Network Component Analysis (NCA)

Falko Krause
M# 3760306

Abstract

The Network Component Analysis is a tool for analyzing high dimensional data of dynamic gene networks. Its functional efficiency can be validated. To apply the NCA certain conditions have to be fulfilled. Furthermore the biological background of applying a NCA in a micro array experiment to analyze gene-trancriptionfactor networks has to be taken into account.

1 Introduction

Gene expression networks can be investigated experimentally by micro array chips. The output of these high throughput experiments is usually multidimensional. To gain knowledge about the regulatory influences of transcription factors the dimensionality of the data must be reduced. This can be done with the help of statistical techniques such as principal component analysis (PCA), singular value decomposition or independent component analysis (ICA). However these techniques are solely mathematical and omit biological facts.

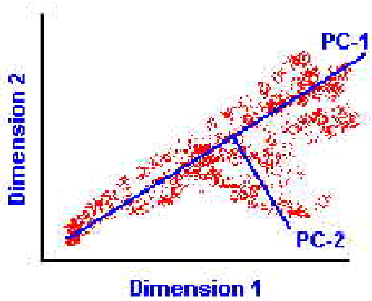

Figure 1: Scatter plot that contains the two main components found by a Principal Component Analysis.

In the PCA the principal components are found by variance maximizing rotation of the original variable space. In the scatter plot of Fig. 1 the x-axis would be rotated so that it approximates the regression line labeled $PC-1$. On this axis then the variance of the dots would be the highest for the whole scatter plot. Further principal components that capture the next highest variance must be orthogonal to the fist principal component. This however must not apply to a biological network.

Furthermore the underlying network structure of a regulatory network is ignored in this technique. Since there is a growing number of transcription factors that can be associated with genes this knowledge is used in the NCA.

2 Network Component Analysis

The input for a NCA is a bipartite graph that expresses the transcriptionfactor gene interactions in the form of a connectivity matrix. This matrix $[A]$ has zeros at the places where no interaction is known and random numbers in all other places. The second input is the expression data from a micro array experiment. It consists of mRNA expression strength values at different time points. It is encoded in the matrix $[E]$. The goal of the NCA is to decompose the matrix $[E]$ as in Eq. 1 where $[A]$ represents the connectivity strength. Connectivity strength in this case means the influence of the transcriptionfactors on the genes they regulate. The Matrix $[P]$ is called the signal matrix and represents the influence of a tranctricptionfactor at a time point

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

To apply a decomposition there are three constraints that have to be fulfilled:

(i) $[A]$ must have full-column rank

(ii) All sub-matrices of $[A]$ must have full-column rank

(iii) $[P]$ must have full row rank

The fist two constraints can be tested a priori the third has to be tested after the decomposition. A illustration of the general scheme of the NCA can be seen in Fig. 2

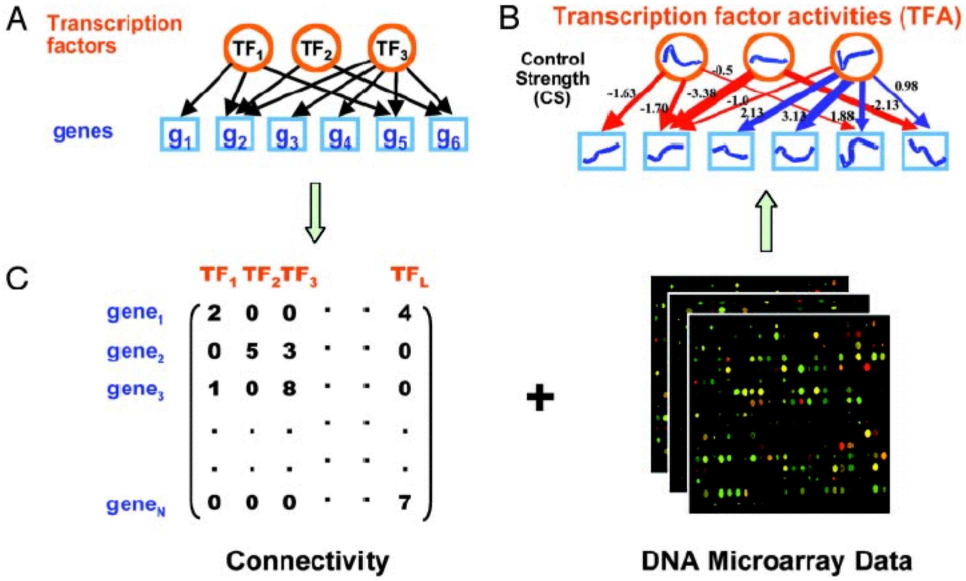

Figure 2: This is the general scheme of a Network Component Analysis.

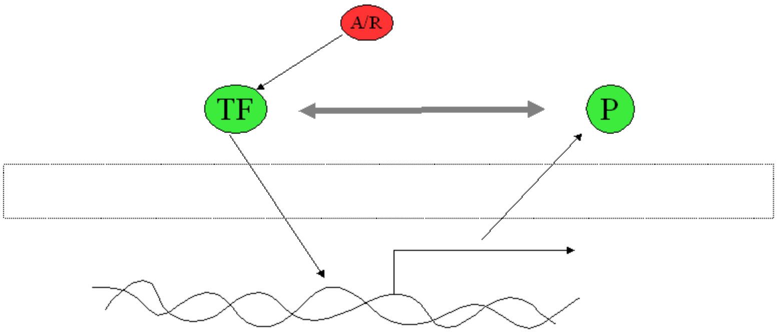

Figure 3: It is important to remember that transcriptionfactor activity is not equal to the concentration of the transcriptionfactor. The activity of transcriptionfactors is influenced by other factors.

3 Boundary Conditions

To apply the NCA to DNA microarray experiments that measure the RNA expression concentration it is essential to understand what the biological influences are that govern RNA expression. There are two main influences on the concentration of RNA in a cell. First the synthesis rate $s(t)$ that changes over time and second the degradation rate α that is a constant. This means the RNA expression at a certain point in time $x(t)$ can be expressed by the following equation $x(t) = s(t) - \alpha x$ where x is the measured concentration.

Even more important is to understand that influence of a transcriptionfactor on RNA expression of a certain gene is not equal to the expression or concentration of the transcription factor. Transcriptionfactor activities are in most cases influenced by ligands or other proteins (like in Fig. 3).

4 Validation

The validation of the NCA was done by an experiment that bears no reference to the study of gene expression networks. In the experiment the absorbance spectra of visible light of seven hemoglobin solutions were analyzed (see Fig. 4). Each solution contained three different components: oxyhemoglobin, methemoglobin, and cyano-methemoglobin. Each component shows a unique absorbance spectra of visible light. The original absorbance spectra was assumed to be unknown as well as the concentration of the components in the solution. To apply the NCA only the information of wich hemoglobin was added to the solution as well as the observed absorbance of the solution was used. The calculation of the NCA yielded the original absorbance spectra of the three hemoglobins used as well as their concentration in each solution.

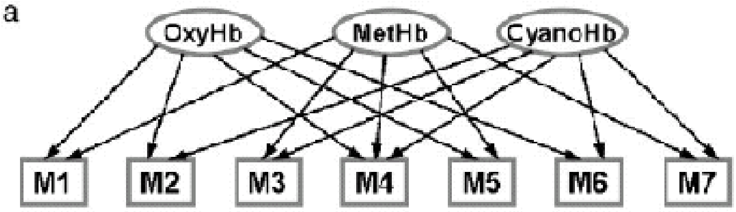

Figure 4: To validate the NCA different solutions of three different hemoglobins were analyzed for their absorbance of visible light . The solutions can be visualized as a bipartite graph.

5 Large Network Experiment

In the large network experiment the temporal influence of transcription factors in *Escherichia coli* during the reorganization from glycolysis to glyconeogenesis as energy source was analyzed. This reorganization in the metabolism of the bacteria was forced by slowly changing the growth media from glucose to acetate as only carbon source. The knowledge of gene-transcriptionfactor interaction in

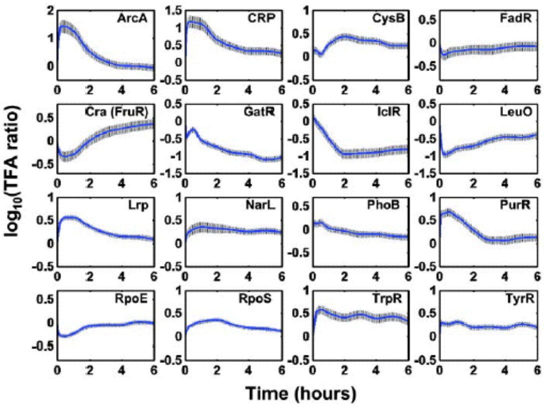

Figure 5: This is a result of NCA that was applied to a research of yeast cultures in changing growth media.

E.coli enabled the construction of a bipartite network graph. The interaction (connectivity) data was obtained from the regulonDB [1]. To proof the feasible of the connectivity data the full cloumn-rank of the whole connectivity matrix as well as the full row-rank of all sub matrices was checked. The NCA was applied and the signal matrix was tested for full row-rank. The NCA returned an approximation of the influence of several transcriptionfactors over time (Fig. 5)

References

[1] A. M. Huerta, H. Salgado, D. Thieffry, and J. Collado-Vides. RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res*, 26(1):55–59, Jan 1998.

[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.

[3] 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.