Abstract

Transcription factors (TFs) constitute key components of cellular transcriptional networks regulating such diverse processes as cell differentiation and proliferation. A regulatory code established via DNA-amino acid interactions thereby allows TFs to identify their short target sites even within the vast genomes of higher eukaryotes. TF-DNA interactions are highly degenerate, permitting a given TF to bind not only to a single sequence but to a broad variety of sites with varying strength. This promiscuity renders the accurate prediction of target genes for TFs a challenging task. Traditional computational solutions to the problem divide the sequence space into binding and non-binding sites. However, the emergence of large scale experimental binding data has highlighted the need for alternative approaches capable of accounting for the gradual binding strength of individual TF-DNA interactions.

The goal of this thesis is to develop a new method (called TRAP) that predicts the binding affinity of a transcription factor to a DNA sequence based on a biophysical model, which avoids binary separation between binding and non-binding sites. Correlating TRAP predictions with measured TF binding affinities thereby resulted in the derivation of a biophysically motivated prescription for generically setting the TRAP parameters. This prescription holds not only for TFs from yeast but also from higher organisms including *Drosophila* and human. The TRAP approach is shown to be both conceptually and practically more powerful than traditional hit based methods and to outperform alternative affinity based approaches that rely on a standard biophysical model.

In order to detect the regulatory association between TFs and entire groups of genes TRAP was embedded into a statistical framework called PASTAA, which analyzes the enrichment of potential TF target genes in user-defined gene sets by applying a series of hypergeometric tests. Using PASTAA for the analysis of sets of tissue specific genes not only recovered a more comprehensive number of experimentally known TF-tissue associations than alternative approaches but also allowed to draw a number of important biological conclusions. For instance, binding signals for tissue specific TFs were found to cluster in proximal promoters largely upstream of the respective transcription start site. The results of the analysis were found to be remarkably robust against changes in the sequence space as well as expression data.