A new algorithm for identifying words in biological sequences

Brenda GARCIA-MAYA and Nikolaos LIMNIOS

Sorbonne University, Université de technologie de Compiègne,

LMAC Laboratory of Applied Mathematics of Compiègne

Compiègne, France

Abstract

Genome sequencing is often compared to a code ruled by four nitrogenous bases Cytosine (C), Guanine (G), Adenine (A) and Thymine (T). In a sense, a genomic sequence is a string of letters in an unknown language. Given a sequence of letters of a fixed alphabet, it is interesting to recognize a word (pattern) within this progression. This word could be, for example, the responsible for the production of a particular protein, a genetic disorder, etc. Nevertheless, to achieve this task over a sequel of a few hundred thousand letters is a very cumbersome job to do by simple inspection. The automatic search is therefore indispensable. Several models have been proposed under the assumption where the sequence is described by a Markov chain. The Markovian hypothesis imposes restrictions on the distribution of the sojourn time in a state, which should be geometrically distributed in only a discrete chain. This is the main drawback when applying Markov chains in real applications. We propose a new algorithm under the hypothesis that the sequence of letter is modeled by a semi-Markov chain. To achieve this objective we use the auxiliary prefix and backward time chain. The corresponding probability distribution, the mean waiting time and the variance are obtained. We give a numerical example from a biological DNA sequence.

Keywords: semi-Markov chain · first occurrence of a word · prefix process · backward process · DNA analysis.