Discovery of Regulatory Motifs in DNA
(*Abstract*)

Esko Ukkonen

Department of Computer Science
University of Helsinki
P.O. Box 68, FI-00014, Finland
esko.ukkonen@cs.helsinki.fi

In biological sequence analysis, representation and discovery of various DNA motifs with a biological function is a central task. In particular, identification of regulatory elements such as the binding sites in DNA for the so-called transcription factors is an important step in the attempts to understand the regulation mechanisms of gene expression. Transcription factors are proteins that may bind to DNA, typically close to transcription start site of a gene. Such a binding may activate or inhibit the transcription machinery of the associated gene. As the regulated gene may again be a transcription factor, such pairwise regulatory relations between genes induce a genome-wide network model for gene regulation.

The possible binding sites of a transcription factor are short DNA segments. The DNA sequences of different sites are close variants of an underlying consensus sequence. For most transcription factors no biophysical model of this variation is currently known. Hence simplified formal representations of binding motifs have been used: a motif is represented as a set of weighted DNA sequences that may occur in the binding site, or a probabilistic Markov model of order 0 or 1 or higher is used. Here Markov model of order 0 is usually called a position weight matrix (PWM). Once we have available training DNA sequences that contain enriched amounts of instances of the motif, we may estimate the motif in our representation class that fits best the training data.

The talk surveys the representations and corresponding discovery algorithms for transcription factor binding motifs. We consider basic motifs for single factors (monomers) as well as composite motifs for pairs of factors (dimers) and for chains of factors. Such chains are models for regulatory modules built of clusters of several factors making together a regulatory complex. We will discuss both the EM algorithm based learning of motifs which is the dominating approach in practice as well as a novel seed-driven approach aiming at faster learning. The role of string algorithms in these methods will also be discussed.

Esko Ukkonen: Discovery of Regulatory Motifs in DNA, p. 1.
Proceedings of PSC 2018, Jan Holub and Jan Žďárek (Eds.), ISBN 978-80-01-06484-9