A novel method for finding oligonucleotide regulatory elements of unknown length in DNA sequences

G.Pavesi (1), G.Mauri (1), G.Pesole (2)

(1) Dipartimento di Informatica, Sistemistica e Comunicazione, Università di Milano-Bicocca
(2) Dipartimento di Fisiologia e Biochimica Generali Università di Milano

Signal finding (pattern discovery) in biological sequences is one of the most challenging problems in computational biology. Indeed, such signals that correspond to oligonucleotide patterns significantly over-represented in a set of functionally equivalent sequences are good candidates of some functional activity in the regulation of gene expression. In particular, we focus on the problem of finding, in a set of unaligned DNA sequences, all the patterns that appear in every sequence of the set. The proposed method allows to search for common approximate patterns, that is, they can present mutations, insertions and deletions. While searching for approximate occurrences of a specific pattern is a problem that can be solved efficiently in different ways, exhaustive enumeration, that is, searching for all the possible patterns up to a given length to find out which ones actually occur in all the sequences is computationally time consuming, and feasible only for very short signals with a limited number of errors. So far, several methods based on different heuristics have been proposed, in order to obtain faster results. In some cases, the search is limited to a subset of the search space, where patterns are selected according to different criteria. The problem is that the signal is sometimes so complicated that can be completely missed simply because the algorithms do not search it. Other approaches try to identify the signal by sampling or analyzing the patterns occurring in the set of sequences. In this case, the performance degrades quickly as the length of the sequences grows, since the real signal is lost in the noise of random spurious signals. The problem can be thus considered to be far from being solved. Moreover, in some cases the exact length of the pattern to be sought must be known in advance and provided as input to the programs. The algorithm we present is based on exhaustive enumeration, and needs in advance only an error ratio e ($0 < e < 1$), such that a pattern of length k can appear with at most ek mutations in the sequences. The algorithm first builds a suffix tree for the set of sequences, and then searches for patterns by performing depth-first traversals of the tree itself. The speed up in our approach is obtained by requiring mutations to be spread more or less uniformly along the pattern. In other words, a pattern of length k can occur with at most ek errors, but at most one error can be found in the first 1/e symbols, at most two errors in the first 2/e, and so on. For example, a valid occurrence of pattern of length 12 with error ratio 0.25 can present at most one mismatch in the first four symbols, two mismatches in the first eight symbols, and three mismatches in all. Although the search is limited to a subset of all the possible approximate occurrences of the patterns, we show how the algorithm can be also applied to general instances of the problem, where no constraints are imposed on the location of mismatches. Moreover, it is possible to give an estimate of the probability of missing a signal based on the number of sequences, the length of the pattern, and the error ratio. Finally we show how the algorithm can be extended to find gapped signals, or to work with different error measures.