On Sparsity as a Criterion in Reconstructing Biochemical Networks [⋆]

Torbjörn E.M. Nordling [*] Elling W. Jacobsen [*]

[*] *Automatic Control, School of Electrical Engineering, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden (e-mail: tn@kth.se, jacobsen@kth.se).*

Abstract: A common problem in inference of gene regulatory networks from experimental response data is the relatively small number of samples available in relation to the number of nodes/states. In many cases the identification problem is underdetermined and prior knowledge is required for the network reconstruction. A specific prior that has gained widespread popularity is the assumption that the underlying network is sparsely connected. This has led to a flood of network reconstruction algorithms based on subset selection and regularization techniques, mainly adopted from the statistics and signal processing communities. In particular, methods based on ℓ_1 and ℓ_2-penalties on the interaction strengths, such as LASSO, have been widely proposed and applied. We briefly review some of these methods and discuss their suitability for inferring the structure of biochemical networks. A particular problem is the fact that these methods provide little or no information on the uncertainty of individual identified edges, combined with the fact that the identified networks usually have a large fraction of false positives as well as false negatives. To partly overcome these problems we consider conditions that can be used to classify edges into those that can be uniquely determined based on a given incomplete data set, those that cannot be uniquely determined due to collinearity in the data, and those for which no information is available. Apart from providing a label of confidence for the individual edges in the identified network, the classification can be used to improve the reconstruction by employing standard unbiased identification methods to the identifiable edges while employing sparse approximation methods for the remaining network. The method is demonstrated through application to a synthetic network in yeast which has recently been proposed for *in vivo* assessment of network identification methods.

Keywords: Network inference, Regularization, Modelling, Sparse networks

1. INTRODUCTION

Interacting genes, proteins and metabolites form dynamical systems controlling cellular processes (Wolkenhauer et al., 2005). The architecture of these systems varies among organisms, cell types, developmental phases, environmental and epigenetic conditions (Huang et al., 2009). While the genetic code provides the blueprint for the system components, it is the context dependent interactions that generate the specific biological function. A key problem in systems biology is the inference of the direct causal interactions underlying a given function (Wolkenhauer et al., 2009). Network reconstruction based on gene expression data obtained from *in vivo* experiments in which the system is perturbed by known disturbances and the resulting gene activities are measured, can reveal all direct causal interactions existing within a set of observed genes (Hecker et al., 2009; Tegnér and Björkegren, 2007; Gardner and Faith, 2005; Goncalves and Warnick, 2008).

It is important to distinguish network inference, as discussed herein, from the problem of constructing predictive models. In particular, a model with good predictive properties does not by any means need to reflect the true structure of the underlying network (Nordling and Jacobsen, 2009). Similarly, in network inference one is primarily interested in the existence and secondarily in the strength of individual interactions, implying that the predictive properties of the overall network model may be almost arbitrarily poor.

To successfully infer the biochemical network that underly a given biological function of interest, two problems must be resolved. First, sufficiently informative data that allow discrimination between models with different network structures must be recorded. Second, the "correct" network model with a structure including only the active gene interactions must be selected based on the recorded data set. During the last decade the second problem, that of reverse engineering biochemical networks based on given data, has received significant interest in the bioinformatics and systems biology communities and a multitude of traditional estimation methods and novel inference algorithms have been adopted and developed; see *e.g.*, the review articles Hecker et al. (2009); Tegnér and Björkegren (2007); Bansal et al. (2007); Gardner and Faith (2005). Concerning the first problem, the availability of sufficiently informative data for structure discrimination, the major focus has so far been on the fact that

[⋆] This work was supported by the Swedish Research Council (VR).

there usually are fewer response samples available than the number of components in the network. This renders the network reconstruction problem underdetermined, and some form of prior knowledge is hence required to obtain a unique estimate of the network structure. Since it is known that most gene regulatory networks are relatively sparsely connected, various variable selection methods from signal processing, statistics and machine learning have been adopted to infer the sparsest network that can explain the available data set. See the above reviews and references therein.

In this paper we address the problem of identifying network structures based on incomplete data sets. We first briefly review some of the most popular subset selection and regularization methods for variable selection and discuss their suitability for reconstruction of biochemical networks. It is important to emphasize that inference of biochemical networks differ from the typical problems faced in statistics and signal processing, from where most methods originate, in the sense that sparsity is a means for determining the "true" network structure and not an aim in terms of predicting or compressing data with good accuracy using as few variables as possible. In this paper we therefore propose to apply criteria based on sparsity only to those parts of the network for which the available response data do not allow unique and statistically significant determination of the direct interactions. To make this feasible we consider conditions that can be used to divide all potential network interactions into three classes: those that can be uniquely identified with statistical significance, those for which a unique solution can not be determined, and those for which no significant information is available in the given data set. The derived conditions also provide some insight into how experiments should be designed to identify the connectivity of specific genes. Based on the derived conditions, we propose a strategy in which we first identify the statistically significant interactions using standard unbiased identification methods, and then identify the remaining network using LASSO regularization, *i.e.* imposing an ℓ_1-penalty on the interactions in this part of the network.

We start the paper by defining the problem and providing a brief overview of methods for subset selection and regularization. We then derive conditions for uniquely determining a given network edge, *i.e.*, a direct interaction between two given genes, from available data. For the purpose of insight, we first consider the problem in a deterministic setting and then extend the derived conditions to a stochastic setting in which the measurements are assumed to be corrupted by noise. Finally, we demonstrate the usefulness of the proposed method by application to a synthetic network in yeast which recently has been proposed for *in vivo* assessment of network identification methods (Cantone et al., 2009). For comparison we also apply the frequently employed LASSO regularization to the same problem.

2. PROBLEM DESCRIPTION

We consider gene regulatory networks that can be described by a system of linear ordinary differential equations (ODEs), *i.e.* a linear state-space model

$$\frac{d\boldsymbol{x}}{dt}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\left(\boldsymbol{p}(t) - \boldsymbol{f}(t)\right) \tag{1a}$$

$$\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{e}(t). \tag{1b}$$

The state vector $\boldsymbol{x}(t) = [x_1(t), ..., x_n(t)]^T$ is here restricted to only include mRNA abundances of the considered genes, and we use the term gene space to refer to the state-space. The designed input or perturbation vector $\boldsymbol{p}(t) = [p_1(t), ..., p_l(t)]^T$ contains all external factors used to perturb the system by changing the experimental conditions, and is possibly corrupted by unknown perturbations or process errors represented by a random vector $\boldsymbol{f}(t) = [f_1(t), \ldots, f_l(t)]^T$. The observed output or response vector $\boldsymbol{y}(t) = [y_1(t), \ldots, y_o(t)]^T$ contains the measurement of the dependent variables, corrupted by random measurement errors $\boldsymbol{e}(t) = [e_1(t), \ldots, e_o(t)]^T$.

If we directly measure the mRNA level of all genes in the network, then the matrix $\boldsymbol{C}$ is diagonal. Similarly, if we directly perturb the rate of transcription, *i.e.* change in mRNA level, of all genes independently, then also the matrix $\boldsymbol{B}$ is diagonal. In this case we can without restrictions scale our perturbations and responses such that $\boldsymbol{B}$ and $\boldsymbol{C}$ are identity matrices and $o = l = n$.

The aim of network reconstruction is to infer the signed structure of the interaction matrix $\boldsymbol{A}$, corresponding to the signed adjacency matrix of the network represented as a directed graph, based on recorded perturbations and response data. Apart from the signed structure, it is usually of interest to also estimate the size of the non-zero elements of $\boldsymbol{A}$, *i.e.* the strength of the interactions.

We limit ourselves to consider the use of steady-state data only, *i.e.* data recorded after applying a linear combination of constant perturbations and measuring the response of the system when it has reached a new steady-state. However, the results we derive can easily be extended to the case with time-series data if one considers a discrete time model, see *e.g.*, Schmidt et al. (2005). Let $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$ denote the matrix of all measured responses to the m combinations of perturbations stored in the column vectors of $\boldsymbol{P} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{E}, \boldsymbol{F} \in \mathbb{R}^{n\times m}$ be unknown noise realizations, then our data model is

$$\boldsymbol{Y} = -\boldsymbol{A}^{-1}\boldsymbol{P} + \boldsymbol{A}^{-1}\boldsymbol{F} + \boldsymbol{E} = \boldsymbol{G}\boldsymbol{P} - \boldsymbol{G}\boldsymbol{F} + \boldsymbol{E}. \tag{2}$$

where $\boldsymbol{G}$ is the steady-state gain matrix corresponding to the inverse of the interaction matrix. Near a particular stable physiological state the GRN hence acts as a linear mapping from the space of all possible perturbations to the corresponding responses of the measured state variables, $\boldsymbol{G} : \mathbb{R}^n \mapsto \mathbb{R}^n$.

The network reconstruction problem corresponds to finding a solution to

$$\boldsymbol{A}\boldsymbol{Y} = -\boldsymbol{P} + \boldsymbol{R}, \tag{3}$$

such that $\boldsymbol{A}$ has a structure which is consistent with the structure of the network that generated the data. Here $\boldsymbol{R}$ is a residual that should reflect the errors in the measurements and perturbations.

We are here primarily concerned with the situation in which the number of experiments m is less than the number of network nodes n, rendering equation (3) underdetermined. The methodology as such is, however, relevant also for cases in which $m \geq n$.

3. SPARSE IDENTIFICATION METHODS

The problem of determining the structure of the interaction matrix $\boldsymbol{A}$ from the observations $\boldsymbol{Y}$ and $\boldsymbol{P}$ corresponds to the problem of variable selection as considered in machine learning, statistics and signal processing. However, as pointed out above, an important difference is that, in the case of biochemical network reconstruction, variable selection serves as a means for detecting physical variable influences when the problem is underdetermined, while in other contexts the aim is usually to reduce model complexity or compress signals.

For the case in which (3) is underdetermined with $m < n$, the ordinary least squares solution will not be unique, and will in general be overfitted by forcing the residual $\boldsymbol{R}$ to zero. Furthermore, it will usually result in a full matrix $\boldsymbol{A}$ not reflecting the structure of the "true" network. To partly overcome these problems when inferring biochemical networks, for which $m < n$ is more the rule than the exception, it has become common practice to utilize the knowledge that biochemical networks usually are sparse, and based on this search for the sparsest solution $\boldsymbol{A}$ that satisfies (3) (Yeung et al., 2002; Tegnér and Björkegren, 2007).

There are a number of available approaches for sparse approximations, of which stepwise regression, or subset selection, methods and regularization corresponding to convex relaxations are the most commonly employed, see *e.g.* Tropp (2006). Due to space limitation we here only provide a brief overview of some of these methods.

Stepwise regression, or stepwise selection, is based on sequentially picking variables and then stopping the selection according to some criteria like the Aikaike Information Criterion. This includes forward selection methods, such as matching pursuit, backward selection, all subsets and least angle regression (LARS). In these methods each row of $\boldsymbol{Y}$, corresponding to the recorded responses of a given gene, is considered as a regressor and the idea is to sequentially pick those regressors that are most correlated with the regressand. While these methods can succeed in finding the sparsest model (Couvreur and Bresler, 2000), it is well known that they also may fail completely (Chen et al., 1998). Furthermore, several of these methods will often fail to select variables that are important but closely correlated with a previously picked variable. The latter problem is dealt with in LARS (Efron et al., 2004), which has been used with some success to identifying gene regulatory networks, in *e.g.* Madar et al. (2010).

A more robust approach to sparse approximation is offered by regularization methods, such as the LASSO formulation proposed in Tibshirani (1996). The idea here is to use optimization methods to determine the sparsest network that can explain the available data. Thus, ideally one seeks to minimize the ℓ_0 norm of the coefficients in $\boldsymbol{A}$ under a constraint on the squared residual of (3), or minimize the squared residual under a constraint on the ℓ_0 norm of $\boldsymbol{A}$

$$\min_{\boldsymbol{A}} \|\boldsymbol{A}\|_{\ell_0} \qquad \min_{\boldsymbol{A}} \|\boldsymbol{AY}+\boldsymbol{P}\|_{\ell_2} \tag{4}$$

$$\text{s.t. } \|\boldsymbol{AY}+\boldsymbol{P}\|_{\ell_2} \le \lambda \qquad \text{s.t. } \|\boldsymbol{A}\|_{\ell_0} \le \frac{1}{\lambda}. \tag{5}$$

Alternatively a weighted unconstrained formulation is used

$$\min_{\boldsymbol{A}} \|\boldsymbol{AY}+\boldsymbol{P}\|_{\ell_2} + \lambda \|\boldsymbol{A}\|_{\ell_0}. \tag{6}$$

These problems are however computationally intractable since the ℓ_0-norm causes a combinatorial explosion and are therefore typically relaxed using the ℓ_1 norm instead. The use of the ℓ_1-norm, as in LASSO, makes the problem convex and computationally attractive even for large scale problems. However, there is no longer any guarantee that the solution will correspond to the sparsest network. Conditions on the network and data for which the solution of the relaxed problem converges to the sparsest network are given in Candes et al. (2006), but these are in practice impossible to evaluate in the case of biochemical networks. A number of variations of the LASSO problem have been proposed, including elastic net methods (Zou and Hastie, 2005) and iterative weighted relaxations (Julius et al., 2009).

Relaxation methods have frequently been employed to infer gene regulatory networks, see *e.g.* Hecker et al. (2009); Tegnér and Björkegren (2007); Gardner and Faith (2005). However, as is clear from many studies, such as the DREAM challenges aimed at benchmarking different reconstruction methods (Marbach et al., 2010), inference of gene regulatory networks usually result in a large fraction of false negatives and false positives (Stolovitzky et al., 2009). This is partly related to the limited information content in the available data, but undoubtedly also to the algorithms employed for inference.

While sparse approximation methods appear attractive for biochemical network reconstruction, for which there is often a lack of data, it is important to point out that several potential pitfalls exist. The first problem is related to the rather obvious fact that the assumption that the sparsest network explaining available response data corresponds to the "true" network does not necessarily hold. The second problem stems from the fact that determination of the sparsest network is a combinatorial problem which can only be solved for small networks, and hence one must resort to greedy search methods or convex relaxations as discussed above. Although there exists theoretical results on when the relaxed problem converges to the original solution, these are hard to evaluate in practice and hence one can almost never guarantee that the methods are able to provide the sparest network. In applications to biochemical networks, one also typically finds that the methods provide a relatively large number of false positives, *i.e.*, identification of edges absent in the "true" network, and false negative, *i.e.*, missed edges present in the "true" network, see *e.g.* Stolovitzky et al. (2009). Finally, the methods provide little or no information on the probability of individual edges being correctly identified.

The are several possible remedies to the above mentioned problems. The first is additional *a priori* information, such as the probability of existence of specific edges, provided that such information is available. Second, one can perform additional experiments to increase the information content in $\boldsymbol{Y}$ and $\boldsymbol{P}$. However, before requesting more information one should analyze the information content in the available data more in detail rather than simply applying brute force methods. In particular, one should consider whether the available data allow for a unique and precise identification of parts of the network, and if this is the case only apply the sparsity criterion to those parts of the network for which sufficient information for a unique identification is lacking. The latter is considered below. We first consider the deterministic case to derive precise conditions for identifiability of individual network edges, and then expand the results to the case with noisy measurements.

4. CONDITIONS FOR INFERENCE FROM NOISE-FREE DATA

For the noise-free case with $m < n$, the network reconstruction problem corresponds to solving the underdetermined system of linear equations $\boldsymbol{AY} + \boldsymbol{P} = \boldsymbol{0}$ such that the structure of $\boldsymbol{A}$ corresponds to the structure of the network that generated the data. This problem may be solved for each row of $\boldsymbol{A}$ independently using the standard regression formulation

$$\underbrace{\boldsymbol{Y}^T}_{\triangleq\boldsymbol{\Phi}}\underbrace{\boldsymbol{A}_j^T}_{\triangleq\boldsymbol{\theta}} = \underbrace{-\boldsymbol{P}_j^T}_{\triangleq\boldsymbol{\xi}}. \tag{7}$$

Here $\boldsymbol{A}_j$ denotes the jth row of the unknown network matrix and $\boldsymbol{P}_j$ denotes the jth row of the known perturbation matrix. We introduce the regressor matrix $\boldsymbol{\Phi}$, such that each regressor $\boldsymbol{\phi}_i$ is the ith row of the output matrix $\boldsymbol{Y}$, the parameter vector $\boldsymbol{\theta} = \boldsymbol{A}_j^T$ and the regressand $\boldsymbol{\xi} = -\boldsymbol{P}_j$ in order to simplify the notation and to make the analogy to standard regression problems clear. Based on (7) we are in a position to apply well known results from linear algebra to deduce properties of the solution.

Our aim is to determine conditions on the data $\boldsymbol{Y}$ and $\boldsymbol{P}$, corresponding to the regressor $\boldsymbol{\Phi}$ and regressand $\boldsymbol{\xi}$, which allows a unique determination of an individual parameter θ_i, corresponding to an element of $\boldsymbol{A}$ representing a single edge in the network. Such a condition can be derived based on results concerning linear independence of the regressor vectors.

Theorem 1. Consider the linear regression problem $\boldsymbol{\Phi\theta} = \boldsymbol{\xi}$ with $\boldsymbol{\Phi} \in \mathbb{R}^{m\times n}$, $\boldsymbol{\theta} \in \mathbb{R}^n$ and $\boldsymbol{\xi} \in \mathbb{R}^m$. Let $\boldsymbol{\Phi}_{j\neq i}$ and be the matrix obtained by removing row i in $\boldsymbol{\Phi}$ and $\boldsymbol{T}_{j\neq i} = \boldsymbol{\Phi}_{j\neq i}\left(\boldsymbol{\Phi}_{j\neq i}^T\boldsymbol{\Phi}_{j\neq i}\right)^{-1}\boldsymbol{\Phi}_{j\neq i}^T$ be the projection matrix onto the linear subspace spanned by $\boldsymbol{\Phi}_{j\neq i}$. Then the coefficient θ_i can be uniquely determined if and only if $(\boldsymbol{I} - \boldsymbol{T}_{j\neq i})\,\boldsymbol{\phi}_i \neq 0$. In particular, $(\boldsymbol{I} - \boldsymbol{T}_{j\neq i})\boldsymbol{\phi}_i\theta_i = (\boldsymbol{I} - \boldsymbol{T}_{j\neq i})\boldsymbol{\xi}$.

Proof. The sufficient condition follows trivially from the observation that if the regressor $\boldsymbol{\phi}_i$ cannot be expressed as a linear combination of all the other regressors, *i.e.*, spans a unique direction in the column space of $\boldsymbol{\Phi}$, then the corresponding direction in the regressand $\boldsymbol{\xi}$ is uniquely described by $\boldsymbol{\phi}_i$. The necessary condition follows from the fact if any subset is linearly dependent, then there exist an infinite number of linear combinations that yield the same projection on $\boldsymbol{\xi}$.

Thus, if a regressor $\boldsymbol{\phi}_i = \boldsymbol{Y}_i^T$, corresponding to the observation vector of gene i, cannot be expressed as a linear combination of all the other gene regressors $\boldsymbol{\phi}_{j\neq i}$, then we say that it is linearly independent and the corresponding coefficients a_{ki} $\forall k$ in column i of $\boldsymbol{A}$ can then be uniquely determined from the observations $\boldsymbol{Y}$ and $\boldsymbol{P}$. If also the corresponding regressand $\boldsymbol{\xi} = \boldsymbol{P}_k^T$ is linearly independent of the gene regressors $\boldsymbol{\phi}_{j\neq i}$, then there exist an edge from gene i to gene k, while if $\boldsymbol{\xi} = \boldsymbol{P}_k^T$ can be expressed as a linear combination of the regressors in $\boldsymbol{\phi}_{j\neq i}$, then we say that it is linearly dependent and conclude that no directed edge exist from gene i to gene k. Finally, if a regressor $\boldsymbol{\phi}_i$ is a null vector then no information exist in the available data concerning the existence of edges to gene i, and one might therefore as well set $a_{ki} = 0$ $\forall k$. In summary, we can based on the available samples for the underdetermined problem classify the network edges into four groups:

(1) existing edges with non-zero weights
$(\boldsymbol{I} - \boldsymbol{T}_{j\neq i})\boldsymbol{\phi}_i \neq \boldsymbol{0}$ and $(\boldsymbol{I} - \boldsymbol{T}_{j\neq i})\boldsymbol{\xi} \neq \boldsymbol{0}$,
(2) non-existing edges
$(\boldsymbol{I} - \boldsymbol{T}_{j\neq i})\boldsymbol{\phi}_i \neq \boldsymbol{0}$ and $(\boldsymbol{I} - \boldsymbol{T}_{j\neq i})\boldsymbol{\xi} = \boldsymbol{0}$,
(3) uncertain edges with possible non-zero weights
$(\boldsymbol{I} - \boldsymbol{T}_{j\neq i})\boldsymbol{\phi}_i = \boldsymbol{0}$,
(4) edges for which no information exist in the available data set
$\boldsymbol{\phi}_i = \boldsymbol{0}$.

Note that sparse approximation algorithms in general do not utilize the above information explicitly. However, since in the deterministic case it is reasonable to force the residual to zero, most algorithms will implicitly utilize the above information since they in principle are based on using the null space of the regressor matrix $\boldsymbol{\Phi}$ to minimize the ℓ_1-norm of the coefficient matrix. This implies that they in the deterministic case typically will identify edges belonging to class (1) and (2) correctly, set edges in class (4) to zero and then minimize the ℓ_1-norm of edges in class (3) using the degrees of freedom offered by the null space of $\boldsymbol{\Phi}$. Thus, in a deterministic setting the main use of the derived conditions is that it provides explicit information to the user on which edges that have been identified with certainty, which edges that are uncertain and which edges that have been set to zero due to lack of information in the available data. However, in a more realistic setting with measurements corrupted by noise, the conditions will prove useful also in the identification procedure since statistical testing can be used to classify the edges according to the above scheme prior to applying stepwise regression or convex optimization. This is discussed in more detail below. We first illustrate the results with a simple example.

We consider a 5-gene network engineered in yeast cells by Cantone et al. (2009) with the specific aim of evaluating network reconstruction algorithms. The network has a known structure

$$\boldsymbol{A} = \begin{pmatrix} -1 & 0 & 1 & -1 & 0 \\ 1 & -1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & -1 & 1 & 0 & -1 \end{pmatrix}, \tag{8}$$

The network is illustrated in Fig. 1. Note that the methods discussed in the paper are relevant for large scale networks with hundreds or thousands of nodes, but that this simple example serves to illustrate the problems and methodologies well. We perturb the 2nd, 3rd and 5th gene (GAL4, SWI5, GAL80) in three independent transcrip-

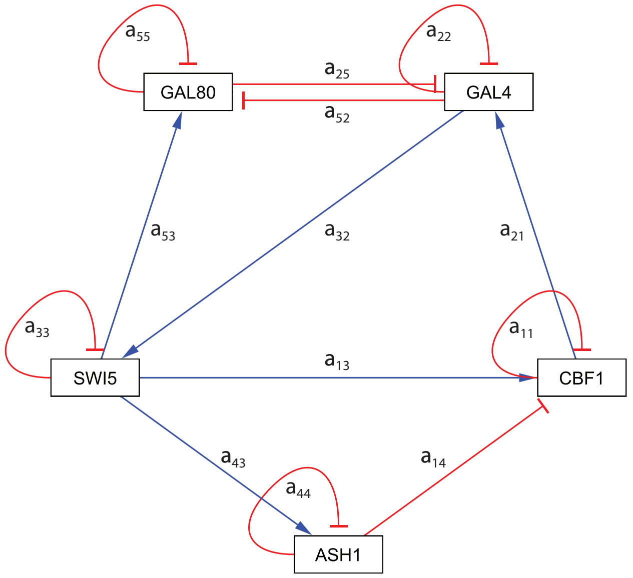

Fig. 1. True IRMA network: Structure of the network with the corresponding element of the interaction matrix $\boldsymbol{A}$ marked on each edge.

tional perturbation experiments and record the response, which gives the regressor matrix

$$\boldsymbol{\Phi} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \\ 0 & -1 & -1 & -1 & 1 \end{pmatrix}, \tag{9}$$

We consider identification of the third row of $\boldsymbol{A}$, corresponding to edges affecting gene 3 (SWI5), with the corresponding regressand.

$$\boldsymbol{\xi} = (\, 0 \ -1 \ 0 \,)^T. \tag{10}$$

Using the conditions in Theorem 1 we find that regressors $\boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_5$ are linearly independent. Thus, the data contain sufficient information to infer the directed edges from gene 2 and gene 5 to gene 3. For regressor $\boldsymbol{\phi}_2$ we find that also $\boldsymbol{\xi}$ is linearly independent of the remaining regressors, and hence the corresponding edge belongs to class (1). For regressor $\boldsymbol{\phi}_5$ we find that $\boldsymbol{\xi}$ is linearly dependent on the remaining regressors, and hence the edge belongs to class (2), *i.e.*, it does not exist. Regressor $\boldsymbol{\phi}_1$ is the null vector and hence belongs to class (4), *i.e.* we cannot deduce any information from the data concerning the existence of an edge from gene 1 to gene 3. Regressors $\boldsymbol{\phi}_3$ and $\boldsymbol{\phi}_4$ are linearly dependent and hence the corresponding edges from genes 3 and 4 to gene 3 belong to class (3), and we need extra information to make a certain decision as to their existence. If we minimize the ℓ_0-norm of the coefficient matrix, using a combinatorial search, we find

$$\boldsymbol{\theta}^{\mathcal{R}} = (\, 0 \ 1 \ 0 \ -1 \ 0 \,)^T, \tag{11}$$

which is as sparse as the true network, but still fails to correctly identify the edges from genes 3 and 4.

In the example above we found that we after 3 experiments had obtained a regressor matrix with two linearly independent regressors. Due to space limitations we do not derive any detailed conditions here on when perturbation experiments on sparse networks yield linearly independent regressors in the deterministic case. However, two obvious cases where one will obtain linearly independent regressors are (i) when one in m experiments obtain response in m genes only using linearly independent perturbation vectors, *i.e.*, the remaining genes show no steady-state response, and (ii) when one perturb the m_i genes which have edges from gene i in m_i linearly independent perturbation experiments. The latter condition was fulfilled for genes 3 and 5 in the example above.

For the case with measurement noise we will in general not obtain any linearly independent regressors until we have performed n linearly independent perturbation experiments, corresponding to the number of genes in the network. However, in this case one should use some form of statistical hypotheses testing to evaluate if the underlying noise-free regressor in fact is linearly independent. This is discussed next.

4.1 Conditions for classification based on noisy data

The conditions in Theorem 1 are based on linear independence of the rows of the response matrix $\boldsymbol{Y}$. In general, a row $\boldsymbol{Y}_i$ is linearly independent of the other rows if the rank of $\boldsymbol{Y}_{j\neq i}$, obtained by removing row $\boldsymbol{Y}_i$, is less than the rank of $\boldsymbol{Y}$. Consider now the case in which we measure the mRNA abundances with some error, here modelled as normally distributed noise $\boldsymbol{E}$ such that

$$\boldsymbol{Y} = \boldsymbol{Y}_0 + \boldsymbol{E}, \quad \boldsymbol{Y}, \boldsymbol{Y}_0, \boldsymbol{E} \in \mathbb{R}^{n\times m} \tag{12}$$

where $\boldsymbol{Y}_0$ denotes the true mRNA abundances. If the elements of the noise matrix $\boldsymbol{E}$ are normally distributed, it will in general be a full matrix with full rank, and with all submatrices also having full rank. This implies that $\boldsymbol{Y}$ in general will share the same properties, and any submatrix obtained by removing a row from $\boldsymbol{Y}$ will hence have rank $r = min(n-1, m)$. Thus, only if $m \geq n$ will the rank drop when removing a row, and hence no regressors will be linearly independent until we have collected as many samples as there are nodes (genes) in the network. However, as shown above, the underlying noise free matrix $\boldsymbol{Y}_0$ will often have linearly independent rows. Hence, we need to employ statistical hypotheses testing to determine if a row of $\boldsymbol{Y}_0$ is linearly independent based on knowledge of $\boldsymbol{Y}$ and some properties of the noise $\boldsymbol{E}$.

There exist a number of statistical methods for determining the rank based on noisy data. We will here employ a simple hypothesis test based on the singular values of the regressor matrix, which works if the separation between the singular values of $\boldsymbol{Y}$ and $\boldsymbol{E}$ is sufficiently large. Consider again equation (12). According to the Mirsky Theorem in Stewart (1990) the root mean square (RMS) of the error in the singular values of $\boldsymbol{Y}$ is bounded by the RMS of the singular values of the error matrix $\boldsymbol{E}$. Thus, one can with some confidence state that $\boldsymbol{Y}_0$ is unlikely to be rank deficient if the RMS of the singular values of the error matrix is less than the smallest singular value of $\boldsymbol{Y}$. However, this condition can be quite conservative and hence not too useful to test for rank deficiency. Assuming that the elements of $\boldsymbol{E}$ are normally distributed with zero mean and variance ϵ^2 and the singular values simple, Stewart (1990) derive an expression for the expected sum of squared errors $\Delta\sigma_i^2$ on the singular values of $\boldsymbol{Y}$

$$\mathbf{E}\left[\sum_i \Delta\sigma_i^2\right] = m\epsilon^2. \tag{13}$$

Thus, to test for linear independence we can formulate a simple hypothesis test based on the singular values of the matrix of interest. We here choose to test the singular values against the threshold $\sqrt{m\epsilon^2}$.

For the 5 gene network in yeast studied above we first consider the same three experiments as before, but now add normally distributed noise with variance $\epsilon^2 = 0.01$ to the measurements. The noise, combined with the fact that $m < n$, implies that none of the resulting regressors are linearly independent. Also, the regressor for gene 1 is not a null vector anymore. However, the variance in gene 1 is not found to be statistically significant and it is determined to be a null vector. Thus, before proceeding we perform an additional experiment in which gene 1 is perturbed, resulting in all regressors being significantly different from zero. The resulting regressor matrix is

$$\boldsymbol{\Phi} = \begin{pmatrix} 0.077 & 0.913 & 0.989 & 1.08 & 0.275 \\ -0.167 & -0.984 & 0.019 & -0.147 & 1.00 \\ -0.132 & -1.05 & -0.906 & -1.05 & 1.01 \\ 1.29 & 1.06 & 1.01 & 1.01 & 0.035 \end{pmatrix}. \tag{14}$$

If the number of singular values below the threshold $\sqrt{m\epsilon^2} = 0.2$ is larger in $\boldsymbol{\Phi}_{j\neq i}$ than in $\boldsymbol{\Phi}$, then we consider regressor $\boldsymbol{\phi}_i$ to be linearly independent. From this condition we find that regressors $\boldsymbol{\phi}_1$, $\boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_5$ are linearly independent. Thus, columns $1, 2$ and 5 of $\boldsymbol{A}$, corresponding to edges from the corresponding genes, can be determined uniquely from the data. Based on a similar rank test with the corresponding rows of $\boldsymbol{\Phi}$ replaced by $\boldsymbol{\xi}$ we determine edges that are identically zero, or non-existent. For the two columns of A with collinear regressors we use the LASSO convex relaxation to determine a sparse solution. The resulting identified network interaction matrix is

$$\hat{\boldsymbol{A}} = \begin{pmatrix} -0.84 & 0 & -0.52 & 0.58 & 0 \\ 0.78 & -1.01 & 0.10 & 0 & -0.86 \\ 0 & 1.26 & 0.28 & -1.52 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -0.99 & 0.54 & 0.58 & -0.92 \end{pmatrix}. \tag{15}$$

By comparing with the true interaction matrix in (8) we see that we have obtained a good fit of the structure as well as the magnitude for columns $1, 2$ and 5. For columns 3 and 4 the LASSO method yields a solution which has sparsity on par with the true network, but with several false positives and negatives. Note that row 4 of the identified matrix is a null vector and that of these only elements $1, 2$ and 5 are identified with statistical significance while elements 3 and 4 are zero since this is the sparsest solution satisfying $\boldsymbol{\xi} = \mathbf{0}$.

For comparison we also employ the Lasso method to the same data, tuning the parameter λ to avoid overfitting while making a sound trade-off between sparsity and residuals. The resulting identified interaction matrix is

$$\hat{\boldsymbol{A}} = \begin{pmatrix} -0.81 & 0.06 & 0 & 0 & 0 \\ 0.67 & -0.82 & 0 & 0 & -0.72 \\ 0 & 1.09 & -1.07 & 0 & 0.12 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -0.91 & 0 & 1.06 & -0.79 \end{pmatrix}. \tag{16}$$

We see that we now get false positives and/or negatives in almost all columns, including those for which the corresponding edges belong to class (1) and (2) in the classification defined above. Note that we did not get a better fit of the structure for any other choice of λ. The two identified networks are shown in Fig. 2.

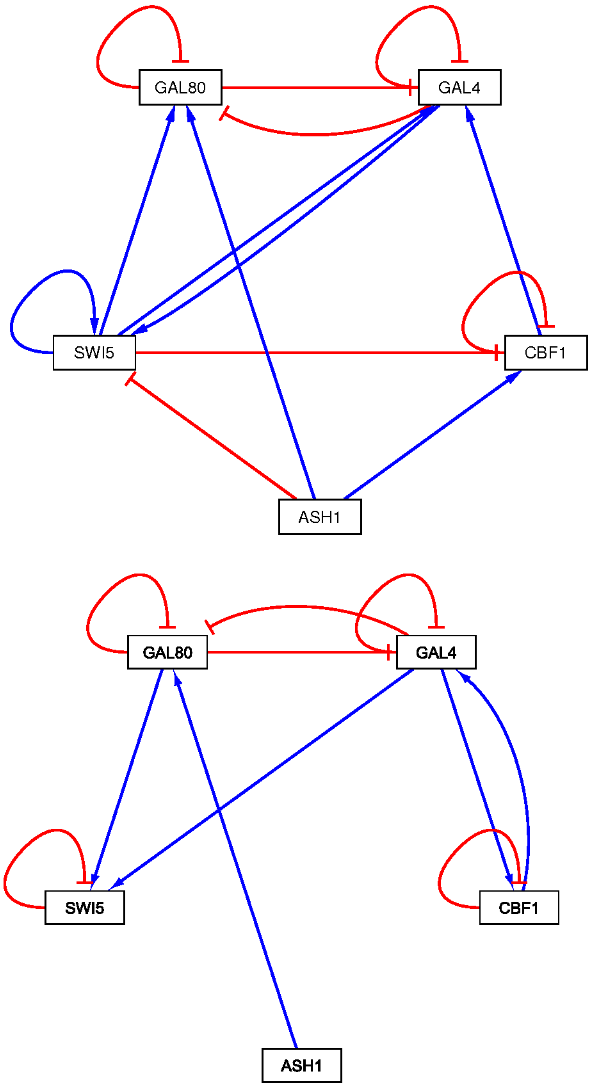

Fig. 2. Identified networks: Inferred structure using the proposed method (top) and lasso (bottom).

In summary we note that the advantages with the method proposed here, in which the network reconstruction is divided into two steps with standard identification of the identifiable parts of the network followed by sparse approximation of the remaining network, partly is an improved inference precision and partly information on the quality of the individual fitted edges.

5. SUMMARY AND CONCLUSIONS

Various methods for sparse approximation has gained popularity for reconstructing biochemical networks, mainly based on the knowledge that biochemical networks typically are sparsely connected combined with the fact that there usually are relatively few samples available for identification. However, the experience is that these methods often yield a relatively large fraction of falsely identified edges. Furthermore, the methods provide little or no information concerning the confidence in which the individual edges have been identified. To partly overcome these problems we have in this paper considered dividing the problem into two steps. In a first step we use standard tools from linear algebra to determine linearly independent regressors, and based on that the edges that can be uniquely determined from the available data. These edges are then fitted using standard unbiased identification methods. In a second step we employ sparse approximation, such as LASSO, to determine a sparse solution for the remaining network. The efficiency of the proposed method was demonstrated on a five gene synthetic network in yeast which has been developed to evaluate network reconstruction methods. Apart from yielding improved results over LASSO applied to the complete network, we highlight that a significant advantage is that the solution is accompanied by a label of confidence for the individual edges in the identified network.

REFERENCES

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Mol Syst Biol*, 3, 78. doi: 10.1038/msb4100120.

Candes, E.J., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2), 489–509. doi: 10.1109/TIT.2005.862083.

Cantone, I., Marucci, L., Iorio, F., Ricci, M.A., Belcastro, V., Bansal, M., Santini, S., di Bernardo, M., di Bernardo, D., and Cosma, M.P. (2009). A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, 137(1), 172–81. doi:10.1016/j.cell.2009.01.055.

Chen, S.S., Donoho, D.L., and Saunders, M.A. (1998). Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific Computing*, 20(1), 33–61.

Couvreur, C. and Bresler, Y. (2000). On the Optimality of the Backward Greedy Algorithm for the Subset Selection Problem. *SIAM Journal on Matrix Analysis and Applications*, 21(3), 797.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, 32(2), 407 – 451.

Gardner, T.S. and Faith, J.J. (2005). Reverse-engineering transcription control networks. *Physics of Life Reviews*, 2(1), 65–88. doi:10.1016/j.plrev.2005.01.001.

Goncalves, J. and Warnick, S. (2008). Necessary and Sufficient Conditions for Dynamical Structure Reconstruction of LTI Networks. *Automatic Control, IEEE Transactions on*, 53(7), 1670–1674. doi: 10.1109/TAC.2008.928114.

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models-a review. *Bio Systems*, 96(1), 86–103. doi: 10.1016/j.biosystems.2008.12.004.

Huang, S., Ernberg, I., and Kauffman, S. (2009). Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. *Seminars in cell & developmental biology*, 20(7), 869–76. doi: 10.1016/j.semcdb.2009.07.003.

Julius, A., Zavlanos, M., Boyd, S., and Pappas, G.J. (2009). Genetic network identification using convex programming. *IET systems biology*, 3(3), 155–66. doi: 10.1049/iet-syb.2008.0130.

Madar, A., Greenfield, A., Vanden-Eijnden, E., and Bonneau, R. (2010). DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator. *PloS one*, 5(3), e9803. doi: 10.1371/journal.pone.0009803.

Marbach, D., Prill, R.J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences of the United States of America*, 107(14), 6286–91. doi:10.1073/pnas.0913357107.

Nordling, T.E.M. and Jacobsen, E.W. (2009). Interampatteness–a generic property of biochemical networks. *IET Syst Biol*, 3(5), 388–403. doi:10.1049/iet-syb.2009.0008.

Schmidt, H., Cho, K.H., and Jacobsen, E.W. (2005). Identification of small scale biochemical networks based on general type system perturbations. *FEBS J*, 272(9), 2141–2151. doi:10.1111/j.1742-4658.2005.04605.x.

Stewart, G.W. (1990). Perturbation Theory for the Singular Value Decomposition. *SVD and Signal Processing, II: Algorithms, Analysis and Applications*, 99–109.

Stolovitzky, G., Kahlem, P., and Califano, A. (2009). The challenges of systems biology. Preface. *Annals of the New York Academy of Sciences*, 1158, ix–xii. doi: 10.1111/j.1749-6632.2009.04470.x.

Tegnér, J. and Björkegren, J. (2007). Perturbations to uncover gene networks. *Trends Genet*, 23(1), 34–41. doi: 10.1016/j.tig.2006.11.003.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58(1), 267–288.

Tropp, J. (2006). Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3), 1030–1051. doi: 10.1109/TIT.2005.864420.

Wolkenhauer, O., Fell, D., De Meyts, P., Blüthgen, N., Herzel, H., Le Novère, N., Höfer, T., Schürrle, K., and van Leeuwen, I. (2009). SysBioMed report: advancing systems biology for medical applications. *IET Syst Biol*, 3(3), 131–136. doi:10.1049/iet-syb.2009.0005.

Wolkenhauer, O., Ullah, M., Wellstead, P., and Cho, K.H. (2005). The dynamic systems approach to control and regulation of intracellular networks. *FEBS Lett*, 579(8), 1846–1853. doi:10.1016/j.febslet.2005.02.008.

Yeung, M.K., Tegnér, J., and Collins, J.J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proc Natl Acad Sci U S A*, 99(9), 6163–6168. doi:10.1073/pnas.092576199.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society B*, 67, 301–320.