

us hope that there is no curse on this subterranean treasury of plants, but that its discovery will evolve into a better understanding of root growth and will open the way to improve the root systems of crops by using genetic engineering.

References

- 1 Bucher, M. (2002) Molecular root bioengineering. In *Plant Roots, the Hidden Half*, 3rd edn, (Waisel, Y. *et al.*, eds), pp. 279–294, Marcel Dekker
- 2 Birnbaum, K. *et al.* (2003) A gene expression map of the *Arabidopsis* root. *Science* 302, 1956–1960
- 3 Scheres, B. *et al.* (2002) Root development. In *The Arabidopsis Book* (Somerville, C.R. and Meyerowitz, E.M., eds), American Society of Plant Biologists (<http://www.bioone.org/bioone/?request=get-toc&issn=1543-8120>)
- 4 Beeckman, T. *et al.* (2001) The peri-cell-cycle in *Arabidopsis*. *J. Exp. Bot.* 52, 403–411
- 5 Davies, P.J. ed. (1995) *Plant Hormones: Physiology, Biochemistry and Molecular Biology* Kluwer Academic Publishers
- 6 Casimiro, I. *et al.* (2003) Dissecting *Arabidopsis* lateral root development. *Trends Plant Sci.* 8, 165–171
- 7 Sabatini, S. *et al.* (1999) An auxin-dependent distal organizer of pattern and polarity in the *Arabidopsis* root. *Cell* 99, 463–472
- 8 Swarup, R. *et al.* (2001) Localization of the auxin permease AUX1 suggests two functionally distinct hormone transport pathways operate in the *Arabidopsis* root apex. *Genes Dev.* 15, 2648–2653

0167-7799/\$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tibtech.2004.04.016

Extracting novel information from gene expression data

Zheng Li and Christina Chan

Department of Chemical Engineering and Material Science, Michigan State University, East Lansing, MI 48824, USA

Data from high throughput technologies, such as DNA microarrays, necessitated the development of new computational methodologies for analyzing the high dimensional information contained within the gene expression data. Liao's group suggested the use of network component analysis to predict transcription factor activities by integrating gene expression data from *Escherichia coli* with known connectivity information between their genes and transcription factors. This introduces an approach for obtaining novel information from gene expression data.

Recent advances in high throughput technologies have generated a plethora of biological information, such as gene expression, protein–protein interaction, and metabolic data. These various types of data capture different aspects of the cellular response to environmental factors and contain information about the underlying regulatory network structure. Much effort has been devoted to analyzing these datasets to reconstruct the regulatory features. The majority of the mathematical approaches, such as clustering [1], independent component analysis (ICA) [2], principal component analysis (PCA) [3], Boolean networks [4], and Bayesian networks [5] have been applied to infer biological information from high dimensional microarray data. These methods typically do not incorporate known regulatory information into the structure of the models. Methods such as clustering, ICA and PCA attempt to identify functionally similar groups of genes, whereas Boolean and Bayesian networks try to uncover the regu-

latory interaction between genes, all without *a priori* information on the regulatory network structure. However, the completion of genome sequences and the recent development of a method for genome-wide location (binding) analysis provides a means for identifying sets of genes that can be bound by each transcription factor [6]. The genome-wide binding analysis involves micro-array analysis of epitope-tagged transcription factors coupled to chromatin immunoprecipitation (ChIP) [6]. The binding data provides information on the presence of a regulator at a promoter site, suggestive of binding ability but not necessarily function. Coupling the information obtained through location (binding) data with gene expression data provides functional information to facilitate the reconstruction of regulatory signals and networks. Lee *et al.* have attempted to apply this approach to detect transcriptional interactions in *Saccharomyces cerevisiae* [7]. In addition, Hartemink *et al.* [8], using expression data alone, found the results to be inconsistent with location data. Therefore, they combined known location data with expression data in a Bayesian network framework to infer unknown genetic regulatory networks involved in *S. cerevisiae* pheromone response. Bar-Joseph *et al.* [9] combined gene expression and binding information for 106 transcription factors of *S. cerevisiae* and identified groups of co-expressed genes to which a set of transcription factors bind and reconstructed their regulatory networks. They identified both established regulatory interactions as well as unexpected interactions that suggest models of transcriptional regulation for further studies. Using similar data modalities, Liao *et al.* [10] developed network component analysis (NCA) to reconstruct regulatory

Corresponding author: Christina Chan (krischan@egr.msu.edu).

Available online 26 June 2004

signals in *Eshcerichia coli*. NCA was used to determine multiple transcription regulator activities from gene expression data and connectivity diagrams derived from location data.

Network component analysis

Typically, a gene is regulated by several different transcription factors and the contribution from each transcription factor to each gene is not easily obtained. NCA offers an approach for predicting the contribution from each transcription factor through its connectivity (or control) strength as well as predicting a dynamic profile of these transcription factor activities. This method was applied to a model system – the glucose to acetate carbon source transition in *E. coli* [11]. NCA determined the activities of multiple transcription factors from the gene expression profile at various time points after the transition from glucose to acetate metabolism. The activities of 16 transcription factors were determined from the expression profile of 100 genes that are regulated by these transcription factors. To confirm the results, the predicted activity of one of these transcription factors, the catabolite repressor protein (CRP), was compared with the measured intracellular cAMP concentrations, because activation of CRP requires the binding of cAMP.

The interactions between transcription factors and the genes they regulate are modeled as a two-layer regulatory network with the transcription factors as the first layer and the genes as the second layer, with unidirectional edges going from transcription factors to genes. Information on connectivity between transcription factors and genes is obtained from genome-wide location data. Each element in the connectivity matrix represents the connectivity strength between a transcription factor and a gene. To initialize the connectivity matrix, the elements corresponding to unconnected edges are set to zero and the elements corresponding to connected edges are set to arbitrary values. Unconnected edges suggest no known interaction between a gene and a transcription factor. The connectivity strength of the connected edges and the transcription factor activities are simultaneously determined by using a two-step least square algorithm to minimize the fitting error between the measured and predicted gene expression levels.

Benefits and limits of NCA

Unlike methods such as ICA or PCA, NCA does not presuppose mutual independence or mutual orthogonality in the underlying network structure. Instead, NCA incorporates known regulatory information into the structure of the models. Typically, transcription factor activities are very difficult to measure directly. NCA provides an approach for predicting their dynamic activities from the gene-expression profiles, which are easily obtained. As connectivity information between transcription factors and genes becomes more readily available and complete, NCA could provide a modeling alternative for determining transcription factor activities,

which in turn would facilitate reconstruction of transcriptional regulatory networks. This approach could be used to distinguish changes in the transcriptional regulatory networks of disease versus normal states and provide a systematic method of identifying the contributions of various transcription factors to the expression of several genes, thus facilitating the identification of 'better' and more comprehensive targets for drug therapies. In addition, NCA has the potential to infer regulatory networks at different levels of the biological system, for example at the metabolic or protein level, both of which are also controlled by regulatory signals.

NCA selects a solvable subspace from the entire space defined by gene expression and transcription factors based upon currently available (known) location data. It is possible to introduce bias into the problem in the process of subspace selection. For example, NCA deletes transcription factors that regulate less than three genes and therefore to satisfy the full column rank requirement, it necessitates the removal of the corresponding genes associated with those transcription factors. This deletion could result in loss of information. As illustrated in Figure 1, if the transcription factor TF_1 known to connect to only two genes was deleted along with $gene_1$ and $gene_3$, the elimination of these genes would also remove their connection to the other transcription factors. Thus, useful information for inferring activities of the transcription factors connected to $gene_1$ and $gene_3$ could be lost. This could affect the accuracy of the model prediction. Similarly, as pointed out by Kao *et al.* [11] it is difficult for NCA to predict the transcription factor activity of the subregulon when a transcription factor forms a regulon that is a subset of other transcription factors. Perhaps as the connectivity information becomes more complete, fewer of these transcription factors would regulate just one or two genes circumventing the aforementioned limitation. Computational cost will then be the limitation.

Finally, the gene regulatory network motif that NCA currently models is the multi-input motif; other motifs such as autoregulation, feed-forward loop, the regulator chain, or interactions between transcription factors [7] are not currently included. Nevertheless, NCA provides an

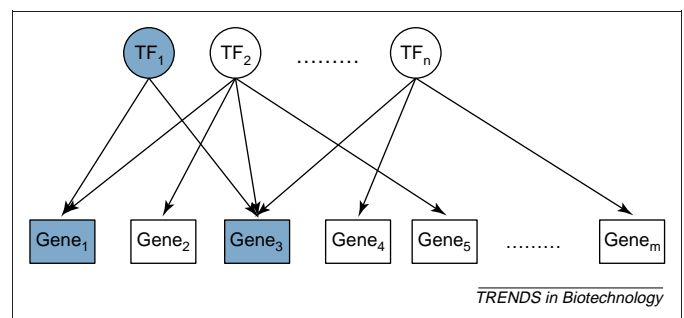


Figure 1. A gene regulatory network composed of n transcription factors and m genes, in which transcription factor TF_1 was deleted along with $gene_1$ and $gene_3$ to satisfy the full column rank requirement. As $gene_1$ and $gene_3$ are also connected to other transcription factors, it is possible that relevant information for inferring transcription factor activities will be lost.

approach for readily predicting transcription factor activities (which are difficult to measure) with some degree of confidence from gene expression data (easy to measure) and known location data.

Complementary techniques to further the applicability of NCA

One of the limiting factors in fully realizing the capabilities of NCA is incomplete location (binding) data. Attempts have been made to address the limited location data, for example Pilpel *et al.* [14] developed an approach that searches for transcription factor binding sites and identifies all genes containing these promoter motifs as opposed to using actual binding data. They coupled this information with gene-expression data to identify synergistic motif combinations in *S. cerevisiae*. This is based on the idea that gene expression in eukaryotes often requires combined activity of two or more transcription factors. From the results of their analysis, they inferred potential interactions between transcription factors and in the transcriptional regulatory network. Indeed, Berman *et al.* searched for clusters of transcription factor binding sites to identify *cis*-regulatory modules in the *Drosophila* genome. Here they based their approach on the assumption that multiple transcription factor binding sites tend to cluster together [15]. They integrated the information with gene expression data to uncover interactions in the transcriptional regulatory network during *Drosophila* development.

An alternative is to use stochastic approaches to address this limitation. For example, Bayesian network analysis, which has been applied to infer the structure of gene regulatory [5,8] and metabolic [12] networks can be applied to reconstruct currently unknown connections between transcription factors and genes. Another possible approach is ICA, which can be used to infer unknown regulatory information from gene-expression data. This method was used to reveal different sets of gene signatures for classifying ovarian cancer [13]. A gene signature defines a group of genes that behave in a similar fashion or are involved in a similar process. Because a gene (or group of genes) may be involved in several different processes and influenced by several transcription factors, the gene(s) may participate in several independent clusters or patterns. ICA could be useful in exploring the possibility that a group of genes are regulated by a single transcription factor or a group of transcription factors [13]. Therefore, it is possible to combine NCA with Bayesian network analysis or with ICA to explore the entire space defined by gene expression and transcription factors. In this scenario, Bayesian network analysis or ICA would be applied first to identify currently unknown transcription factor–gene

interactions. Subsequently, NCA would be applied to obtain the activity profiles of the transcription factors in the entire transcription factor–gene space.

In summary, NCA provides an approach for identifying transcription factor activities from gene expression data coupled with location data for the first time. The applicability of NCA might be further extended when more information on location data becomes either experimentally or theoretically available.

Acknowledgements

This work is supported in part by the National Science Foundation (BES 0222747 and BES 0331297) and the Whitaker Foundation. L. Z. is supported in part by the Center for Biological Modeling: Quantitative Biology Interdisciplinary Graduate Research Award at Michigan State University.

References

- 1 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868
- 2 Liebermeister, W. (2002) Linear modes of gene expression determined by independent component analysis. *Bioinformatics* 18, 51–60
- 3 Alter, O. *et al.* (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10101–10106
- 4 Liang, S. *et al.* (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing*, pp. 18–29
- 5 Pe'er, D. *et al.* (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics* 17 (Suppl 1), S215–S224
- 6 Ren, B. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science* 290, 2306–2309
- 7 Lee, T.I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799–804
- 8 Hartemink, A.J. *et al.* (2002) Combining location and expression data for principled discovery of genetic regulatory network models. *Pacific Symposium on Biocomputing 2002, Kauai, HI, United States, Jan. 3–7*, World Scientific Press. pp. 437–449
- 9 Bar-Joseph, Z. *et al.* (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.* 21, 1337–1342
- 10 Liao, J.C. *et al.* (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. U. S. A.* 100, 15522–15527
- 11 Kao, K.C. *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, 641–646
- 12 Li, Z. and Chan, C. (2004) Inferring pathways and networks with a Bayesian framework. *FASEB J.* 18, 746–748
- 13 Martoglio, A.M. *et al.* (2002) A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. *Bioinformatics* 18, 1617–1624
- 14 Pilpel, Y. *et al.* (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* 29, 153–159
- 15 Berman, B.P. *et al.* (2002) Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl. Acad. Sci. U. S. A.* 99, 757–762