Linear time-varying models can reveal non-linear interactions of biomolecular regulatory networks using multiple time-series data

Jongrae Kim\textsuperscript{1,2}, Declan G. Bates\textsuperscript{2,3}, Ian Postlethwaite\textsuperscript{2,3}, Pat Heslop-Harrison\textsuperscript{2,4} and Kwang-Hyun Cho\textsuperscript{5,*}

\textsuperscript{1}Department of Aerospace Engineering, University of Glasgow, Glasgow G12 8QQ, \textsuperscript{2}Systems Biology Lab, \textsuperscript{3}Department of Engineering, \textsuperscript{4}Department of Biology, University of Leicester, Leicester, LE1 7RH, UK and \textsuperscript{5}Department of Bio and Brain Engineering and KI for the BioCentury, Korea Advanced Institute of Science and Technology (KAIST), 335 Gwahangno, Yuseong-gu, Daejeon 305-701, Republic of Korea

Received on January 16, 2008; revised on March 7, 2008; accepted on March 22, 2008
Advance Access publication March 26, 2008
Associate Editor: Martin Bishop

\textbf{ABSTRACT}

\textbf{Motivation:} Inherent non-linearities in biomolecular interactions make the identification of network interactions difficult. One of the principal problems is that all methods based on the use of linear time-invariant models will have fundamental limitations in their capability to infer certain non-linear network interactions. Another difficulty is the multiplicity of possible solutions, since, for a given dataset, there may be many different possible networks which generate the same time-series expression profiles.

\textbf{Results:} A novel algorithm for the inference of biomolecular interaction networks from temporal expression data is presented. Linear time-varying models, which can represent a much wider class of time-series data than linear time-invariant models, are employed in the algorithm. From time-series expression profiles, the model parameters are identified by solving a non-linear optimization problem. In order to systematically reduce the set of possible solutions for the optimization problem, a filtering process is performed using a phase-portrait analysis with random numerical perturbations. The proposed approach has the advantages of not requiring the system to be in a stable steady state, of using time-series profiles which have been generated by a single experiment, and of allowing non-linear network interactions to be identified. The ability of the proposed algorithm to correctly infer network interactions is illustrated by its application to three examples: a non-linear model for cAMP oscillations in Dictyostelium discoideum, the cell-cycle data for Saccharomyces cerevisiae and a large-scale non-linear model of a group of synchronized Dictyostelium cells.

\textbf{Availability:} The software used in this article is available from http://sbie.kaist.ac.kr/software
Contact: ckh@kaist.ac.kr
Supplementary information: Supplementary data are available at Bioinformatics online.

\section{INTRODUCTION}

Biomolecular regulatory networks can contain complicated non-linear interactions amongst many different molecules, where the result is not simply the sum of multiples of the inputs. Functional regulation in such systems involves multiple interactions between inputs and intermediates, forming a non-linear network. Depending on the number of different interactions involved, such networks can be modelled using stochastic equations, ordinary differential equations, or undirected/directed graphs with nodes and edges (Bornholdt, 2005). The problem of inferring biomolecular networks from time-series expression profiles so as to obtain a quantitative understanding of the underlying biomolecular mechanisms is a fundamental one for Systems Biology, and has been geared up by the ever increasing amount of high-throughput experimental data from real biomolecular networks (Cho et al., 2007).

Many different methods for inferring biomolecular networks have been proposed in the recent literature. They can roughly be divided into two main categories. The first uses perturbation experiments about a steady state to generate a Jacobian matrix which is then used to infer the network interactions (Bansal et al., 2006; Cho et al., 2005; Khodolenko et al., 2002; Schmidt et al., 2005; Sontag et al., 2004; Tegner et al., 2003). The perturbation-based approach requires several assumptions to be made about the system, such as the existence of a steady state and the possibility of conducting experiments which achieve just the right level of perturbation in the system. This means that such approaches are not suitable for application to biomolecular regulatory networks which do not have a stable steady state, such as those which display oscillatory dynamics, e.g. limit cycles. It is also well known that the size of the perturbation is critical for the success of these methods. If the perturbation is too large, the resulting kinetics are no longer near the original steady state and the inferred network may display completely different characteristics. On the other hand, if it is too small, then the resulting data may not contain sufficient information for the algorithm to infer the network interactions correctly. In addition, each additional perturbation experiment entails additional time and effort, and in some cases, multiple experiments may not even be possible.

The second category of inference methods employ probabilistic approaches, such as Bayesian methods, which uncover the network interactions using causal relations between the states. However, many such approaches can only infer network structures correctly when there are no feedback loops in the
networks, i.e. the resulting networks are directed acyclic graphs (Chen et al., 2006; Friedman, 2004). Although dynamical Bayesian methods can overcome these problems, the resulting problem formulation needs to be solved using global optimization algorithms such as greedy search, simulated annealing and genetic algorithms (Yu et al., 2004), which are computationally expensive. This problem was partially addressed in Zhao et al. (2006) where dynamical Bayesian networks and probabilistic Boolean networks are combined to reduce the parametric search space for inferring the network. However, the amount of calculation required still quickly expands as the number of nodes increases, thus dividing large networks into several relatively small networks.

Since most biomolecular networks are inherently non-linear and the average interactions of individual molecules generally behave in a continuous fashion, it would seem appropriate to formulate such problems as a non-linear system identification problem using ordinary differential equations. However, unlike the identification problem for linear time-invariant systems, non-linear system identification problems are ill-posed in general, because not only the coefficients in the system but also the system structure must be estimated. System identification using linear time-invariant models, on the other hand, simply involves finding an optimal set of coefficients for a given model, since the model structure is fixed in advance. In the literature, there are many well-developed theoretical results for the linear system identification problem (Ljung, 1999). Any model of the system derived using purely linear approaches will, however, not be able to accurately represent the true non-linear behaviour of the real biomolecular network. For example, many biomolecular regulatory networks produce robust and stable limit cycle dynamics, circadian rhythms (Goldbeter, 1995), cAMP oscillation in Dictyostelium discoideum (Laub and Loomis, 1998), Ca$^{2+}$ oscillations (Dellen et al., 2005). Such oscillatory dynamics cannot be produced by purely linear time-invariant systems, and therefore the underlying network can never be accurately identified using linear time-invariant models.

In previous work, Cho et al. (2006) proposed a simple and highly effective algorithm to infer biomolecular networks from just one set of time-series data using phase portrait analysis, where the phase portrait is a state trajectory on a plane spanned by the state variables. Because of its simplicity it can be applied to very large-scale networks, but non-linear actions still cause difficulty for this approach as it is based purely on the use of linear models. In addition, any self interactions in the network cannot be identified using this approach. In order to address the above issues, in this article we adopt linear time-varying systems as our model for inferring biomolecular networks. Although linear time-varying systems are still in a linear form, they have a much richer range of dynamic responses than linear time-invariant systems. Hence, a wider range of time-series expression profiles, including oscillatory trajectories, can be approximated by such models.

Using linear time-varying models, the network inference problem is formulated as an optimization problem. The cost, which is the sum of the error between the measurement data and the states generated by the model, is minimized using an optimization algorithm and the optimal set of coefficients for the model is obtained. To the authors’ knowledge, this is the first time that the use of linear time-varying systems for inferring biomolecular networks has been proposed. Moreover, to generate the cost function for the optimization algorithm, we use only time-series expression profiles that may be obtained by a single experiment. To reduce the possible number of solutions for the network, we propose an algorithm based on random perturbations, where the perturbations are numerical ones to the measurement data. Previous efforts to address this problem have attempted to exploit the sparsity property of biomolecular networks to reduce the number of alternative solutions. While many biomolecular networks including gene networks, protein–protein interactions, and metabolic networks are indeed sparsely connected (Gardner et al., 2005; Yeung et al., 2002), the levels of sparsity will vary from network to network, and moreover, measurement data from biological experiments will include noise whose statistical properties are not generally known. In addition, it is impossible to infer all interactions correctly using a single set of data. Therefore, robust inferring algorithms must give consistent answers for lightly perturbed measurements. In our approach, phase portraits are used to quantify the relative effect of each element in the network, and based on the results from several random perturbation the inconsistent network interactions are identified and declared as unidentifiable nodes. Inferring is then performed for the remaining consistent network connections.

Three biomolecular network systems are used as examples in the article for testing the proposed inferring algorithm. The first is an in silico example, the Laub–Loomis model for D.discoideum cAMP oscillation (Laub and Loomis, 1998). This model captures many of the most important features of several classes of biomolecular networks that oscillate with constant frequencies. The second is an in vivo example, the cell cycle of the budding yeast cell, Saccharomyces cerevisiae, for which we use the actual experimental data for inferring the cell-cycle network (Spellman et al., 1998). The third is a large size network example, a group of D.discoideum cells. Note that the biomolecular network underlying cAMP oscillations in D.discoideum cannot be inferred from any of the current perturbation-based approaches found in the literature, since, during this particular phase of its development it does not have steady states but displays limit cycle oscillations. Hence, the Jacobian, which gives the local network interactions, is not constant but periodically time varying (Kim et al., 2006; Ma and Igieas, 2002), and there is currently no method that can identify a time-varying Jacobian from time-series data. For the second example, the cell-cycle data is generated from time-series expression profiles of gene levels for which there are no experimental perturbation data provided. Hence, perturbation-based approaches are also not applicable for this system.

2 METHODS

2.1 Approximating non-linear interactions using linear time-varying models

The dynamics of most biomolecular regulatory networks arise from complex biomolecular interactions which are non-linear, and can be written as follows:

$$\frac{dx(t)}{dt} = f(x_1(t), x_2(t), \ldots, x_{n-1}(t), x_n(t))$$

for $i = 1, 2, \ldots, n - 1, n$, where $dx/dt$ is the time derivative and $f(\cdot)$ is a function that describes the dynamical interactions on $x_i(t)$ from $x_1(t), x_2(t), \ldots, x_{n-1}(t)$ and $x_n(t)$. If $f(\cdot)$ and $x_i(t)$ increase and decrease...
in a synchronous fashion, i.e. \( f(x(t)) \) increases or decreases as \( x(t) \) increases or decreases, it is said that \( x(t) \) activates \( f(x(t)) \). On the other hand, if \( f(x(t)) \) and \( x(t) \) increase and decrease in an asynchronous fashion, i.e. \( f(x(t)) \) increases or decreases as \( x(t) \) increases or decreases, it is said that \( x(t) \) inhibits \( f(x(t)) \). Consider the following \( p \)-number of experimental data points:

\[
\hat{x}(t_k) = x(t_k) + v(t_k)
\]

for \( i = 1, 2, \ldots, n - 1, n \) and \( k = 1, 2, \ldots, p - 1, p \), where the measurement \( \hat{x}(t_k) \) is corrupted by some measurements noise \( v(t_k) \) and \( t_k \) is the sampling time. Typically, in experiments the sampling interval \( t_{k+1} - t_k \) is not necessarily the same for all \( k \) and the statistical properties of the noise are also generally unknown.

The estimation of \( f(x) \), which involves fitting the time profile of the states and finding the structure of the function, is an ill-posed problem. On the other hand, if the model is assumed to be linear time invariant as follows:

\[
\frac{dx(t)}{dt} = \sum_{j=1}^{n} a_{ij} x_j(t)
\]

for \( i = 1, 2, \ldots, n - 1, n \), then the constant coefficients, \( a_{ij} \), may be estimated and the problem is well-posed. In this case, however, the linear model may not be a good fit for the experimental data, which has been generated from non-linear network interactions. A typical non-linear phenomenon that cannot be approximated by a linear time-invariant model is a limit cycle. A limit cycle can be viewed as a set of stable points, i.e. whenever the states of the system are perturbed away from the limit cycle, they immediately converge back into the limit cycle. To make the estimation problem well-posed while preserving the ability of the candidate model to closely-fit the data, we employ a linear time-varying model, which is given by

\[
\frac{dx(t)}{dt} = \sum_{j=1}^{n} a_{ij}(t) x_j(t)
\]

for \( i = 1, 2, \ldots, n - 1, n \), where \( a_{ij}(t) \) is a time-varying function. We further simplify the estimation problem by limiting the rate of change of \( a_{ij}(t) \) with time. This is reasonable, since the measurement frequency of any biological experiment is limited and therefore only information up to a certain frequency in the data can be correctly uncovered from the measurements. In this case, \( a_{ij}(t) \) can be written as a finite sum of Fourier series as follows (Proakis and Manolakis, 2006):

\[
a_{ij}(t) = a_{ij} \sin(\omega t + \phi_j) + \beta_j
\]

for \( i, j = 1, 2, \ldots, n - 1, n \), where \( a_{ij}, \omega, \phi_j, \) and \( \beta_j \) are the constants to be determined. \( \beta_j \) represents the linear part of the interactions and the sinusoidal term approximates any non-linear terms in the interactions. If needed, more sinusoidal terms can easily be included to more closely approximate the non-linearities, at the cost of increasing the computational burden for the optimization algorithm.

### 2.2 Optimization problem and reducing the problem dimension and measurement noise effects

By using the linear time-varying model, the following optimization problem can be formulated:

\[
\min_{a_{ij}, \beta_j, \omega, \phi_j, \check{x}(t_k)} J_i = \frac{1}{\max_{i} [\check{x}(t_k) \leq \check{x}(t_k) \geq \check{x}(t_k)]} \sum_{k=1}^{n} [x(t_k) - \check{x}(t_k)]^2
\]

for \( i = 1, 2, \ldots, n - 1, n \), subject to (4), where \( \check{x}(t_k) \) is the numerically perturbed measurement, which will be defined later, and \( x(t) \) is the solution of (4). For a fixed \( i \), the number of parameters to be estimated is \( 3n + 2 \) including the initial condition of (4), \( x(t_0) \). The only computational limitation on the size of problems which may be solved by our algorithm is that arising from the optimization algorithm chosen to solve (6). In this article, we show that a simple simplex search method implemented in MATLAB (Mathworks, 2007) is able to cope with even relatively large-scale problems. However, if the number of nodes is greater than several hundred, optimization algorithms which have been specifically developed for problems involving large numbers of decision variables would be more appropriate, e.g. the simultaneous perturbation algorithm proposed in Spall (2003).

Note that the cost function is normalized by the maximum value of the measurements of each state. The optimization problem is formulated separately for each \( dx_i(t)/dt \) in order to reduce the rate of increase in the number of parameters to be estimated as the dimension of \( x(t) \) increases. If the problem is formulated for all \( x_i(t) \), the number of parameters increases according to \( 3^m + n + 1 \). The price to be paid for reducing the number of parameters to be computed in this way is the increased effect of noise, since in order to solve (4) all the states except \( x_i(t) \) have to be interpolated from the measurements and this will necessarily introduce the direct effect of noise on the estimate. To reduce this effect some elements of the noise could be filtered out before the data are used in the interpolation, e.g. by using Principal Component Analysis (Sanguinetti, et al., 2005), or the noisy measurements could be replaced by the solution of the differential equation, (4). After the optimal solution is obtained for (6), if the optimal cost is smaller than a certain bound, for example 10% of the maximum of the measurements, the measurements are replaced by the solution of the differential equations, under the assumption that the model gives less noisy data without deteriorating the original measurements significantly. As the optimization problem is solved from \( x_1 \) to \( x_n \), more measurements may be replaced. At the final stage, all measurements except the measurements for \( x_n \) could be replaced by the filtered states. To remove the unbalanced noise effect, the same procedure is repeated in the opposite direction, i.e. starting from \( x_{n-1} \) to \( x_1 \) since the earlier states may be affected more by the noise. The cost bound should be a bigger number when the measurement noise is expected to be large. However, if the measurement is too noisy, then the inferring results cannot be good. On the other hand, if the measurement noise is mild, the high frequency noise can be removed by replacing the measurement by the model output. The 10% bound is the reasonable level for removing the mild noise effect and it is fixed to 0.1.

Note that (6) is a very flexible expression. It can cope with cases where the sampling time is not evenly distributed, and weighting can also be used in the cost when the error bar at each sampling time is given. The optimization problem is, however, non-linear, and hence it may have many local solutions. If biologically plausible ranges for the parameters are known, these can be used in choosing a better initial guess for the optimization, in order to improve the chances of finding the globally optimal solution. Initial values for the parameters may also be chosen by inspecting the finite difference magnitude of \( \check{x}(t_k) \), since, although the data are corrupted by noise, the rate of change of \( x(t) \) should still not be very different from the magnitude of the finite difference. The initial guess for \( \omega \) can be obtained by calculating the dominant frequency of the measurement data using Fourier transforms (Proakis and Manolakis, 2006).

### 2.3 Measuring the strength of interactions using phase portraits

The phase portrait of a system is a trajectory of states drawn on a plane. Phase portraits have long been used as a useful graphical tool in dynamical systems theory. Usually, a 2D phase portrait is drawn from two states and several dynamic properties such as the location and the stability of equilibriums, the region of attraction of equilibrium points, etc. are inspected (Khalil, 2002). The idea of inferring biomolecular interactions using phase portraits was first presented in Cho et al. (2006). In the article, the slope index of two states, \( x(t) \) and
\( x_i(t) \), was defined as the average slope of the curve drawn by \( x_i(t) \) and \( x_i(t) \) on the phase portrait. By inspecting the sign of the slope for two given time points, i.e. \( [x_i(t_{k+1}) - x_i(t_k)]/[x_i(t_{k+1}) - x_i(t_k)] \), it can be decided whether the slope is tilted to the right or left. The average value is then calculated from multiple points on the curve. Depending on the sign of the slope index, it can hence be deduced whether \( x_i(t) \) activates or inhibits \( x_i(t) \).

In this article, we adapt this approach to the case of time-varying models. The slope index is now defined using \( x_i(t) \) and \( a_i(t)x_i(t) \), since we are considering the direct effect of \( x_i(t) \) on the time derivative of \( x_i(t) \). The slope index, \( SI(\cdot, \cdot) \), is therefore defined as follows:

\[
SI[x_i(t), a_i(t)x_i(t)] := \frac{1}{m-1} \sum_{k=1}^{m-1} \text{sign} \left( \frac{a_i(t_{k+1}x_i(t_{k+1}) - a_i(t_k)x_i(t_k))}{x_i(t_{k+1}) - x_i(t_k)} \right),
\]

where \( m \) is a positive integer \( > 1 \), a bigger value of \( m \) would produce a more accurate approximation, and \( \text{sign}(\cdot) \) is the sign function which takes values of \(-1, 0, +1\), depending on the sign of the argument. By plotting \( a_i(t)x_i(t) \) with respect to \( x_i(t) \) in the phase plane, which is illustrated in Figure 1, the slope index can be seen as the slope of the dashed line passing through the phase portrait. A positive SI means that when \( x_i(t) \) increases or decreases, the time derivative of \( x_i(t) \) increases or decreases in a synchronous fashion, hence, we can say \( x_i(t) \) activates \( x_i(t) \). Correspondingly, a negative SI means that \( x_i(t) \) inhibits \( x_i(t) \).

While the slope index indicates the type of reaction, i.e. activation or inhibition, the following indicates the magnitude of the reaction, i.e. how much \( x_i(t) \) affects \( x_i(t) \):

\[
\rho_{ij} := \max_{\forall i=1} a_i(t)x_i(t) - \min_{\forall i=1} a_i(t)x_i(t)
\]

for \( i, j = 1, 2, \ldots, n \). As shown in Figure 1, for example, since \( \rho_{12} \) is significantly smaller than \( \rho_{11} \), the effect of \( x_1(t) \) on \( x_2(t) \) or \( dx_2(t)/dt \) could be negligible compared to the effect of \( x_1(t) \) on \( dx_1(t)/dt \).

\( SI(\cdot, \cdot) \) and \( \rho_{ij} \) are multiplied by each other and the result is defined as the interaction strength (IS) matrix, whose \( (i,j) \)th row and \( j \)th column element is given by

\[
IS(i,j) := \rho_{ij} \cdot SI[x_i(t), a_i(t)x_i(t)].
\]

The \((i,j)\)th element of IS is defined by the multiplication of the influence of \( x_i(t) \) on \( dx_j(t)/dt \), i.e. \( \rho_{ij} \), and the average slope. Hence, the interaction strength is strongest when both of these quantities have a large magnitude.

### 2.4 Inferring the biomolecular interactions using random perturbations

The network interactions could be inferred by inspecting the sign of the elements of the IS matrix. However, there is a fundamental difficulty in inferring biomolecular interaction networks. The chances of finding the correct combinations of \( a_i(t) \) values to represent the network are very low because of the presence of multiple solutions. Even if an algorithm shows a high success ratio for a specific example, this may simply have happened by chance and the same algorithm may show bad performance in a different example. In addition, it is clear that the set of multiple solutions cannot be reduced to one unique solution when just using one set of time-series profiles, except for some special cases. Hence, an inferring algorithm, when used with one set of data, must decide whether a specific connection can be identifiable. If it is not identifiable, it is declared as an unknown. If it is identifiable, then it is inferred further whether it is activation, inhibition or no connection. To do this, artificial noise is added into the measurements as follows:

\[
\dot{x}_i(t_k) = \dot{x}_i(t_k) + w_i(t_k)
\]

where \( w_i(t_k) \) is a Gaussian noise with zero mean and standard deviation \( \sigma \).

For several different perturbations, the optimization problem, (6), is solved. The slopes for \( dx_j(t)/dt \) by \( x_i(t) \) and \( x_j(t) \) are positive and negative, respectively. On the other hand, the slope for \( dx_i(t)/dt \) by \( x_j(t) \) is almost zero. How much each state affects \( dx_j(t)/dt \) is indicated by \( \rho_{11}, \rho_{12} \) and \( \rho_{13} \).

For \( i = 1, 2, \ldots, n \), and \( k = 1, 2, \ldots, p - 1 \), where \( w_i(t_k) \) is the zero-mean Gaussian distribution with the standard deviation (SD), \( \sigma_w \). For several different perturbations, the optimization problem, (6), is solved. The larger \( \sigma_w \), the larger the terms in the IS matrix will fluctuate. The amount of the fluctuation shows the level of robustness of the interaction to be inferred. To obtain the mean and the SD of each term in the IS matrix with artificial perturbations in the measurements, the optimization problem is solved \( N \) times with different random artificial noises.

To remove the interactions that are not easy to identify, four steps of filtering are performed to construct the interaction matrix (IM) based on the mean and the SD of the IS matrices obtained from the \( N \) optimizations with different sets of random perturbation of the measurements. That is, for each optimization process, an IS matrix is obtained. Since the measurements are perturbed randomly for each optimization process, the \( N \) IS matrices may be slightly or significantly different from each other. The higher the variation is, the less the confidence is. This filtering procedure allows us to find the connections that cannot be decided consistently for a given measurement dataset, or which are too weak for it to be said that any directional connection exists. Firstly, all connections having large SDs are declared as being unidentifiable.

\[
\text{If } \sigma[IS(i,j)] > (1 - \gamma_s) \max_{\forall i,j \in [1,2,\ldots,n]} \sigma[IS(i,j)], \quad \text{set IM}(i,j) = 0,
\]

where \( \gamma_s \) is the SD and \( \gamma_r \) is a threshold in \([0, 1]\). As \( \gamma_r \) approaches 1, more terms are declared as unidentifiable. Secondarily,

\[
\text{If } IS(i,j) < \gamma_w \max_{\forall i,j \in [1,2,\ldots,n]} IS(i,j), \quad \text{set IM}(i,j) = 0,
\]

where \( \gamma_w \) is the mean value and \( \gamma_m \) is a threshold in \([0, 1]\). As \( \gamma_m \) approaches 1, more nodes are declared as being disconnected. Thirdly,

\[
\text{If } \text{sign}[IS(i,j) + \gamma_s \sigma[IS(i,j)] \neq \text{sign}[IS(i,j) - \gamma_s \sigma[IS(i,j)]], \quad \text{set IM}(i,j) = 0,
\]

where \( \gamma_s \) is the SD and \( \gamma_r \) is a threshold in \([0, 1]\).
As $\gamma_i$ approaches to 1, more terms are declared as unidentifiable. Finally, the remaining elements of the IM are set to the sign of the corresponding elements of IS matrix and the interactions are inferred as follows:

- if IM($i, j$) is empty, it is declared as unidentifiable,
- if sign(IM($i, j$)) > 0, $x_i(t)$ activates $x_j(t)$,
- if sign(IM($i, j$)) < 0, $x_i(t)$ inhibits $x_j(t)$ and
- if IM($i, j$) is zero, there is no direct interaction, from $x_i(t)$ to $x_j(t)$.

The pseudo-code of the algorithm is given in the Supplementary Material.

3 RESULTS

The algorithm is tested using two small size examples: a Dictyostelium discoideum cAMP oscillations model and the Saccharomyces cerevisiae cell-cycle data, and one large size example: 15 synchronized D.cerevisiae cells. Dictyostelium and Saccharomyces were also used to test an inferring algorithm based on the use of epistasis analysis in Van Driesse et al. (2005) and St Onge et al. (2007). Further details about the method of epistasis analysis can be found in Huges (2005). In the following results, the optimization problem, (6), was solved using a simplex search method in the MATLAB (Mathworks, 2007).

3.1 In silico example: Dictyostelium discoideum cAMP oscillations

Dictyostelium discoideum cells signal each other by emitting stable oscillations of cAMP at the beginning of the aggregation phase of their development. The oscillations continue during chemotaxis towards higher gradients of cAMP concentration as Dictyostelium aggregate to survive. A model for the biomolecular network underlying the stable oscillations in cAMP was postulated by Laub and Loomis (1998). The corresponding non-linear differential equations are given in the Supplementary Material.

The Laub–Loomis model exhibits a stable limit cycle, i.e. whenever the states are perturbed, they immediately converge back into the limit cycle. This type of trajectory cannot be modelled using a linear time-invariant system. In fact, the Jacobian along the limit cycle is periodically time varying, and hence the estimation of the Jacobian is also a very difficult problem. Here, without any perturbation experiment, a total of 16 measurements for each state are generated with a sampling time interval of 2 min, where the measurement noise $v(t_k)$ is the zero-mean Gaussian distribution with the variance equal to the 5% of the mean value of $\{x(t_1), x(t_2), \ldots, x(t_p)\}$. Note that the statistical properties of $v(t_k)$ are unknown in general. Some measurement examples are shown in Figure S1 (Supplementary Material). Using this set of measurements, the non-linear system is assumed to be in the linear time-varying form given by (4) with the coefficients (5) and the optimization problem, (6), is then solved with the numerically perturbed measurement given by (10). The artificial perturbation $w(t_k)$ is added, which is the zero-mean Gaussian and the variance equal to the 5% of the mean value of $\{\delta x(t_1), \delta x(t_2), \ldots, \delta x(t_p)\}$. After the optimization problem is solved, the amount of effect from each $x(t)$ to $x(t)$, i.e. $\gamma_{ij}$, is identified. This procedure is repeated 10 times ($N=10$) and the inferring procedure is performed with 10-IS matrices.

The variation in the success ratios with respect to the thresholds are shown in Figure 2 after performing 100 Monte Carlo simulations with a different set of measurements for each simulation. The first column shows the success ratio between the number of nodes that are inferred as activation and the nodes that are indeed activation. The second column shows the success ratio for the inhibition and the third column shows the success ratio for no connection. Each row gives the mean, the worst and the best results for the 100 Monte Carlo simulations. The worst is the 3rd bound, which is three times the SD of the success ratios for the Monte Carlo simulations, and the corresponding threshold values guarantee that the success ratio is higher than the worst bound with the probability 99.73%. The true positive and negative success ratios from the dynamic Bayesian approaches are 22% and 76%, where activation and inhibition are not distinguished but only the existence of connections are (Cho et al., 2006). These boundaries are indicated by the solid lines in the first row. For a wide range of the threshold values, the success ratios could be higher than those generated by the Bayesian approaches.

3.2 In vivo example: Saccharomyces cerevisiae cell cycle

The cell cycle, in which a cell duplicates its contents and divides into two, is an essential activity of all life and involves some highly elaborate regulatory mechanisms. The cell cycle is divided into four phases, which are genome duplication (S-phase), nuclear division (mitosis or M-phase) and two gap phases (G1- and G2-phase) (Bähler, 2005). In this example, S.cerevisiae cell-cycle microarray data is used to infer the underlying biomolecular network interactions (Spellman, 1998). The measurements include a total of 17 gene-expression levels with 16 samples each, and the complete network interactions are shown in Figure S2 (Supplementary Material). This network is chosen because their interactions are well studied (Bähler, 2005). The samples are from a Factor-based synchronization experiment. As noted by Friedman (2004), this is a very difficult problem, due to the fact that we are attempting to infer the interaction networks solely from the gene-expression levels, without any information on protein activity levels. This is a problem because many activations and inhibitions are performed through post-transcriptional protein modifications. Such information is, however, usually unavailable from current measurement technology, as in the case of the cell-cycle data used here. This represents a fundamental limitation when biomolecular networks are to be inferred from gene-expression levels only (Friedman, 2004).

The inferring is performed for the given experimental data, where $N=10$ and $w(t_k)$ is the zero-mean Gaussian with the variance equal to 5% of the mean of $\{\delta x(t_1), \delta x(t_2), \ldots, \delta x(t_p)\}$. The inferring result with $\gamma_s = 0.6$ and $\gamma_a = 0.18$ is shown in Figure S2 (Supplementary Material) and a part of the inferring matrix is shown in Table 1. The true activation/inhibition and the true negative ratios for different threshold values are shown in Figure S3 (Supplementary Material). The figure is generated with 50 Monte Carlo simulations of the algorithm for the same measurement set. On average, the true activation ratio is above a 20% success ratio for a wide region of the threshold values. The true inhibition ratio has a smaller region where the ratio is above 20%. On the other hand, because of the sparse property of the
We presented a novel algorithm for inferring biomolecular networks, which uses linear time-varying models. The algorithm uses phase portraits to construct a systematic filtering method to reject the inconsistent elements in the network, based on numerical perturbations. The algorithm was tested on three examples: the Laub–Loomis model for *D.discoideum* cAMP oscillation, the cell cycle of budding yeast cell, *S.cerevisiae* and the group of *Dictostelium* cells. The success ratios are sensitive to two threshold values in the algorithm, which are for finding nodes that are not connected or inconsistent with the perturbation. If the first threshold is too small, most of interactions will be considered to be connected. On the other hand, if it is too large, most of them will be considered to be disconnected. Hence, ideally, the threshold for finding no connection has to be between 0.1 and 0.5 given that in general biological networks are not fully connected. The value of the threshold for rejection depends on the richness of the given network, the true negative ratio remains high for most of the threshold values.

### 3.3 In silico large size example: multiple Dictyostelium discoideum cells cAMP oscillations

To test the algorithm for large size problems, the extended model for *Dictyostelium* cAMP oscillations is used (Kim *et al.*, 2007). The model includes the effect of interactions between multiple cells (See Supplementary Material). For 15 cells, the number of states is 105 (15 × 7) and the number of edges to be inferred is 11025 (105 × 105). The algorithm calculation time for $N = 10$ is about 64 hours in Windows XP, Intel Core2 Quad CPU, 2.66 GHz and 2.75 GB of RAM and the simplex method is used for solving the optimization problem. The calculation time will be reduced significantly by using some parallel computing technique and/or randomization approach for the optimization. The inferring results is shown in Figure S4 (Supplementary Material). The true activation and the true inhibition rate are lower compared to the previous examples as we expected since the number of possible network connections are increased. However, the true negative ratio stays high since the network satisfies the sparse property, which can be applied for many biomolecular networks.

### 4 DISCUSSION

The incorrect inferring results are given in grey and the correct ones are given in bold. The matrix is also used for designing some perturbation experiments. For example, since the first and the fifth columns have the most number of inconclusive elements, we may want to perturb either Ace2 or Cln1 to maximize the possible information to be obtained.

<table>
<thead>
<tr>
<th>Destination</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ace2</td>
<td>Ø</td>
</tr>
<tr>
<td>Clb1</td>
<td>0</td>
</tr>
<tr>
<td>Clb2</td>
<td>Ø</td>
</tr>
<tr>
<td>Clb6</td>
<td>Ø</td>
</tr>
<tr>
<td>Cln1</td>
<td>0</td>
</tr>
</tbody>
</table>

Fig. 2. Success ratios of the inferring algorithm for true activation, true inhibition and true negative with respect to two threshold values, $\gamma_s$ and $\gamma_m$, are shown in three different metrics, i.e. the mean, the worst and the best performance, where the worst is given by the 3 σ bound. The results are obtained from 100 Monte Carlo simulations of the algorithms for the given measurement data. The true negative ratio is relatively less sensitive to the threshold values because of the sparse nature of the interaction network. The solid lines on the first row are the boundary of success ratio of the Bayesian algorithm for the best rate. To display the results more clearly, the boundary lines are shown in the first row rather than the third row.

Table 1. A part of the IM is shown for $\gamma_s = 0.6$ and $\gamma_m = 0.18$, where 0 denotes no connection, 1 denotes activation, −1 denotes inhibition and Ø denotes no conclusion.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/10/1286/177382 by guest on 16 December 2018
measurement data and the number of types of genes. It should be inversely proportional to the richness of the data and proportional to the number of types of genes. Since there is no clear way to quantify the richness of signals, the threshold can in practice be decided purely based on the number of types of genes, since the greater the number of types of genes, the higher is the number of possible network connections. Hence, the threshold must be larger as the number of types of genes increases. This can be clearly seen in three examples considered in this article. Dictyostelium has seven different type of molecules while Saccharomyces has 17 and the group of Dictyostelium has 105. As shown in the figures, the highest success ratios for Saccharomyces and the group of Dictyostelium are achieved with a higher value of the rejection bound than was seen for Dictyostelium.

The proposed algorithm has many practical advantages and potential extensions. First, the error bars or noise in the measurement can be weighted appropriately when the cost function is formulated. Second, the measurements do not need to be evenly spaced in time. Third, the algorithm can infer not only feedback paths but also self activations or inhibitions in networks. Fourth, the number of unknowns to be estimated by the optimization algorithm does not increase exponentially with the number of nodes in the network. Fifth, only time-series expression profiles are used and no perturbation experiments are required. Sixth, inferring results can be used to design some optimal perturbation experiments. Finally, the linear time-varying models employed by the algorithm can effectively capture many non-linear interactions, which cannot be approximated by linear time-invariant models, including limit cycle oscillations.

ACKNOWLEDGEMENTS

This study was initiated during a visit of the the first author to the K.-H.Cho’s Systems Biology Laboratory in Korea. The first author appreciates the hospitality and many helpful and interesting discussions with the lab members.

Funding: This work was carried out under UK EPSRC Research Grant GR/S61874/01 and BBSRC Research Grant BB/D015340/1. It was supported by a Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MOST) (M10503010001-07N030100112) and also supported by the Korea Ministry of Science and Technology through a Nuclear Research Grant (M20708000001-07B0800-00110) and the 21C Frontier Microbial Genomics and Application Center Program.

Conflict of Interest: none declared.

REFERENCES


