BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 16 2010, pages 2020–2028
doi:10.1093/bioinformatics/btq325

Systems biology

Advance Access publication June 16, 2010

Causal relationship inference for a large-scale cellular network

Tong Zhou[1,2,†,*], and Ya-Li Wang[1,†]

[1]Department of Automation and [2]Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing, 100084, China

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Cellular networks usually consist of numerous chemical species, such as DNA, RNA, proteins and small molecules, etc. Different biological tasks are generally performed by complex interactions of these species. As these interactions can rarely be directly measured, it is widely recognized that causal relationship identification is essential in understanding biological behaviors of a cellular network. Challenging issues here include not only the large number of interactions to be estimated, but also many restrictions on probing signals. The purposes of this study are to incorporate power law in cellular network identification, in order to increase accuracy of causal regulation estimations, especially to reduce false positive errors.

Results: Two identification algorithms are developed that can be efficiently applied to causal regulation identification of a large-scale network from noisy steady-state experiment data. A distinguished feature of these algorithms is that power law has been explicitly incorporated into estimations, which is one important structural property that most large-scale cellular networks approximately have. Under the condition that parameters of the power law are known and measurement errors are Gaussian, a likelihood maximization approach is adopted. The developed estimation algorithms consist of three major steps. At first, angle minimization between subspaces is utilized to identify chemical elements that have direct influences on a prescribed chemical element, under the condition that the number of direct regulations is known. Second, interference coefficients from prescribed chemical elements are estimated through likelihood maximization with respect to measurement errors. Finally, direct regulation numbers are identified through maximizing a lower bound of an overall likelihood function. These methods have been applied to an artificially constructed linear system with 100 elements, a mitogen-activated protein kinase pathway model with 103 chemical elements, some DREAM initiative *in silico* data and some *in vivo* data. Compared with the widely adopted total least squares (TLS) method, computation results show that parametric estimation accuracy can be significantly increased and false positive errors can be greatly reduced.

Availability: The Matlab files for the methods are available at http://bioinfo.au.tsinghua.edu.cn/member/ylwang/Matlabfiles_CNI.zip

Contact: tzhou@mail.tsinghua.edu.cn

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

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Received on January 19, 2010; revised on May 27, 2010; accepted on June 11, 2010

1 INTRODUCTION

Inferring causal relationships of a cellular network is one of the fundamental problems in understanding cell behaviors. Major reasons behind extensive attentions to this problem appear to lie upon two facts. One is that cellular networks are generally of a large scale and there are a huge number of causal interactions to be estimated. The other is that there are many restrictions on probing signals in biochemical experiments. For example, the perturbations must be non-negative and some chemical elements may not be directly affected, and so on. These make the problem of estimating a cellular network much more difficult than identifying an industrial process (Akutsu *et al*., 1999; Andrec *et al*., 2005; Barabasi and Oltvai, 2004; Gardner and Faith, 2005).

With significant development of high-throughout technologies and proteomics analysis methods, several approaches have now been proposed for unraveling the interactions of a cellular network, such as Boolean network methods (Akutsu *et al*., 1999; Shmulevich *et al*., 2002; Zheng and Kwoh, 2004), Bayesian network methods (Ferrazzi *et al*., 2007; Perrin *et al*., 2003), partial correlation analysis (de la Fuente *et al*., 2004), differential equation-based time series analysis (Bansal *et al*., 2006; Sontag, 2008) and so on. However, when applying these methods to large-scale cellular networks, several difficulties may arise (Cantone *et al*., 2009; Prill *et al*., 2010). For example, with the number increment of chemical elements in a network, computational burden of most of the methods increases exponentially. To overcome this difficulty, some methods limit the maximum direct regulation number, but this may restrict application ranges of the method itself and give rise to a problem of selecting principal chemical elements. On the other hand, statistical methods such as partial correlation analysis rely on a variety of pairwise correlation measures. This means that an indirect effect may be wrongly recognized as a direct one, and therefore leads to a high false positive rate.

In addition, a 'top-down' approach is proposed in Kholodenko *et al*. (2002) and Andrec *et al*. (2005) for causal regulation inference from steady-state concentration changes of chemical species. This strategy is based on the total differential formula and total least square (TLS) estimations. Afterwards, the results have been extended to time series data, quasi-steady-state data, etc. (Sontag, 2008), and the 'non-direct effect' condition on experiment designs has been significantly weakened (Berman *et al*., 2007). Our computation experiences, however, show that when data length is of a moderate size, it is very rare that an identified regulation

coefficient is near zero. Recalling that a large-scale system usually has a sparse structure, this means that TLS estimates usually give incorrect information in case that a chemical element has no direct effect on another chemical element. In other words, there may exist significant false positive errors in TLS-based estimates.

On the other hand, it is widely recognized that in cellular network identification, distinguishing direct and indirect regulations is of great biological significance (Barabasi and Oltvai, 2004; Cantone *et al*., 2009; Prill *et al*., 2010). In actual identifications, however, experimental data alone are rarely sufficient to obtain a statistically sound model. To make things worse, it may even be possible that the identification problem is under-determined if direct regulations exist between every two chemical elements. These imply that in this identification problem, compared with false negative errors, it may be much more difficult to reduce false positive errors. To overcome this difficulty, various interesting methods have been suggested, such as to incorporate qualitative knowledge, to restrict the maximum number of non-zero regulatory inputs, to penalize the sum of direct regulation strengthes, etc. (Cantone *et al*., 2009; Chang *et al*., 2008; Gardner *et al*., 2003; Prill *et al*., 2010).

In this article, we investigate possibilities of utilizing structural information of a large-scale system to reduce both computational complexities and false positive errors. As pointed out in Barabasi and Oltvai (2004), most large-scale cellular networks obey approximately a so-called power law.[1] This observation gives us some important a priori information about the structure of cellular networks. However, to the best of our knowledge, there are few researches so far that have explicitly explored possibilities of utilizing this structural property in cellular network identification. In this article, we incorporate this property into some estimation algorithms, in order to facilitate the inference of direct causal regulations from noisy steady-state experiment data. More specifically, a power law is incorporated into a likelihood function to give an estimate about the direct regulation number to a chemical element. When the number of direct regulations is fixed, angle minimization between subspaces spanned by two groups of measurement data vectors is adopted to identify chemical elements that provide direct regulations, while the regulation coefficients are estimated through likelihood maximization with respect to measurement errors. Compared with a combinatorial optimization-based algorithm, an attractive characteristic of the suggested methods is that their computational complexity increases only linearly with the increment of species number.

The suggested algorithms are compared with the widely adopted TLS method, on an artificially generated linear network with 100 elements, a mitogen-activated protein kinase (MAPK) signaling network model with 103 proteins (Schoeberl *et al*., 2002), some *in silico* data of the DREAM initiative (Prill *et al*., 2010) and some *in vivo* data of Gardner *et al*. (2003) and Cantone *et al*. (2009). Computation results show that parametric estimation accuracy can be significantly increased and false positive errors can be greatly reduced.

[1]It has been recently argued in Clauset *et al*. (2009) that distributions such as log-normal and stretched exponential ones may also be appropriate in describing the sparseness of protein interaction networks, and it may be more appropriate to use other distributions to characterize the structure of a metabolic network. Our attentions, however, are focused on the power law distribution. The reasons are that only two parameters are required to describe this distribution, and the so-called parsimonious principle is widely adopted in system identification (Ljung, 1999). On the other hand, the suggested algorithms can be easily modified to situations in which the number of direct regulations obeys other stochastic distributions.

2 METHODS

2.1 A general description of the identification procedure

In this article, causal relations are inferred from steady-state concentration changes of chemical species in a cellular network. This strategy is based on the total differential of a nonlinear function which has also been used in Kholodenko *et al*. (2002) and Andrec *et al*. (2005). A brief derivation is given in Appendix A of the Supplementary Material for relations among measurements of species concentrations and direct regulations, which extends the results of Kholodenko *et al*. (2002) and Andrec *et al*. (2005) from single perturbation to multiple perturbations. Assume that there are n chemical elements in a cellular network and in the j-th experiment, some external perturbations are added on a chemical element which do not directly change the concentration of this species, then, the following relation can be established approximately for a cellular network when the dynamics of its species concentrations can be described by a set of ordinary nonlinear differential equations,

$$\sum_{k=1,k\neq i}^{n} r_{ik}R_{kj} \approx R_{ij}, \qquad j=1,2,\ldots,m \tag{1}$$

Here, r_{ik} stands for the direct influences of the k-th chemical element on the i-th chemical element around the perturbed equilibrium, while $R_{kj}|_{k=1}^{n}$ can be obtained from measured species concentrations. More specifically, a positive/negative r_{ik} means that there is an activation/repression effect from the k-th chemical element to the i-th chemical element. If $r_{ik}=0$, then, it is regarded that there are no direct influences from the k-th chemical element to the i-th chemical element.

Assume that m experiments have been performed. Denote vectors $[r_{i1}\, r_{i2} \ldots r_{i,i-1}\, r_{i,i+1} \ldots r_{in}]^T$ and $[R_{1j}\, R_{2j} \ldots R_{i-1,j}\, R_{i+1,j} \ldots R_{nj}]$, respectively by x and R_j. Moreover, define matrix A and vector b, respectively, as $A=col(R_j|_{j=1}^{m})$ and $b=col(R_{ij}|_{j=1}^{m})$, in which $col(\cdot)$ represents stacking matrices/vectors in the brackets from top to bottom. Then, the relation of Equation (1) can be compactly expressed as

$$Ax \approx b \tag{2}$$

The problem discussed in this article is to identify vector x under the condition that both matrix A and vector b are provided. A distinctive characteristic of this problem is that measurement errors exist in both matrix A and vector b. On the other hand, when n is large, it is now well known that the distribution of the number of non-zero elements of vector x obeys approximately the so-called power law. More precisely, let n_i represent the number of chemical elements that have direct influences on a randomly chosen chemical element in a cellular network, and $\mathbf{Pr}\{\cdot\}$ the probability of the occurrence of a random event. Then, there exist a positive number γ and a positive integer $k_{\min}$, such that[2]

$$\mathbf{Pr}\{n_i=k\}=\begin{cases} ck_{\min}^{-\gamma} & 1\leq k\leq k_{\min} \\ ck^{-\gamma} & k_{\min}<k\leq n \end{cases} \tag{3}$$

[2]It is worthwhile to point out that in structural analysis for a large-scale network, investigations are usually focused on large n_i's (Clauset *et al*., 2009). As there is generally no statistical information about the distribution of small n_i's, it is reasonable to assume that all of them have an equal probability to occur. This treatment makes the methods suggested in this article also applicable to identification of small- or moderate-sized networks, as well as consistent with the method proposed in Gardner *et al*. (2003), which restricts the number of non-zero direct regulations.

in which $c=\left[k_{\min}^{1-\gamma}+\sum_{k=k_{\min}+1}^{n}k^{-\gamma}\right]^{-1}$. The purposes of this article are to incorporate the above structural information into cellular network identification and to investigate its capabilities in estimation accuracy improvements.

2.2 Identification algorithm

As a first step toward incorporating the so-called power law into cellular network identification, the following two assumptions are adopted.

- Represent measurement errors of matrix A and vector b, respectively, by ε_A and ε_b. It is assumed that elements of $[\varepsilon_A\ \varepsilon_b]$ are independent of each other and have an identical normal distribution $\mathbf{N}(0,\sigma^2)$ with a known σ.
- Parameters $k_{\min}$ and γ are known for the power law.

When these pieces of information are available, a natural approach for causal regulation identification from experiment data is likelihood maximization. More specifically, let $\#(\cdot)$ represent the number of non-zero elements of a matrix or vector. Then, the likelihood function of measurement errors and the direct regulation number, denote it by $L(\varepsilon_A,\varepsilon_b,k)$, can be written as follows.

$$L(\varepsilon_A,\varepsilon_b,k)=(2\pi\sigma^2)^{-mn/2}\mathbf{Pr}\{n_i=k\}e^{-\frac{\mathbf{tr}\{[\varepsilon_A\ \varepsilon_b]^T[\varepsilon_A\ \varepsilon_b]\}}{2\sigma^2}} \tag{4}$$

$$\text{subject to: } (A-\varepsilon_A)x=b-\varepsilon_b \text{ and } \#(x)=k \tag{5}$$

in which $\mathbf{tr}(\cdot)$ denotes the trace of a square matrix.

Note that m, n, $k_{\min}$, γ and σ are assumed to be known. Recalling that $\mathbf{ln}(\cdot)$ is a monotonically increasing function over $(0,\infty)$, it is obvious that the above maximization problem is equivalent to minimizing a cost function $l(\varepsilon_A,\varepsilon_b,k)$ under the conditions of Equation (5). Here,

$$l(\varepsilon_A,\varepsilon_b,k)=\frac{\mathbf{tr}\{[\varepsilon_A\ \varepsilon_b]^T[\varepsilon_A\ \varepsilon_b]\}}{2\sigma^2}+\begin{cases}\gamma\mathbf{ln}(k_{\min}) & 1\le k\le k_{\min}\\ \gamma\mathbf{ln}(k) & k_{\min}<k\le n\end{cases} \tag{6}$$

While the above cost function is physically significant, both discrete and continuous variables are included in its optimization. Currently, it appears difficult to derive an analytic expression for the optimal k, ε_A and ε_b, as well as to develop a globally/locally convergent optimization algorithm. In order to obtain an estimate about a cellular network, the following three major steps are adopted in this article.

- For a fixed k, angle minimization between two subspaces is utilized to determine positions of the non-zero elements of vector x.
- Under the condition that positions of non-zero elements are prescribed, vector x is estimated through minimizing the first term of the cost function $l(\varepsilon_A,\varepsilon_b,k)$ with respect to measurement errors ε_A and ε_b.
- The number of non-zero elements of vector x is obtained through a numerical search which minimizes an upper bound of the cost function $l(\varepsilon_A,\varepsilon_b,k)$.

These three steps are investigated, respectively, in the following subsections.

2.2.1 Position determination for direct regulations When the number of non-zero elements of vector x, denote it by k, is given, there are in principle C_{n-1}^k possibilities to locate these non-zero elements. Here, C_{n-1}^k denotes the combinatorial number of selecting k elements from the set $\{1,2,\ldots,n-1\}$. Note that for a fixed k, C_{n-1}^k increases exponentially with the increment of n. This implies that when a large-scale system is under investigation, in other words, when n is large, it is usually computationally intractable to consider all these combinations in order to find the optimal locations of the non-zero elements. As a matter of fact, according to our experiences, computation time is currently prohibitive if four direct regulations should be searched for a chemical element within a network having more than 100 species. On the other hand, it is clear from Equation (5) that when there are only k non-zero elements in vector x, then, only k columns of matrix A are used to fit the experiment data of vector b. Denote the i-th column of matrix A from the left by a_i, the i-th element of vector x from the ceiling by x_i, $i=1,2,\ldots,n-1$. Moreover, assume that the j_α-th element of vector x has been determined to be non-zero, $\alpha=1,2,\ldots,k$. Define matrix $\tilde{A}$ and vector $\tilde{x}$, respectively, as $\tilde{A}=[a_{j_1}\ a_{j_2}\ldots a_{j_k}]$ and $\tilde{x}=[x_{j_1}\ x_{j_2}\ldots x_{j_k}]^T$. Then, the first constraint of Equation (5) can be rewritten as

$$[\tilde{A}-\varepsilon_{\tilde{A}}]\tilde{x}=b-\varepsilon_b \tag{7}$$

in which $\varepsilon_{\tilde{A}}=[\varepsilon_{a_{j_1}}\ \varepsilon_{a_{j_2}}\ldots\varepsilon_{a_{j_k}}]$, and ε_{a_j} represents measurement errors of vector a_j, $j=1,2,\ldots,n-1$.

Note that under this situation, there are no longer restrictions on ε_{a_j} whenever $j\neq j_\alpha$, $\forall\alpha=1,2,\ldots,k$. According to the definition of the cost function $l(\varepsilon_A,\varepsilon_b,k)$, the corresponding optimal ε_{a_j} is the zero vector. It can therefore be declared that

$$\begin{aligned}&\min_{s.t.\,(A-\varepsilon_A)x=b-\varepsilon_b \text{ and } \#(x)=k}\mathbf{tr}\{[\varepsilon_A\ \varepsilon_b]^T[\varepsilon_A\ \varepsilon_b]\}\\ &=\min_{s.t.\,(\tilde{A}-\varepsilon_{\tilde{A}})\tilde{x}=b-\varepsilon_b}\mathbf{tr}\{[\varepsilon_{\tilde{A}}\ \varepsilon_b]^T[\varepsilon_{\tilde{A}}\ \varepsilon_b]\}\end{aligned} \tag{8}$$

The minimization problem on the right-hand side of the above equation is widely known as the TLS, and has been well settled. In fact, the following results have been established for a long time (Gloub and Van Loan, 1989; Van Huffel and Vandewalle, 1991).

$$\min_{s.t.\,(\tilde{A}-\varepsilon_{\tilde{A}})\tilde{x}=b-\varepsilon_b}\mathbf{tr}\{[\varepsilon_{\tilde{A}}\ \varepsilon_b]^T[\varepsilon_{\tilde{A}}\ \varepsilon_b]\}=\underline{\sigma}^2([\tilde{A}\ b]) \tag{9}$$

in which $\underline{\sigma}(\cdot)$ stands for the minimal singular value of a matrix. Note that every singular value of a matrix is non-negative. Therefore, the optimal non-zero element position determination problem can be mathematically expressed as

$$\min_{j_\alpha\in\{1,2,\ldots,n-1\},\alpha=1,2,\ldots,k}\underline{\sigma}([\tilde{A}\ b]) \tag{10}$$

While the above results are elegant, they cannot be directly applied to the optimal position determination of the non-zero elements of vector x, as the corresponding minimization problem is still a combinatorial optimization one for which it is generally hard to find a globally or locally convergent algorithm with polynomial computational complexities. On the other hand, let range(b) and range($\tilde{A}$) denote, respectively, the subspaces spanned by vector b and vectors of matrix $\tilde{A}$. Then, it can be proved that there is an upper bound of $\underline{\sigma}([\tilde{A}\ b])$ that is proportional to the sine of the half of the angle between range($\tilde{A}$) and range(b). More precisely, we have the next theorem, while its proof is deferred to Appendix B of the Supplementary Material.

THEOREM 1. *Let $\tilde{\theta}_k$ represent the angle between range($\tilde{A}$) and range(b), $||b||_2$ the Euclidean norm of vector b. Then,*

$$\underline{\sigma}([\tilde{A}\ b])\le 2||b||_2\sin\frac{\tilde{\theta}_k}{2} \tag{11}$$

In addition, define matrix $\hat{A}$ and vector $\hat{b}$, respectively, as $\hat{A}=A(A^TA)^{-1/2}$ and $\hat{b}=b(b^Tb)^{-1/2}$, and denote the j-th column of matrix $\hat{A}$ from the left by $\hat{a}_j$, $j=1,2,\ldots,n-1$. Then, we have the following results about subspace angle minimization, while their proof is given in Appendix C of the Supplementary Material.

THEOREM 2. *Let θ_k represent the angle between the subspace spanned by vectors $\hat{a}_{j_\alpha}|_{\alpha=1}^k$ and the subspace spanned by $\hat{b}$. Moreover, let $j_\alpha^{[opt]}$, $\alpha=1,2,\cdots,k$, stand for the positions of the elements of vector $\hat{A}^Tb$ with the first k biggest magnitudes. Then,*

$$\{j_\alpha^{[\text{opt}]}|_{\alpha=1}^k\}=arg\min_{j_1,j_2,\ldots,j_k}\theta_k \tag{12}$$

From these two theorems, it can be seen that if the minimization of the minimal singular value of $[\tilde{A}\ b]$ is replaced by the minimization of θ_k, then, the optimal $j_\alpha|_{\alpha=1}^k$ can be expressed analytically, and therefore the problem of the exponential increment of computation complexities is avoided.

On the other hand, let ψ_i, $1 \le i \le n-1$, denote the i-th row element of $A^T b$. Then, it can be shown that

$$\sin^2 \tilde{\theta}_k \le 1 - \frac{1}{\bar{\sigma}^2(A)||b||_2^2} \sum_{\alpha=1}^{k} \psi_{j_\alpha}^2 \tag{13}$$

A derivation of this inequality is given in Appendix D of the Supplementary Material. Therefore, through selecting $j_\alpha|_{\alpha=1}^k$ that maximizes $\sum_{\alpha=1}^k \psi_{j_\alpha}^2$, an upper bound of $\tilde{\theta}_k$, and therefore an upper bound of $\underline{\sigma}([\tilde{A}\ b])$, has been minimized.

It is worthwhile to point out that although relations between θ_k and $\tilde{\theta}_k$ are still not very clear, computation experiences show that compared with maximization of $\sum_{\alpha=1}^k \psi_{j_\alpha}^2$, minimization of θ_k generally leads to better estimation performances in cellular network identifications. This is illustrated by examples in the next section. However, theoretical reasons for this phenomenon are still under investigations.

2.2.2 Estimation of regulation coefficients When locations of non-zero elements of vector x have been determined, their values can be directly obtained through singular value decomposition (Gloub and Van Loan, 1989; Van Huffel and Vandewalle, 1991). More specifically, we have the following results.

THEOREM 3. *Assume that $\underline{\sigma}(\tilde{A}) > \underline{\sigma}([\tilde{A}\ b])$. Let $v = [v_1\ v_2 \dots v_{k+1}]^T$ denote the right singular vector of matrix $[\tilde{A}\ b]$ with respect to its minimal singular value, and $\tilde{x}^{[opt]}$ the optimal vector $\tilde{x}$ corresponding to the left hand side minimization problem of Equation (9). Then, $\tilde{x}^{[opt]} = -\frac{1}{v_{k+1}}[v_1\ v_2 \dots v_k]^T$.*

When the condition $\underline{\sigma}(\tilde{A}) > \underline{\sigma}([\tilde{A}\ b])$ is not satisfied, analytical forms are still available for the optimal solution of the aforementioned minimization problem, but the expressions are more complicated. An interested reader is recommended to refer to Gloub and Van Loan (1989), and Van Huffel and Vandewalle (1991) for details.

2.2.3 Determination of the number of direct regulations In the above estimations, it is assumed that the number of direct regulations is known for a species in the cellular network. In actual applications, this is generally not the case. To have an estimate of this number, the cost function of Equation (6) should be minimized. This minimization problem, however, is not mathematically tractable currently. To overcome this difficulty, on the basis of Equation (9), an estimate of the direct regulation number is obtained through minimizing a cost function $J(k)$ which is defined as follows.

$$J(k) = \frac{\underline{\sigma}^2\left(\left[a_{j_1^{[opt]}}\ a_{j_2^{[opt]}} \dots a_{j_k^{[opt]}}\ b\right]\right)}{2\sigma^2} + \begin{cases} \gamma \mathbf{ln}(k_{\min}) & k \in [1, k_{\min}] \\ \gamma \mathbf{ln}(k) & k \in (k_{\min}, n] \end{cases} \tag{14}$$

Obviously,

$$J(k) \ge \min_{\varepsilon_A,\ \varepsilon_b} l(\varepsilon_A, \varepsilon_b, k) \quad \text{subject to Equation (5)} \tag{15}$$

Therefore, minimization of $J(k)$ has an explanation of minimizing an upper bound of the cost function $l(\varepsilon_A, \varepsilon_b, k)$, which is equivalent to maximizing a lower bound of the likelihood function $L(\varepsilon_A, \varepsilon_b, k)$.

Note that σ, $k_{\min}$ and γ are prescribed positive numbers. This means that although it appears difficult to obtain an analytic form for the optimal k, this optimum can be obtained through linear searches for which many methods are available. On the other hand, it is obvious that the first term of the above cost function is a non-negative and monotonically decreasing function of k, while the second term is a non-negative and monotonically increasing function of k. It can be expected that $J(k)$ usually only has one local minimum, which has been confirmed by extensive computation experiences.

From the above analyses, the following algorithms are suggested in this article for identifying direct effects in a large-scale cellular network.

Algorithm I for Cellular Network Inference:

(1) Initialize vector x and $J(0)$, respectively, as $x = 0$ and a large positive number, for example, $J(0) = 10^{100}$.

(2) Compute matrix $\hat{A}$ and vector $\hat{A}^T b$. Denote the i-th row element of $\hat{A}^T b$ by y_i, and assume that $|y_{j_1}| \ge |y_{j_2}| \ge \dots \ge |y_{j_{n-1}}|$.

(3) Construct matrix $\tilde{A}_k$ as $\tilde{A}_k = [a_{j_1}\ a_{j_2} \dots a_{j_k}]$. Compute the value of $J(k)$ which is defined as

$$J(k) = \frac{\underline{\sigma}^2([\tilde{A}_k\ b])}{2\sigma^2} + \begin{cases} \gamma \mathbf{ln}(k_{\min}) & 1 \le k \le k_{\min} \\ \gamma \mathbf{ln}(k) & k_{\min} < k \le n \end{cases}$$

(4) If $J(k) < J(k-1)$, replace k by $k+1$ and repeat the above Step 3. If $J(k) \ge J(k-1)$, go to the next step.

(5) Perform singular value decomposition of matrix $[\tilde{A}_{k-1}\ b]$. Denote its right singular vector associated with its minimum singular value $\underline{\sigma}([\tilde{A}_{k-1}\ b])$ by v.

(6) Let v_i represent the i-th row element of vector v. Replace the j_i-th row element of vector x by $-\frac{v_i}{v_k}$, $i = 1, 2, \dots, k-1$.

Algorithm II for Cellular Network Inference: this algorithm is completely the same as that of Algorithm I, except that in the second step, matrix $\hat{A}$ is replaced by matrix A.

3 RESULTS

To illustrate the effectiveness of the developed inference algorithms, we compare them with the TLS method in this section using five examples, which include an artificially constructed large-scale linear system, a MAPK pathway model of ordinary nonlinear differential equations, some *in silico* datasets from the DREAM initiative (Prill *et al*., 2010), and some *in vivo* datasets of Gardner *et al*. (2003) and Cantone *et al*. (2009).

Note that in predicting the behaviors of a network, not only structure of the network, but also regulation directions and strengthes among its elements, play important roles. In addition to some widely adopted specifications in network inferences, such as ROC (receiver operating characteristics) curve, PR (precision recall) curve, AUROC (area under a ROC curve), AUPR (area under a PR curve), PPV (positive predictive value), Se (sensitivity), FP (false positive) rate (Cantone *et al*., 2009; Gardner and Faith, 2005; Prill *et al*., 2010), etc., the following three specifications are also adopted in this article, which reflect the rate of false sign errors and parameter estimation accuracies, respectively.

To be more specific, consider a cellular network consisting of n chemical elements. Let x_{ij}, $i, j = 1, 2, \dots, n$, denote the actual direct effect of the j-th chemical element on the i-th chemical element, and $\hat{x}_{ij}$ its estimate. Moreover, let $\mathbf{N}$ and $\mathbf{P}$ represent the total numbers of $x_{ij} = 0$ and $x_{ij} \ne 0$, respectively. Furthermore, define FP (false positive number), FN (false negative number), FS (false sign number), LA (low accuracy number) and AA (acceptable accuracy number), respectively, as follows.

- FP: the total number of the case that $x_{ij} = 0$ but $|\hat{x}_{ij}| > \rho_a$;
- FN: the total number of the case that $x_{ij} \ne 0$ but $|\hat{x}_{ij}| \le \rho_a$;
- FS: the total number of the case that $x_{ij}\hat{x}_{ij} < 0$;
- LA: the total number of the case that $x_{ij} \ne 0$ and $\hat{x}_{ij} \ne 0$, but $\left|(\hat{x}_{ij} - x_{ij})/x_{ij}\right| > \rho_r$;
- AA: the total number of the case that $x_{ij} = 0$ and $|\hat{x}_{ij}| \le \rho_a$, as well as the case that $x_{ij} \ne 0$, $\hat{x}_{ij} \ne 0$ and $\left|(\hat{x}_{ij} - x_{ij})/x_{ij}\right| \le \rho_r$.

Here, ρ_a and ρ_r represent acceptable magnitude bounds, respectively, for absolute and relative parametric estimation errors. Their values may vary with applications.

With the above numbers, the following three specifications are defined, which are used in this article to evaluate the effectiveness of a network identification algorithm.

- FS rate (rate of false sign errors): FS/P;
- LA rate (rate of estimates with low accuracy): LA/P;
- AA rate (rate of estimates with acceptable accuracy): AA/(P+N).

Moreover, note that the sum of the numbers of false negative estimates and true positive estimates equals to P. We therefore have that $\mathrm{PPV}=\frac{\mathrm{P-FN}}{\mathrm{P-FN+FP}}$, $\mathrm{Se}=1-\frac{\mathrm{FN}}{\mathrm{P}}$ and $\mathrm{FP\,rate}=\frac{\mathrm{FP}}{\mathrm{N}}$.

3.1 Example 1: an artificial linear network

In this subsection, the suggested algorithms are tested on an artificially constructed network in which all the adopted assumptions are completely satisfied. More precisely, a network with 100 elements are utilized which obeys the power law with $k_{\min}=1$ and $\gamma=2.5$. The number of direct regulations, as well as regulation coefficients, to a network element, are randomly generated. Moreover, 500 trails have been performed to investigate the statistical properties of the estimates.

The numerical simulation settings are as follows. Denote the matrix consisting of the actual interactions among network elements by X_0. Every column of X_0 is independently generated according to the next three steps.

- At first, number of non-zero elements of a column is randomly generated according to the power law with $k_{\min}=1$ and $\gamma=2.5$. Assume it to be k.
- Locations of non-zero elements are determined by the Matlab m-file *randperm.m* for random permutations. That is, elements of the set $\{1,2,\ldots,100\}$ are at first randomly permuted, and then the first k elements are adopted as the numbers of the rows in this column with a non-zero entry. Denote them by $j_\alpha|_{\alpha=1}^{k}$.
- The entry of the j_α-th row of this column is generated independently according to a uniform distribution over $[-2,-\rho_a]\bigcup[\rho_a 2]$, $\alpha=1,2,\ldots,k$. All the other entries are assigned to be zero.

On the other hand, a $10^3\times10^2$ dimensional real matrix A_0 is randomly generated whose entries are independent and uniform samples from [1, 5]. Moreover, two $10^3\times10^2$ dimensional matrices ε_A and ε_B are generated with independent random samples of a normal distribution N(0, 4). Finally, matrices A and B are generated as $A=A_0+\varepsilon_A$ and $B=A_0X_0+\varepsilon_B$.

After the production of matrices A and B, every column of X_0 is estimated independently on the basis of the data of matrix A and that of the corresponding column of matrix B. Estimation performances are evaluated using the aforementioned specifications with $\rho_a=10^{-5}$ and $\rho_r=5\%$.

In Table 1, both the empirical expectations and the empirical standard deviations (SDs) are reported for the aforementioned specifications, on the basis of 500 independent simulation trails. In Figure 1, the empirical expectations of the corresponding ROC and PR curves are shown, together with their empirical SDs.

Table 1. Estimation Performances for a linear network with 100 species

	Algorithm I		Algorithm II		TLS	
	Mean	SD	Mean	SD	Mean	SD
FP rate	0.0184	0.0027	0.0163	0.0028	0.9999	3.8540×10^{-5}
FS rate	0.0231	0.0160	0.0206	0.0128	0.2145	0.0410
LA rate	0.5459	0.0639	0.3929	0.0496	0.9465	0.0190
AA rate	0.9666	0.0044	0.9696	0.0048	0.0010	0.0003
PPV	0.4101	0.0393	0.4050	0.0398	0.0181	0.0032
Se	0.6992	0.0802	0.6052	0.0767	1.0000	0.0000
AUROC	0.8457	0.0408	0.7977	0.0390	0.6857	0.0291
AUPR	0.6038	0.0954	0.4337	0.0875	0.0329	0.0126

*$\sigma=2$; elements of A_0 are from [1, 5].

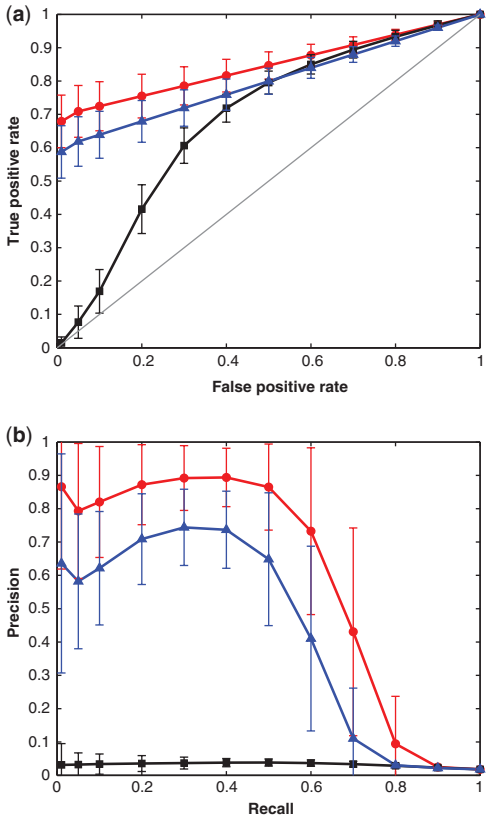

Fig. 1. Averaged ROC and PR curves of Example 1 with a fixed noise SD $\sigma=2.0000$. (**a**) Averaged ROC curve. (**b**) Averaged PR curve. Filled circle, Algorithm I; filled triangle, Algorithm II; filled square, TLS.

To investigate influences of noise strengthes on estimation performances, the above simulations have been repeated with different noise level σ. In actual computations, the value of σ is taken from the set of {0.01, 0.10, 0.22, 0.50, 0.71, 1.00, 1.25, 1.50, 1.55, 1.60, 1.73, 1.90, 2.00, 3.00, 3.50, 4.00, 5.00, 6.00}. Figure 2 shows variations of the empirical expectations of PPV and FP rate of the estimates, together with their SDs. Variations of other specifications, such as AUROC, AUPR, etc., are provided in Appendix E of the Supplementary Material.

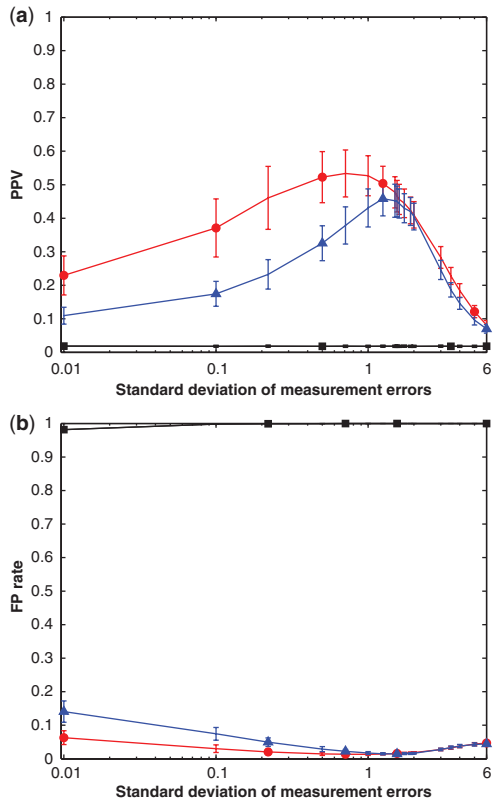

Fig. 2. Variations of the averaged PPVs and FP rates of Example 1 with respect to noise level. (**a**) Averaged PPV. (**b**) Averaged FP rate. Filled circle, Algorithm I; filled triangle, Algorithm II; filled square, TLS.

The above numerical experiments have been repeated when entries of matrix A_0 take independent and uniform random samples from $[-5,-1]\bigcup[1,5]$. The corresponding results are given in the aforementioned appendix of the Supplementary Material.

From these results, it is obvious that the suggested methods have distinguished advantages over TLS estimations in both false positive errors and parametric estimation accuracy. On the other hand, recall that in a TLS estimate, an entry very rarely takes a value near 0. It is not surprising to see that with respect to the rate of false negative errors, which is equal to $1-\text{Se}$ by definitions, the TLS method has an almost perfect performance. Moreover, compared with the rate of false positive errors of the TLS estimate, the rate of false negative errors of the suggested methods takes a much smaller value. Furthermore, although compared with Algorithm I, Algorithm II performs better sometimes in reducing both the rate of false sign errors and the rate of low accuracy parametric estimates, the differences are not very significant. In addition, when elements of A_0 take values from $[1,5]$, Algorithms I always outperforms Algorithm II with the specifications Sensitivity, AUROC and AUPR. However, when the set $[1,5]$ is replaced by $[-5,-1]\bigcup[1,5]$, Algorithm II is slightly superior to Algorithm I in general.

It is worthwhile to point out that if only AUROC and AUPR are adopted in performance assessment, then, when measurement errors are not very significant, the above simulations show that the TLS method performs better than both Algorithms I and II. This is not very surprising, as in the extreme situation, that is, when $\sigma=0$, Equation (2) is replaced by $Ax=b$ for a linear network, but there is no guarantee that minimization of θ_k or $\tilde{\theta}_k$ always leads to actual positions of non-zero elements of vector x.

3.2 A nonlinear MAPK pathway model

In actual applications, it can rarely be expected that assumptions adopted in algorithm developments are perfectly satisfied. In this subsection, the suggested estimation algorithms are tested on a MAPK pathway model, which consists of 103 chemical elements and is described by a set of first-order ordinary nonlinear differential equations. These equations take completely the same form as that of Equation (S1) of the Supplementary Material. This model is originally built in Schoeberl *et al*. (2002) and capable of explaining many biological observations. The structure of the model is given in Figure S4 of Appendix F of the Supplementary Material.

Causal interactions among chemical elements of this network are identified around one of its steady states or stable equilibria. To apply the suggested algorithms, parameters for the power law are required. To obtain this information and evaluate estimation accuracies, the Jacobian matrix of the nonlinear function vector $[f_i(x_k|_{k=1}^{103},p_k|_{k=1}^{247})]|_{i=1}^{103}$ is at first computed at the selected stable equilibrium $x^{[s]}$, which is further used to calculate the actual interactions among chemical elements using Equation (S3) in the Supplementary Material. Based on these information, parameters of the power law are estimated through counting the number of non-zero r_{ij} with a fixed i, $i,j=1,2,\ldots,103$; and fitting the logarithm of the corresponding empirical probabilities. Using this method, $\hat{\gamma}=0.8000$ and $\hat{k}_{\min}=1$ are obtained. On the other hand, using the likelihood maximization-based method of Clauset *et al*. (2009), $\hat{\gamma}=1.4718$ and $\hat{k}_{\min}=1$ are obtained. Figure S5 of Appendix F of the Supplementary Material shows the calculated empirical probabilities and the estimated power laws.

It is worthwhile to mention that although there are 103 chemical elements in the MAPK pathway model, some of the species do not have direct influences on their own concentration change speeds, that is, for these species, $\frac{\partial f_i(x,p)}{\partial x_i}\Big|_{x=x^{[s]}}=0$. In this case, the interference parameter r_{ij} of Equation (S3) in the Supplementary Material is not well defined, and therefore the suggested methods, as well as the TLS method, cannot be applied. On the other hand, numerical simulations show that for some species with self-concentration influences, the first-order Taylor series approximation of Equation (S6) in the Supplementary Material is not valid, and therefore invalidates applications of all linearization-based estimation methods. This happens when $O(||\delta_x^{[s]}||_2^2,||\delta_p||_2^2)$ is of the same order of or greater than the magnitude of $\frac{\partial f_i(x,p)}{\partial x_i}\Big|_{x=x^{[s]}}\delta_{x_i}^{[s]}$. In fact, using the definition of r_{ij}, Equation (S9) in the Supplementary Material can be rewritten as

$$\sum_{k=1,k\neq i}^{n} r_{ik}R_{kl}=R_{il}\left[1+\kappa_i(x^{[s]},\delta_x^{[s]},\delta_p)\right],\qquad l=1,2,\ldots,m \quad (16)$$

in which $\kappa_i(x^{[s]},\delta_x^{[s]},\delta_p)=O(||\delta_x^{[s]}||_2^2,||\delta_p||_2^2)/\left(\frac{\partial f_i(x,p)}{\partial x_i}\Big|_{x=x^{[s]}}\delta_{x_i}^{[s]}\right)$. Simulations show that at the selected stable equilibrium, for some species, $\kappa_i(x^{[s]},\delta_x^{[s]},\delta_p)$ has a magnitude >0.5; and in the worst case, this magnitude is very close to 10^8. Under these situations, the approximation of Equation (S9) in the Supplementary Material is no longer valid.

Table 2. Estimation performances for a MAPK pathway model

	Algorithm I		Algorithm II		
	$\hat{\gamma}=0.8000$	$\hat{\gamma}=1.4718$	$\hat{\gamma}=0.8000$	$\hat{\gamma}=1.4718$	TLS
FP rate	0.1949	0.1748	0.6275	0.5913	0.9986
FS rate	0.1005	0.1096	0.2146	0.1963	0.3105
LA rate	0.2785	0.2785	0.4703	0.4475	0.6621
AA rate	0.7806	0.7995	0.3692	0.4033	0.0207
PPV	0.1706	0.1865	0.0701	0.0725	0.0575
Se	0.6575	0.6575	0.7763	0.7580	1.0000
AUROC	0.7765	0.7807	0.7012	0.6982	0.7745
AUPR	0.4804	0.4806	0.1468	0.1443	0.2143

$N=3592$, $P=219$.

Based on these considerations, 37 species are finally chosen whose steady-state concentration change can be appropriately approximated by the first-order Taylor series and whose concentration change speed can be directly influenced by its own concentration.

In data generations, kinetic parameters $p_j|_{j=1}^{247}$ and initial values of $x_k|_{k=1}^{103}$ are changed in a way similar to that of Andrec *et al*. (2005) and Kholodenko *et al*. (2002). That is, when direct influences on the i-th species are to be estimated, only the values of these p_j, $j\in 1,2,\ldots,247$, are permitted to be changed which do not explicitly alter the value of the nonlinear function $f_i(x,p)$. More specifically, an appropriate p_j is selected together with 8∼12 x_ks that are, respectively, changed to $0.9999\times\alpha_j p_j$ for all the simulated time and $0.9999\times\beta_k x_k$ at the initial time. Here, both α_j and β_k are independent and uniform random samples from [0.9, 1]. Steady-state concentration of every species in the network is calculated before and after a perturbation using the toolbox Simulink of the commercial software MATLAB. To every calculated relative concentration change at the steady states, that is, $\delta_{x_i}^{[s]}/x_i^{[s]}$, a random number is added which is independently generated according to the normal distribution with mean 0 and SD 10^{-5}.

Estimation performances with $\rho_a=10^{-8}$ and $\rho_r=5\%$ are shown in Table 2 when experiment data length takes an integer between 142 and 146, while Figure 3 shows the corresponding ROC and PR curves. The actual data length varies slightly with the number of direct regulations from external perturbations on the chemical element under investigation. From this table, it is clear that the TLS method is superior to the methods of this article in false negative errors. On the other hand, in both false positive errors and parameter estimation accuracy, the methods of this article have prominent advantages. Moreover, in either false sign errors or estimates with large relative errors, both Algorithms I and II are also better than the TLS method. Furthermore, except Sensitivity, Algorithm I outperforms Algorithm II with respect to all the other adopted specifications. These results are consistent with those obtained in the previous linear network simulation example. However, when AUROC and AUPR are considered, the TLS method has a better performance than Algorithm II.

It is also clear from these estimation results that although curve fitting and likelihood maximization lead to very different estimates for parameters of the power law, these differences do not lead to significant changes in network estimation performances, no matter whether Algorithm I or II is used. This may imply that these two estimation procedures are robust against estimation errors of these parameters. In fact, several other artificial $\hat{k}_{\min}$ and $\hat{\gamma}$ have also been utilized, and consistent observations have been obtained.

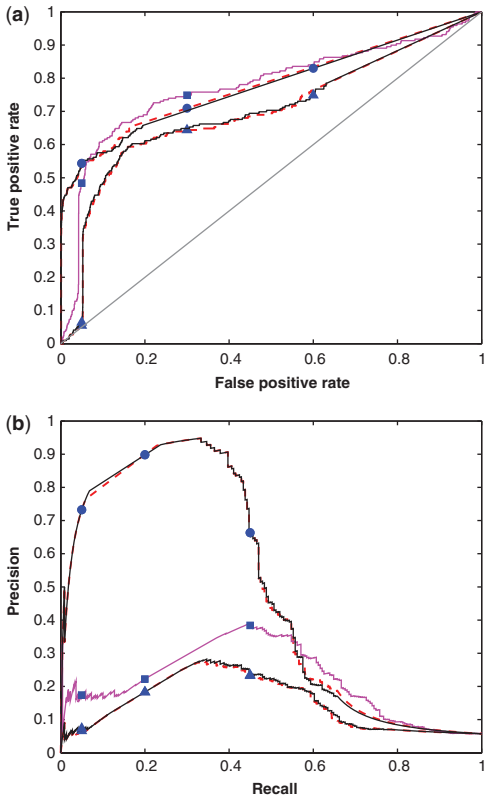

Fig. 3. ROC and PR curves in the MAPK network inference. (**a**) ROC curve. (**b**) PR curve. Filled circle, Algorithm I; filled triangle, Algorithm II; filled square, TLS. (Continuous line: $\hat{\gamma}=0.8000$ and $\hat{k}_{\min}=1$; dashed line: $\hat{\gamma}=1.4718$ and $\hat{k}_{\min}=1$.)

In addition, a detailed analysis of the false negative errors of Algorithm I with the curve fitting-based parametric estimates of the power law, shows that there are totally 75 direct regulation coefficients that have been incorrectly estimated as zero. With respect to these regulations, the results of the corresponding TLS estimates are as follows.

- Forty-one of them have a sign different from its actual value.
- In the remaining thirty-four estimates, only one of them has a relative error belonging $[-5\%, 5\%]$. As a matter of fact, most of them have a relative error with a magnitude $>100\%$.

Consistent phenomena have also been observed for the false negative estimates of Algorithm I with the likelihood maximization based estimates of the power law parameters, as well as those of Algorithm II. These results are summarized in Table S2 of the Supplementary Material.

To investigate convergence properties of the suggested algorithms, the species numbered 11 in Schoeberl *et al*. (2002), that is, (EGF-EGFRi)2, is selected for which its direct regulation coefficients from other species are estimated with the data length m varying from $m=20$ to $m=10^4$. In these investigations,

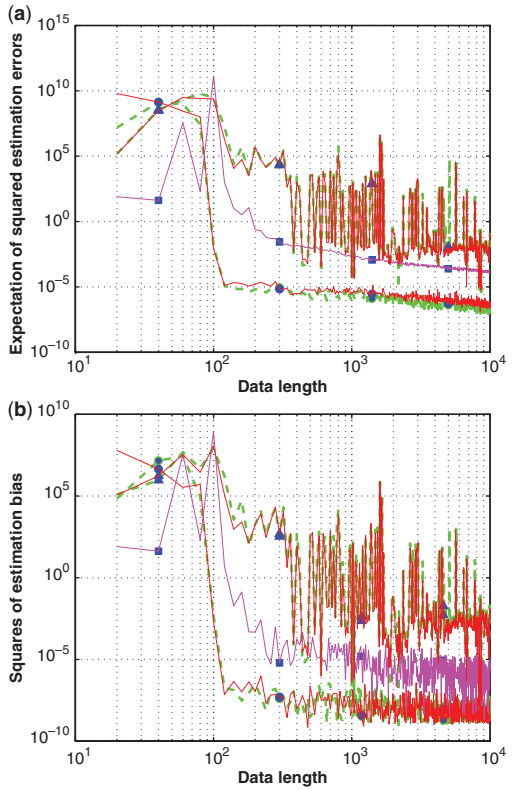

Fig. 4. Variations of parametric estimation error and estimation bias with respect to data length increment. (**a**) Expectation of the squares of estimation errors. (**b**) Squares of estimation bias. Filled square, Algorithm I; filled triangle, Algorithm II; filled square, TLS. (continuous line: $\hat{\gamma}=0.8000$ and $\hat{k}_{\min}=1$; dashed line: $\hat{\gamma}=1.4718$ and $\hat{k}_{\min}=1$.)

500 samples are taken for the data length m which equally distributes over $[20, 10^4]$. Moreover, for every prescribed data length m, 100 simulations are performed to calculate empirical mean square errors and empirical bias of the estimates, which are widely adopted in assessing performances of an algorithm in system identification and state estimations (Akutsu *et al.*, 1999; Gardner and Faith, 2005; Ljung, 1999). Definitions of these two specifications are given in Appendix F of the Supplementary Material.

Calculated values of these two specifications are, respectively, shown in Figure 4a and b.[3] From this figure, it is clear that for both of the TLS method and the methods proposed in this article, estimation error and estimation bias converge to zero. But if convergence rate should also be considered, which is again an important index in performance assessment of estimation algorithms, then, it can be declared that Algorithm I has obvious advantages. But performances of Algorithm II are poor.

As false positive error reduction is very essential in cellular network estimations, errors are also computed for the estimated number of direct regulations. In Figure S6 of Appendix F of the Supplementary Material, these errors are shown for every simulation with Algorithms I and II when the data length m increases from 20 to 10^4. Clearly, for Algorithm I, this error also monotonically decreases in magnitude with the increment of data length. But for Algorithm II, the magnitude of this error does not decrease appreciably even if the data length becomes very large. This implies that when Algorithm I is adopted, false positive errors can also be asymptotically reduced. This is significantly different from the TLS method, for which zero elements can rarely be obtained in a corresponding estimate.

[3]Note that when the data length is <102, Equation (2) is under-determined. This implies that results for the TLS estimate in these figures are valid *only* when the data length is ≥ 102. On the other hand, it can be simply proved that when Algorithm I or II is adopted, the optimal estimate for the number of direct regulations never exceeds the number of rows of matrix $\tilde{A}_k$. This means that the aforementioned under-determinedness problem happens neither to Algorithm I nor II.

3.3 Application to other datasets

To evaluate performances of the developed algorithms, they have also been applied to the *in silico* steady-state dataset of the Size 100 subchallenges of DREAM3 and DREAM4 in the DREAM initiative (Prill *et al.*, 2010). The results are given in Appendix G of the Supplementary Material.

From these results and those available on the DREAM project web site, it may be concluded that, Algorithm I/Algorithm II/TLS would have been placed 4th/5th/12th in the DREAM3 subchallenge among the 22 participated teams with a score of 32.0813/14.1991/2.7557, and 16th/17th/18th in the DREAM4 subchallenge among the 19 participated teams with a score of 9.4248/5.8975/0.8830. However, it should be emphasized that these results are not directly comparable because the reported participants were completely blind to the structures and dynamics of the networks.

Note that the top-5 teams of DREAM3 utilized both steady-state data and time series data. Moreover, it is clear now that these two kinds of data provide complementary information about network structures (Prill *et al.*, 2010). It appears safe to declare that Algorithm I works well with the DREAM3 Size 100 sub-challenges. In addition, our estimation procedures have significantly lower computational complexities. To be more specific, for the best performer of these subchallenges, it is reported that approximately 78 h have been consumed to obtain an estimate with a high-end cluster. But for both Algorithms I and II, <2 s were required even if a personal computer is utilized, which is equipped with 2.00 GB RAMs and two 2.40 GHz Intel(R) Core(TM) Quad CPUs.

However, performances of both Algorithms I and II are still very poor for the DREAM4 Size 100 subchallenges. This may possibly be due to that the adopted assumptions on noises ε_A and ε_b have been seriously deteriorated.

On the other hand, except Sensitivity, both Algorithms I and II outperform the TLS method in all these subchallenges. Moreover, except in some specifications such as FS rate and FP rate, etc., Algorithm II is slightly superior to Algorithm I, Algorithm I outperforms Algorithm II in almost all the other adopted specifications. In addition, when ROC curve is utilized to assess estimation performances, both Algorithms I and II perform better than random predictions in all the attacked subchallenges, but in most of the DREAM4 subchallenges, performances of the TLS method are very close to random predictions. In the extreme case, that is, for the network Net1, the TLS method performs even worse than random predictions.

To investigate the flexibility of the suggested methods in cellular network identifications, they have also been applied to the *in vivo* data of the SOS test network (Gardner *et al.*, 2003) and that of the IRMA (*in vivo* 'benchmarking' of reverse-engineering and modeling

approaches) network (Cantone *et al*., 2009). The SOS test network and the IRMA network have, respectively, only 9 and 5 genes. Generally, the power law cannot be applied to networks with a size of this order. But when the distribution of Equation (3) is utilized, we found that it is still efficient in improving performances of TLS estimates. The details of the estimation results are provided in Appendix H of the Supplementary Material. In these cases, as the size of the involved network is not very large, it is feasible to consider all the possible positions of non-zero elements of vector x when the number of direct regulations is prescribed. To see the conservativeness of the upper bound of Equation (14), the corresponding estimation results have also been included in the Supplementary Material.

The results of Appendix H of the Supplementary Material show that performances of the suggested methods are close but slightly better than the reported estimates. On the other hand, combinatorial optimization-based estimate does not provide significant estimation performance improvements, although its computation time is almost 10 times longer than that of both Algorithms I and II.

In the above estimations, only curve fitting-based estimates for the power law parameters are utilized. But it has also been found that final estimation performances do not change appreciably even if the likelihood maximization-based estimates for these parameters are utilized.

4 CONCLUSION

In this article, possibilities have been investigated for incorporating the so-called power law into identification of direct causal regulations in a large-scale cellular network. Under the assumption that measurement errors are independently and normally distributed, a likelihood maximization approach is suggested. On the basis of relations between subspace angles and matrix singular values, two procedures are suggested to estimate these interactions.

These methods are applied to an artificially constructed linear large-scale network, a MAPK pathway model, some recently developed performance assessment networks and the *in vivo* data set of a nine-gene network and a five-gene network. Comparisons with the widely adopted TLS method show that the suggested methods have distinguished advantages on both reduction of false positive errors and improvement of parametric estimation accuracy. Moreover, one of the supposed algorithms has a much faster convergence speed when either estimation error or estimation bias is considered.

While most of the reported results are encouraging, the developed estimation procedures are still far from satisfaction of practical application requirements. This has been made very clear by the unsatisfactory performances with the sub-challenges of DREAM4. As further researches, it is interesting to investigate parameter estimation from experiment data for the power law, simultaneous identification of direct causal regulations for multiple species, appropriate statistical descriptions of noise and measurement errors, as well as extending the procedures to estimations with other kinds of experiment data, such as quasi-steady-state data, time series data with non-uniform sampling, etc.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to Profs Q.C. Zhao and S. Li, Tsinghua University, China, for various invaluable discussions.

Funding: National Natural Science Foundation of China (grant numbers 60625305, 60721003, in parts); 973 Program of China (grant number 2009CB320602, in parts).

Conflict of Interest: none declared.

REFERENCES

Akutsu,T. *et al.* (1999) Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac. Symp. Biocomput.*, **4**, 17–28.

Andrec,M. *et al.* (2005) Inference of signaling and gene regulatory networks by steady-state perturbation experiments: structure and accuracy. *J. Theor. Biol.*, **232**, 427–441.

Bansal,M. *et al.* (2006) Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, **22**, 815–822.

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

Berman,P. *et al.* (2007) Randomized approximation algorithms for set multicover problems with applications to reverse engineering of protein and gene networks. *Discrete Appl. Math.*, **155**, 733–749.

Cantone,I. *et al.* (2009) A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, **137**, 172–181.

Chang,R. *et al.* (2008) Quantitative inference by qualitative semantic knowledge mining with Bayesian model averaging. *IEEE Trans. Knowl. Data Eng.*, **20**, 1587–1599.

Clauset,A. *et al.* (2009) Power-law distributions in empirical data. *SIAM Rev.*, **51**, 661–703.

de la Fuente,A. *et al.* (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, **20**, 3565–3574.

Ferrazzi,F. *et al.* (2007) Bayesian approaches to reverse engineer cellular systems: a simulation study on nonlinear Gaussian networks. *BMC Bioinformatics*, **8** (Suppl. 5), S2.

Gardner,T.S. *et al.* (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102–105.

Gardner,T.S. and Faith,J.J. (2005) Reverse-engineering transcription control networks. *Phys. Life Rev.*, **2**, 65–88.

Gloub,G.H. and Van Loan,C.F. (1989) *Matrix Computation*, 2nd edn. The John Hopkins University Press, Baltimore.

Kholodenko,B.N. *et al.* (2002) Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proc. Natl Acad. Sci. USA*, **99**, 12841–12846.

Ljung,L. (1999) *System Identification: Theory for the User*, 2nd edn. Prentice Hall PTR, Upper Saddle River, New Jersey.

Perrin,B.E. *et al.* (2003) Gene networks inference using dynamic Bayesian networks. Bioinformatics, **19** (Suppl. 2), II138–II148.

Prill,R.J. *et al.* (2010) Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS One*, **5**, e9202.

Schoeberl,B. *et al.* (2002) Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat. Biotechnol.*, **20**, 370–375.

Shmulevich,I. *et al.* (2002) Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, **18**, 261–274.

Sontag,E. (2008) Network reconstruction based on steady-state data. *Essays Biochem.*, **45**, 161–176.

Van Huffel,S. and Vandewalle,J. (1991) *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM, Philadelphia.

Zheng,Y. and Kwoh,C.K. (2004) Reconstruction Boolean networks from noisy gene expression data. *Int. Conf. Control Autom. Robot. Vis.*, **4**, 58–72.