CHAPTER 6

Model-Based Inference of Transcriptional Regulatory Mechanisms from DNA Microarray Data

Harmen J. Bussemaker*

Abstract

The development of DNA microarray technology has made it possible to monitor the mRNA abundance of all genes simultaneously (the transcriptome) for a variety of cellular conditions. In addition, microarray-based genomewide measurements of promoter occupancy (the occupome) are now available for an increasing number of transcription factors. With this data and the complete genome sequence of many important organisms, it is becoming possible to quantitatively model the molecular computation performed at each promoter, which has as input the nuclear concentration of the active form of various regulatory proteins (the regulome) and as output a transcription rate, which in turn determines mRNA abundance. In this chapter, we describe how our group has used multivariate linear regression methods to: (i) discover cis-regulatory elements in upstream regulatory regions in an unbiased manner; (ii) infer a regulatory activity profile across conditions for each transcription factor; and (iii) determine whether the mRNA expression level of a gene whose promoter is occupied by a particular transcription factor is truly regulated by that factor, through integrated modeling of expression and promoter occupancy data. Together, these results show model-based analysis of functional genomics data to be a versatile conceptual and practical framework for the elucidation of regulatory circuitry, and a powerful alternative to the currently prevalent clustering-based methods.

Introduction

The recent development of high-throughput genomics technologies has had a major impact on the gene expression regulation field. It has become feasible to study the cell from a systems point of view, as a network of interacting genes and their protein products. The genomes of many important model organisms, as well as that of *Homo sapiens*, have been sequenced.[1-4] This has given rise to the development of DNA microarrays as a tool for monitoring the mRNA transcript abundance of all genes in a cell simultaneously,[5,6] and more recently for performing genomewide profiling of the occupancy of noncoding DNA by transcription factors (TFs) using ChIP[7,8] or DamID.[9] The genomewide mRNA expression pattern is commonly referred to as the "transcriptome", while we here propose to refer to the set of genomewide TF occupancies (the terms "binding data" and "location data" are less accurate, in our opinion) as the "occupome".

*Harmen J. Bussemaker—Department of Biological Sciences, and Center for Computational Biology and Bioinformatics, Columbia University, New York, New York, U.S.A.
Email: Harmen.Bussemaker@columbia.edu

Discovering Biomolecular Mechanisms with Computational Biology,
edited by Frank Eisenhaber.

For higher organisms only a small fraction of the genome codes for proteins. The function of the remaining noncoding DNA is largely unknown. It is widely believed that the complexity of an organism crucially depends on the way its genes interact. The unexpectedly low number of protein encoding genes found in the human genome supports this view.[10] However, our understanding of the molecular mechanisms underlying the control of gene expression by regulatory proteins such as transcription factors that bind to noncoding DNA is still very limited, especially for higher eukaryotes. Through integrated analysis of mRNA expression data, transcription factor occupancy data and genome sequence, advances are likely to be made in the mechanistic understanding of the genomewide regulatory network.[11,12]

Two Classes of Tools for Finding Motifs from Expression Data

Knowing the mRNA expression level and the promoter sequence of each gene makes it possible to use computational methods to identify cis-regulatory elements (CREs). Among the various tools used to link gene expression data to cis-regulatory motifs, a fundamental distinction between two classes can be made:

- *Feature enrichment scoring methods ("Class A")* define a subset of genes of interest based on expression data (e.g., all genes upregulated above noise, or the output of a hierarchical clustering algorithm) and subsequently analyze the promoter regions of these genes for overrepresentation of specific sequence patterns.
- *Expression-based feature scoring methods ("Class B")* first define a subset of genes based on a expression-independent feature (e.g., genes whose promoter region contains a specific motif) and subsequently score this feature by comparing the average expression level of the genes containing the feature to the genomewide distribution of expression levels.

Most currently used motif-finding approaches belong to Class A.[13,14] The regression-based methods developed by our laboratory, discussed below, can all be viewed as belonging to Class B. By combining the signals of multiple genes on the microarray, Class B tools have greatly enhanced statistical power to detect differential expression at the level of multi-gene pathways. A change in activity for given transcription factor may be scored as highly significant even if no single gene controlled by that factor shows a change in mRNA expression above noise level.

A similar distinction can be made among methods that aim to combine functional annotation information from Gene Ontology[15] with gene expression data: set enrichment scoring using the cumulative hypergeometric distribution[16] belongs to class A, while methods that score the average expression of genes in each GO category[17,18] belong to class B, and are therefore more sensitive.

REDUCE: Motif-Based Regression Analysis of the Transcriptome

As a first step towards the goal of "reverse engineering" the cell-wide regulatory network from large functional genomics data sets, we recently developed a motif-based regression analysis method named REDUCE, an acronym for "regulatory element detection using correlation with expression".[19] Class A motif finding tools rely on the clustering of genes based on their expression profile across a large number of experimental conditions. By contrast, REDUCE fits a simple model for transcriptional control to a *single* genome-wide expression pattern measured using DNA microarrays. It not only identifies cis-regulatory elements (CREs) in noncoding DNA, but also infers changes in the nuclear concentration of the transcription factors that bind these CREs. Another unique feature of REDUCE is that it naturally takes into account the combinatorial nature of gene expression regulation by allowing multiple factors to control each gene in a unique way, defined by its promoter sequence. Several other groups have adopted and extended our model-based approach.[20-22]

At the core of REDUCE is the following linear model for transcriptional regulation:

$$A_g^{predicted} = \sum_{m \in M} F_m N_{mg}. \qquad (1)$$