Extraction of recurrent motifs synthetic genomic sequences via dictionary-based compression

Menconi G (1), Dionisi F (2), Marangoni R (3)

(1) Department of Mathematics, University of Bologna, Bologna.
(2) ProteoGen Bio S.r.l., Pisa.
(3) Department of Informatics, University of Pisa, Pisa.

Motivation

Linguistic analysis of symbol sequences has a natural application to genomic sequence analysis. The large extent of biological databases paves the way for both a large-scale query of recurrent words in complete genomes and a selective search of important motifs strictly related to determined functional meaning or specific for a gene family. Words should be selected in order to be reliable and faithful as to linguistics as to biology. From a technical point of view, the methods of motif extraction should be sufficiently fast and computationally light also to allow an on-line fruition to be set.

Methods

The proposed method is based on the use of CASToRe, a dictionary-based compression algorithm of the Lempel-Ziv family. The algorithm selects a dictionary by exact matches and parses the input sequence in some variable-length recurrent words. The algorithm CASToRe is a very efficient complexity detector and it was already successfully used in genome clustering and coding sequence identification in Prokaryotic genomes. The input sequence is parsed in subwords belonging to the final dictionary relative to the sequence. A weight function is defined on the dictionary and a score is assigned to each word, based on its occurrence and length. Eventually, the words with highest score are collected in the so-called "set of interesting words": intuitively, they are the longest words occurring repeatedly along the input sequence. Preliminary applications of such approach on real eukaryotic genomic sequences have produced hard to interpret results, due to the high complexity of eukaryotic genomes. To bypass this problem, we focused our interest on synthetic sequences, generates by simulating specific symbol frequency distributions, and investigating the behaviour of the dictionary words on these artificial conditions.

Results

We generated several data sets of: periodic, Bernoullian and noise-perturbed periodic binary strings, in order to test the validity of the method. As a first conclusion, the score selects the more distinctive patterns within the sequence; for instance, the only interesting word in periodic sequences is the period pattern, while in coding genomic sequences the set of interesting words is mainly made of codons. Going into deeper details on the performed analysis, for each data set we have taken under consideration the distribution of word length and word occurrence within each sequence of the set. Moreover, we have studied the weight function against word length. For what concerns periodic sequences, we have also focused on the dependence of the average word score on the period length; in the case of the Bernoullian sequences, the same index was studied with respect to the characteristic probability p; furthermore, we concentrated on how the noise intensity affects the results on some periodic strings.

Contact email: marangon@di.unipi.it