

Reconstructing Hidden Regulatory Layers by Network Component Analysis: Theory and Application

Riccardo Boscolo, Chiara Sabatti, James C. Liao, and Vwani P. Roychowdhury

Abstract

The authors recently introduced a framework, named Network Component Analysis (NCA), for the reconstruction of the dynamics of transcriptional regulators activities from gene expression assays. In this paper, our goal is to characterize NCA as a general purpose network and signal reconstruction technique: given only the noisy output signals of a multi-dimensional linear system and certain a-priori knowledge about the connectivity among the inputs and the outputs, the method is capable of reconstructing both the input signals as well as the unknown connectivity coefficients. In particular, the following aspects of NCA are investigated: I) The sufficient conditions on the a-priori connectivity information (required for successful reconstructions via NCA) are made less stringent, allowing easier verification of whether a network topology is identifiable, as well as extending the class of identifiable systems. II) We show that the two-stage least square iterative procedure used in NCA identifies stationary points of the likelihood function, under gaussian noise assumption. III) A framework for the simultaneous reconstruction of multiple regulatory sub-networks is introduced, thus overcoming one of the limitations of the original formulation of the decomposition, occurring for small sample size data. A set of monte carlo simulations with synthetic data suggest that the approach is indeed capable of accurately reconstructing regulatory signals when these are the input of large-scale networks that satisfy the suggested identifiability criteria, even under fairly noisy conditions. The sensitivity of the reconstructed signals to inaccuracies in the hypothesized network topology is also investigated. The results obtained in [1] by applying NCA to experimental gene expression measurements of the bacterium *Escherichia coli* are extended to the

Riccardo Boscolo and Vwani P. Roychowdhury are with the Electrical Engineering Department, University of California Los Angeles (e-mail: {riccardo,vwani}@ee.ucla.edu).

Chiara Sabatti is with the Department of Human Genetics and Statistics, University of California Los Angeles (e-mail: csabatti@mednet.ucla.edu).

James C. Liao is with the Department of Chemical Engineering, University of California Los Angeles (e-mail: liaoj@ucla.edu).

reconstruction of multiple regulatory sub-networks with partially overlapping sets of transcriptional regulators.

Index Terms

System identification, Biology and genetics, Network models, Data analysis.

I. INTRODUCTION

Recent advances in biotechnology have resulted in the introduction of high-throughput techniques for the measurement of biological signals. An example of such technologies is DNA microarray assays [2], which allow simultaneous monitoring of the expression levels of several thousand genes in an organism. Such increase in the amount of data made available to biologists has driven a parallel effort aimed at developing information processing tools required to analyze such data sets. Accordingly, the last few years have witnessed a rapid increase in the introduction of new statistical tools and associated computational methodologies for the analysis as well as modeling of biological systems [3]–[5].

The task of extracting information about the structure and dynamics of intra- and inter-cellular processes from these large-scale data sets has, however, proven to be difficult, and the reasons are quite apparent: the observed signals are the outputs of a complex stochastic dynamical system involving a considerable number of hidden factors. The approaches that have been adopted so far can be broadly classified into two categories, namely, non-parametric and parametric. In the non-parametric case, no generative model for the signals is assumed, and inferences about intracellular mechanisms are based on different measures of dependencies (or lack thereof) among the signals themselves. For example, works based on clustering of genes using their expression profiles [6] and on extracting regulatory information using Bayesian statistics [7] fall under this category. While these methods have been applied successfully to the problem of finding patterns of co-expression between genes, it is well known that such methods are limited in the fact that they do not provide means for including specific biological modeling assumptions that could be derived from available a-priori knowledge on the system under study.

In a parametric approach, on the other hand, an a priori set of generative models is assumed and the parameters of these models are estimated from the data set. One such example is provided in [8], where the experiment design involves measurements of the outputs of a targeted pathway in

response to perturbations in the inputs, and the objective concerns the determination of causative links among the input and output signals. A linearized state-space model for the underlying mechanism is assumed and an overall sparsity constraint on the connectivity is imposed, *viz.* each output signal is influenced by at most K out of N perturbed inputs. The inference procedure is then centered around the issue of selecting a sparse connectivity pattern that best fits the data.

In many experiments one does not have the capability of selectively and precisely perturbing a significant number of the input signals. Instead, one has access to only a subset of the system output signals measured under different conditions, which have the overall effect of perturbing the physiological state of the cells in unidentified ways. In DNA microarray experiments, for example, one can monitor the expression level of genes, which are modulated by a hidden set of transcriptional regulatory mechanisms, including inter-cellular and intra-cellular signaling, co-activation mechanisms, and competitive binding, just to mention a few. It is unclear how one might infer characteristics of the hidden mechanisms and signals from only the expression data. Even if one were to assume a linearized model for the interactions among the activated transcription factors and the genes, one will be left with the intractable problem of simultaneously estimating the parameters of the linearized system (*i.e.* the input-output connectivity and the related strengths), as well as the hidden regulatory signals. Hence, one needs to impose further constraints on the linearized model to make the inverse problem solvable.

In [1], [9] the authors introduced a data decomposition technique, named Network Component Analysis (NCA), which uses a certain type of a-priori knowledge about the connectivity pattern among the input and output signals, in order to reconstruct both the network input signals and the strengths of its connections, when only the output signals are accessible. In the case of transcriptional regulation, the a-priori knowledge on transcription factor (TF) binding sites affinity provides information about whether a particular TF-gene regulatory link is significant or not. For example, if the activated form of a TF is known not to bind significantly to the binding sites in the promoter region of a gene, then one can set the corresponding parameter in the linearized model to zero. Data on potential TF-gene interactions (or absence thereof) is readily available for several prokaryotes as well as eukaryotes in the form of publicly available databases. Thus, the NCA decomposition technique effectively combines available structural data with the measurement of the outputs of the system (in this case the gene expression levels), in order to estimate certain hidden quantities of the regulatory network, namely the activity levels of the major transcriptional

regulators, as well as the relative control strength they exercise on different genes. It is worthwhile to note that while well-known linear decomposition techniques, such as Principal Component Analysis (PCA) [10] or Independent Component Analysis [11], [12] have found application in the analysis of certain gene expression and other biological data sets [13]–[15], such methods are not suitable for the problem of recovering hidden mechanisms, as defined in this paper. Rather than using a priori biological knowledge to make the decomposition problem solvable, these methods impose mathematical/statistical constraints on the input signals (*viz.*, orthogonality for PCA and statistical independence for ICA) and yield potentially dense connectivity patterns, both of which are not representative of the underlying biological system.

In this paper, we derive certain novel results on the estimation problem associated with NCA and demonstrate its wide applicability as a generalized decomposition technique. In particular, we analyze systematically for the first time the performance characteristics of the NCA decomposition as a function of the measurement error as well as of the complexity of the regulatory network, in the ideal scenario when the model linearity is not violated. The results obtained in a large-scale set of monte carlo simulations demonstrate indeed the efficacy of the method even in those cases when the modeling assumptions are subject to noise perturbations.

In Section III, we analyze the sufficient conditions for identifiability, *i.e.*, when does one have enough prior knowledge, in terms of the absences of links in the networks, so that one can perform the intended decomposition. We derive a new and reduced set of sufficient conditions for system identifiability. We then establish a statistical framework for NCA and show that the iterative method introduced in [9] for estimating the regulatory signals and the connection strengths from the observed data is equivalent to a particular case of Maximum Likelihood (ML) estimation technique (Section IV). In Section V, we provide a systematic study of the performance of NCA by simulating different types of transcriptional networks with synthetically generated input data, under various noise levels, in a large-scale settings. Based on the newly derived set of conditions, we introduce an approach which aims at overcoming certain limitations of the original formulation, due to either insufficient sample size or incomplete connectivity information. In particular, we demonstrate how from the same set of data, one can consistently estimate multiple sub-networks. For example, in [1] we reported the application of NCA to a single regulatory sub-network involving 11 transcriptional factors and 100 genes of the bacterium *Escherichia Coli* (*E.coli*). Using the same data set, we show how the parameters of several

overlapping subnetworks can be consistently estimated, therefore extending the analysis to several major transcriptional regulators of the organism under study (a total of 37 TFs controlling 237 genes).

II. THE MODEL

For a given biological system, we can describe the non-linear relationship between a set of L unknown input signals $\{p_1, \dots, p_L\}$ and a set of N measurable output signals $\{e_1, \dots, e_N\}$ as follows:

$$e_n(t_m) = \mathcal{F}(\alpha_1, \dots, \alpha_K; p_1(t_m), \dots, p_L(t_m)), \quad n = 1, \dots, N, \quad m = 1, \dots, M, \quad (1)$$

where $\alpha_1, \dots, \alpha_K$ is a set of parameters of the non-linear model, and t_1, \dots, t_M is a discrete set of time points for which the system dynamics are assumed to be in *quasi steady-state*. Let us now consider the following linear approximation of (1):

$$e_n(t_m) = \sum_{l=1}^L a_{nl} p_l(t_m) + \gamma_n(t_m), \quad n = 1, \dots, N, \quad m = 1, \dots, M, \quad (2)$$

where $\gamma_n(t_m)$ is an error term that incorporates both model inaccuracies and measurement noise. Equation (2) can be expressed in a matrix form as follows:

$$\boxed{E = AP + \Gamma}, \quad (3)$$

where $E = [e_n(t_m)]$ (size: $N \times M$), $A = [a_{nl}]$ (size: $N \times L$), $P = [p_l(t_m)]$ (size: $L \times M$), and $\Gamma = [\gamma_n(t_m)]$ (size: $N \times M$). A linear model of the type described by equations (3) can be effectively visualized as a bi-partite network (similar to those depicted in Fig. 1) where the input layer is associated to a set of hidden regulatory signals which are mapped to the output layer (the measurable output signals). The strengths of the connections in the network are measured by the parameters a_{nl} .

This type of linear networks finds application in several fields (*e.g.* telecommunications, signal processing, statistical learning), where different degrees of partial knowledge on the network inputs or parameters might be available. In the special case where both the input variables and

the connectivity strengths are unknown, the solution space of (3) becomes infinite dimensional, unless further constraints are imposed on the estimation problem. Principal Component Analysis (PCA) [10] and Independent Component Analysis [11] provide two examples of valid solutions to the problem, by requiring the unknown input signals to be orthogonal or statistically independent, respectively.

The authors recently introduced a framework, named Network Component Analysis (NCA) [1] [9], for modeling gene transcriptional regulation networks (in quasi-steady state conditions) through an input-output relationship of the type defined by (3), where the unknown transcriptional factor activities $\{p_1, \dots, p_L\}$ are the hidden nodes in the network, the gene expression levels $\{e_1, \dots, e_N\}$ are the measurable outputs, and the parameters a_{nl} measure the control strength of each regulator protein for each promoter binding site. NCA overcomes a fundamental problem affecting the applicability of available decomposition frameworks such as PCA or ICA to biological data, *viz.* the fact that these are based on specific assumptions on the statistical properties of the reconstructed signals (orthogonality or statistical independence), which do not generally hold for actual regulatory signals, and therefore tend to produce results which are difficult to interpret from a biological standpoint. Moreover, such techniques do not explicitly provide means for including information on known regulatory interactions, *i.e.* constrains on the connectivity matrix A .

The main idea behind NCA is to derive a biologically meaningful solution of (3) by exploiting the available a-priori knowledge on known regulatory interactions (generally available for several prokaryotes as well as eukaryotes, in the form of transcription factor binding affinity to different promoter regions), avoiding at the same time imposing further constraints on the estimated time-courses of the transcriptional factor activities.

III. NETWORK IDENTIFIABILITY CRITERIA

The available information regarding each transcription factor binding affinity to different promoter regions will translate into a set of constraints on the network connectivity matrix A [9]:

$$a_{ij} = \begin{cases} 0 & \text{if TF } j \text{ does not regulate gene } i \\ + & \text{if TF } j \text{ positively regulates gene } i \\ - & \text{if TF } j \text{ negatively regulates gene } i \end{cases} \quad (4)$$

The relational constraints defined by (4) impose a specific structure on the matrix A . The following definition formalizes this concept:

Definition 1 (Regulatory Pattern): Given a matrix A , and a set $\mathcal{R}_0 \subset \mathcal{Z}^2 = \{(i, j) : \forall i, j\}$, we say that A is characterized by the *regulatory pattern* \mathcal{R}_0 if and only if:

$$a_{ij} \equiv 0 \quad \text{for } (i, j) \notin \mathcal{R}_0 \quad (5)$$

The identifiability of (3) will, in general, depend on the specific regulatory pattern \mathcal{R}_0 . For example, consider the case when $N = L$ and \mathcal{R}_0 defines the set of diagonal matrices A . The resulting system has trivially a unique solution if the magnitude of the elements in the diagonal of A is assigned. The same conclusion would apply when considering an arbitrary permutation of a diagonal matrix as regulatory pattern \mathcal{R}_0 . In general, it is reasonable to assume that the larger the number of zero elements in A , the more likely it is that the set of parameters satisfying model (3) is uniquely determined.

This intuition can be formalized by showing that the special case consisting of arbitrary permutation of diagonal matrices is not the only one where the solution of (3) is unique when scaling is taken into consideration. In order to generalize such property, let us introduce the following additional definitions:

Definition 2 (Scaling Property): Given two matrices $A \in \mathcal{R}_0$ and P , and a matrix $T \triangleq AP$ (size: $N \times M$), we will say that the decomposition of T given by the pair (A, P) is *essentially unique* if and only if all pairs (\tilde{A}, \tilde{P}) , such that $\tilde{A}\tilde{P} = T$ can be expressed as follows:

$$\begin{aligned} \tilde{A} &= AX^{-1}, \\ \tilde{P} &= XP, \end{aligned} \quad (6)$$

where X is an arbitrary non-singular diagonal matrix.

The definition identifies (without proving its existence at this stage) a class of matrix pairs (A, P) , whose product can be decomposed only as scaled versions of the original matrices.

The next definition aims at establishing a fundamental property of the connectivity pattern associated with a given matrix A , by identifying a class of regulatory networks for which each transcriptional regulator has an independent role, *i.e.* its regulatory function is unique and cannot be replaced by a combination of the regulatory effects due to the other transcriptional regulators.

Definition 3 (Non-Redundant Connectivity Pattern): Given a matrix $A \in R_0$ (size: $N \times L$, $N \geq L$), the associated connectivity pattern will be referred to as non-redundant, if and only if each matrix obtained from A by arbitrarily selecting one of its columns and removing those rows corresponding to the non-zero elements of the selected column has rank equal to $L - 1$.

It is straightforward to prove that if a matrix A satisfies Definition 3, it is also full column rank. The following theorem, which was introduced and proved in [9], will be used as a starting point to demonstrate a fundamental property of the decomposition which is formulated in Theorem 2:

Theorem 1 (NCA Decomposition): Given two matrices A (size: $N \times L$) and P (size: $L \times M$), define their matrix product as $T \triangleq AP$ (size: $N \times M$). If the following hypotheses are satisfied:

- i) A is characterized by a non-redundant regulatory pattern \mathcal{R}_0 (as in Definition 3)
- ii) P is full row rank.

Then, for any matrices \tilde{P} (size: $L \times M$), \tilde{A} (size: $N \times L$), such that $\tilde{A} \in \mathcal{R}_0$, and $\tilde{A}\tilde{P} = T$, it is always possible to find a diagonal non-singular matrix X (size: $L \times L$), *s.t.*:

$$\begin{aligned}\tilde{A} &= AX^{-1} \\ \tilde{P} &= XP\end{aligned}$$

Therefore, the NCA decomposition is essentially unique.

Theorem 1 can be used to prove the following useful result, which identifies a simple set of sufficient conditions on the topology of the network defined by \mathcal{R}_0 :

Theorem 2 (Generalized Identifiability): Consider a set of L linearly independent input signals $\mathbf{p} = \{p_1, \dots, p_L\}$, and a set of N output signals $\mathbf{b} = A\mathbf{p}$ obtained through a random mixing matrix $A \in \mathcal{R}_0$ (size: $N \times L$). If the following hypotheses are satisfied,

- i) $N \geq L$
- ii) Each column of A has at least $L - 1$ zeros,
- iii) None of the columns of A has a set of non-zero entries which is the subset of the non-zero entries of any other column,

then the input signals can be reconstructed up to arbitrary scaling from the output signals $\{b_1, \dots, b_N\}$. Analogously, the mixing matrix A can be reconstructed up to arbitrary scaling of its columns.

The proof is provided in Appendix A. The identifiability conditions of Theorem 2 have the following straightforward interpretation from a graph theory perspective. First, each node in the regulatory layer cannot be connected to more than $N - L + 1$ nodes in the output layer (where L and N are the number of nodes in the input and output layer, respectively). Moreover, when considering the set of nodes connected to an arbitrary node in the regulatory layer, this set cannot be a subset of the children of another regulator node. Equivalently, when a node in the regulatory layer is considered, and all the nodes it connects to are removed, none of the remaining nodes in the network should become disconnected as a result.

The theorem provides a set of sufficient conditions on the regulatory pattern \mathcal{R}_0 that can be easily tested for compliance. Fig.1 shows two examples of regulatory patterns characterized by the same number of nodes as well as by the same number of connections (the edges of the network). The regulatory network shown in Fig.1(a) is not identifiable due to the regulatory pattern of the transcriptional regulator corresponding to the first column, which violates hypothesis (iii). The network in Fig.1(b), on the other hand satisfies all the identifiability conditions. Notice that the adjacency matrices of both networks are full column rank, thus showing that set of identifiability conditions cannot be reduced to more conventional properties of linear systems.

IV. REGULATORY NETWORK ESTIMATION FRAMEWORK

In [9], the authors proposed a two-step iterative least square algorithm in order to estimate the unknown parameters (A, P) , subject to the connectivity constraints. In this section, we derive a maximum likelihood (ML) estimation framework for (3), and we show that, under gaussian noise assumptions, the stationary points of the likelihood function must satisfy the normal equations of the associated weighted least-square criterion.

A. Maximum Likelihood Estimation

Let us consider the noisy linear model defined in (3), where the connectivity matrix A defines a regulatory pattern which satisfies the hypotheses of Theorem 2:

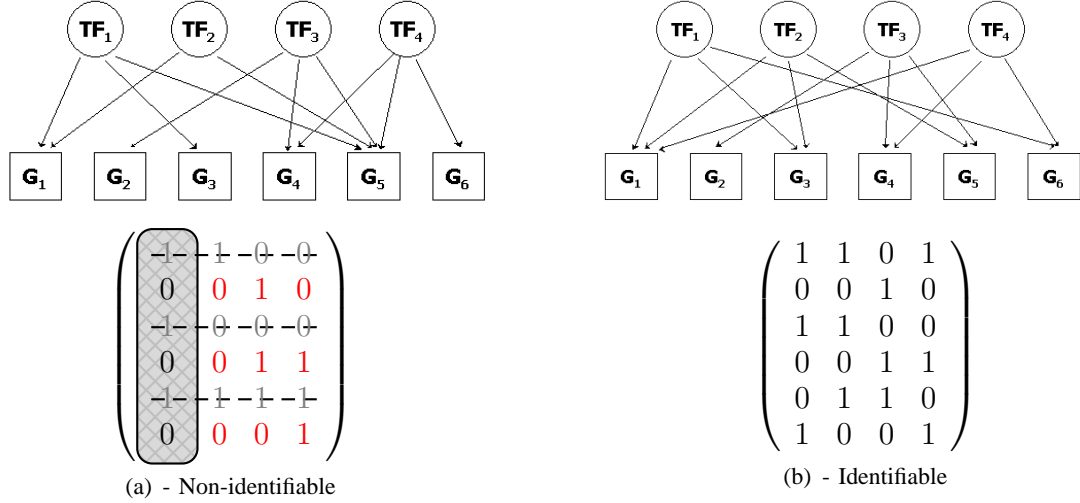


Fig. 1. An example of non-identifiable regulatory pattern (on the left) and an example of an identifiable one (on the right) are shown. Notice that the system matrices of both networks are full column rank.

$$E = AP + \Gamma, \quad A \in \mathcal{R}_0, \quad (7)$$

where Γ is a matrix of measurements errors which are assumed to be zero-mean gaussian and independent. Since we are assuming a gaussian independent noise characteristic, the negative log-likelihood function [16] associated with (7) has the following simplified expression:

$$L(A, P) = \sum_{n=1}^N \sum_{m=1}^M \frac{(e_{nm} - \mathbf{a}_n^{(r)} \mathbf{p}_m^{(c)})^2}{\sigma_{\gamma_{nm}}^2}, \quad (8)$$

where $\mathbf{a}_n^{(r)}$ is the n th row of the matrix A , $\mathbf{p}_m^{(c)}$ is the m th column of the matrix P , and

$$E[\gamma_{ij}\gamma_{mn}] = \begin{cases} \sigma_{\gamma_{nm}}^2 & \text{if } i = m \text{ and } j = n \\ 0 & \text{otherwise} \end{cases}, \quad (9)$$

where γ_{mn} is the (m, n) element of the noise matrix Γ . The quantities $\sigma_{\gamma_{nm}}^2$ are assumed to be known experimentally. When multiple samples are available for the measurement data $e_{nm,t}$ we can simply consider $e_{nm} = E[e_{nm,t}]$, *i.e.* their ergodic averages. The minimum of the negative likelihood function can be computed by setting its gradient equal to zero:

$$\nabla L(A, P) = \left[\frac{\partial L}{\partial a_{ij}} \quad \frac{\partial L}{\partial p_{kq}} \right]^T = 0. \quad (10)$$

We have from (8):

$$\frac{\partial L(A, P)}{\partial a_{ij}} = -2 \sum_{m=1}^M \frac{p_{jm}(e_{im} - \mathbf{a}_i^{(r)} \mathbf{p}_m^{(c)})}{\sigma_{\gamma_{im}}^2}, \quad (11)$$

for $i = 1, \dots, N$, and $j = 1, \dots, L$. Also:

$$\frac{\partial L(A, P)}{\partial p_{kq}} = -2 \sum_{n=1}^N \frac{a_{nk}(e_{nq} - \mathbf{a}_n^{(r)} \mathbf{p}_q^{(c)})}{\sigma_{\gamma_{nq}}^2}, \quad (12)$$

for $k = 1, \dots, L$ and $q = 1, \dots, M$. The components of the gradient vector have a straightforward interpretation. Notice, in fact, that (for $i = 1, \dots, N$):

$$\left[\sum_{m=1}^M \frac{p_{jm} e_{im}}{\sigma_{\gamma_{im}}^2} \right]_{j=1, \dots, L} = \begin{bmatrix} \mathbf{e}_i^{(r)} C_{\gamma_i^{(r)}}^{-1} \mathbf{p}_1^{(r)T} \\ \vdots \\ \mathbf{e}_i^{(r)} C_{\gamma_i^{(r)}}^{-1} \mathbf{p}_L^{(r)T} \end{bmatrix}^T = \mathbf{e}_i^{(r)} C_{\gamma_i^{(r)}}^{-1} P^T, \quad (13)$$

$$\left[\sum_{m=1}^M \frac{p_{jm} \mathbf{a}_i^{(r)} \mathbf{p}_m^{(c)}}{\sigma_{\gamma_{im}}^2} \right]_{j=1, \dots, L} = \begin{bmatrix} \mathbf{a}_i^{(r)} P C_{\gamma_i^{(r)}}^{-1} \mathbf{p}_1^{(r)T} \\ \vdots \\ \mathbf{a}_i^{(r)} P C_{\gamma_i^{(r)}}^{-1} \mathbf{p}_L^{(r)T} \end{bmatrix}^T = \mathbf{a}_i^{(r)} P C_{\gamma_i^{(r)}}^{-1} P^T, \quad (14)$$

with $C_{\gamma_i^{(r)}}$ given by:

$$C_{\gamma_i^{(r)}} = E[\gamma_i^{(r)T} \gamma_i^{(r)}], \quad i = 1, \dots, N, \quad (15)$$

where $\gamma_i^{(r)}$ is the i th row of the noise matrix Γ . The remaining components of the gradient satisfy (for $q = 1, \dots, M$):

$$\left[\sum_{n=1}^N \frac{a_{nk} e_{nq}}{\sigma_{\gamma_{nq}}^2} \right]_{k=1, \dots, L} = \begin{bmatrix} \mathbf{a}_1^{(c)T} C_{\gamma_q^{(c)}}^{-1} \mathbf{e}_q^{(c)} \\ \vdots \\ \mathbf{a}_L^{(c)T} C_{\gamma_q^{(c)}}^{-1} \mathbf{e}_q^{(c)} \end{bmatrix} = A^T C_{\gamma_q^{(c)}}^{-1} \mathbf{e}_q^{(c)}, \quad (16)$$

$$\left[\sum_{n=1}^N \frac{a_{nk} \mathbf{a}_n^{(r)} \mathbf{p}_q^{(c)}}{\sigma_{\gamma_{nq}}^2} \right]_{k=1, \dots, L} = \begin{bmatrix} \mathbf{a}_1^{(c)T} C_{\gamma_q^{(c)}}^{-1} A \mathbf{p}_q^{(c)} \\ \vdots \\ \mathbf{a}_L^{(c)T} C_{\gamma_q^{(c)}}^{-1} A \mathbf{p}_q^{(c)} \end{bmatrix} = A^T C_{\gamma_q^{(c)}}^{-1} A \mathbf{p}_q^{(c)}, \quad (17)$$

with:

$$C\boldsymbol{\gamma}_q^{(c)} = E[\boldsymbol{\gamma}_q^{(c)}\boldsymbol{\gamma}_q^{(c)T}], \quad q = 1, \dots, M, \quad (18)$$

where $\boldsymbol{\gamma}_q^{(c)}$ is the q th column of the noise matrix Γ . Hence, it holds that:

$$\frac{\partial L(A, P)}{\partial a_{ij}} = 0 \implies \mathbf{a}_i^{(r)} P C \boldsymbol{\gamma}_i^{(r)} P^T = \mathbf{e}_i^{(r)} C \boldsymbol{\gamma}_i^{(r)} P^T, \quad i = 1, \dots, N, \quad (19)$$

and:

$$\frac{\partial L(A, P)}{\partial p_{kq}} = 0 \implies A^T C \boldsymbol{\gamma}_q^{(c)} A \mathbf{p}_q^{(c)} = A^T C \boldsymbol{\gamma}_q^{(c)} \mathbf{e}_q^{(c)}, \quad q = 1, \dots, M. \quad (20)$$

Equations (19) and (20) show that the gradient of (8) is zero when the pair (A, P) satisfies the normal equations of the weighted least-squares criterion in the case when P is given, and in the case when A is given, respectively.

V. SIMULATION RESULTS

In order to evaluate the performance of the proposed decomposition, we conducted several experiments both with synthetic data and with real hybridization data measured during whole genome microarray assays of the bacterium *Escherichia Coli K12*. Furthermore, the sensitivity of the method to inaccuracies in the hypothesized connectivity pattern was also investigated.

A. Synthetic Data with Additive Noise

The goal of this set of simulation experiments is to assess the efficacy of the method in reconstructing the network dynamics in a large-scale settings when both the model linearity and identifiability assumptions hold strictly, with the exception that the outputs are perturbed by additive noise. This test aims at establishing the performance characteristics of the NCA decomposition as a function of the measurement error as well as of the complexity of the regulatory network, in the ideal scenario when the model linearity is not violated.

In a first set of simulation experiments, expression level data was generated synthetically based on the model defined in (3). We considered two different examples of connectivity topology. The first (named *Network A*) consists of a network of 321 genes and 25 transcriptional factors, with a fan-in (the number of TFs controlling each gene) ranging between 1 and 6, for a total of 661 regulatory interactions. This first example aims at mimicking a real transcriptional regulatory

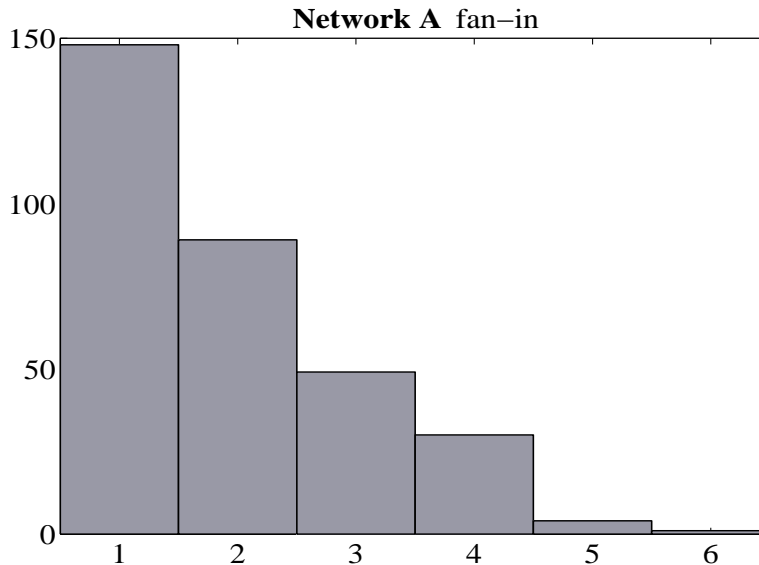


Fig. 2. Histogram of the fan-in (number of TFs controlling each gene) of *Network A*

network, where most binding sites are in general associated to less than three regulatory factors. A histogram plot of the fan-in of the resulting network is shown in Fig. 2.

The second example (*Network B*) comprises a similar number of genes and TFs (400 and 25 respectively), but with a larger average network fan-in (*cf.* Fig. 3), resulting in an increased total number of connections in the network (2,475). In general denser connectivity matrices will result in harder estimation problems, both because the total number of variables is larger and also because of the increase in redundancy in the regulation mechanism.

The synthetic time-courses of the transcriptional factor activities were generated keeping in mind one of the key aspect of the proposed decomposition, *viz.* its capability of reconstructing the regulatory signals without requiring specific assumptions on their statistics. Therefore, three different sets of transcriptional factor activities (for each of the two networks) were synthetically generated, characterized by an increasing degree of statistical dependence between the various TFs profiles. The first set consists of 25 nearly uncorrelated zero-mean gaussian signals, while the remaining two sets were generated by linearly mixing additional sets of gaussian signals until a certain pre-specified level of statistical dependency between the signals was achieved.

The condition number ξ of the signal matrix P (defined as the ratio between its largest and its smallest singular values) was adopted as an overall measure of statistical correlation

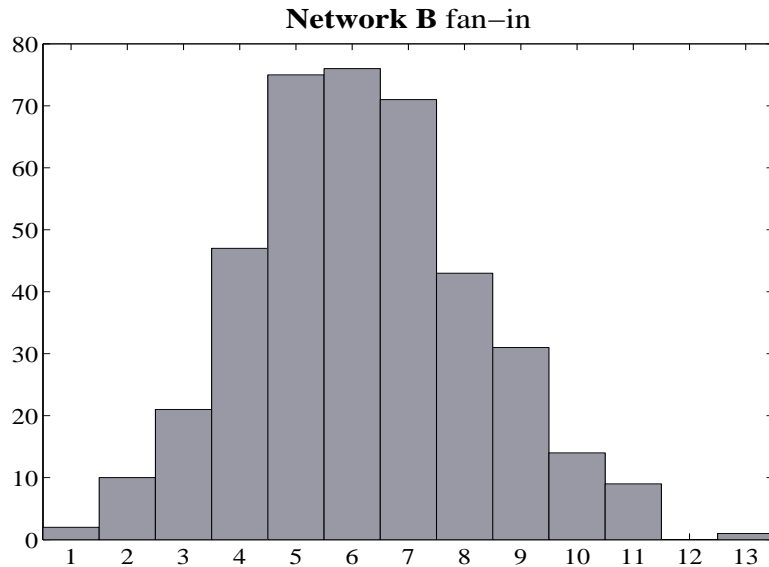


Fig. 3. Histogram of the fan-in (number of TFs controlling each gene) of *Network B*

between the input signals of the regulatory network. The condition number for each set of TFs profiles is shown in Table I, along with the results that were obtained when attempting the reconstruction of 1,000 different regulatory networks for each combination of network topology and set of transcriptional factor activities. Different levels of additive noise were also considered for comparison. In all cases, the mean square error was used as a measure of accuracy in the reconstruction of both the control strengths (A) and the TF activity profiles (P). The accuracy in the fit of the synthetic expression data matrix is also reported.

The results shown in Table I confirm the method's capability of reconstructing the network parameters, even in those cases for which the driving signals are strongly correlated. In particular, the reconstruction error becomes arbitrarily small when the additive noise is zero, in all cases. Moreover, the reconstruction accuracy appears to slowly deteriorate for increasing levels of measurement noise, as well as for increasing levels of correlation between the regulatory signals, thus suggesting that the algorithm is robust to moderate model perturbations.

B. Synthetic Data with Inaccuracies in the Connectivity Topology

The accuracy of the a-priori knowledge on the regulatory interactions between input nodes and output nodes in the network plays a key role in NCA. The connectivity topology for

TABLE I

RESULTS OBTAINED BY APPLYING NCA TO TWO SYNTHETIC TRANSCRIPTIONAL REGULATORY NETWORK ARCHITECTURES (EACH SIMULATED 1,000 TIMES WITH DIFFERENT PARAMETERS). THE SIMULATION EXPERIMENTS INCLUDED THREE DIFFERENT SETS OF TRANSCRIPTIONAL REGULATOR ACTIVITY PROFILES, CHARACTERIZED BY VARIOUS DEGREES OF MUTUAL STATISTICAL DEPENDENCE (HERE COLLECTIVELY DESCRIBED BY THE CONDITION NUMBER ξ OF THE MATRIX P). DIFFERENT LEVELS OF ADDITIVE GAUSSIAN NOISE WERE ALSO CONSIDERED. THE RESULTS ARE SHOWN IN TERMS OF MEDIAN MEAN-SQUARE-ERROR (MSE), AND 90-PERCENTILE OF THE MSE (*i.e.* IN 90% OF THE SIMULATIONS, A MSE SMALLER THAN THE ONE REPORTED WAS OBSERVED).

Dataset	Noise Level	A mse		P mse		Data fit	
		median	90%	median	90%	median	90%
Network A							
$\xi = 5.713$ (uncorrelated)	0%	<1e-10	<1e-9	<1e-9	<1e-9	<1e-010	<1e-9
	5%	0.0252	0.0266	0.0651	0.0679	0.0497	0.0521
	10%	0.0357	0.0378	0.0925	0.0962	0.0708	0.0740
	20%	0.0508	0.0540	0.1311	0.1364	0.1004	0.1054
$\xi = 187.28$ (moderately correlated)	0%	<1e-10	<1e-9	<1e-9	<1e-9	<1e-10	<1e-9
	5%	0.0260	0.0278	0.0661	0.0690	0.0504	0.0530
	10%	0.0371	0.0394	0.0937	0.0977	0.0717	0.0750
	20%	0.0529	0.0563	0.1329	0.1387	0.1019	0.1065
$\xi = 2,639.4$ (strongly correlated)	0%	<1e-9	<1e-9	<1e-9	<1e-8	<1e-9	<1e-9
	5%	0.0289	0.0309	0.0851	0.0891	0.0649	0.0682
	10%	0.0411	0.0441	0.1208	0.1265	0.0921	0.0965
	20%	0.0588	0.0629	0.1713	0.1796	0.1310	0.1376
Network B							
$\xi = 4.735$ (uncorrelated)	0%	2.2e-10	3.6e-10	4.6e-10	7.4e-10	3.3e-10	5.1e-10
	5%	0.0454	0.0467	0.0500	0.0515	0.0943	0.0968
	10%	0.0644	0.0661	0.0710	0.0732	0.1338	0.1371
	20%	0.0917	0.0942	0.1016	0.1047	0.1910	0.1956
$\xi = 211.19$ (moderately correlated)	0%	8.5e-10	1.3e-9	7.4e-10	1.2e-9	4.0e-10	6.3e-10
	5%	0.0563	0.0583	0.0251	0.0259	0.0459	0.0472
	10%	0.0805	0.0831	0.0358	0.0370	0.0654	0.0672
	20%	0.1154	0.1194	0.0514	0.0534	0.0936	0.0962
$\xi = 1,574.0$ (strongly correlated)	0%	9.9e-10	1.8e-9	2.1e-9	3.6e-9	9.3e-10	1.7e-9
	5%	0.0725	0.0762	0.0740	0.0774	0.1322	0.1365
	10%	0.1041	0.1095	0.1061	0.1101	0.1887	0.1942
	20%	0.1534	0.1625	0.1542	0.1616	0.2717	0.2806

TABLE II

RECONSTRUCTION ACCURACY FOR DIFFERENT LEVELS OF INACCURACY IN THE DESCRIPTION OF THE NETWORK
TOPOLOGY

Connectivity Errors	A mse		P mse		Data fit	
	median	90%	median	90%	median	90%
2%	0.0938	0.1299	0.0889	0.0914	0.1750	0.2020
5%	0.1138	0.1998	0.0897	0.1096	0.1891	0.2446
10%	0.1473	0.2289	0.0925	0.1551	0.2145	0.2687

actual biological regulatory networks is conventionally built from different sources of a-priori information on the interaction between different chemical species. For example, in the case of transcriptional networks this kind of information is largely available in the literature for several organisms, and it is often accessible through publicly available databases. Nonetheless, one must take into account the fact that experimentally derived interaction patterns are often prone to errors.

In this second simulation experiment, we considered the connectivity architecture of *Network A* from the previous set of simulations, and we evaluated the effect of introducing arbitrary errors in the hypothesized connectivity topology. Once the synthetic data was generated according to the true network connectivity pattern (with 10% additive gaussian noise), we systematically introduced random errors in the connectivity matrix initial guess, by randomly selecting a subset of gene-TF pairs and adding a connection if one was not present, or otherwise removing it if one was present. The procedure was selectively applied to 2%, 5% and 10% of the connections in the network¹.

The results we obtained are shown in Table II. The measured mean square error in the reconstruction of the connectivity matrix A is clearly larger than in the case when the correct topology was considered, and tends to increase monotonically with the percentage of wrong connectivity assumptions. However, the reconstructed profiles of the transcriptional factor activities are in general less affected by the errors in the topology, even though the consistency of the

¹Notice that the percentages are expressed in terms of the number of connections in the network, not in term of the total number of elements in the A matrix.

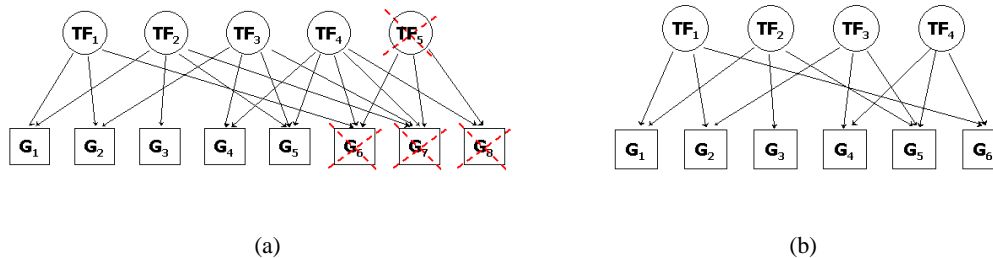


Fig. 4. The figure shows an example of non-NCA-compliant topology (a), as well as the selection of an identifiable sub-network (b), obtained by removing TF₅ and the genes that are associated to it.

reconstruction worsens as witnessed by the larger values of the 90% percentile. This phenomenon is easily explained by considering that the profile of a given transcriptional regulator is estimated from a large set of gene expression level time-courses: the least-squares criterion tends to favor the regulatory interactions that better explain the data, thus discarding misleading connections in the regulatory pattern.

In general, the behavior of the proposed decomposition in those situations when erroneous connectivity assumptions are made cannot be fully predicted in advance. Factors that will influence its performance are, for example, the noise in the measurement data, the number of wrong connectivity assumptions, as well as their location, and the degree of correlation between the regulatory signals. However, for a given network topology, and a set of measurement data characterized by a known noise level, one can attempt to establish the sensitivity of the reconstructed network dynamics by conducting a series of simulation experiments of the type described in this section, therefore assessing the consistency of the results when different patterns of error in the hypothesized network architecture are introduced.

C. Selection of Identifiable Regulatory Sub-networks in *E.coli*

When networks of the type described by model (3) are built upon available biological connectivity information, a fundamental problem may arise due to the fact that the identifiability conditions described in Theorem 2 are not always satisfied. This might be due to either an insufficient number of data samples collected during the experimental stage (when $M < L$ the linear independence of the input signals is necessarily violated), or to the topology of the regulatory network not satisfying the hypotheses of Theorem 2. A straightforward solution to

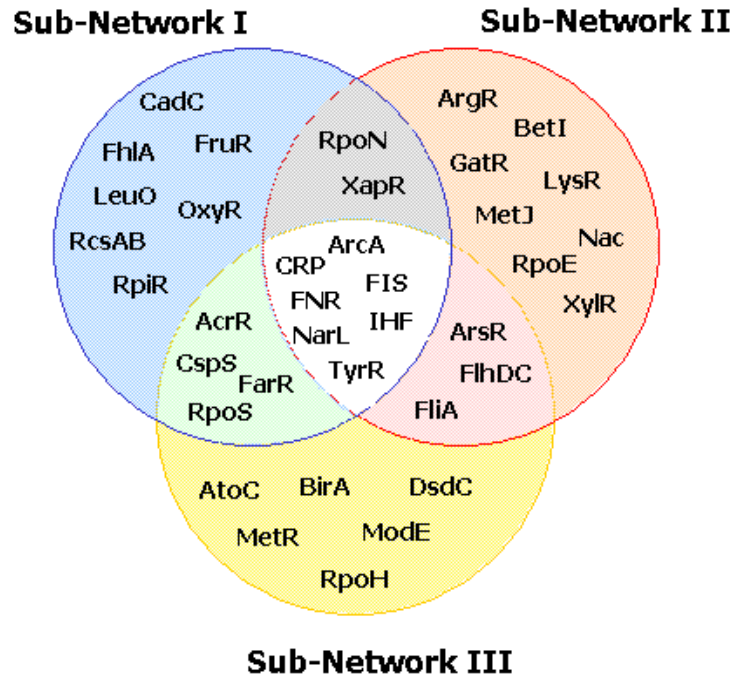


Fig. 5. The figure shows the three regulatory subnetworks of *E. coli* analyzed with NCA. The three subnetworks are characterized by partially overlapping sets of transcriptional regulators.

the problem is obtained by increasing the number of sample points collected or the number of genes assayed, respectively, until the conditions are satisfied.

When such solution is not viable, one can still select a subset of the factors involved in the regulation of the genes assayed in the experiment. The genes affected by the transcriptional regulators that have been excluded must be pruned accordingly. Consider for example the network in Fig.4(a): it is easy to verify that the topology of this network violates the hypotheses of Theorem 2. Therefore, although the complete system is not identifiable, it is nonetheless possible to focus on the transcriptional regulators TF_1 to TF_4 and build an identifiable subnetwork by selecting the subset of genes which are not regulated by TF_5 .

The primary challenge, when applying NCA to real data is therefore to identify a suitable set of active genes and transcriptional regulators, which not only are significant from a biological standpoint, but also satisfy the identifiability criteria required by the decomposition. The connectivity information in the case of the bacterium *E. coli* was derived from the publicly available binding site affinity information provided in the RegulonDB [17] database which included 120

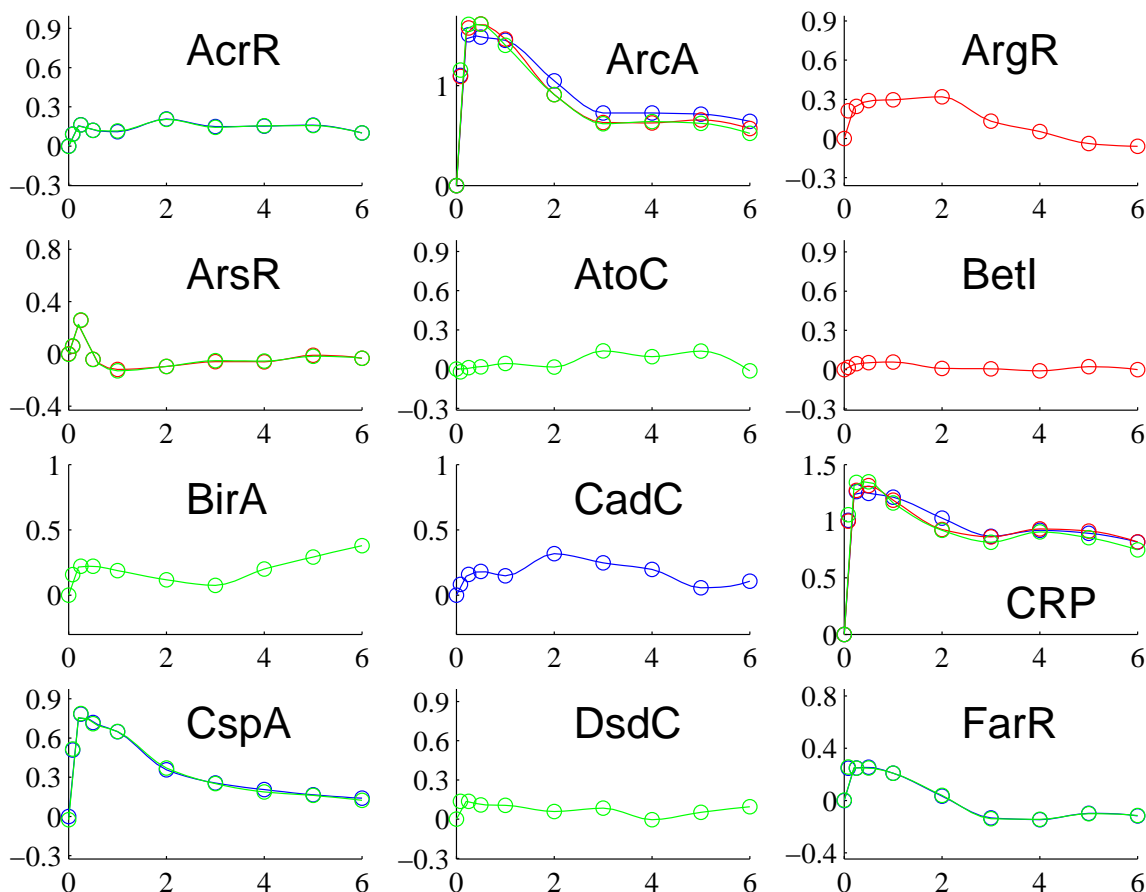


Fig. 6. Transcription factor activities of the bacterium *E.coli* during growth in a medium transition, estimated from three regulatory subnetworks by NCA analysis. Each color (red, green or blue) is associated to one of the three sub-networks. The time axis is expressed in hours.

regulatory proteins and 833 genes.

In [1], a microarray assay experiment is described during which a total of 25 time-points were collected during the growth of the organism in a medium transition (from glucose to acetate). In this experiment, the sample size represented the critical factor in determining the largest possible identifiable transcriptional sub-network. The analysis of the dataset with NCA resulted in the reconstruction of a single regulatory subnetwork including 16 key transcriptional factors whose activity profiles were reconstructed along with their control strength on a total of 100 genes.

The problem of finding all possible subsets of transcriptional regulators that are NCA estimation compliant can be shown to be a combinatorial NP-hard problem. However, by taking advantage of the formulation of the identifiability conditions given in Theorem 2, we devised a

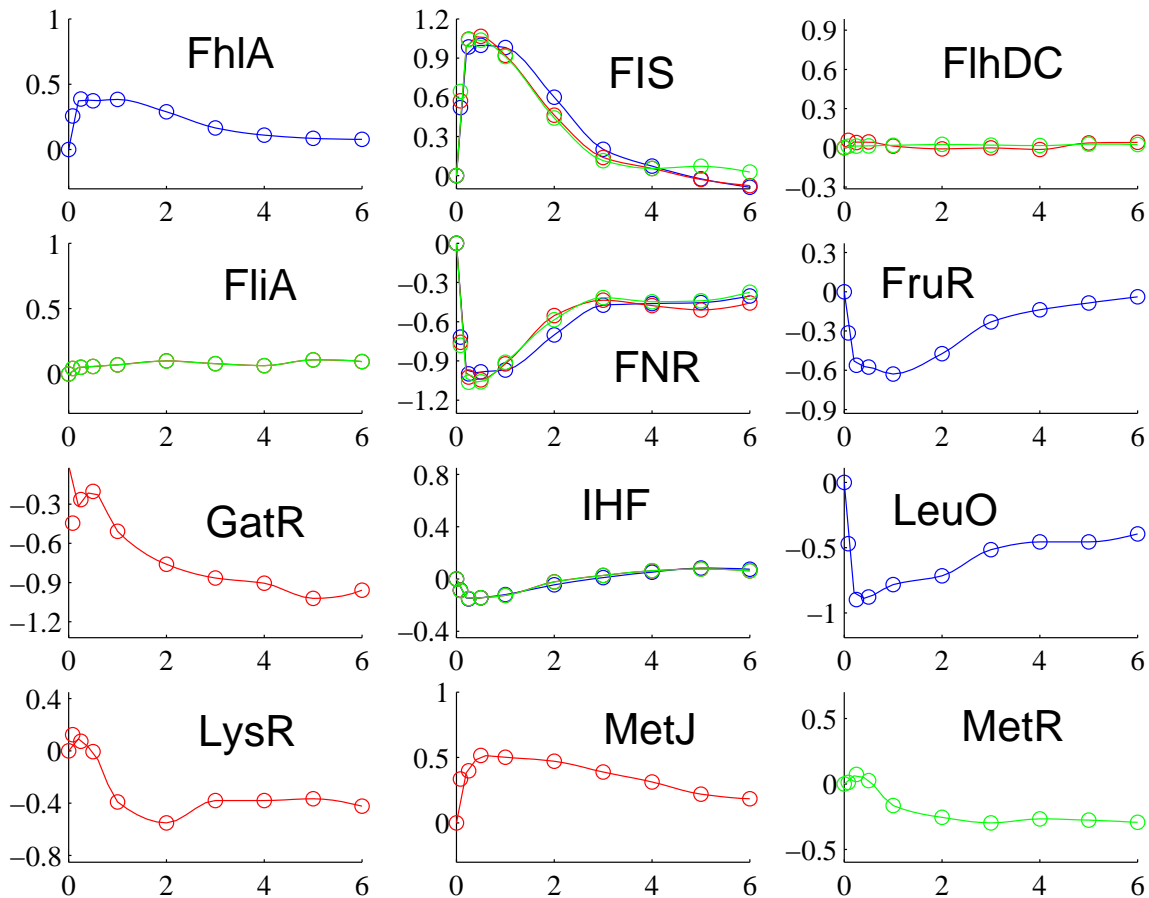


Fig. 7. Transcription factor activities of the bacterium *E.coli* during growth in a medium transition, estimated from three regulatory subnetworks by NCA analysis. Each color (red, green or blue) is associated to one of the three sub-networks. The time axis is expressed in hours.

simple heuristic which provides an efficient and reliable alternative to the combinatorial approach. In general, for large regulatory networks it is the violation of hypothesis (iii) together with the limited sample size ($M \geq L$ is required to ensure the linear independency between the regulatory signals) that prevents the applicability of NCA. Therefore, one can start by randomly selecting $K \leq M$ out of the L regulatory nodes and check whether hypothesis (iii) is satisfied by the subnetwork obtained by considering the selected regulatory nodes as well as all the genes connected to them but not connected to any one of the pruned regulators. When such initial choice violates the identifiability conditions, the TFs selection is updated iteratively according to the following heuristics. Among the pruned TFs the one with the largest out-degree connectivity

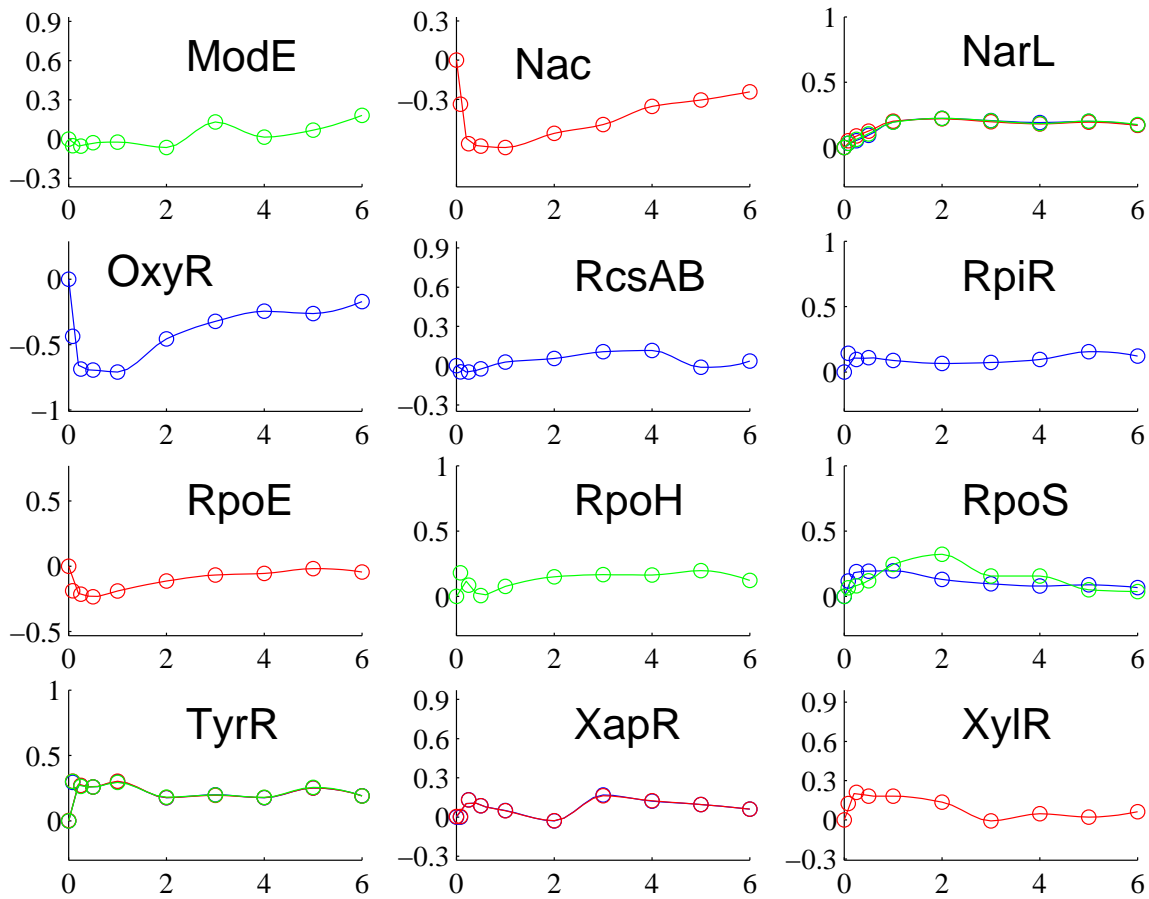


Fig. 8. Transcription factor activities of the bacterium *E.coli* during growth in a medium transition, estimated from three regulatory sub-networks by NCA analysis. Each color (red, green or blue) is associated to one of the three sub-networks. The time axis is expressed in hours.

is selected to replace one of the regulators responsible for the violation of hypothesis (iii) in the current selection. The latter can be either chosen at random or according to its out-degree connectivity. Simulation experiments show that the approach is in general capable of identifying a suitable subnetwork within a few iterations (in general less than 10).

For the dataset described in [1], and by following the procedure described above, three significant transcriptional sub-networks were identified for a total of 37 unique TFs (shown in Fig.5) and 237 genes. The reconstructed time-courses of the regulators are shown in Fig. 6,7, and 8. These results show a significant consistency in the estimation of the transcriptional factor activities across different sets, thus demonstrating the applicability of a simultaneous

reconstruction of overlapping regulatory sub-networks. In particular, the activities of additional major regulators such as FIS, FNR and IHF were reconstructed, thus providing further insight on key regulation pathways in *E.coli*.

For example, the estimated activity profile (*cf.* fig. 7) of FIS (a DNA binding protein involved in DNA replication specific processes) confirms previous experimental evidence [18] that the activity of FIS rapidly increases within 15-30 minutes of a nutritional upshift and then slowly levels off when cells begin to grow.

VI. DISCUSSION

The results presented in Section V demonstrate that NCA is indeed capable of accurately reconstructing regulatory signals when these are the input of networks that satisfy certain topological criteria, even under fairly noisy conditions. In particular, the method succeeds in separating the input signals even when these are highly-correlated, thus overcoming a major limitation of standard exploratory techniques such as PCA or ICA. Moreover, when investigating the sensitivity of the algorithm to inaccuracies in the hypothesized network topology, we observed that the method is still capable of providing a faithful reconstruction of the input signals, as long as the relative number of inaccurate connections is small when compared to the total number of connections. This result is of particular interest, since the a-priori knowledge on the network topology is often derived from experimental data which is prone to errors.

The simplified set of identifiability conditions derived in Theorem 2, provide not only a straightforward method for testing a network topology for NCA compliance, but also suggest a strategy for selecting candidate identifiable sub-networks from the overall topology. We demonstrated the applicability of the sub-network selection method in Section V. Starting from regulatory protein binding site affinity information for the bacterium *E.coli*, we built a diagram of the overall connectivity topology which relates promoter regions to transcriptional regulators. The identifiability of such initial network is unattainable both because of the limited sample size available in the experimental data and also because several regulators violate hypothesis (iii) of Theorem 2. By applying the approach discussed in Section V, we were able to identify three NCA compliant sub-networks, with overlapping sets of transcriptional regulators. For such set of regulators, the consistency in the profiles reconstructed starting from different sub-networks provided compelling preliminary evidence that the parallel reconstruction of the profiles of the

regulators across different transcriptional sub-networks is viable.

APPENDIX

A. Proof of Theorem 2

Theorem 2 can be demonstrated by showing that its hypotheses are equivalent to those of Theorem 1 with probability that goes to one, when the matrix A is a random matrix constrained by the regulatory pattern \mathcal{R}_0 . From the linear independence assumption of the input signals $\{p_1, \dots, p_L\}$, derives that the second hypothesis of of Theorem 2 is necessarily satisfied.

Hypotheses (i) and (ii) of the theorem are necessary conditions in order for the regulatory pattern \mathcal{R}_0 to be non-redundant. In fact, when they are violated, then there must be at least one column of A for which the matrix obtained by removing all the rows corresponding to its non-zero entries will have less than $L - 1$ rows, thus violating the rank condition of Def. 3.

We can therefore proceed by showing that when hypothesis (iii) of the theorem is also satisfied, the identifiability conditions of Theorem 1 are satisfied with probability one over the set of all possible values assumed by the non-zero entries of the matrix A . Select an arbitrary column r of A , and in order to simplify the notation, consider a permutation of the rows of A such that the first $N - K$ entries of column r are arbitrary non-zero real numbers, and the last K entries are all zeros (see figure 9). K must be greater or equal than $L - 1$ because of hypotheses (i) and (ii). The resulting $K \times (L - 1)$ sub-matrix consisting of the last K rows of A and all of its columns but the selected one, must be full column rank in order to satisfy the property of Definition 3. When the entries $a_{ij} \notin \mathcal{R}_0$, ($i = N - K + 1, \dots, N$, $j = 1, \dots, L$, $j \neq r$) are samples drawn independently from the continuous distribution of a random variable, this sub-matrix is full column rank with probability one, as long as none of its columns has all zero entries induced by the connectivity pattern. The only case in which such condition is violated is when the non-zero entry pattern of one or more columns of A is a sub-set of the non-zero entry pattern of another column of A . Such case would indeed violate hypothesis (iii), thus proving the theorem.

B. Properties of the ML Estimator

In this section, we demonstrate certain properties of the ML estimate, defined as the optimum of the cost function (8). In particular, we will show that the biases of the estimates of A and P

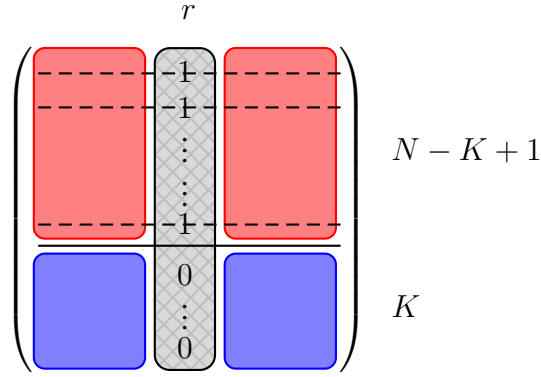


Fig. 9. The figure shows how to build from A the $K \times (L-1)$ sub-matrix (in blue), whose rank must satisfy Def.3, for a given column r .

are mutually related.

From (19), the ML estimate of the i th row of A is given by:

$$\tilde{\mathbf{a}}_i^{(r)} = \mathbf{e}_i^{(r)} C_{\gamma_i^{(r)}}^{-1} \tilde{P}^T \left(\tilde{P} C_{\gamma_i^{(r)}}^{-1} \tilde{P}^T \right)^{-1}, \quad i = 1, \dots, N, \quad (21)$$

where \tilde{P} is the ML estimate of P , whose columns can be computed as:

$$\tilde{\mathbf{p}}_q^{(c)} = \left(\tilde{A}^T C_{\gamma_q^{(c)}}^{-1} \tilde{A} \right)^{-1} \tilde{A}^T C_{\gamma_q^{(c)}}^{-1} \mathbf{e}_q^{(c)}, \quad q = 1, \dots, M. \quad (22)$$

Given that:

$$\mathbf{e}_i^{(r)} = \mathbf{a}_i^{(r)} P + \gamma_i^{(r)}, \quad E[\gamma_i^{(r)}] = 0, \quad i = 1, \dots, N, \quad (23)$$

Thus, we have:

$$E[\tilde{\mathbf{a}}_i^{(r)}] = \mathbf{a}_i^{(r)} E \left[P C_{\gamma_i^{(r)}}^{-1} \tilde{P}^T \left(\tilde{P} C_{\gamma_i^{(r)}}^{-1} \tilde{P}^T \right)^{-1} \right]. \quad (24)$$

Similarly, since:

$$\mathbf{e}_q^{(c)} = A \mathbf{p}_q^{(c)} + \gamma_q^{(c)}, \quad E[\gamma_q^{(c)}] = 0, \quad q = 1, \dots, M, \quad (25)$$

it holds that:

$$E[\tilde{\mathbf{p}}_q^{(c)}] = E \left[\left(\tilde{A}^T C_{\gamma_q^{(c)}}^{-1} \tilde{A} \right)^{-1} \tilde{A}^T C_{\gamma_q^{(c)}}^{-1} A \right] \mathbf{p}_q^{(c)}. \quad (26)$$

Therefore:

$$E[\tilde{\mathbf{a}}_i^{(r)}] \rightarrow \mathbf{a}_i^{(r)} \iff E[\tilde{\mathbf{p}}_q^{(c)}] \rightarrow \mathbf{p}_q^{(c)}. \quad (27)$$

REFERENCES

- [1] K.C. Kao, Y.-L. Yang, R. Boscolo, C. Sabatti, V.P. Roychowdhury, and J.C. Liao. Determination of multiple transcription regulator activities in escherichia coli using network component analysis. *Proceedings of the National Academy of Sciences (PNAS)*, 101(2):641–646, 2004.
- [2] A.M. Campbell and L.J. Heyer. *Discovering Genomics, Proteomics, and Bioinformatics*. Benjamin/Cummings, 2002.
- [3] M.B. Eisen, P.T. Spellman, P.O. Brown, , and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, December 1998.
- [4] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 3:281–285, July 1999.
- [5] G. Getz, E. Levine, and E. Domany. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. USA*, 97:12079–12084, October 2000.
- [6] T.R. Hughes *et al.* Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, July 2000.
- [7] E. Segal, M. Shapira, A. Regev, D. Pe’er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–176, June 2003.
- [8] T.S. Gardner, D. di Bernardo, D. Lorenz, and J.J. Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301:102–105, July 2003.
- [9] J.C. Liao, R. Boscolo, Y.-L. Yang, L.M. Tran, C. Sabatti, and V.P. Roychowdhury. Network-enabled reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences (PNAS)*, 100(26):15522–15527, 2003.
- [10] J.E. Jackson. *A User’s Guide to Principal Components*. Wiley-Interscience, New York, 1991.
- [11] S. Roberts and R. Everson, editors. *Independent Component Analysis : Principles and Practice*. Cambridge University Press, 2001.
- [12] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, Inc., New York, 2001.
- [13] N.S. Holter, M. Mitra, A. Maritan, M. Cieplak, J.R. Banavar, and N.V. Fedoroff. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proceedings of the National Academy of Sciences (PNAS)*, 97(15):8409–8414, 2000.
- [14] N.S. Holter, A. Maritan, M. Cieplak, N.V. Fedoroff, and J.R. Banavar. Dynamic modeling of gene expression data. *Proceedings of the National Academy of Sciences (PNAS)*, 98(4):41693–1698, 2001.
- [15] W. Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60, 2002.
- [16] S.R. Eliason. *Maximum Likelihood Estimation : Logic and Practice*. Sage Publications, Newbury Park, California, 1993.
- [17] H. Salgado *et al.* Regulondb (version 4.0): transcriptional regulation, operon organization and growth conditions in *escherichia coli* k-12. *Nucleic Acids Research (Database issue)*, 32:D303–D306, 2004.
- [18] O. Ninnemann, C. Koch, and R. Kahmann. The *e.coli* fis promoter is subject to stringent control and autoregulation. *EMBO Journal*, 11(3):1075–1083, March 1992.