Wie Zellen Denken WS05/06

Large Networks: Network Component Analysis

Christian Ehrlich
MaNr. 3760306

21. February 2006

Abstract

A method will be presented, which does the decomposition of high-dimensional data into low-dimensional signals which are driven through an interacting network. Traditional method, like Principle Component Analysis (PCA) or Independent Component Analysis (ICA), which perform the same task, do not take network topology information into account and need to assume biological unjustified statistical properties. Therefore the resulting decomposition does only represent a phenomenological model for the observed data and does not necessarily contain biologicaly meaningful information. Network Component Analysis is the method which will show to be a powerful tool in decomposing real biological data, like DNA microarray, in their regulatory signals over time points and their connectivity strength between the regulatory units and the output layer of the underlying network.

1 Introduction

Bacteria respond to a change in environmental condition through a variety of sensor proteins which eventually relay the signal to DNA binding proteins which will modulate transcription. These DNA binding proteins, or transcription factors (TFs), can be quantified by analytic methods such as DNA microarray. These technics supply a high dimensional description of the cell which is typically the end product of low dimensional regulatory signals driven trough an interacting network.

In the past couple of years a variety of analytic method have been developed in order to decompose the regulatory signals, e.g. Principle Component Analysis (PCA), Independent Component Analysis (ICA). Even though these technics have shown to be a vital instrument in order to understand regulatory processes, they have not been designed to address the hidden dynamics reconstruction problem. They discard available information of the underlying network and they need to assume mutual orthogonally and statistical independence of regulatory signals.

Here we will describe a method which relies on the *a priori* known network topology and does not need to make assumptions on orthogonally or independents. We will describe the principle of Network Component Analysis (NCA) in detail and we will also give a brief example of a real life application of NCA.

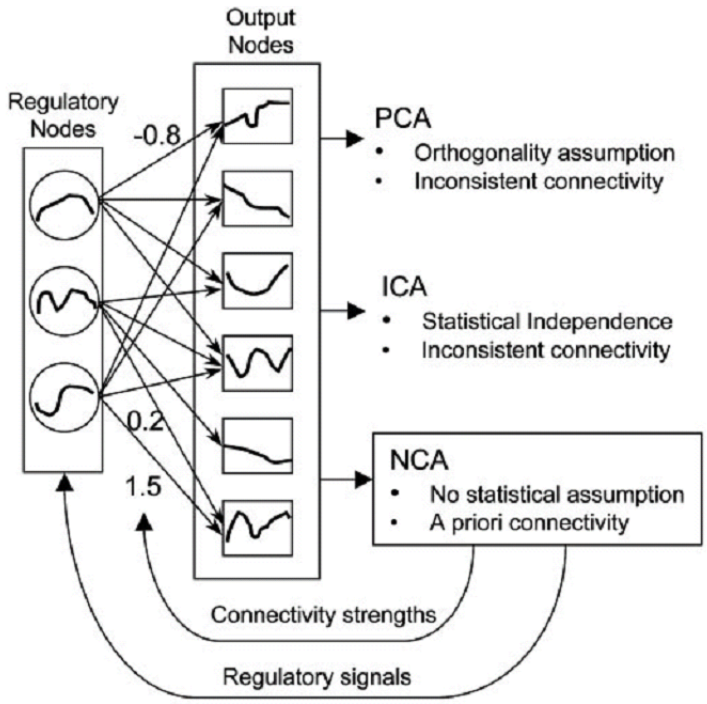

Figure 1: A regulatory network in which the output nodes are controlled by the strength of the input units. (Taken from [2])

2 Network Component Analysis

The bases for the NCA is the data retrieved from the DNA microarray. The obtained data points are organized in a matrix $[E]$(size N x M) which holds M samples (or time points) of N output variables (such as the expression ratio of transcripts). In order to reconstruct the underlying network, $[E]$ needs to be decomposed into

$$[E] = [A][P] \tag{1}$$

$[P]$ (size L x M) which consists of L regulatory signals over M time points (note that usually L is much smaller than M) and $[A]$ (size N x L) which encodes the connectivity strength between the regulatory units and the output layer (Fig. 1).

The decomposition shown in Eq. 1 is an inverse problem for which no unique solution exist without further assumptions. This can be demonstrated [2] by introducing a non singular matrix $[X]$ (size L x L) such that

$$[E] = ([A][X])([X^{-1}][P]) = [\hat{A}][\hat{P}] \tag{2}$$

Thus without further constrains, $[E]$ can not be uniquely decomposed according to Eq. 1. Therefore we need to introduce criteria that do allow a unique decomposition while not making assumptions on the statistical nature of the problem.

3 Criteria for NCA

Considering Eq.2, multiple $[A]$s and $[P]$s can be found that reconstruct $[E]$ equally well. Nevertheless, by applying certain constrains on the connectivity of $[A]$, it can be shown that the matrix $[X]$ can only be diagonal (see Appendis of [2] for proof). Furthermore, when $[A]$ has full-column rank and $[P]$ has full row rank, the decomposition can be solved uniquely. In other words, Eq. 2 represents all possible solutions (see Appendix of [2] for proof). Under these constrains, Eq. 1 can be satisfactorily solved up to a scaling factor (represented by the matrix $[X]$ in Eq. 2).

In summary, the constrains necessary to perform NCA are

(i) $[A]$ (connectivity matrix) must have full-column rank. So that all regulatory signals are represented.

(ii) All sub-matrices of $[A]$ must have full-column rank. Otherwise the network topology does not allow a unique identification of regulatory signals.

(iii) $[P]$ must have full row rank. Meaning that no regulatory signal can be expressed as a linear combination of two other signals.

If these criteria are satisfied, the experimentally obtained data matrix $[E]$ can be decomposed into the connectivity matrix $[A]$ and the signal matrix $[P]$. $[A]$ will contain the estimated connectivity strength on each edge, whereas $[P]$ will represent the regulatory signals of each node (Fig. 1).

In order to test these three criteria, an initial connectivity matrix $[A]$ (size N x L) is constructed from the known network topology. An entry a_{ij} is set to zero, if no connection from regulatory unit j to the output node i exist in the model. All other values are set to a random non zero number. After construction of the initial matrix $[A]$, the first criteria is tested. If $[A]$ has full-column rank, submatrixes are generated by removing one regulatory unit j and all output nodes $i_0, ..., i_n$ that are influenced by j. The second criteria is satisfied, if and only if, all submatrixes have rank equal to $L - 1$.

Remember, L was the number of regulatory signals.
A priori testing of the third criteria is not possible, but the criteria implies, that L (the number of regulatory nodes) must be smaller than M (the number of data points). If $L < M$, the matrix $[P]$ is very likely to have full row rank for real biological data. In any case it is required to check matrix $[P]$ for full row rank, after NCA is performed.

Fig. 2 gives an easy example of two network, that have an identical number of regulatory units and output nodes. But because of the connection pattern of R_3, the network in Fig 2a is identifiable, whereas the network shown in Fig. 2b is not.

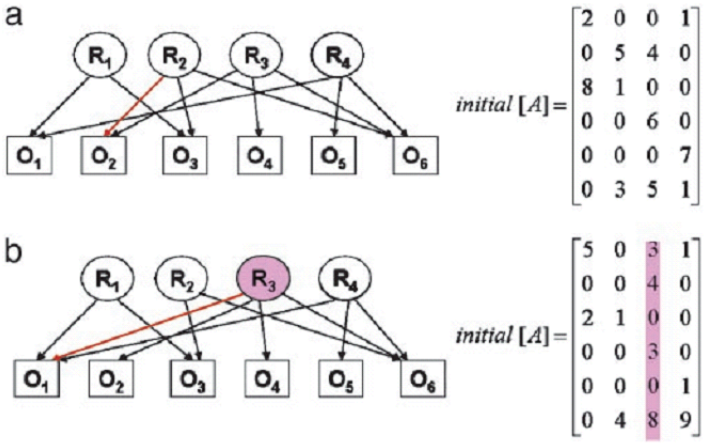

Figure 2: An identifiable network (a) and an unidentifiable network (b). The network differences are shown in red. (Taken from [2])

4 Method for NCA

After the identifiability of the network has been verified, the regulatory signals $[P]$ and the connectivity strength $[A]$ can be reconstructed by the following procedure. First of all the known network topology is modeled in $[A]$ by setting all entries to zero for which no edge between the regulatory unit and the output node exists. All other entries are initialized with an non zero value. Unfortunately when dealing with biological data the signals are noisy, and an exact solution is not possible. However, the solution can be approximated, in this case by a least-square error measure.

The following objective function is minimized:

$$min||[E] - [A][P]||^2, s.t. A \epsilon Z_0, \tag{3}$$

where Z_0 is the know network topology.

Note that the minimization is an iterative process, in which the values of $[A]$ and $[P]$ are alternately optimized. This procedure always ends in an global optimum (see [2] for proof).

5 Experimental Validation of NCA

In order to assure the reliability of NCA, a test case was constructed in which the solution was known. A network of seven hemoglobin solutions where prepared. Each solution contains a combination of the three components: oxyhemoglobin, methemoglobin, and cyano-hemoglobin. For details of preparation see Appendix of [2]. The absorbance spectra where measured between 380 and 700nm with 1-nm increments resulting in the matrix [Abs] of size 7 x 321. According to Beer-Lambert law, the measured absorbance spectra can be decomposed by

$$[Abs] = [C][\epsilon] \tag{4}$$

where the rows of $[Abs]$ are the absorbance spectra of each solution at each wavelength, the columns of the connectivity matrix $[C]$ are the concentration of each component, and the rows of $[\epsilon]$ are the spectra of pure components. The known network topology is shown in Fig. 3a.

From the results in Fig. 3b it is clear that NCA gives the best reconstruction of the true spectra. Especially for cyano-hemoglobin the results obtained by PCA or ICA are misleading and might result in a misinterpretation of the data.

6 Conclusions

This article has shown a method for decomposing experimental data into their connectivity strength and regulatory signals while including known facts about the underlying network structure. Additionally the NCA does not make any statistical assumptions of orthogonality or independence like other methods (e.g. PCA, ICA). Therefore it represents a procedure that is based on biological relevant assumptions. As shown above, the NCA represents a powerful tool for reconstructing regulatory network parameters. For further applications of NCA see [2] or [1].

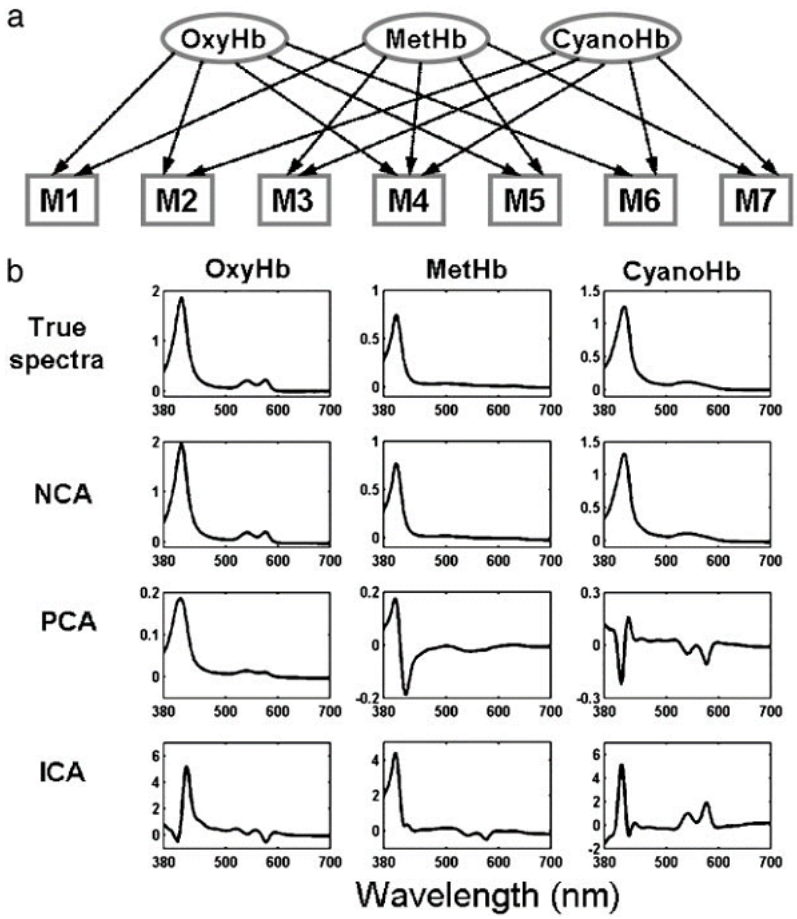

Figure 3: Experimental validation of NCA using absorbance spectra of hemoglobin solutions. (Taken from [2])

References

[1] Katy C Kao, Young-Lyeol Yang, Riccardo Boscolo, Chiara Sabatti, Vwani Roychowdhury, and James C Liao. Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis. *Proc Natl Acad Sci U S A*, 101(2):641–646, Jan 2004.

[2] James C Liao, Riccardo Boscolo, Young-Lyeol Yang, Linh My Tran, Chiara Sabatti, and Vwani P Roychowdhury. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci U S A*, 100(26):15522–15527, Dec 2003.