Linear matrix inequalities approach to reconstruction of biological networks

C. Cosentino, W. Curatola, F. Montefusco, M. Bansal, D. di Bernardo and F. Amato

Abstract: The general problem of reconstructing a biological interaction network from temporal evolution data is tackled via an approach based on dynamical linear systems identification theory. A novel algorithm, based on linear matrix inequalities, is devised to infer the interaction network. This approach allows to directly taking into account, within the optimisation procedure, the a priori available knowledge of the biological system. The effectiveness of the proposed algorithm is statistically validated, by means of numerical tests, demonstrating how the a priori knowledge positively affects the reconstruction performance. A further validation is performed through an in silico biological experiment, exploiting the well-assessed cell-cycle model of fission yeast developed by Novak and Tyson.

1 Introduction

Last decade has witnessed a tremendous boost in interdisciplinary research aimed at application of engineering techniques to biological problems. A considerable challenge is represented by the modelling and simulation of biological networks at the molecular level: gene regulatory, metabolic, signal transduction and protein–protein interaction networks provide an unbounded field of application for mathematicians, systems theorists and bioinformaticians. The reason for this renewed appeal has to be chased in the new goals achieved in the field of biotechnology (like cDNA microarrays and oligonucleotide chips [1, 2]) that make it possible experimental measurements of the temporal evolution of large complex biological systems.

Identifying and predicting expression profiles and interrelations between genes can yield positive outcomes both in terms of biological knowledge and disease treatment: abnormal expression of a certain gene may represent a marker for a disease (if the gene is involved in cell cycle, the disease is typically a cancer) and an accurate in silico model can provide valuable information on how to bring back the system to a normal operative condition.

The goal of our research is to devise algorithms able to identify the interaction network between genes from gene expression data. Several approaches have been exploited in the literature to tackle the challenge, comprising graph theory, Boolean networks, linear, nonlinear and piecewise-linear differential equations, Bayesian networks, stochastic equations, Petri nets and others (see [3–6] and references therein for further details). Conceivably, the most suitable techniques for reconstructing the network structure from experimental temporal data seem to be those based on dynamical models, for example, systems of differential/difference equations.

A promising approach in the field of biological modelling is that based on a particular class of hybrid systems, namely piecewise affine (PWA) systems [7, 8]. These systems benefit from the well-established techniques from linear systems theory, but are also suitable for approximating the behaviour of nonlinear systems, including the presence of multiple steady states.

An appealing feature of PWA approaches is their inherent capability to describe networks exhibiting a time-varying topology. Luscombe *et al*. [9], indeed, have shown that a correct approach to the comprehension of biological networks should not linger on statistical analysis, but rather consider the dynamical changes in the topology and understand how they are related to the different biological phases and to external stimuli.

The methods presented here belong to the family of inference methods based on dynamical linear systems identification theory. Hence, it can also be exploited in the general framework of PWA modelling.

Several authors have proposed techniques based on linear regression [10] algorithms for the identification of linear models interpolating time-series measurements [11–19].

Besides the technical differences in the details of the inference algorithms, all these approaches build upon the common assumption that biological networks exhibit loose connectivity, that is, the number of connections per node is much lower than the total number of nodes. More precisely, in the last years there have been increasing evidences that many biological networks, including metabolic [20], protein–protein interaction [21] and transcriptional networks [22], as well as many other genomic properties [23], exhibit a scale-free topology [24], that entails a power-law distribution of the connectivity degree. Such assumption can be exploited to sieve the family of models interpolating the experimental data set, looking for the one whose structure better represents the real biological network. Examples of a mathematical translation of such strategy are the minimum weight solutions to linear equations problem [11], or the maximum parsimony criterion used in the framework of graph theory [25].

doi:10.1049/iet-syb:20060054

Paper first received 28th July 2006 and in revised form 31st January 2007

C. Cosentino , W. Curatola , F. Montefusco and F. Amato are with School of Computer and Biomedical Engineering, Università degli Studi Magna Græcia di Catanzaro, Viale Europa, Campus di Germaneto, Catanzaro 88100, Italy

M. Bansal and D. di Bernardo are with Telethon Institute for Genetics and Medicine, Via P. Castellino 111, Napoli 80131, Italy

E-mail: carlo.cosentino@unicz.it

Another important observation derived from experiments is that large sets of genes exhibit similar expression profiles. This property is usually exploited in order to obtain reduced order models, by clustering many genes into a single node/state of the network.

A great effort has to be addressed towards the exploitation of preexisting biological knowledge, which can yield a significant improvement to the reconstruction methods, as will be shown in subsequent sections: Given a set of experimental measurements, there are typically many possible models and network topologies that can equivalently fit the data from a mathematical viewpoint; therefore to maximise the chances to correctly reconstruct the original network, it is essential to take into account all the information already available about its structure.

Although this has been recognised by several authors as a major goal [5, 26], to the best of our knowledge there is not yet any method based on differential/difference equations really achieving this aim. To some extent, prior biological knowledge can be incorporated in certain reconstruction techniques based on probabilistic models, like Bayesian networks [27] or probabilistic Boolean networks [28]. On the other hand, these formalisms exhibit several shortcomings, for example, Bayesian networks are based on acyclic graphs and therefore not suitable for describing feedback mechanisms, whereas Boolean models can only provide a coarse-grained reproduction of the system dynamics. An interesting extension of static probabilistic models is represented by the dynamic Bayesian networks formalism [29, 30], which allows the presence of cyclic (non-contemporaneous) interactions, at the expenses of a greater complexity, by considering a different network for each time step. It is worth recalling that the problem of learning the structure of a Bayesian network is NP-hard [31], which severely hampers practical applications of such methods to real-world large systems.

Other algorithms, instead, use the known links to posteriori evaluate different solutions and choose the optimal ones [30], albeit this approach is not equivalent to a direct optimisation over the family of model networks with a priori fixed known links, and thus it can just yield sub-optimal solutions.

In this regard, the main contribution of the present work is a technique that allows to a priori and explicitly includes preexisting knowledge about the network structure into the reverse engineering process, by defining and solving a convex constrained optimisation problem in the framework of linear matrix inequalities (LMIs) [32].

In order to evaluate the potentialities of the proposed reconstruction technique, it should be applied, at a first stage, to a model system like a simple monocellular organisms, for example, yeasts, for which there is a huge availability of experimental data and also extensive prior knowledge about transcriptional regulatory networks. To this aim, the technique proposed in this work is evaluated by applying it to reconstruct the regulatory network of the fission yeast cell cycle.

Given two nodes, they are said to be adjacent if one directly influences the expression of the other. Obviously, the network topology can create non-direct connections, for example, if A is a transcription factor for B and B inhibits the expression of C then also A can be considered, indirectly, an antagonist of C. However, the main goal is to draw a map of adjacent nodes, from which it is possible to reconstruct all the interaction pathways. Furthermore, the approach presented here aims at identifying also the type of interaction, that is, if a node has a positive (promoting) or negative (inhibiting) effect on the expression level of a given adjacent.

A statistical evaluation of the proposed method has been performed by means of randomly generated numerical systems. Moreover, the method has been validated via the well-known in silico model of fission yeast by Novak and Tyson (NT) [33, 34]. Indeed, looking at the equations of the NT model, it is possible to draw a graph of the interactions between the states/nodes of the system. At the same time, the NT model is used to generate a set of virtual experimental data that is used by the proposed algorithm to infer the interaction network. The inferred network can be compared with the network directly derived from the equations of the NT model, which thus serves as a benchmark.

2 Preliminaries and problem statement

2.1 Networks, graphs and dynamical systems

The interaction network we aim to reconstruct can be described via the formalism of directed graphs (or digraphs). Mathematically, a digraph is an ordered pair of sets $G := (V, A)$, where V is a set of vertices and A is a set of ordered pairs (called arcs) of vertices [35, p. 223]. If a weight (i.e. a real scalar) is assigned to every arc, the result is a weighted digraph. A compact way to define a weighted digraph of n nodes and m arcs is in the form of an $n \times n$ matrix, having m non-zero coefficients.

Looking at the problem from the system theorist point of view, the dynamical behaviour of a biological network can be described, at least for small excursions of the state variables from some equilibrium point, by means of linear systems, made up of ordinary differential equations in the continuous-time case, or difference equations in the discrete-time case (see [14–19] and the references therein).

Let us consider a discrete-time system, composed by a system of n difference equations, in the n scalar state variables x_i, $i = 1, \ldots, n$,

$$x(k+1) = Ax(k) \tag{1}$$

where

$$x(k) = (x_1(k), \ldots, x_n(k)^{\mathrm{T}}, \quad A \in \mathbb{R}^{n \times n} \tag{2}$$

Note that the dynamics of x_i at step $k+1$ is influenced by that of x_j at step k if $a_{ij} \neq 0$. Moreover, the type (i.e. promoting or inhibiting) and extent of this influence can be associated, respectively, to the sign and magnitude of the coefficient. In view of these considerations, it is possible to state that the dynamic matrix of a linear system can be considered as a representation of the interaction network between the entities associated to the state variables of the system. In this respect, the problem of reconstructing the network topology can be rephrased as the problem of identifying a dynamical system starting from experimental measurements. The advantage of recasting the problem in the framework of linear system identification relies on the existence of well-established and computationally appealing techniques, which can be tailored to the specific application.

2.2 LMI problem statement

Assume that h data points (number of experiments) are available, then

$$\Theta := (x(h) \ \cdots \ x(1)) = A\Omega \tag{3}$$

where

$$\Omega := (x(h-1) \ \cdots \ x(0)) \tag{4}$$

Our aim is to reconstruct the matrix $\boldsymbol{A}$ from the experimental values $x(k)$, $k = 0, \ldots, h$. This problem can be transformed mathematically into the problem of minimising the norm of $\Theta - \boldsymbol{A}\Omega$, thus we state the following problem.

Problem 1: Given the set of data $x(k)$, $k = 0, \ldots, h$, and the associated matrices Θ and Ω, find

$$\begin{aligned} &\min_{A \in \mathbb{R}^{n \times n}} \varepsilon \\ &\text{s.t.} \qquad (\Theta - \boldsymbol{A}\Omega)^{\mathrm{T}}(\Theta - \boldsymbol{A}\Omega) < \varepsilon I \end{aligned} \tag{5}$$

Note that condition (5) is quadratic in the unknown matrix variable $\boldsymbol{A}$. In order to obtain a linear optimisation problem, we turn it into the equivalent LMI

$$\begin{pmatrix} -\varepsilon I & (\Theta - \boldsymbol{A}\Omega)^{\mathrm{T}} \\ (\Theta - \boldsymbol{A}\Omega) & -I \end{pmatrix} < 0 \tag{6}$$

The equivalence between (5) and (6) is readily derived by applying the following lemma.

Lemma 1: Let $M \in \mathbb{R}^{n \times n}$ be a square symmetric matrix partitioned as

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{12}^{\mathrm{T}} & M_{22} \end{pmatrix} \tag{7}$$

assume that M_{22} is non-singular and define the Schur complement of M_{22}, $\Delta := M_{11} - M_{12}M_{22}^{-1}M_{12}^{\mathrm{T}}$. The following statements are equivalent:

(i) M is positive (negative) definite;
(ii) M_{22} and Δ are both positive (negative) definite.

Proof: Recall that M is positive (negative) definite iff

$$\forall x \in \mathbb{R}^n, \quad x^{\mathrm{T}}Mx > 0 \ (< 0)$$

moreover it can be decomposed as [36, p. 14]

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{12}^{\mathrm{T}} & M_{22} \end{pmatrix} = \begin{pmatrix} I & M_{12}M_{22}^{-1} \\ 0 & I \end{pmatrix} \times \begin{pmatrix} \Delta & 0 \\ 0 & M_{22} \end{pmatrix} \begin{pmatrix} I & M_{12}M_{22}^{-1} \\ 0 & I \end{pmatrix}^{\mathrm{T}}$$

The latter is a congruence transformation [37, p. 568], which does not modify the sign definiteness of the transformed matrix; indeed, $\forall x \in \mathbb{R}^n$ and $\forall C, P \in \mathbb{R}^{n \times n}$,

P positive (negative) definite $\Rightarrow x^{\mathrm{T}}C^{\mathrm{T}}PCx = z^{\mathrm{T}}Pz > 0 \ (< 0)$

Therefore M is positive (negative) definite iff M_{22} and Δ are both positive (negative) definite. □

Problem 1 with the inequality constraint in the form (6) is a generalised eigenvalue problem [32, p. 10], and can easily be solved through off-the-shelf efficient numerical algorithms, such as those implemented in the Matlab LMI Toolbox [38].

2.3 Constraints specification

As discussed in Section 1, it is important to devise a method to force the optimisation algorithm to take into account pre-existing knowledge about the network topology. As the sign and magnitude of a coefficient can be interpreted as the weight of the edge between two nodes of the network, the goal can be achieved by forcing some of the optimisation variables (i.e. the coefficients of A) to be zero and other ones to be non-zero and to have a predetermined sign. In order to directly include them in the optimisation problem, also these constraints have to be specified as a set of LMIs. For example, if one wants to constrain a_{ij} to be positive (negative), it can be done by adding the following inequality to the LMIs set

$$u_i^{\mathrm{T}}Au_j + u_j^{\mathrm{T}}A^{\mathrm{T}}u_i > 0 \ (< 0)$$

where $u_k := [0 \ \cdots \ 1 \ \cdots \ 0]^{\mathrm{T}}$ is a column vector of zeros, with a 1 at the kth position. A coefficient can be forced to zero, instead, by simply removing the corresponding variable from the optimisation problem and setting to zero the corresponding entry in the $\boldsymbol{A}$ matrix.

3 Inference algorithm

In the previous section, the mathematical machinery to identify a linear system from experimental data by means of LMIs has been illustrated. Moreover, a method has been given to assign the structure (partially or totally) of the system to be identified. These are the nuts and bolts of the inference algorithm that will be described in the present section.

3.1 Premise

Let us assume that, through some identification algorithm, it is possible to find a model that exhibits low estimation error with respect to the experimental data set. Is that model actually representative of the original network?

It is not possible to answer this question just on the basis of the experimental measurements, because the same data set could be fitted by many different models. Hence, the inference of the original network is possible only if some information about its topology is a priori known and such information can be exploited within the inference process.

At present, most published inference techniques for biological networks are based on the assumption that the network is sparsely connected, with a low average degree of connectivity. This assumption is implicit in the scale-free hypothesis that has been widely accepted during the last years among biologists and bioinformaticians.

This feature is also exploited in our method and it is translated in mathematical terms by assuming that the dynamical matrix $\boldsymbol{A}$ of system (1) is sparse.

However, different from other methods in literature, the proposed algorithm requires neither knowledge of the average degree of connectivity nor knowledge of the degree of distribution. This issue is very important when dealing with networks exhibiting scale-free or hierarchical topology, as recently shown by Wildenhain and Crampin [39]. They have proven, indeed, that the algorithms making use of a preassigned number of connections per node undergo severe performance degradation when applied to networks with scale-free or hierarchical topology, with respect to the results obtained with random topology networks. On the contrary, the performance is almost independent of the topology, when the number of connections is adaptively computed for each node by iteratively applying some pruning algorithm.

This is also our approach: The iterative procedure starts from a fully connected network, then the edges are subsequently pruned according to a maximum parsimony criterion. The pruning algorithm terminates when the estimation error exceeds an assigned threshold.

The following basic idea underpins the pruning algorithm: Given the identified (normalised) connectivity matrix, the edges connecting non-adjacent (in the original network) nodes have lower weights than the others. This is a straightforward mathematical translation of the reasonable assumption that indirect interactions are weaker than direct ones.

In addition to the loose connectivity assumption, our algorithm is also capable of directly exploiting information about some specific interactions that are a priori known, taking into account both the direction of the influence and its type (promoting or repressing).

3.2 Implementation

The reconstruction algorithm is structured as follows:

Step 1: A first system is identified by solving Problem 1, and adding all the known sign constraints of the form (8).

Step 2: Let a $\boldsymbol{A}_{(k)}$ be the matrix computed at the kth step; in order to compare the values of the identified coefficients, the matrix has to be normalised, so let us define the normalised matrix $\tilde{\boldsymbol{A}}_k = \{\tilde{a}_{ij}\}_{i,j=1,\ldots,n}$, where

$$\tilde{a}_{ij} = \frac{a_{ij}}{\| a_{*j} \| \| a_{i*} \|}$$

is obtained by dividing the original value by the norms of its row, a_{*j}, and column, a_{i^*}.

Step 3: The normalised matrix is analysed to choose the coefficients to be nullified at the next identification step; various rules can be adopted at this step in order to define the threshold below which a given coefficient is nullified. Good performance has been obtained setting, for each coefficient, two thresholds proportional to the mean values on its row and column. Then the element is nullified only if its absolute value is lower than the minimum of the two thresholds. This rule reflects the idea that an arc is a good candidate for elimination if its weight is low compared to the other arcs arriving to and starting from the same node.

Step 4: After choosing the coefficients to be nullified, a new LMI problem is casted, eliminating the corresponding optimisation variables, and a new solution is computed.

Step 5: The evolution of the identified system is compared with the experimental data: If the estimation error exceeds a prefixed threshold, then the algorithm stops, otherwise another iteration starts from Step 2.

The algorithm requires tuning two optimisation parameters: (1) The threshold value used in the pruning phase, which affects the number of coefficients eliminated at each step; (2) The upper bound defining the admissible estimation error, which determines the algorithm termination. The first parameter influences the connectivity of the final reconstructed network: the greater its value, the lower the number of connections, thus it will be named specificity parameter. The algorithm terminates when either it does not find new arcs to remove or the estimation error becomes too large.

4 Performance evaluation

A first evaluation of the proposed algorithm is conducted using data sets generated through in silico experiments. The models used to generate such data are linear systems, with randomly generated sparse and stable 10×10 $\boldsymbol{A}$-matrices. The number of non-zero off-diagonal elements on each row/column varies from 0 to 4, the average sparsity coefficient is equal to 0.82, thus the connectivity matrix has an average of 16 non-zero off-diagonal elements. Note that the elements on the diagonal of the matrix are always assumed to be non-zero, as the variation of a state variable in a dynamical system typically depends also on its current value. Therefore the diagonal elements are not considered in the computation of the performance indices.

It is important to remark that the algorithm is not based on any a priori assumption about the connectivity degree distribution. Therefore the evaluation will be conducted using networks with a random topology and low average number of connections per node. Future works will be aimed at studying whether and to what extent the performance are affected by a scale-free or hierarchical topology, following the guidelines of [39].

In order to allow comparison with other inference techniques, the performance indices will be computed with respect to three different classes of network: (1) undirected, (2) directed and (3) weighted directed. In the following, we will also refer to the latter class with the term 'signed', because in the context of topological reconstruction, the interest is not in the value of the coefficient itself, but rather in the sign.

4.1 Performance indices

Independent of the classification given above, the original and reconstructed networks are always described by the corresponding connectivity matrices, which correspond, in our case, to the dynamical matrices of the underlying models. Therefore different definitions are needed for the positive and negative occurrences in the three cases.

The following occurrences are considered in the computation of the performance indices, based on the original and reconstructed connectivity matrices.

1. True positives (TP):

   Undirected. The elements (i, j) and/or (j, i) are non-zero in the original matrix, and at least one of the two is non-zero in the reconstructed matrix.
   Directed. The element (i, j) is non-zero both in the original and in the reconstructed matrix.
   Signed. The element (i, j) is non-zero both in the original and in the reconstructed matrix and the signs are the same.

2. True negatives (TN):

   Undirected. The element (i, j) and (j, i) are zero both in the original and in the reconstructed matrix.
   Directed/signed. The element (i, j) is missing both in the original and in the reconstructed matrix.

3. False-positives (FP):

   Undirected. The elements (i, j) and/or (j, i) are non-zero in the reconstructed matrix and they are both null in the original one.
   Directed. The element (i, j) is non-zero in the reconstructed matrix and null in the original one.
   Signed. The element (i, j) is non-zero in the reconstructed matrix and it either is null or has a different sign in the original one.

4. False-negatives (FN):

   Undirected. The elements (i, j) and/or (j, i) are non-zero in the original matrix and they are both missing in the reconstructed one.

Table 1: Performance indices

Specificity	Sensitivity	PPV
$\frac{TN}{TN+FP}$	$\frac{TP}{TP+FN}$	$\frac{TP}{TP+FP}$

Directed. The element (i, j) is non-zero in the original matrix and zero in the reconstructed one.
Signed. The element (i, j) is non-zero in the original matrix and it either is null or has a different sign in the reconstructed one.

The performance indices are defined in Table 1. The 'specificity' is the fraction of false occurrences that are unveiled by the test. The 'sensitivity' is the fraction of positive occurrences that are unveiled by the test. Finally, the positive predictive value (PPV) measures the reliability of positive occurrences. An ideal inference technique should obtain a score equal to 1 on each these indices.

Note that, according to the definitions above, in signed networks a sign mismatch (SM) produces the occurrence of both a false-positive and a false-negative. This yields, in the case of signed networks, to a definition of the specificity index that is slightly different from the usual statistical meaning. In the presence of SMs, indeed, the specificity value will be less than 1 even when the algorithm correctly recognises all the non-existing connections.

4.2 Statistical evaluation

In this section, the inference power of the algorithm is assessed statistically, by evaluating the average reconstruction performance over a large number of test networks with ten nodes and random structure. The effect of a priori knowledge on the reconstruction performance is evaluated by repeating the experiments assuming different numbers of a priori known links (0, 20 and 40% of the original connections).

In order to evaluate how the performance is affected by experimental factors, the tests are repeated adding increasing noise levels on the measurements. White noise with 10 and 30% standard deviation (with respect to the measured value) has been added to the data. Data sets of different lengths (10 points and 25 points) are considered. Finally, some tests on larger networks (25 and 50 nodes) have been performed, to investigate whether the method can be effectively scaled-up.

In order to have a relative measure of the inference power, the results are compared with those obtained by a random reconstruction algorithm. The performance of the random algorithm depends on the value assigned to the positive probability parameter, ρ, which defines the probability of assigning a non-zero value to the generic element (i, j) of the connectivity matrix. The complement to 1 of ρ is the negative probability parameter $\nu := 1 - \rho$. Large values of ν yield large specificity values, whereas small values entail greater sensitivity. The performance curve of the random algorithm, generated by varying ν in the range [0, 1], is reported on each plot. The formulae for computing the average random performance can be derived from the definitions given in Section 4.1 by applying straightforward probability calculus; they are reported in Table 2, where σ represents the sparsity coefficient of the connectivity matrix (fraction of null elements) and $\eta := 1 - \sigma$. Such relations can be easily generalised to account for the presence of a priori known links. The prior knowledge percentage measures how many of the original connections are exploited as prior knowledge by the algorithm in each experiment.

Table 2: Probability of occurrences with a random reconstruction algorithm

	UNDIR	DIR	SIGNED
TP	$(1-\nu^2)(1-\sigma^2)$	$\rho\eta$	$\frac{1}{2}\rho\eta$
TN	$\nu^2\sigma^2$	$\nu\sigma$	$\nu\sigma$
FP	$(1-\nu^2)\sigma^2$	$\rho\sigma$	$\rho\sigma+\frac{1}{2}\rho\eta$
FN	$\nu^2(1-\sigma^2)$	$\nu\eta$	$\nu\eta+\frac{1}{2}\rho\eta$

ρ is the positive probability, ν is the negative probability equal to $1 - \rho$, σ is the sparsity coefficient, $\eta = 1 - \sigma$

Figs. 1–12 show the reconstruction performance for different experimental conditions. Each result is the average performance over a set of 500 networks. The performance for every experimental condition is represented in two diagrams, reporting specificity against sensitivity and PPV against sensitivity. A solid line on each subplot represents the average performance, in the same experimental conditions, of the random reconstruction algorithm. Note that, for each reconstruction experiment, we can tune the specificity parameter of the algorithm in order to increase or decrease the specificity value; the results in each subplot are obtained using different values of such parameter. For a given specificity value, the cross represents the mean sensitivity value and the bar represents the standard deviation. Similarly, the second subplot shows the PPV average and the standard deviation for each sensitivity value. The figure title summarises the experimental conditions (tp is number of time points/samples used in that experiment).

The results in Figs. 1–3 show that the reconstruction performance improves significantly when introducing a priori knowledge: The mean value, indeed, approaches the point (1, 1), and the standard deviation decreases. The results above have been obtained from noisy experimental data (10% noise level). Then, in order to analyse the robustness of the algorithm against measurement noise, the results obtained with 0 and 30% noise levels have been compared. From Figs. 4 and 5, we can conclude that the performance degradation is fairly limited. Subsequently, the effect of decreasing the number of samples has been investigated. A comparison between the results in Figs. 4 and 6 reveals that, also in this case, the performance degradation is acceptable.

It is interesting to note that the inference power of the algorithm seems to be greater in the case of signed networks than in the directed and undirected ones. In the former case, indeed, the distance between the performance curve obtained by the algorithm and that obtained by random

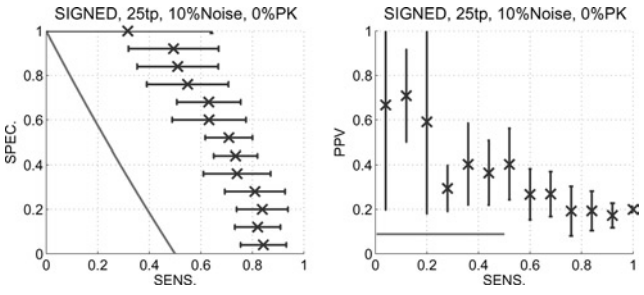

Fig. 1 *Signed networks reconstruction performance with 0% of a priori knowledge*

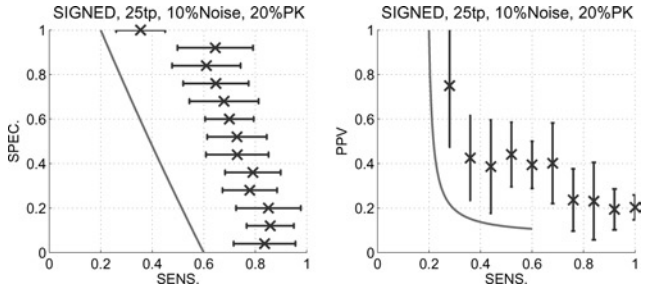

Fig. 2 *Signed networks reconstruction performance with 20% of a priori knowledge*

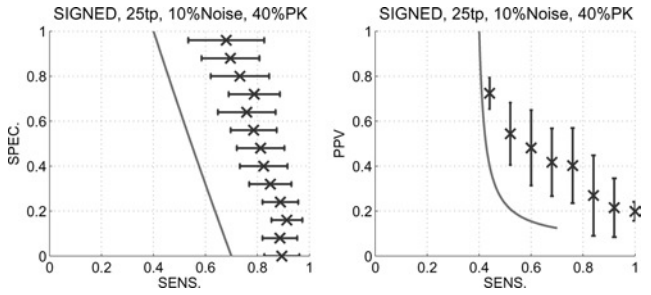

Fig. 3 *Signed networks reconstruction performance with 40% of a priori knowledge*

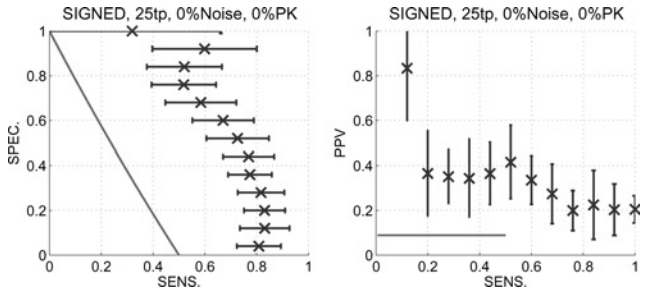

Fig. 4 *Signed networks reconstruction performance with 0% of noise*

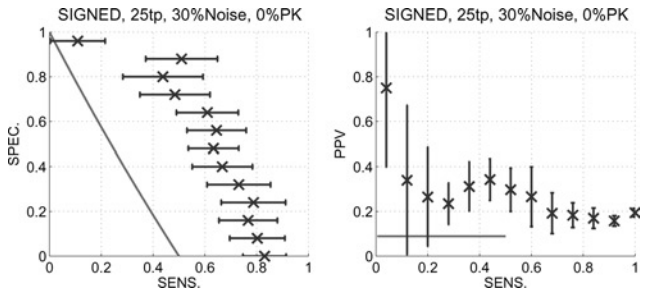

Fig. 5 *Signed networks reconstruction performance with 30% of noise*

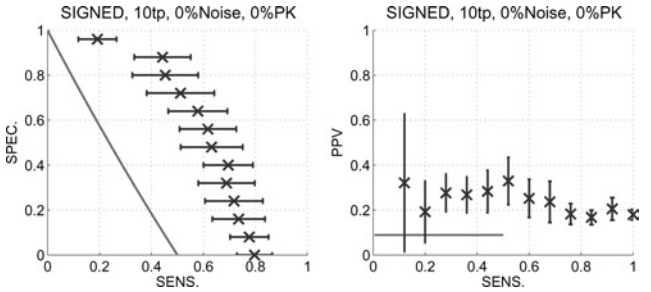

Fig. 6 *Signed networks reconstruction performance with reduced number of time points*

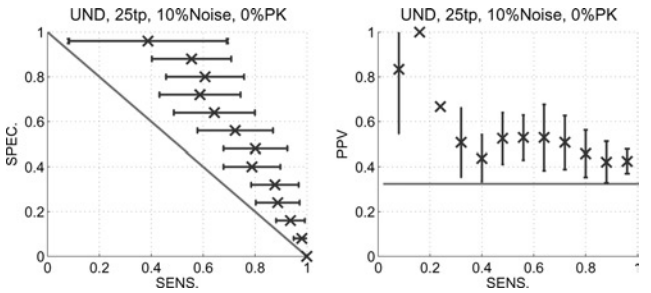

Fig. 7 *Undirected networks reconstruction performance with 0% of a priori knowledge*

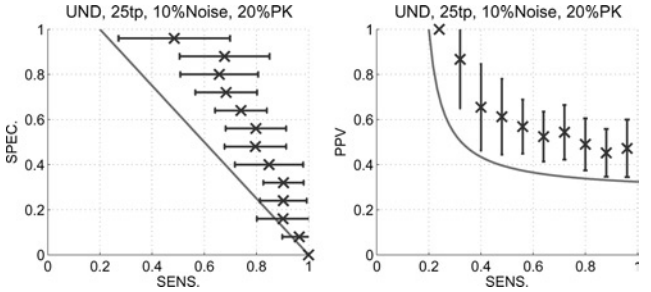

Fig. 8 *Undirected networks reconstruction performance with 20% of a priori knowledge*

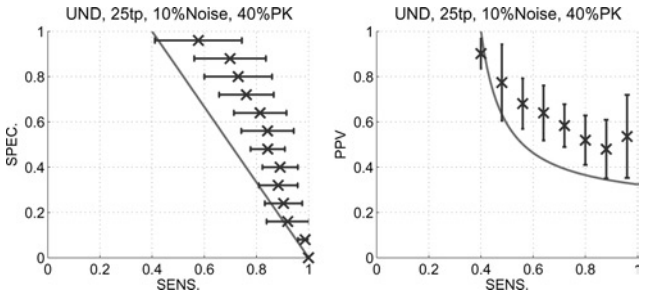

Fig. 9 *Undirected networks reconstruction performance with 40% of a priori knowledge*

reconstruction is larger than in the latter cases, as it is possible to deduce from Figs. 7–12.

Finally, some tests on larger networks (25 and 50 nodes) have been conducted, to demonstrate that the algorithm can be effectively scaled-up. On a standard pc endowed with an AMD Athlon 2.14 GHz processor, a single test on a network of 10 nodes takes about a couple of seconds; the computational time increases to about 1 min with 25 nodes, and about 50 min with 50 nodes. An exact prediction of the computational time is not possible, as it depends on the number of iterations and of constraints imposed at each optimisation step. The performance results are in line with those obtained with networks of 10 nodes, and therefore they have not been reported.

4.3 Comparison with existing techniques

As discussed in Section 1, among the several techniques for biological network inference existing in literature, those based on Bayesian networks (BN) have proven to be

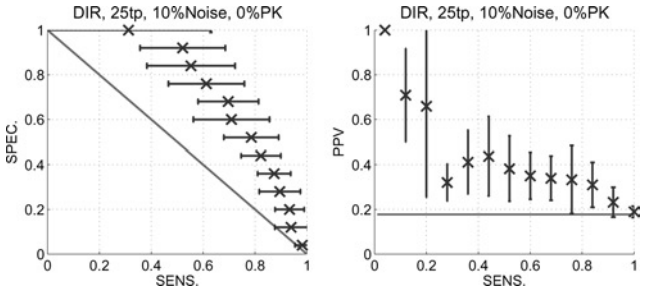

Fig. 10 *Directed networks reconstruction performance with 0% of a priori knowledge*

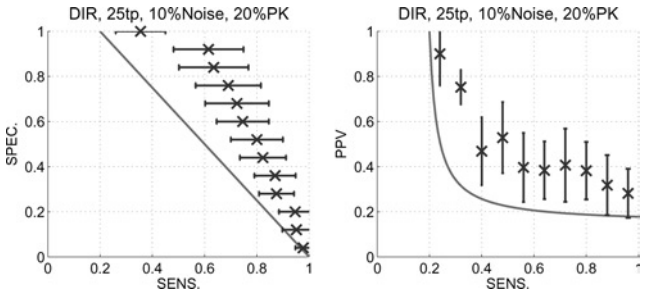

Fig. 11 *Directed networks reconstruction performance with 20% of a priori knowledge*

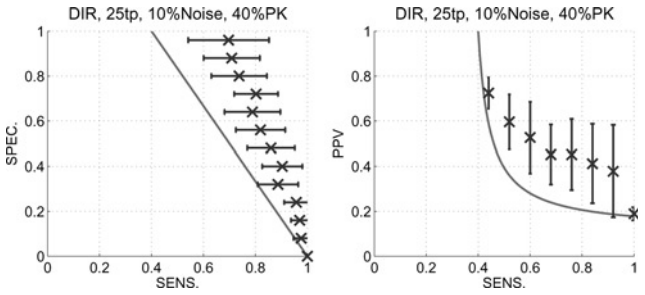

Fig. 12 *Directed networks reconstruction performance with 40% of a priori knowledge*

particularly successful in the last years. One of the reasons of this appeal has to be attributed to the fact that this theoretical framework allows to take into account the a priori biological knowledge in an elegant and principled manner. Moreover, thanks to the extension brought by the dynamic Bayesian networks (DBN) theory, this framework can also be exploited in the reconstruction from time-series measurements. On the other hand, a well-known drawback of using BN-based techniques is the need for huge data sets, in order to obtain good results. Although this is not a severe limitation for some applications, in other cases real experimental data set consist of just few time-point samples, especially when dealing with gene regulatory networks (e.g. microarray data). This is a fundamental point and a remarkable difference between DBN methods and the algorithms based on linear regression, like the one proposed here. In general, the latter class yields better performance when dealing with small data sets, and this is confirmed by the results presented in Section 4.2 are satisfactory even using a small number of time points. In order to evaluate the suitability of DBN methods in the same experimental conditions, we used the software BANJO (BAyesian Network inference with Java Objects), a tool developed by Hartemink and coworkers [40], which is based on a generalisation of the DBN approach, and is aimed to inferring signed directed networks with feedback loops from time-series data.

The average performance indices, computed over a set of 20 randomly generated networks of 10 nodes, are reported in Table 3. The computed sensitivity/specificity indices are very poor in several cases. Indeed, in the majority of the experiments, BANJO predicts a lot of connections with zero influence score, returning a large number of false-negatives. In other cases the sensitivity is higher, but the specificity is rather low, that means that there is a large number of false-positives. In summary, the results obtained in this set of experiments are almost comparable with those of a random reconstruction algorithm. Unfortunately, we did not succeed in doing tests with a priori knowledge using the current software version 1.0.5.

Other successful approaches include those based on mutual information theory [41]. However, this framework is not suitable for network inference from time-series experimental data, because it deals with steady-state values; therefore it is not possible to perform a comparison with our method.

5 Biological example

The proposed algorithm has been assessed in the previous section by means of numerical experiments using linear models. A further step toward in vitro application has been performed by testing the algorithm on a data set generated by a nonlinear model of a real biological system. The simulation model considered in this section is that of cell-cycle regulation in fission yeast, devised by Novak and Tyson [33, 34]. It has been chosen in view of the availability of detailed literature and because it is widely known and

Table 3: Reconstruction performance with BANJO

	10 samples			25 samples		
	PPV	Sens.	Spec.	PPV	Sens.	Spec.
no noise						
Und.	0.375	0.689	0.341	0.362	0.217	0.788
Dir.	0.235	0.501	0.563	0.263	0.145	0.895
Sign.	0.282	0.062	0.965	0.163	0.041	0.943
10% noise						
Und.	0.374	0.715	0.314	0.463	0.246	0.804
Dir.	0.221	0.493	0.537	0.373	0.171	0.901
Sign.	0.096	0.028	0.952	0.305	0.057	0.953

Indices are averaged over a set of 20 networks, with average connectivity degree 0.2 and random topology

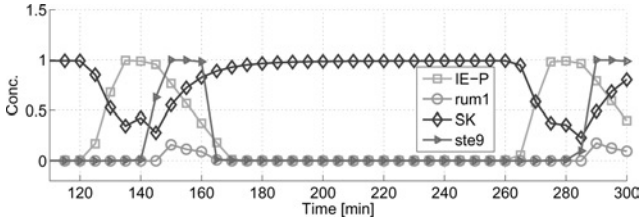

Fig. 13 *Data set generated by the NT model*

well assessed in the scientific community as an awesome example of biomathematical modelling.

Fission yeast is a unicellular eukaryote, whose cell division cycle is regulated by the mitosis promoting factor (MPF), a complex between a cyclin-dependent protein kinase, CdK1 (also known as Cdc2) and cyclin Cdc13. The termination of the mitosis process is regulated by an anaphase promoting complex (APC) that degrades cyclins, thereby permitting cells to reenter G1 phase (the initial phase of cell cycle).

To work properly, APC requires the presence of auxiliary proteins, Slp1 and Ste9, which can be inactivated (through phosphorylation) by Cdk/cyclin complexes and activated (through dephosphorylation) by the phosphatases that oppose Cdk/cyclin complexes. The activity of MPF is also opposed by a stoichiometric inhibitor, Rum1, that can also be inactivated by MPF.

The NT model, for the sake of simplicity, lumps all the starter kinases into a single entity, SK, and assumes that MPF inhibits the synthesis of SK by phosphorylating its transcription factor, TF. Moreover, the activation of Slp1/APC is not performed directly by MPF, but through an intermediary enzyme, named IE in the model. For more detailed information about the model and the biochemical processes occurring during the cell cycle, the reader is referred to [34] and references therein.

From the brief description given above, it is possible to sketch the structure of the corresponding cell-cycle regulatory network. Such network provides a biologically significant benchmark for the proposed reconstruction algorithm. In silico experiments have been conducted by numerically solving the NT model equations, reported in [34], and sampling the resulting curves to obtain 25 experimental observations, which constitute the input to the inference algorithm. Figs. 13 and 14 show the time courses during an entire cycle; however, note that the points used to infer the network are only those belonging to the S/G2 phase, namely the time points between 150 and 260 min. Moreover, a 10% white noise has been added to the data before running the inference algorithm. There are two main motivations underlying the choice of the data window: (1) The model devised by Novak and Tyson is not representative of the entire regulatory network of the cell cycle, but it is mainly focused on the mechanisms

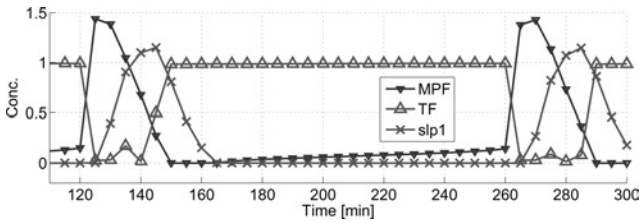

Fig. 14 *Data set generated by the NT model*

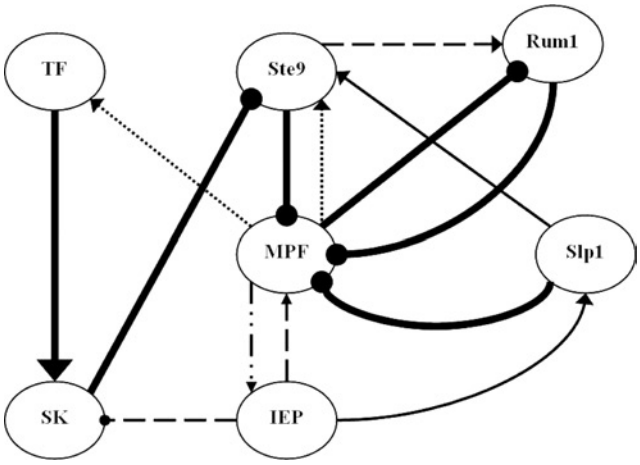

Fig. 15 *Reconstructed regulatory network from NT model*

Solid arc:= TP; thickset solid arc:= TP used as a priori knowledge; dotted arc:= SM; dashed arc:= FP; dashed-dotted arcs:= FN
Arc termination indicates the type of influence predicted by the algorithm: arrow:= promoting; circle:= inhibitory

that make the cell progress through S/G2 phase to finally enter the mitosis phase. (2) The algorithm is based on linear models, and therefore it can work only on a data set, or subintervals of it, exhibiting linear behaviour. The identification of a suitable subinterval in this specific application has been driven by the biological knowledge of the process. The cell cycle is classically divided into four phases, thus the most natural choice is to have an identification interval corresponding to each phase. Concerning the general case, a technique for the optimal splitting into linear subintervals, for modelling through PWA systems, has been presented by the authors in [42].

The test has been conducted by assuming prior knowledge of six interactions (the a priori known edges are drawn with thickset solid arcs). In this case, the number of possible unknown interactions, that the biologist would have to test through in vitro experiments, is equal to 36. The reconstructed network is shown in Fig. 15 and the corresponding performance indices are reported in Table 4. By applying the proposed technique, we have unravelled four interactions out of the remaining five, and for two of them also the correct type of influence has been found. On the other hand, the algorithm has returned three false-positives and one false-negative. Note, however, that the interaction between MPF and IEP has been unveiled, even though with wrong direction; thus, the total number of interaction experiments to perform in vitro, relying on these results, would be equal to 6, instead of the original 36. A more realistic strategy would be the following: (1) run the algorithm; (2) test the interactions suggested by the algorithm by means of in vitro experiments; (3) when a new interaction is found through in vitro experiments, start again from step (1), including the new acquired knowledge.

Table 4: Reconstruction of the cell-cycle regulatory network in fission yeast: performance indices

	Spec.	Sens.	PPV
Und.	0.83	1	0.82
Dir.	0.90	0.91	0.77
Sign.	0.85	0.73	0.62

6 Conclusions

In this work, a novel approach to the reconstruction of interaction networks has been illustrated, with particular attention to its application in the biological domain. The problem tackled consists of identifying the original network structure starting from time-series measurements of the expression levels of each node. A significant hurdle is represented by the fact that, given a set of experimental data, one can, in general, find many interpolating models exhibiting similar temporal evolution. In view of this fact, a good interpolation of the experimental data set does not guarantee that the identified network structure actually matches the original one.

The probability to reconstruct the original network intuitively increases if it is possible to take advantage of the preexisting knowledge from the system expert domain; this means finding a solution with a partially assigned structure. Such a goal has been pursued by recasting the problem in the framework of LMIs.

The proposed algorithm has been statistically validated by means of numerical tests. Subsequently, it has been applied to a more challenging case study, that is, the regulatory network of cell cycle in fission yeast, by exploiting a well-assessed in silico model for data set generation and results evaluation. The tests have yielded satisfactory results; especially, they have proven that the exploitation of a priori known interactions yields a significant improvement.

It is worth reminding that the proposed technique has to be fitted in the more general framework of PWA approaches to modelling of biological systems. In this view, it can be applied also to nonlinear systems by suitably segmenting the analysis of the experimental data into subintervals exhibiting quasi-linear behaviour, as it has been done in the case of fission yeast cell cycle. On the other hand, the identification of the quasi-linear subintervals is not straightforward in the generic case, and is likely to be itself the subject of optimisation procedures to be devised.

A future improvement of the present technique will definitely concern the possibility to reconstruct networks with a predefined connectivity degree distribution. On the other hand, further tests will be conducted, to analyse to what extent the network topology affects the reconstruction performance, following the guidelines of [39].

7 References

1 Brown, P.A., and Botstein, D.: 'Exploring the new world of the genome with DNA microarrays', *Nat. Genet.*, 1999, **21**, (suppl.), pp. 33–37

2 Lipschutz, R.J., Fodor, S.P.A., Gingeras, T.R., and Lockhart, D.J.: 'High density synthetic oligonucleotide arrays', *Nat. Genet.*, 1999, **21**, (suppl.), pp. 20–24

3 Bower, J.M., and Bolouri, H.: 'Computational modelling of genetic and biochemical networks' (MIT Press, Cambridge, MA, 2001)

4 Kitano, H.: 'Foundations of systems biology' (MIT Press, Cambridge, MA, 2001)

5 de Jong, H.: 'Modeling and simulation of genetic regulatory systems: a literature review', *J. Comput. Biol.*, 2002, **9**, pp. 67–103

6 d'Alché-Buc, F., and Schachter, V.: 'Modeling and simulation of biological networks'. Proc. Int. Symp. on Applied Stochastic Models and Data Analysis, Brest, France, May 2005

7 de Jong, H., Gouzé, J.-L., Hernandez, C., Page, M., Sari, T., and Geiselmann, J.: 'Qualitative simulation of genetic regulatory networks using piecewise-linear models', *Bull. Math. Biol.*, 2004, **66**, pp. 301–340

8 de Jong, H.: 'Dynamical modeling of biological regulatory networks', *BioSystems*, 2006, **84**, pp. 77–80

9 Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., and Gerstein, M.: 'Genomic analysis of regulatory network dynamics reveals large topological changes', *Nature*, 2004, **431**, pp. 308–312

10 Nelles, O.: 'Nonlinear system identification' (Springer, Berlin, 2001), pp. 36–37

11 Chen, T., He, H.L., and Church, G.M.: 'Modeling gene expression with differential equations'. Proc. Pacific Symp. on Biocomputing, Hawaii, USA, January 1999, pp. 29–40

12 Weaver, D.C., Workman, C.T., and Stormo, G.D.: 'Modeling regulatory networks with weight matrices'. Proc. Pacific Sympo. on Biocomputing, Hawaii, USA, January 1999, pp. 112–123

13 Stephen Yeung, M.K., Tegnér, and Collins, J.J.: 'Reverse engineering gene networks using singular value decomposition and robust regression', *Proc. Natl. Acad. Sci. USA*, 2002, **99**, pp. 6163–6168

14 Gardner, T.S., di Bernardo, D., Lorenz, D., and Collins, J.J.: 'Inferring genetic networks and identifying compound mode of action via expression profiling', *Science*, 2003, **301**, pp. 102–105

15 Sontag, E., Kiyatkin, A., and Kholodenko, B.: 'Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data', *Bioinformatics*, 2004, **20**, pp. 1877–1886

16 Cho, K.-H., Choo, S.-M., Wellstead, P., and Wolkenhauer, O.: 'A unified framework for unraveling the functional interaction structure of a biomolecular network based on stimulus-response experimental data', *FEBS Lett.*, 2005, **579**, pp. 4520–4528

17 di Bernardo, D., Gardner, T.S., and Collins, J.J.: 'Robust identification of large genetic networks'. Proc. Pacific Symp. on Biocomputing, Hawaii, USA, Jan 2004, pp. 486–497

18 di Bernardo, D., Thompson, M.J., Gardner, T.S., Chobot, S.E., Eastwood, E.L., Wojtovich, A.P., Elliott, S.J., Schaus, S.E., and Collins, J.J.: 'Chemogenomic profiling on a genomewide scale using reverse-engineered gene networks', *Nat. Biotech.*, 2005, **23**, pp. 377–383

19 Schmidt, H., Cho, K.-H., and Jacobsen, E.W.: 'Identification of small scale biochemical networks based on general type system perturbations', *FEBS J.*, 2005, **272**, pp. 2141–2151

20 Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabasi, A.L.: 'The large-scale organization of metabolic networks', *Nature*, 2000, **407**, pp. 651–654

21 Wagner, A.: 'The yeast–protein interaction network evolves rapidly and contains few redundant duplicate genes', *Mol. Biol. Evol.*, 2001, **18**, pp. 1283–1292

22 Featherstone, D.E., and Broadie, K.: 'Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network', *Bioessays*, 2002, **24**, pp. 267–274

23 Luscombe, N.M., Qian, J., Zhang, Z., Johnson, T., and Gerstein, M.: 'The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties', *Genome Biol.*, 2002, **3**, research0040.1–research0040.7

24 Barabási, A.-L., and Albert, R.: 'Emergence of scaling in random networks', *Science*, 1999, **286**, pp. 509–512

25 Nakhleh, L., Jin, G., Zhao, F., and Mellor-Crumm, J.: 'Reconstructing phylogenetic networks using maximum parsimony'. Proc. IEEE Computational Systems Bioinformatics Conf., Stanford, CA, USA, August 2005, pp. 93–102

26 Segal, E., Taskar, B., Gasch, A., Friedman, N., and Koller, D.: 'Rich probabilistic models for gene expression', *Bioinformatics*, 2001, **17**, (Suppl. 1), pp. S243–S252

27 Friedman, N., Linial, M., Nachman, I., and Pe'er, D.: 'Using Bayesian networks to analyze expression data', *J. Comput. Biol.*, 2000, **7**, pp. 601–620

28 Shmulevich, I., Dougherty, E.R., Kim, S., and Zhang, W.: 'Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks', *Bioinformatics*, 2002, **18**, pp. 261–274

29 Perrin, B.E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., and d'Alché-Buc, F.: 'Gene networks inference using dynamic Bayesian networks', *Bioinformatics*, 2003, **19**, (Suppl. 2), pp. ii138–ii148

30 Husmeier, D.: 'Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks', *Bioinformatics*, 2003, **19**, pp. 2271–2282

31 Chickering, D.M.: 'Learning Bayesian networks is NP-complete' in Fisher, D., and Lenz, H.J. (Eds.): 'Learning from data: artificial intelligence and statistics' (Springer, New York, 1996, vol. 5), pp. 121–130

32 Boyd, S., El Ghaoui, L., Feron, E., and Balakrishnan, V.: 'Linear matrix inequalities in system and control theory' (SIAM, Philadelphia, PA, 1994)

33 Novak, B., and Tyson, J.J.: 'Modeling the control of DNA replication in fission yeast', *Proc. Natl. Acad Sci. USA*, 1997, **94**, pp. 9147–9152

34 Novak, B., Pataki, Z., Ciliberto, A., and Tyson, J.J.: 'Mathematical model of the cell division cycle of fission yeast', *Chaos*, 2001, **11**, pp. 277–286

35 Kocay, W., and Kreher, D.L.: 'Graphs, algorithms, and optimization' (Chapman & Hall/CRC, Boca Raton, FL, 2004)
36 Zhou, K.: 'Essentials of robust control' (Prentice-Hall, Upper Saddle River, NJ, 1998)
37 Meyer, C.D.: 'Matrix analysis and applied linear algebra' (SIAM, Philadelphia, PA, 2000)
38 Gahinet, P., Nemirovski, A., Laub, A.J., and Chilali, M.: 'LMI control toolbox' (Mathworks, Inc, Natick, MA, 1995)
39 Wildenhain, J., and Crampin, E.J.: 'Reconstructing gene regulatory networks: from random to scale-free networks', *IEE Proc., Syst. Biol.*, 2006, **153**, pp. 247–256
40 Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J., and Jarvis, E.D.: 'Advances to Bayesian network inference for generating causal networks from observational biological data', *Bioinformatics*, 2004, **20**, pp. 3594–3603
41 Liang, S., Fuhrman, S., and Somogyi, R.: 'REVEAL, a general reverse engineering algorithm for inference of genetic network architectures'. Proc. Pacific Symp. on Biocomputing, Hawaii, USA, January 1998, pp. 3–18
42 Amato, F., Bansal, M., Cosentino, C., Curatola, W., and di Bernardo D.: 'Modelling the cell cycle of fission yeast by means of piecewise linear systems'. Proc. IEEE Conf. on Control Applications, Munich, *Germany*, October 2006, pp. 3301–3305