Abstract

Despite the fact that each cell in an organism has the same genetic information, it is possible that cells fundamentally differ in their function. The molecular basis for the functional diversity of cells is governed by biochemical processes that regulate the expression of genes. Key to this regulatory process are proteins called transcription factors that recognize and bind specific DNA sequences of a few nucleotides. Here we tackle the problem of identifying the binding sites of a given transcription factor. The prediction of binding preferences from the structure of a transcription factor is still an unsolved problem. For that reason, binding sites are commonly identified by searching for overrepresented sites in a given collection of nucleotide sequences. Such sequences might be known regulatory regions of genes that are assumed to be coregulated, or they are obtained from so-called ChIP-seq experiments that identify approximately the sites that were bound by a given transcription factor. In both cases, the observed nucleotide sequences are much longer than the actual binding sites and computational tools are required to uncover the actual binding preferences of a factor. Aggravated by the fact that transcription factors recognize not only a single nucleotide sequence, the search for overrepresented patterns in a given collection of sequences has proven to be a challenging problem.

Most computational methods merely relied on the given set of sequences, but additional information is required in order to make reliable predictions. Here, this information is obtained by looking at the evolution of nucleotide sequences. For that reason, each nucleotide sequence in the observed data is augmented by its orthologs, i.e. sequences from related species where the same transcription factor is present. By constructing multiple sequence alignments of the orthologous sequences it is possible to identify functional regions that are under selective pressure and therefore appear more conserved than others. The processing of the additional information exerted by ortholog sequences relies on a phylogenetic tree equipped with a nucleotide substitution model that not only carries information about the ancestry, but also about the expected similarity of functional sites.

As a result, a Bayesian method for the identification of transcription factor binding sites is presented. The method relies on a phylogenetic tree that agrees with the assumptions of the nucleotide substitution process. Therefore, the problem of estimating phylogenetic trees is discussed first. The computation of point estimates relies on recent developments in Hadamard spaces. Second, the statistical model is presented that captures the enrichment and conservation of binding sites and other functional regions in the observed data. The performance of the method is evaluated on ChIP-seq data of transcription factors, where the binding preferences have been estimated in previous studies.