Machine learning methods for transcription data integration

D. T. Holloway M. Kon C. DeLisi

Abstract

Gene expression is modulated by transcription factors (TFs), proteins that generally bind to DNA adjacent to coding regions and initiate transcription. Each target gene can be regulated by more than one TF, and each TF can regulate many targets. For a complete molecular understanding of transcriptional regulation, researchers must first associate each TF with the set(s) of genes that it regulates. Here we present a summary of completed work on the ability to associate 104 TFs with their binding sites using support vector machines (SVM), a classification algorithm based in statistical learning theory.

We use several types of genomic datasets to train classifiers for TF binding prediction in the yeast genome. These include motif matches, subsequence counts, motif conservation, functional annotation, and expression profiles. A simple weighting scheme varies the contribution of each type of genomic data when building a final SVM classifier, which we evaluate on known binding sites taken from ChIP-chip experiments[1, 2], the Transfac Database[3], and the Yeast Proteome Database[4].

The SVM algorithm works best when all datasets are combined, producing 73% coverage of known interactions, with the percentage of true predictions at almost 0.9. New ideas and preliminary work for improving SVM classification on biological data are also discussed.