PhD Proposal MAP-I

Analysis of DNA sequences through compression-based complexity profiles

Supervisor: Armando J. Pinho (ap@ua.pt)

Research Unit: Instituto de Engenharia Electrónica e Telemática de Aveiro (IEETA)

1 Motivation

The complexity profile of a DNA sequence is a numerical sequence of the same length that indicates a measure of the predictability of each DNA base. Complexity profiles are important because they allow, for example, looking for repetitive structures inside a chromosome or across several chromosomes. These structures are often associated with regulatory functions of DNA. Moreover, the complexity profiles can also be used in finding evolutionary distances and, therefore, in the construction of phylogenetic trees.

The analysis of the huge amounts of genomic data that are continuously been generated puts a number of challenging problems to several research areas and, particularly, to the areas of computational biology and bioinformatics. One of these challenges is related with the problem of generating complexity profiles of DNA sequences in an efficient way, because these sequences might be as large as entire genomes (for example, the human genome is composed of about 3000 million bases).

The theory behind these complexity profiles goes back to the works of several researchers in the 60's, such as Solomonoff, Kolmogorov, Chaitin and Wallace et al., and is tightly related to the area of data compression. The profiles are obtained using compression algorithms, because the size of the bitstream generated by a compression algorithm can be viewed as an upper bound of the Kolmogorov complexity of the compressed object. The problem is that the existing DNA compression algorithms are too time demanding. For example, the best performing DNA compression techniques, such as NML-1 or XM, could take hours for compressing a single human chromosome.

2 Objectives

During the course of our ongoing work on DNA data compression, we have obtained several encouraging and important results related to DNA modeling based on finite-context models (also known as Markov chain models). Probably the most important of those results is the finding that DNA can be much better represented by Markov models than what it was previously believed. Although still not as compression effective as NML-1 or XM, the finite-context models have been already able to produce preliminary complexity profiles that are almost identical to those produced by the much more time consuming XM algorithm. In fact, the difference in time requirements is overwhelming. For example, compressing the human chromosome number 2 with the techniques based on finite-context models takes less than ten minutes in a 1.66 GHz laptop computer. This DNA sequence has about 240 million bases.

In this project, it will be studied the full potential of using compression-based complexity profiles for the analysis of long DNA sequences. This will be done in two major, although related, research topics. One is intra-species analysis, where the major goal is to explore the information conveyed by the complexity profiles for locating and classifying repetitive structures occurring inside a chromosome or across several chromosomes of the same species. The other topic is inter-species analysis, where the main goal is to use the complexity information for computing evolutionary distances among the species. In both topics, finite-context modeling will play a key role.