Abstract

A cell's gene regulatory network refers to the coordinated switching on and off of genes by regulatory proteins that bind to non-coding DNA. Studying the behavior of gene regulatory networks using high-throughput genomic data, like gene expression data from microarrays and DNA sequence data, has become one of the central problems in computational biology. Most work in this area has focused on learning structure from data – for example, finding clusters of potentially co-regulated genes, or building a graph of putative regulatory "edges" between genes – and has been used to generate qualitative hypotheses about regulatory networks.

Instead of adopting the structure learning viewpoint, our focus is to build predictive models of gene regulation, i.e., models that allow us to make accurate quantitative predictions on new or held-out experiments (test data). In our approach, we learn a prediction function for the regulatory response of genes, using a boosting algorithm to enable feature selection from a high-dimensional search space while avoiding overfitting. In particular, we generate motifs representing putative regulatory elements whose presence in the promoter region of a gene, coupled with activity of a regulator in an experiment, is predictive of differential expression. We combine this information into a global predictive model for gene regulation. In experiments for the environmental stress response in yeast, our method is able to make accurate predictions about which genes will be up- or down-regulated in held-out (test) microarray experiments, discover true regulatory elements, and suggest interpretable biological hypotheses about regulatory mechanisms.