Robust dynamical network reconstruction

Ye Yuan[1], Guy-Bart Stan[2], Sean Warnick[3] and Jorge Gonçalves[1]

Abstract— **Motivated by biological applications, this paper addresses the problem of network reconstruction from data. Previous work has shown necessary and sufficient conditions for network reconstruction of noise-free LTI systems. This paper assumes that the conditions for network reconstruction have been met but here we additionally take into account noise and unmodelled dynamics (including nonlinearities). Algorithms are therefore proposed to reconstruct dynamical (Boolean) network structure from time-series (steady-state) data respectively in presence of noise and nonlinearities. In order to identify the network structure that generated the data, we compute the smallest distances between the measured data and the data that would have been generated by particular Boolean structures. Information criteria and optimisation techniques are introduced to strike a balance between our definition of distance and model complexity and we show how doing so enables us to recover the true structure of the network. We conclude with biologically-inspired network reconstruction examples which include noise and nonlinearities.**

I. Introduction

One of the fundamental interests in systems biology is the discovery of the specific biochemical mechanisms that explain the observed behaviour of a particular biological system [1]. In particular, we consider the problem of reconstructing the network structure (thereby uncovering the underlying mechanisms responsible for the observed behaviour) from input and partially measured output data of a dynamical system (observed behaviour). The biological network reconstruction problem challenges come from the necessity to deal with noisy and partial measurements (in particular, the number of hidden/unobservable nodes and their position in the network is unknown) taken from a nonlinear and stochastic network.

Even in the ideal situation where the underlying network is assumed to be linear time-invariant (LTI) and the measurements are assumed to be non-noisy, it can be shown that, due to partial observability, this problem is unsolvable using classical system identification techniques [2]. In particular, identification of the system transfer function (obtained, for example, using classic system identification approaches) is useless to solve the network structure reconstruction problem since transfer functions do not contain sufficient information for that purpose (counter-examples can be found in [4]).

[1]Control Group, Department of Engineering, University of Cambridge. [2]Department of Bioengineering and the Centre for Synthetic Biology and Innovation, Imperial College London. [3]Information and Decision Algorithms Laboratories, Computer Science Department, Brigham Young University.

Corresponding author: jmg77@cam.ac.uk (Jorge Gonçalves). This work was supported in part by EPSRC grant number EP/G066477/1, AFRL FA8750-09-2-0219 and Microsoft Research through the PhD Scholarship Program. Guy-Bart Stan gratefully acknowledges the support of the EPSRC Centre for Synthetic Biology and Innovation at Imperial College, London.

In the literature, full observation of all system states is typically assumed to be able to reconstruct the network [7], [15], [9], [10], [16], [19]. However, when there exists hidden states in the network (typically this is the real case if you are considering to reconstruct unknown biological network), these method will be ill-posed and might lead to wrong network structure (examples can be found in [2]). Motivated by this, we are focusing on the effect of hidden states in the network that we are aiming to reconstruct.

Based on this latter observation, a new representation for LTI systems, called dynamical structure functions was introduced in [2]. Dynamical structure functions capture information at an intermediate level between transfer function and state space representation (see Figure 1). Specifically, dynamical structure functions not only encode structural information at the measurement level, but also contain some information about hidden states. Based on the theoretical results presented in [2], we proposed some guidelines for the design of an experimental data-acquisition protocol which allows the collection of data containing sufficient information for the network structure reconstruction problem to become solvable. In particular, we have shown that if nothing is known about the network, then the data-collection experiments must be performed as follows:

(A.1) for a network composed of p measured species, the same number of experiments p must be performed;

(A.2) each experiment must independently control a measured species, i.e., control input i must first affect measured species i.

If the experiments are not performed in this way the network cannot be reconstructed, and any network structure fits the data equally well (e.g. a fully decoupled network or a fully connected network). If biologists have already some information about the network, as it is usually the case, then these conditions can be relaxed as explained in [2].

Using dynamical structure functions as a mean to solve the network reconstruction problem, the following aspects need to be considered (see Figure 1):

First (see (A) in Figure 1), the properties of a dynamical structure function and its relationship with the transfer function associated with the same system need to be precisely established (this was done in [2]).

Second (see (B) in Figure 1), an algorithm for constructing a minimal order state-space representation consistent with an obtained dynamical structure function needs to be developed (this was done in [3]). Using this last set of results, an estimation of the minimal number of hidden nodes that needs to be considered in the state space realisation can be obtained. In the context of biology, this helps understand the

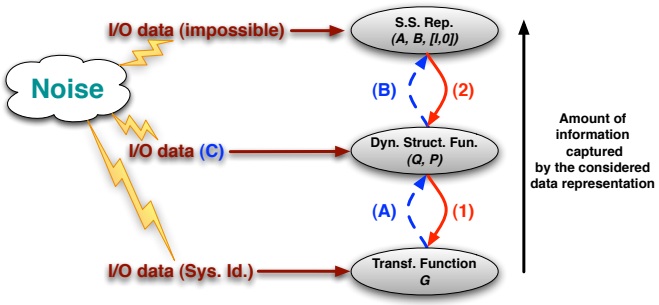

Fig. 1. Mathematical structure of the network reconstruction problem using dynamical structure functions. Red arrows mean "uniquely determine", blue arrows indicate our work.

minimal number of unmeasured molecules in a particular pathway.

Third (see (C) in Figure 1), an efficient method is developed to reconstruct networks in the presence of noise and nonlinearities. We assume that the conditions for network reconstruction presented above in (A.1) and (A.2) have been met. In our approach, we use the same information as traditional system identification methods, i.e., input-output data. However, with our method, steady-state (resp. time-series data) can be used to reconstruct the Boolean (resp. dynamical network) structure of the system.

The structure of the paper is as follows. In Section II, dynamical structure functions are defined and fundamental results concerning their usefulness in the network reconstruction problem are stated. Section III presents the main results of the paper, i.e., robust network reconstruction from input-output data in the presence of noise and nonlinearities. Finally, we conclude the paper with biologically-inspired network reconstruction examples in Section IV.

A. Notation

For a matrix $A \in \mathbb{C}^{M\times N}$, $A_{ij} \in \mathbb{C}$ denotes the element in the i^{th} row and j^{th} column while $A_j \in \mathbb{C}^{M\times 1}$ denotes its j^{th} column. For a column vector α, $\alpha[i]$ denotes its i^{th} element. We define $e_r^T = [0,\ldots,0,1_{r^{th}},0,\ldots,0] \in \mathbb{R}^{1\times N}$. I denotes the identity matrix.

II. Dynamical Structure Functions and Network Reconstruction

In [2] we introduced the notion of dynamical structure functions and showed how they can be used to obtain necessary and sufficient conditions for network reconstruction. For the sake of clarity and completeness, we state these previously obtained results here without proofs. We refer the interested reader to [2] for the corresponding proofs.

Consider a nonlinear system $\dot{\bar{x}} = f(\bar{x},\bar{u},w_1)$, $\bar{y} = h(\bar{x},w_2)$ with n states $\bar{x}$, p measured states $\bar{y}$, $n-p$ hidden states $\bar{z}$ (potentially a large number of them), m inputs $\bar{u}$, and noise w_1, w_2. The system is linearised around an equilibrium point (a point such that $f(\bar{x}^*,\bar{u}^*,0)=0$), and it is assumed that inputs and noise do not move the states too far from the equilibrium point so that the linearised system is a valid approximation of the original nonlinear system. The linearised system can be written as $\dot{x} = Ax + Bu$, $y = Cx$, where $x = \bar{x} - \bar{x}^*$, $u = \bar{u} - \bar{u}^*$ and $y = h(\bar{x},0) - h(\bar{x}^*,0)$. The transfer function associated with this linearised system is given by $G(s) = C(sI-A)^{-1}B$.

Partition the linearised system as follows

$$\begin{aligned} \begin{bmatrix} \dot{y} \\ \dot{z} \end{bmatrix} &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u \\ y &= \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} \end{aligned} \tag{1}$$

where $x = [y^T \ z^T]^T \in \mathbb{R}^n$, is the full state vector, $y \in \mathbb{R}^p$ is a partial measurement of the state (we assume $p > 1$), z are the $n-p$ "hidden" states, and $u \in \mathbb{R}^m$ is the control input. We restrict our attention to situations where output measurements constitute partial state information, i.e., $p < n$. Taking the Laplace transforms of the signals in (1), solving for Z, and substituting into the Laplace transform of the first equation of (1) of yields $sY = WY + VU$, where $W = A_{11} + A_{12}(sI - A_{22})^{-1}A_{21}$ and $V = A_{12}(sI - A_{22})^{-1}B_2 + B_1$. Let D be the matrix composed of the diagonal elements of W and write $(sI-D)Y = (W-D)Y + VU$. Then $Y = QY + PU$ where

$$Q = (sI-D)^{-1}(W-D) \ \text{ and } \ P = (sI-D)^{-1}V \tag{2}$$

Given the system (1), we define the *dynamical structure function* of the system to be (Q,P).

It can be shown that $G = (I-Q)^{-1}P$. From this last equation, it can be seen that the dynamical structure function of a system contains more minformation than the transfer function, and less information than the state-space representation [2]. We can then conclude that, with no other information about the system, dynamical or Boolean reconstruction is not possible. In particular, this shows that the use of criteria such as sparsity or decoupledness to guide our selection of a candidate network structure can be misleading. Thus, if we are to use these kinds of criteria, they must be firmly justified a priori.

Proposition 1: [2] Given a $p \times m$ transfer function G, dynamical structure reconstruction is possible from partial structure information if and only if $p-1$ elements in each column of $\begin{bmatrix} Q & P \end{bmatrix}^T$ are known that uniquely specify the component of (Q,P) in the nullspace of $\begin{bmatrix} G^T & I \end{bmatrix}$.

The importance of this result is that it identifies exactly what information about a system's structure, beyond knowledge of its transfer function, must be obtained to be able to recover the structure without appeal to a priori assumptions, such as sparsity, or parsimony, etc.

Corollary 1: [2] If $m = p$, G is full rank, and there is no information about the internal structure of the system, Q, then the dynamical structure can be reconstructed if each input controls a measured state independently, i.e., without loss of generality, the inputs can be numbered such that P is diagonal. Moreover, $H = G^{-1}$ characterises the dynamical structure as follows

$$Q_{ij} = -\frac{H_{ij}}{H_{ii}} \ \text{ and } \ P_{ii} = \frac{1}{H_{ii}}. \tag{3}$$

Proposition 2: [3] Given a dynamical system (1) and the associated dynamical structure functions (Q,P) constructed as explained above, the following conditions must hold

$$A_{11} - \text{diag}\{A_{11}\} = \lim_{s\to\infty} sQ(s); \quad (4)$$

$$B_1 = \lim_{s\to\infty} sP(s). \quad (5)$$

III. Robust network structure reconstruction

In this section, we consider the problem of robustly reconstructing dynamical network structures. Data are obtained from input-output measurements of a noisy nonlinear system. From this type of data we aim to find the internal network structure Q associated with the linearised system (1).

For simplicity of exposition, we assume that there is no *a priori* information on the internal network structure Q. The results still follow if some *a priori* information about Q is available, and such information can typically be used to relax the experimental protocol according to Proposition 1. Hence, data are collected according to the measurement protocol described in the introduction:
(1) the number of distinct data-collection experiments is the same as the number of measured species. This in particular implies that $u(t), y(t) \in \mathbb{R}^p$;
(2) each input u_i controls first the measured state y_i so that P is a diagonal matrix ($p \times p$). To average out the noise, data-collection experiments are repeated N times.

In the following two subsections (III-A and III-B), we propose two approaches for estimating the dynamical structure function (Q,P) from measured input-output data. The first approach is indirect and involves estimating the transfer function G followed by computing (Q,P) from G. Since some information is lost in the process of estimating G, we consider a second approach where (Q,P) is directly estimated from data (without estimating first G). Concerning the type of input-output data collected, we first consider time-series input-output data and then the special case where only steady-state data are available.

A. Dynamical network reconstruction from identified transfer functions

Consider a transfer matrix $G(s)$ estimated from noisy data. According to Corollary 1, if G is full rank there is a unique Q and diagonal P satisfying $(I-Q)G = P$. Since G is an approximation of the actual system, Q and P will typically be mere approximations of the actual dynamical structure function. Moreover, due to noise and unmodelled dynamics, it is likely that Q does not even have the correct Boolean structure.

The main idea to solve the network reconstruction problem from noisy data is the following. For p measured states, Q has p^2-p unknowns. We want to quantify the *distance* from G (or directly from the measured data) to all possible Boolean structures (and there are 2^{p^2-p} of them). Some of such distances will be large revealing that the corresponding Boolean structures are likely not the correct structures while other will be small making them candidates for the correct structure.

Definition 1: A Boolean mapping $b : Q(s) \to b(Q)$, where $b(Q)$ is a Boolean matrix with the same dimension as the transfer matrix Q and $\forall i,j\ \{b(Q)\}_{i,j} = 0$ if $Q_{ij}(s) = 0$ for all s, otherwise, $\{b(Q)\}_{i,j} = 1$.

For a given p, there are 2^{p^2-p} possible Boolean networks B_k (remember that $Q(s)$ has zeros on the diagonal and therefore $b(Q)$ will always have zeros on the diagonal) which can be ordered using the index $k = 1,...,2^{p^2-p}$.

Definition 2: A Boolean structure $\mathscr{B}_k$ corresponding to a Boolean network B_k is defined as follows: $\{Q(s) : b(Q) = B_k\}$.

The distance from G to the Boolean structure $\mathscr{B}_k$ is defined as the smallest perturbation Δ to G (measured in some norm) so that the perturbed system G_Δ belongs to the set of transfer functions $\tilde{G}$ such that $Q \in \mathscr{B}_k$, where Q is obtained from $(I-Q)\tilde{G} = P$. Finding the distance from G to a Boolean structure $\mathscr{B}_k$, gives us a quantitative information about how much we would need to perturb G (or the data) to obtain a new system transfer function for which the associated Q corresponds to the considered Boolean structure, i.e., for which $Q \in \mathscr{B}_k$.

There are many possible approaches to define such "smallest perturbations", including several uncertainty models and norms to choose from. This choice is key to obtain a convex minimisation problem. In order to obtain a convex minimisation problem, we consider the output (could also be input) feedback uncertainty model. In this framework, the "true" system is given by $(I+\Delta)^{-1}G$, where Δ represents unmodelled dynamics, including nonlinearities, and noise.

Based on this choice of dynamic uncertainty, the problem is defined as follows. Given a particular Boolean structure $\mathscr{B}_k$, the objective is to minimise $\|\Delta\|$, in some norm, such that Q obtained from $(I+\Delta)^{-1}G = (I-Q)^{-1}P$ has the desired Boolean structure, i.e. $Q \in \mathscr{B}_k$.

We can rewrite the above equation as $\Delta = GP^{-1}(I-Q) - I$. So, we are looking to minimise $\|GP^{-1}(I-Q) - I\|$ over $Q \in \mathscr{B}_k$ and P diagonal. Since P is diagonal, its inverse P^{-1} is also diagonal. Define a new matrix $X = P^{-1}(I-Q)$ whose diagonal is the diagonal of P^{-1} and for which the off diagonal elements are given by $P_{ii}^{-1}Q_{ij}$. Since $Q \in \mathscr{B}_k$ this imposes structural constraints on X, i.e., some off-diagonal $X_{ij} = 0$. These zero X_{ij} correspond to those Q_{ij} which are equal to zero (since $X_{ij} = P_{ii}^{-1}Q_{ij}$ for $i \neq j$).

Definition 3: For all k, define $\mathscr{X}_k \triangleq \{X(s) : b(X) = B_k + I_p\}$, where I_p is identity matrix of dimension p. $X(s) \in \mathscr{X}_k$ if the following conditions hold

(i) when $i \neq j$, $X_{ij}(s) = 0$ if $B_{k,ij} = 0$, i.e., the (i,j) element B_k is 0; all other $X_{ij}(s)$ are free variables;
(ii) when $i = j$, $X_{ii}(s)$ is a free variable.

Using Defintion 3, the distance from G to a particular Boolean structure $\mathscr{B}_k$ can be written as $\alpha_k = \inf_{X\in\mathscr{X}_k} \|GX - I\|^2$ which is a convex minimisation problem with a careful choice of a norm.

Next, we show that this problem can be casted as a least squares optimisation problem. If we use the norm defined by $\|\Delta\|^2 =$ sum of all $\|\Delta_{ij}\|_2^2$, where $\|\cdot\|_2$ stands as the $\mathscr{L}_2$-norm over $s = j\omega$, then using the projection theorem [12]

the problem reduces to

$$\begin{aligned}\alpha_k = \inf_{X\in\mathscr{X}_k} \|GX-I\|^2 &= \inf_{X\in\mathscr{X}_k} \sum_i \|GX_i - e_i\|_2^2 \\ &= \sum_i \inf_{Y_i} \|A_iY_i - e_i\|_2^2 \\ &= \sum_i \|A_i(A_i^*A_i)^{-1}A_i^*e_i - e_i\|_2^2,\end{aligned}$$

where X_i is the i^{th} column of $X \in \mathscr{X}_k$, Y_i is a column vector composed by the free (i.e., nonzero) elements of X_i, A_i is obtained by deleting the j^{th} column of G when the corresponding element $X_i(j)$ is 0 for all j, and $(\cdot)^*$ denotes transpose conjugate. The infimum is achieved by choosing $X_i = (A_i^*A_i)^{-1}A_i^*e_i$, and $A_i^*A_i$ is always invertible since G is full rank in Corollary 1.

If experiments are repeated N times (as they should) and we obtain a transfer function G^i for each experiment, then the above analysis still follows simply by forming a higher dimensional matrix $G = \left[\left(G^1\right)^T \ \cdots \ \left(G^N\right)^T\right]^T$.

B. Dynamical network reconstruction directly from time-series data

The previous sections used a two-step approach in which system identification was first used to estimate a transfer function from measured input-output data and then, in a second step, the identified transfer function was used to obtain a dynamical structure function representation of the system which is optimal in terms of a particular metric. This section proposes a method which allows identification of the optimal dynamical structure function representation directly from the measured input-output data. The advantage of this direct network structure reconstruction from data is that no information is lost during the initial transfer function identification stage.

Due to the equivalence between dynamical uncertainty perturbations, we are free to chose, without loss of generality, the type of uncertainty perturbation that best suits our needs. For the direct method, instead of a feedback uncertainty as was considered in the previous section, the uncertainty perturbation we are considering here is the additive dynamic uncertainty on the output, i.e., $Y = G_\Delta(U+\Delta)$. In this case, we think about the "distance" in terms of how much we need to change the input (data) to fit a particular Boolean structure. Since $G_\Delta = (I-Q)^{-1}P = X^{-1}$, the equality $Y = G_\Delta(U+\Delta)$ can be written as

$$\Delta = XY - U,$$

where $X \in \mathscr{X}_k$, for some particular Boolean network k. Recall that structural constraints in Q can be imposed directly on X from the equality $X = P^{-1}(I-Q)$. We can therefore use system identification theory for non-causal autoregression models under the structural constraints to identify X (which might be non-causal). In this case, the distance is defined as the maximum likelihood of the estimation problem.

C. Penalising connections

Information criteria: The above methodology suffers from a crucial weakness: there are several Boolean structures with distances smaller or equal than the distance to the "true" network. Indeed, the extra degrees of freedom of the fully-connected network allow the corresponding distance α_k to be the smallest of all. This is similar to the noisy data over-fitting problem encountered in system identification where the higher the order of the transfer function, the better the fit. Obviously, if we only focus on noisy data best fit, eventually we end up fitting noise and so a large system order is not typically a good choice. Therefore, a compromise has to be struck.

If the true network has l non-existent connections (l off-diagonal elements in Q are zero) and the data are non-noisy, then there are $2^l - 1$ different networks that have a smaller or equal distance (due to the additional degrees of freedom provided by the extra connections). When noise is present, then the "true" network will typically have an optimal distance similar to these other l networks. The question of how to find the "true" network thus arises. With repeated experiments, small enough noise (i.e., large enough signal-to-noise ratio) and negligible nonlinearities, the optimal distances of those l networks are comparable, and they are typically much smaller than those of the other networks. To try to reveal the "true" network, one can strike a compromise between network complexity (in terms of number of connections) and data fitness by penalising extra connections. The are several methods to strike this compromise. Here, we introduce methods known as Akaike's information criterion (AIC) [5], or some of its variants such as AICc (which is AIC with a second order correction for small sample sizes), and the Bayesian information criterion (BIC) [13].

The AIC-type approach is a test between models - a tool for model selection. Given a data set, several competing models may be ranked according to their AIC value, with the one having the lowest AIC being the best. From the AIC value one may typically infer that the best models are in a tie and the rest are far worse, but it would be arbitrary to assign a value above which a given model is rejected [13]. The AIC value for a particular Boolean network B_k is defined as:

$$AIC_k = 2L_k - \ln \alpha_k, \tag{6}$$

where L_k is the number of (non-zero) connections in the Boolean network B_k and α_k is the optimal distance based on this parameter constraint.

Reconstruction with the zero norm: To directly take the number of connections into account in the formulation of the optimisation problem, we redefine the optimisation problem as:

$$DO = \inf_{X\in\mathscr{X}_k} \left(\|XY-U\|^2 + \beta\|X\|_0\right), \tag{7}$$

where β is a parameter balancing data-fitting and model complexity (i.e., the number of non-zero connections). In (7), $\|X\|_0$ denotes the number of nonzero element in the matrix X, and is known as the zero norm. Notice that this minimisation problem can be equivalently written as (if we

choose the 2-norm):

$$\inf_{X\in\mathscr{X}_k}\left(\|XY-U\|_2^2+\beta\|X\|_0\right)$$
$$=\sum_i \inf_{X_i}\left(\|X_i^T Y-U_i^T\|_2^2+\beta\|X_i^T\|_0\right), \qquad (8)$$

where X_i^T is the i^{th} row of $X\in\mathscr{X}_k$ and U_i^T is the i^{th} row of U.

Directly solving such optimisation problem is in general NP-hard. A frequently discussed approximation for steady-state data can be found in [10]. Moreover, since there are p independent optimisations in eq. (8), we can choose different β_i for each i. Alas, there is no clear rule for selecting β to balance optimally the two terms in eq. (8). How to calculate the correct β without information a priori is a currently open problem [10].

D. Boolean network reconstruction from steady-state data

So far we have assumed that time-series data are available. Frequently, however, experimentation costs and limited resources only permit steady-state measurements. In addition, with steady-state measurements it is typically possible to perform a larger number of experiments for the same time, effort and cost. As shown below, most of the connectivity of the network together with the associated steady-state gains (and the associated positive or negative sign) can still be reconstructed from steady-state data. However, no dynamical information will be obtainable. In other words, for most cases we can still recover the Boolean network from steady-state data.

Assume that after some time of maintaining the control input concentrations at a constant value, the measured outputs y have converged to a steady-state value. This is equivalent (if the system is stable or quasi-stable [1]) to assume that we can obtain $G(0)$, i.e., $G(s)$ evaluated at $s=0$. $(I-Q(s))G(s)=P(s)$ evaluated at $s=0$ becomes $(I-Q(0))G(0)=P(0)$. From this equation, all of the results given in Section III-A and III-B follow provided that no element of $G(s)$ has a system zero [14] at 0. In that case, a nonzero element in the obtained Boolean network indicates the existence of a causal relationship between the corresponding pair of nodes while a zero element indicates the absence of such relationship.

IV. Biologically-inspired examples

This section illustrates with one example[1] the theoretical results presented in the previous section. We consider the application of our method to the reconstruction of the underlying dynamic network responsible for chemotaxis in *Rhodobacter sphaeroides*. The network is represented in Figure 2(a) (see [6], [11] for a detailed explanation of this model and its biological interpretation). It involves 10 species dynamically interacting through a complex set of interconnections. As an illustrative example of the application of our method, we consider the case where steady-state data are collected from 3 species only: Y_3^p, Y_6^p and the "motor" (circled in red in Figure 2(a)). As a proof of concept of the type of results that our method allows to obtain, we generate data for these 3 species based on simulations of the nonlinear ordinary differential equation model proposed in [6]. Gaussian noise with variance 0.04 is then added to the collected data to simulate measurement noise in the data set. These noisy data are then fed to our reconstruction algorithm in order to assess its ability to recover the correct network structure.

[1]Due to space limitation, we only put one example here, other examples can be found, for example, in [4]

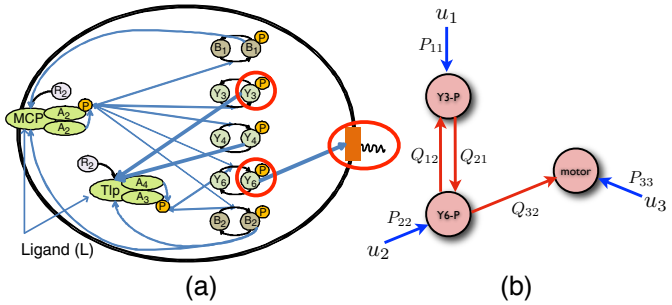

Fig. 2. (a) Network representing the dynamical interaction between the 10 species believed to be responsible for the chemotactic response of *Rhodobacter sphaeroides*. We assume that only species Y_3^p, Y_6^p and "motor" are measured (circled in red). (b) Network connecting the measured states only.

We follow our prescribed experimental protocol where data are collected for each measured species when a step is imposed on the corresponding input while the other inputs are zero. Again, for simplification, only steady-state data (approximated as the final value reached over the considered time interval) are used. Based on the complete network given in Figure 2(a), the correct network that we should aim to recover is presented in Figure 2(b).

Boolean structure	α_k	AICc
[0 0 0 0 0 0]	1.93	7.18
⋮		
[0 1 1 1 1 1]	0.157	29.7
[1 0 0 1 0 0]	0.0309	0.786
[1 1 0 1 0 0]	0.0306	5.56
⋮		
[1 1 0 1 1 1]	0.0287	24.6
[1 0 1 1 0 0]	0.0274	5.22
[1 1 1 1 0 0]	0.0271	12.4
⋮		
[1 1 1 1 1 1]	0.0252	48.2

TABLE I

The binary values in the table are arranged according to the following order $[Q_{21}\ Q_{31}\ Q_{12}\ Q_{32}\ Q_{13}\ Q_{23}]$. The red row indicates the Boolean network obtained as a result of AICc method, while the blue row indicates the true one.

Computing the corresponding distances and AICc values for all the $2^6=64$ possible Boolean networks (Table I), we observe that the network with the smallest AICc (Figure 4(e)) is not the correct network in Figure 2(b). This is not because the method failed but because of the very low signal to noise

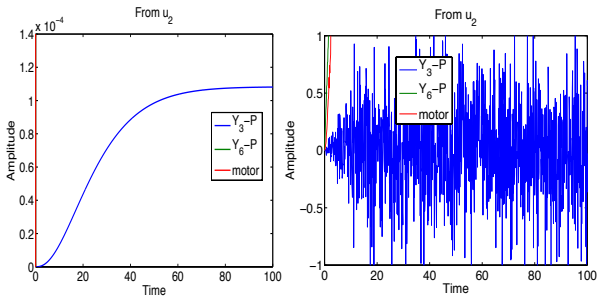

Fig. 3. Left and right are the deterministic and stochastic concentrations of Y_3^p, respectively, in response to a step input in u_2. Note that the amplitude without noise (left) is much weaker than with noise (right), and so the signal information is lost.

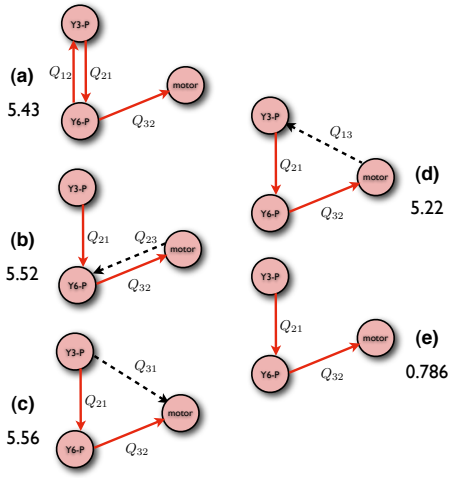

Fig. 4. Candidate networks with AICc values. Black arrows represent spurious connections.

ratio ($0.003 << 1$) observed in the measurements (it is called interampattaness in [18]) when a step is imposed on u_2 (see Figure 3). Y_6^p has a very small influence on Y_3^p since the pathway from Y_6^p to Y_3^p includes a reversible reaction with very small rate constant.

The next set of smallest values of AICc in Table I consists of 4 networks, including the true one. The corresponding candidate networks for the reconstruction are represented in Figure 4. These five candidate networks can then be further discriminated by performing additional and more precise experiments, with reduced noise and increased amplitude of the input step signals (if possible) to help differentiate them.

V. Conclusion and Future Work

This paper proposes a new network reconstruction method in the presence of noise and nonlinearities based on dynamical structure functions. The key idea is to find minimal distances between the existent data and the the data required to obtain particular Boolean networks. The method was illustrated with two biologically-oriented examples. They showed that, even in the presence of nonlinearities and considerable noise, network reconstruction was possible. Eventually, when the signal to noise ratio was too small, reconstruction was no longer possible, but that is true irrespective of the method used.

Obviously, the method has limitations with respect to nonlinearities. With stronger nonlinear terms the method will eventually fail. For example, network reconstruction for oscillatory systems is still an open problem. However, when applied to the reconstruction of various equilibrium point models given in the literature, we observed that reconstruction was always possible when the signal-to-noise ratio of the measured data was not too small (far less than 1).

References

[1] E.D. Sontag, "Network reconstruction based on steady-state data", Essays in Biochemistry, 45:161-176, 2008.

[2] J. Gonçalves and S. Warnick, "Necessary and sufficient conditions for dynamical structure reconstruction of LTI networks", IEEE Transactions on Automatic Control, vol. 53, 2008.

[3] Y. Yuan, G.-B. Stan, S. Warnick and J. Gonçavles, "Minimal dynamical structure realisations with application to network reconstruction from data," in revision, IEEE Transaction on Automatic Control, 2010.

[4] Y. Yuan, G. Stan, S. Warnick and J. Gonçavles, "Robust network reconstruction from data," To appear, Automatica, 2011.

[5] A. Hirotugu, "A new look at the statistical model identification," IEEE Transactions on Automatic Control 19 (6): 716-723, 1974.

[6] M. Roberts, E. August, A. Hamadeh, P. Maini, P. McSharry, J. Armitage and A. Papachristodoulou, "A model invalidation-based approach for elucidating biological signalling pathways, applied to the chemotaxis pathway in R. sphaeroides", BMC Systems Biology, vol. 3., issue 105, 2009.

[7] C. Cosentino, W. Curatola, F. Montefusco, M. Bansal, D. Bernard and F. Amato, "Linear matrix inequalities approach to reconstruction of biological networks," IET system biology, vol. 1, 2007.

[8] L. Ljung. *System Identification–Theory for the User*. Prentice Hall, 1999.

[9] S. Nelander, W. Wang, B. Nilsson, Q. She, C. Pratilas, N. Rosen, P. Gennemark, C. Sander, "Models from experiments: combinatorial drug perturbations of cancer cells," Mol. Syst. Biol., 2008.

[10] A. Julius, M. Zavlanos, S. Boyd, and G. Pappas, "Genetic Network Identification using Convex Programming," IET Systems Biology, 3(3):155-166, May 2009.

[11] G. Wadhams and J. Armitage, "Making sense of it all: Bacterial chemotaxis," Nat Rev Mol Cell Biol 2004, 5:10241037.

[12] N. Young, *An introduction to Hilbert space*. Cambridge university press, 1988.

[13] K. Burnham and D. Anderson, *Model Selection and Inference - A practical information-theoretic approach*. Springer-Verlag, 1998.

[14] K. Zhou, J. Doyle and K. Glover, "*Robust and Optimal Control*," Prentice Hall, 1996.

[15] M. Bansal and D. di Bernardo, "Inference of gene networks from temporal gene expression profiles," IET systems biology, 2006.

[16] T. Gardner, D. di Bernardo, D. Lorenz and J. Collins, "Inferring genetic networks and identifying compound mode of action via expression profiling," Science, 2003.

[17] C. Ward, E. Yeung, T. Brown, B. Durtschi, S. Weyerman, R. Howes, J. Goncalves, H. Sandberg, and S. Warnick, "A Comparison of Network Reconstruction Methods for Chemical Reaction Networks," Foundations of Systems Biology and Engineering, 2009.

[18] T. Nordling and E. Jacobsen, " Interampatteness–a generic property of biochemical networks." IET Syst Biol, 2009.

[19] I. Cantone, L. Marucci, F. Iorio, M. Ricci, V. Belcastro, M. Bansal, S. Santini, M. di Bernardo, D. di Bernardo, M. Cosma, "A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches," Cell, Volume 137, Issue 1, 172-181, 26 March 2009 .