Abstract

The large number of genome sequencing projects in progress and the resulting increase in the volume of uncharacterized data has motivated the search for more precise and efficient computacional methods for identifying the structures that compose the DNA of living beings. In particular, due to its great importance, the search for protein coding regions has been the focus of research for at least twenty years. Coding regions carry in its nucleotides the information necessary to the cellular structures to produce proteins, fundamental component of most living organisms.

The identification of coding regions in DNA sequences is still a difficult problem since the complex cellular mechanisms involved in the process of protein production are not completely known.

In this dissertation, we have developed a statistical method for the identification of protein coding regions. The method is based on Bayes's theorem applied to strings of k consecutive DNA bases, where k is a parameter specified by the user. To compute the conditional and a priori probabilities needed by Bayes's theorem, we use certain hypotheses on the independence of codons and bases, and on the minimum size of coding and non-coding regions, that reduce the computational cost and the size of probability tables. In performed tests the proposed method has presented promising results.

Keywords: Protein coding sequences, Bayes's theorem, pattern recognition, bioinformatics.