Pattern Discovery in Unaligned Biological 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

In the last few years, the growing amount of data generated by the biological community (i.e. genomes, proteomes) has brought under the spotlight the problem of developing efficient and reliable tools for the analysis of the data. Among the issues arising in this field, pattern discovery is one of the most challenging problems both from the computational and from the biological point of view.

In its most general form, the problem can be described as follows. Given a set of functionally related sequences, find all patterns that appear (possibly in different but similar enough forms) in each sequence, or at least in a significant number of sequences of the set. Those patterns could correspond to the regions of the sequences responsible for their function, and could be used later for the functional annotation of newly determined sequences.

Many different methods and computational tools have been devised for this task, each one trying to find an optimal trade off between the accuracy of the results and the efficiency (time and space required) of the algorithm.

The main aim of this talk is to provide a survey of the most widely used methods, trying to point out which ones can be considered the most suitable according, for example, to the type and size of the input sequences, to the degree of approximation allowed for patterns, to the size of patterns sought, and so on.