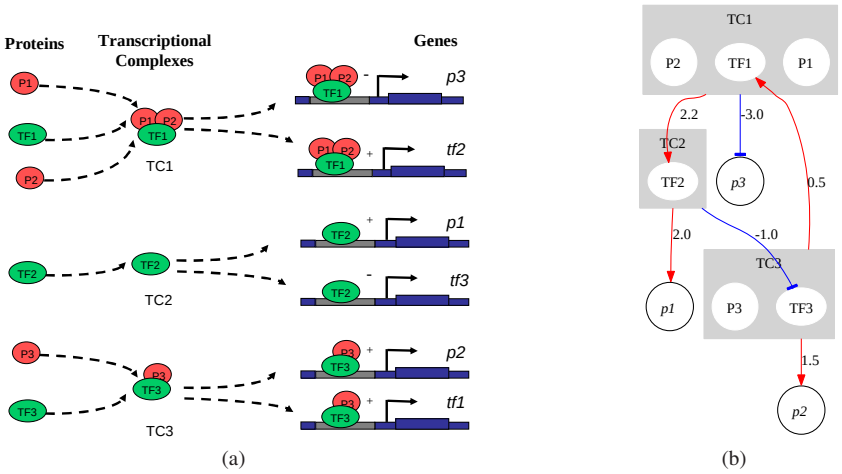

Fig. 2. A simulated transcriptional regulatory network. Genes $p1$, $p2$, $p3$, $tf1$, $tf2$, and $tf3$ synthesize proteins P1, P2, P3, TF1, TF2, and TF3 respectively. The red arrows in the figure indicate repression while the blue arrows indicate activation. (a). Transcription processes (b). Transcriptional regulatory network.

Table 2. The mixture composition inferred by TRNinfer for the hemoglobin data. The values in parenthesis are the inferred composition proportions

Mixtures	OxyHb	MethHb	CyanoHb
M1	0.1(0.12)	0.9(0.79)	0.0
M2	0.5(0.48)	0.0	0.5(0.42)
M3	0.0	0.9(0.77)	0.1(0.07)
M4	0.1(0.11)	0.8(0.70)	0.2(0.07)
M5	0.3(0.27)	0.7(0.59)	0.0
M6	0.2(0.2)	0.0	0.8(0.74)
M7	0.0	0.5(0.44)	0.5(0.48)

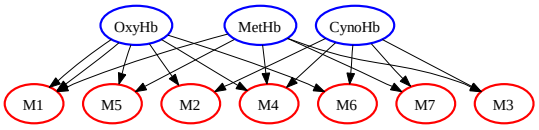

Fig. 3. The mixing diagram of the seven hemoglobin solutions from three pure components.

can be viewed as steady state data. The result of our method on this dataset is summarized in Table 2. Clearly, TRNinfer can identify the components in each of mixed hemoglobin solution. The inferred compositions of each component are near the true one but with a certain error. This is because the mixed absorbance spectra in experiment data are not exactly equal to the linear combination of the composition proportions of three components and the absorbance spectra of each pure solution. Provided that this linear combination exactly holds, our method can not only identify the components but also exactly infer the corresponding composition proportions (see Table 5 in Supporting Materials).

3.2 Experimental data

In this section, we apply TRNinfer to experimental data. Since the transcription activity of a TF is approximated by the corresponding transcription complex (TC) based on the expression levels of individual genes through the law of mass action, we need to collect data of TCs. In the budding yeast *S. cerevisiae*, ChIP-chip experiments have been utilized to elucidate the binding interactions between 6270 genes and 113 preselected TFs (Lee *et al.* (2002)). By checking yeast protein complexes in MIPS (Mewes *et al.* (2002)), we found that 26 TFs are in protein transcription complexes. The number is not so significant mainly because the current deposition of protein complexes is highly incomplete. With the increasing deposition of information, more protein complexes will be available. Among these 26 TFs, some are related to yeast cell cycle (Spellman *et al.* (1998)) and some are related to polyphosphate metabolism in *S. cerevisiae* (Ogawa *et al.* (2000)). We report the results of TRNinfer on these two datasets.

3.2.1 Yeast cell cycle data We first tested our method using gene expression data for cell cycle studies in *S. cerevisiae* (Spellman *et al.* (1998)). According to the ChIP experiments (Lee *et al.* (2002)), there are 11 TFs that are known to be related to cell-cycle regulation, among which 5 TFs are in 4 different transcription complexes (TCs). The details of these TFs and their TCs are given in Table 6 in Supporting Materials. Except these 5 TFs, we selected 8 genes that are closely related to cell cycle based on their function information. According to the gene expression data from Spellman *et al.* (1998), we generated 4 datasets with the number of time points 18, 17, 24, and 14 respectively.

TRNinfer solved the above time-course datasets within five seconds by using LP model (10) and we obtained a transcription regulatory connection matrix which characterizes the interaction between target genes and TFs (or TCs)(see Table 7 of Supporting Materials). The corresponding transcription regulatory network is shown in Fig. 4, where the bold edges indicate that the regulatory relations confirmed by documented information and the thin edges are potential regulations. The unconfirmed inferred regulations are denoted by dotted edges. From Fig. 4, clearly most regulatory

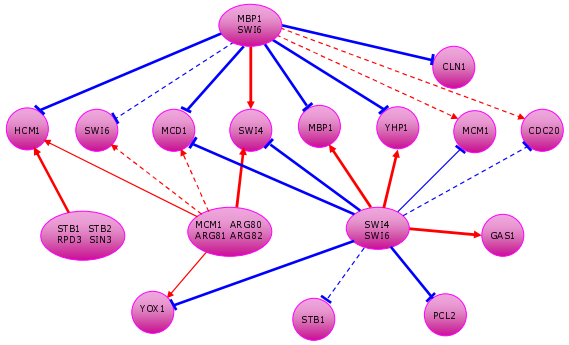

Fig. 4. The inferred yeast cell cycle transcriptional regulatory network. The red arrows in the figure indicate repression while the blue arrows indicate activation.

relations inferred by our method can be confirmed (Teixeira *et al.* (2006)). In this transcriptional network, DNA-binding proteins SWI4 and SWI6 are transcription cofactors, forming a complex to regulate transcription at the G1/S transition. They are involved in meiotic gene expression, localization regulated by phosphorylation and potential Cdc28p substrate (Teixeira *et al.* (2006)). SWI6 and MBP1 are believed to involve in a same regulatory module by many literature (Bar-Joseph *et al.* (2003); Wu *et al.* (2006)). MBP1 forms a complex with SWI6 that binds to MluI cell cycle box regulatory element in promoters of DNA synthesis genes. The inferred yeast cell cycle transcription network demonstrates that the proposed method can provide not only direct relations between TFs and target genes but also the interactions between co-regulating proteins and target genes.

We introduced a p-value formula used in Husmeier (2003) to evaluate the significance of the predicted interactions between TFs and target genes:

$$P = 1 - \sum_{k=0}^{N} \binom{n}{k} p^k (1-p)^{n-k} \qquad (14)$$

where n is the number of all possible interactions between TFs and target genes, N is the confirmed or potential regulations among the predicted ones, k is the number of predicted interactions, and p is the probability that a random edge is a true network edge. In this way, the inferred cell cycle transcriptional network is significant ($P < 0.005$), whereas the average p-value of the inferred network on multiple permuted data sets (the expression level of each gene is permuted with respect to the experiment conditions or time points) is 0.49. If we directly use the mRNA level of a TF as TF activity and ignore the effects of cooperative proteins in transcription complex, the p-value of the inferred network is 0.35.

In order to justify the benefit of considering transcription complexes, we conducted comparative experiments for three methods: LP method based on transcription complexes, LP method based on only mRNA levels of TFs and SVD method based on mRNA levels of TFs (Yeung *et al.* (2002)). Sensitivity and specificity are used to evaluate the inference results. Owing to the scarcity of gold standard information (true network), we treated the regulations with supporting evidences in YEASTRACT (Teixeira *et al.* (2006)) as true edges and all other regulations as true negatives. The results are plotted in Figure 5(a), where the curves are obtained by setting different thresholds on the inferred connection matrix. From Figure 5(a), we can see that considering transcription complexes can make the inference more accurate. The worse performances of LP_mRNA and SVD_mRNA come from two aspects: one is that there may exist bias in the concrete sensitivity and specificity values since all unsupported regulations are treated as false positives which may not be true. Further biological experiments are needed to confirm them. Another is that, in this example, four TFs are all in transcription complexes, which means that these TFs recruit other cooperative proteins in transcription processes, so only using the mRNA levels of TFs as their activity profiles will lead to big deviation. We also conducted similar experiments like Husmeier (2003), i.e. added extra artificial TF complexes to test the robustness of our method. For this data set, two artificial TF complexes (i.e. increased 2×13=26 possible edges) whose activity profiles were randomly generated. The computational result indicates that the original confirmed edges are almost not affected and only a very small portion of the added possible edges are inferred as false positives, which demonstrates the robustness of our method (see Figure 8(a) in Supporting Materials).

Table 3. The p-values of the periodicity for some TFs related with cell cycle

TFs	Experiment conditions	Expression	Activity
MBP1	alpha0min ∼alpha119min	0.525	0.003
SWI4	alpha0min ∼alpha119min	0.0064	0.00019
SWI6	alpha0min ∼alpha119min	0.367	0.00019
SWI4	cdc1510min ∼ cdc15290min	0.132	0.01
SWI6	cdc1510min ∼ cdc15290min	0.024	0.01

To test the possibility of approximating TFAs by the expression levels of individual genes in TCs, we checked the periodicity of the activity levels of the TFs because it is believed that the activities of TFs related to cell cycle tend to be periodic. The activity profiles of MBP1 (TC1: 510.190.70) and SWI4/SWI6 (TC4: 510.190.60) are plotted in Fig. 7 of Supporting Materials. The activity profiles of these regulation factors show highly periodic patterns which are consistent with common biological knowledge. In contrast, the individual gene expression profiles of MBP1, SWI4 and SWI6 are not periodic. This fact can be confirmed by Fisher's g-test in Wichert *et al.* (2004). Table 3 lists the p-values of the periodicity for the activity profiles and the gene expression profiles of these genes which indicate that their activity profiles are more periodic than the corresponding gene expression profiles. Although the periodicity of the activity profiles for MCM1 (TC2:510.190.120) and STB1 (TC3:510.190.150) is not so significant, it is still more significant than their gene expression profiles in terms of p-value. These results indicate that the activity of a TF cannot be well represented by its mRNA level (Boulesteix and Strimmer (2005); Liao *et al.* (2003)) but depends on the expression levels of all genes in the transcription complex. Therefore, those methods which only use the mRNA level of one TF or constituent gene are not biologically reasonable and likely to misexplain important transcriptional regulatory relations.

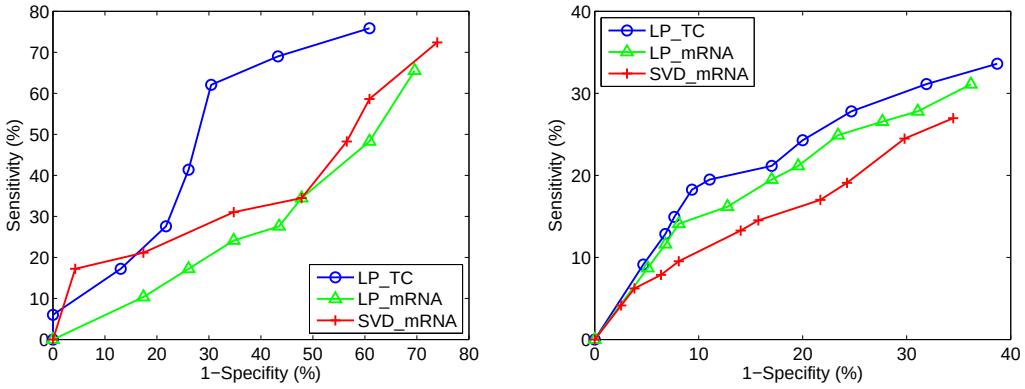

Fig. 5. The comparison results of LP method based on transcription complexes (LP_TC), LP method based on only mRNA levels of TFs (LP_mRNA) and SVD method based on mRNA levels of TFs (SVD_mRNA). (a) on yeast cell cycle data set; (b) on yeast polyphosphate metabolism data set.

3.2.2 Polyphosphate metabolism data Now, we test TRNinfer using gene expression data for polyphosphate metabolism in *S. cerevisiae* (Ogawa *et al.* (2000)). Among the TFs related to polyphosphate metabolism verified by the ChIP experiments (Lee *et al.* (2002)), there are 14 TFs in 9 different transcription complexes. The details of these TFs and their TCs are given in Table 8 of Supporting Materials. This gene expression data have totally 8 conditions. Among the genes in this dataset, those with change of 2-fold up or down in at least two time points of the expression levels are believed to be closely related to polyphosphate metabolism. In such a way, totally 64 genes (including 14 TFs) form a test data.

TRNinfer solved this non-time-course data set within one second by using LP model (12) and we obtained a transcription regulatory network with 106 links. Since the network is large, only a part with three TFs is shown in Fig. 6, where the types of edges have similar meanings to those in cell cycle TRN. The genes with same functions are denoted by the same color and the functions of the genes with gray color are unknown yet. By the formula (14), the inferred polyphosphate metabolism transcriptional network is significant ($P < 0.002$)), whereas the average p-value of the inferred network on multiple permuted data sets is 0.33, and the p-value of the inferred network is 0.32 by using the mRNA level of a TF as activity profile. In this transcriptional network, RTG1 and RTG3 are bHLH/Zip proteins and transcription cofactors. RTG3 forms a complex with RTG1 to activate the retrograde (RTG) and TOR pathways (Teixeira *et al.* (2006)). Again the proposed method can also provide the interactions between coregulating proteins and target genes in addition to the direct relations between TFs and target genes.

The comparative results of three methods mentioned in above subsections on this data set are plotted in Figure 5(b), which again demonstrates the higher inference accuracy of LP method with considering transcription complex activities. The low sensitivity and specificity values of three methods are because they all inferred a smaller size network with fewer edges than the 'true' network. Similarly, three artificial TF complexes (i.e. increased 3×64=192 possible edges) were added to test the robustness of our method.

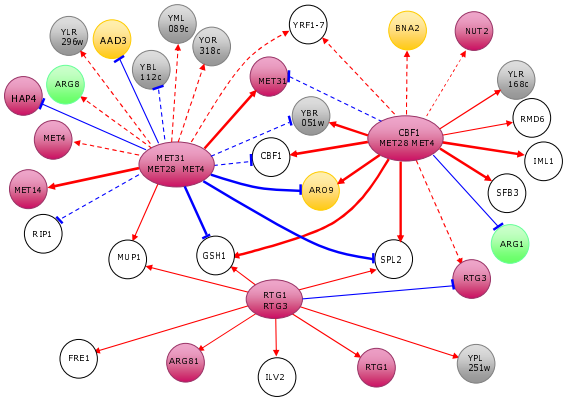

Fig. 6. The inferred transcriptional regulatory network for polyphosphate metabolism. The red arrows in the figure indicate repression while the blue arrows indicate activation. The nodes in the same color indicate that the genes have the same biological function.

The computational result (see Figure 8(b) in Supporting Materials) again demonstrates that our method is robust to false positives.

4 CONCLUSION AND DISCUSSION

Revealing the direct relationships between the target genes and TFs with coregulators is essential for understanding the complex regulatory mechanisms in cellular systems. A TF usually cannot facilitate the transcription without the cooperation of other proteins, so the activity of a TF is not simply determined by its mRNA but depends on the expression levels of all genes in the transcription complex. In this paper, we proposed a method for inferring transcriptional regulatory networks by transcriptional complexes based on mass action law, with major features as follows.

- The proposed method takes into account protein cooperation in transcription process and infers TRNs without the requirement of experimentally measuring the activities of TFs.

- The inference process is formulated as a linear programming problem which ensures global solution can be efficiently found and makes our method scales up well to large-scale networks (e.g. with over hundreds or even thousands of genes).
- We can identify the direct regulations not only between TFs and target genes but also between coregulating proteins and target genes.
- The model can deal with multiple gene expression datasets from different experiments and be easily extended to accommodate the nonlinear formalism of transcription process.

We tested our method by using both simulated data and several experimental data in yeast. Extensive results illustrated the effectiveness of the proposed method on predicting transcription regulatory relationships. Currently, the number of known transcription complexes is small, especially for organisms except yeast. With the rapid development of biotechnologies, more datasets will be available for us which can enhance the significance of the proposed method.

A limitation in our approach is that we ignore post-translational modifications in transcription processes. As is well known to us, post-translational modifications such as phosphorylation, protein degradation may affect the activity of a TF. Current studies on the effects of post-translational modifications are only in the initial stages, and it is still a difficult task to quantify the effects of these biological events (Tootle and Rebay (2005)). Although using the mRNA levels of all constituent genes in a TC is one step forward, more elaborated formulations are needed in the future. Furthermore, the mechanisms of competitive binding and combinatorial control between multiple TFs will also affect TF activity levels. Integrating detailed cooperation processes of TFs in gene regulation will make the reconstruction of TRN more realistic.

In the inference process of transcriptional network, in order to avoid obvious false positives and improve the inference accuracy, generally we require that there is at least one candidate transcriptional complex to be included in data set for each gene. If such information is unavailable, we can filter out those genes according to the weak correlations between $\dot{X}$ and all TFAs among A or biological function annotations as a preprocessing procedure.

ACKNOWLEDGEMENT

The authors are grateful to the anonymous referees for their valuable comments and suggestions. This research work is supported by JSPS under JSPS-NSFC collaboration project and the Ministry of Science and Technology, China, under grant No.2006CB503905.

REFERENCES

Banerjee, N. and Zhang, M.Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.*, **31**, 7024-7031.

Bar-Joseph,Z., Gerber,G.K., Lee,T.I., et al. (2003) Computational discovery of gene modules and regulatory networks. *Nat Biotechnol*, **21**, 1337-42.

Basso,K., Margolin,A., Stolovitzky,G., Klein,U., Dalla-Favera,R. and Califano,A. (2005) Reverse engineering of regulatory networks in human B cells. *Nat Genet.,* **37**, 382–390.

Beal,M.J., Falciani,F., Ghahramani,Z., Rangel,C. and Wild,D.L. (2005) A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics,* **21**, 349-356.

Boulesteix,A.L., Strimmer,K. (2005) Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. *Theoretical Biology and Medical Modelling*, **2**:23 doi:10.1186/1742-4682-2-23.

Charron,F., Paradis,P., Bronchain,O. et al. (1999) Cooperative interaction between GATA-4 and GATA-6 regulates myocardial gene expression. *Molecular and Cellular Biology*, 4355-4365.

Chen,L. and Aihara,K. (2002) Stability of genetic regulatory networks with time delay. *IEEE Trans. on Circuits and Systems-I,* **49**, 602-608.

Crampina,E.J., Schnella,S., McSharry,P.E. (2004) Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. *Progress in Biophysics & Molecular Biology*, **86**, 77-112.

de Boor, C. *A Practical Guide to Splines*. New York: Springer-Verlag Press, 1978.

de Jong, H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, 9(1), 67-103.

Eisbacher,M. Holmes,M.L., Newton,A. et al. (2003) Protein-protein interaction between Fli-1 and GATA-1 mediates dynergistic expression of megakaryocyte-dpecific henes through cooperative DNA Binding. *Molecular and Cellular biology*, p. 3427-3441.

Harbison,C.T., Gordon,D.B., Lee,T.I., et al. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99-104.

Hartemink,A.J. (2005) Reverse engineering gene regulatory networks. *Nature Biotechnology,* **23**, 554–555.

Hughes,T.R., Marton,M.J., Jones,A.R. et al. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109-126.

Husmeier,D. (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics,* **19**, 2271-2282.

Iyer,V.R., Horak,C.E., Scafe,C.S., Botstein,D., Snyder,M., Brown,P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533-538.

Liao,J.C., Boscolo,R., Yang,Y.L., Tran LM, Sabatti, C., Roychowdhury,V.P. (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci USA*, **100**, 15522-15527.

Lee TI, Rinaldi NJ, Robert F, et al. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799-804.

Mewes,H.W., Frishman,D., Guldener,U. et al. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, **30**, 31-34.

Ogawa,N., DeRisi,J., Brown,P.O. (2000) New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis. *Mol Biol Cell.* **11**, 4309-4321.

Reményi, A., Schöer, H. R., Wilmanns, M. (2004) Combinatorial control of gene expression. *Nature Structure & Molecular Biology*, 11, 812-815.

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifing regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, **34**, 166-176.

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D., Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, **9**, 3273-97.

Srividhya,J., Crampin,E.J., McSharry,P.E., Schnell,S. (2007) Reconstructing biochemical pathways from time course data. *Proteomics*, **7**, 828-838.

van Steensel,B., Delrow,J., Bussemaker,H.J. (2003) Genomewide analysis of Drosophila GAGA factor target genes reveals contextdependent DNA-binding. *Proc Natl Acad Sci USA*, **100**, 2580-2585.

Teixeira,M.C., Monteiro,P., Jain,P. et al. (2006) The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae, *Nucl. Acids Res.*, 34, D446-D451.

Tootle, T.L., Rebay, I. (2005) Post-translational modifications influence transcription factor activity: a view from the ETS superfamily. *Bioessays*, 27(3), 285-298.

Veitia, R. A., (2003) A sigmoidal transcriptional response: cooperativity, synergy and dosage effects. *Biol. Rev.*, 78, 149-170.

Wang,Y., Joshi,T., Zhang,X-S., Xu,D., and Chen,L. (2006) Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics*, **22**, 2413-2420.

Wichert,S., Fokianos,K., Strimmer,K. (2004) Indentifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, **20**:5-20.

Wu,W.S., Li,W.H., Chen,B.S. (2006) Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle. *BMC Bioinformatics*, **7**, 421.

Xiong,M., Li,J., Fang,X. (2004) Identification of genetic networks. *Genetics*, **166**, 1037-1052.

Yeung,M.K.S., Tegner,J. and Collins,J. (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA*, 99, 6163-6168.