PROBABILISTIC METHODS TO SEARCH FOR REGULATORY ELEMENTS IN SETS OF COREGULATED GENES

Over the last years, emerging technologies, like microarrays and high-throughput sequencing, have boosted the number of available biological data. This wealth of data demands for new algorithms that can find relevant information in these data. It is here that bioinformatics enters the arena. At this intersection between molecular biology and informatics we can situate this dissertation. In this dissertation we introduce a suite of tools to search for potential transcription factor binding sites based on a probabilistic sequence model and sets of coregulated genes. We have approached the problem from two different angles: supervised and unsupervised.

To tackle the unsupervised problem, we have extended the original Gibbs sampling algorithm for motif finding. First we have introduced higher-order background models to better discriminate between true motifs and background noise. Second, we have used the probabilistic framework to estimate the number of instances of a motif in a sequence. This has resulted in **MotifSampler**. Through a thorough analysis of the influence of the different parameters on the performance of the algorithm we have gained insight in the behavior of our implementation. This analysis has allowed us to present a motif finding procedure that can be applied in real biological examples. In this dissertation we discus in detail the full scale analysis of four different data sets: 33 sequences with the G-box binding sites in plants, 10 regulons in yeast, the σ^{54} factor in prokaryotes and four clusters of co-expressed genes in the yeast cell cycle. The great diversity of these examples nicely illustrates the capabilities and limitations of our methodology. The most important result is that a well designed species-specific background model significantly improves the performance of the motif finding algorithm especially when an high level of noise is present in the data set.

Based on the probabilistic sequence model, we have also implemented **MotifScanner** to search for instances of known motifs in DNA sequences. This is the supervised approach. Again we study in detail the influence of the parameters on the number of instances retrieved. The proposed methodology turns out to be more robust to parameter chances than a classical position-weight matrix scoring scheme. If a set of matrices is available, it is also possible to screen a set of promoter sequences and assess the statistical significance of the number of instances found. As a reference we use the expected number of instances found in all promoter sequences in the genome. The examples in yeast show that this approach is applicable but that it is limited by the quality of the motif models present in the database.

Finally, we discus the implementation of **INCLUSive**, an integrated web-based platform for the analysis of microarray data. The analysis starts with Adaptive Quality-Based Clustering of gene expression measurements. This results in a several clusters of genes that have a similar expression profile. The next step is to look at the sequences of the genes in such a cluster, more specifically at the promoter region of these genes. Therefore, we have implemented an upstream sequence retrieval system to locate the intergenic region on the genomic DNA. The selected sequences can be entered at the web interfaces of both the MotifSampler and the MotifScanner. To illustrate the applicability of our methods, we refer to the work of other researchers who have found specific motifs with MotifSampler in their sets of coregulated genes.

Promotor: Prof. dr. ir. Bart De Moor

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTY of ENGINEERING
DEPARTEMENT ELECTRICAL ENGINEERING
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

UDC: 681.3*F2:575.113