533b Mixed-Integer Reformulations of Network Component Analysis

Eric Yang, Joseph Vitolo, Charles Roth, and Ioannis (Yannis) P. Androulakis

DNA microarray technologies measure in a high-throughput fashion the output of complex networked systems driven by convoluted regulatory signals. Low dimensional representations of these signals are usually derived based on the statistical analysis of the measured signals. However, these approaches by and large ignore the underlying network structures and provide decompositions based on phenomenological models that do not necessarily contain biologically meaningful signals.

Recently, a promising approach, termed Network Component Analysis (NCA), was presented (Liao, Boscolo et al. 2003; Kao, Yang et al. 2004). The method uncovers hidden regulatory signals from output of networked systems, much like gene regulatory architectures. The multidimensional expression data are organized in a matrix format E(NxM), where N are the monitored expression levels and M the time points at which observations are made. NCA seeks to reconstruct the observed signals E in a lower-dimensional space according to the decomposition [E]=[A][P], where the matrix P(LxM) consists of L regulatory signals such that L <

1. A must be full-column rank

2. each column of A must have at least L-1 zeroes

3. P must have full-row rank

The primary challenge when applying NCA to real data is to identify a suitable set of active genes and transcriptional regulators, which satisfy the NCA identifiability criteria and remain biologically significant. The presence of many potential regulatory sites on each target gene makes it difficult to satisfy conditions 1 & 2, and condition 3 can only be checked after the fact.

We propose an alternative formulation for the aforementioned estimation problem, mixed integer NCA (miNCA), which circumvents these limitations. The main advantages over NCA analysis are:

(a) miNCA allows for the identification of the least complex regulatory network(s) that describe the experimental data, which is important in order to avoid the problem of over-fitting.

(b) miNCA allows for the determination of an ensemble of possible solutions, thus permitting the analysis of multiple regulatory structures that can be further analyzed and characterized based on additional data.

In addition to the A and P matrices, we introduce a set of binary variables y(i,j) denoting the existence or absence of each possible regulatory connection. We impose explicitly all the identifiability criteria, and the formulation is defined parametrically in the number of active allowable connections, NC, in the network. The proposed mixed integer, non-linear optimization is solved successively to allow the determination of multiple solutions enable via the introduction of successive integer cuts (Biegler, Grossmann et al. 1997).

The miNCA formulation is powerful in that it allows the optimal derivation of multiple regulatory structures of increased complexity. An upper bound to the value of NC is imposed by the constraint requiring a maximum number of non-zeroes per column of A.

We demonstrate the methodology by using use temporal transcirptional data from (Almon, DuBois et al. 2003) which measured the transcriptional profiling of liver cells in response to methylprednisolone (MPL) sodium succinate (a corticosteroid). The superset of possible regulatory elements is identified using the ModelInspector utility of the Genomatix software suite (http://www.genomatix.de) and the miNCA procedure identifies multiple minimal sets of alternative NCA-compatible structures. Finally, we evaluate the alternative regulatory structures based on their ability to rationalize the phenotypic observations and identify conserved components of the regulatory architectures based on the analysis of the multiple solutions. The efficient solution of the optimization problem suggests that the methodology will be readily applicable to complex regulatory networks.

References

Almon, R. R., D. C. DuBois, et al. (2003). "Gene arrays and temporal patterns of drug response: corticosteroid effects on rat liver." Funct Integr Genomics 3(4): 171-9.

Biegler, L. T., I. E. Grossmann, et al. (1997). Systematic Methods of Chemical Process Design, Prentice Hall.

Kao, K. C., Y. L. Yang, et al. (2004). "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): 641-6.

Liao, J. C., R. Boscolo, et al. (2003). "Network component analysis: reconstruction of regulatory signals in biological systems." Proc Natl Acad Sci U S A 100(26): 15522-7.