Investigating some attributes of periodicity in DNA sequences via semi-Markov modeling

Pavlos Kolias[1] and Aleka Papadopoulou[1]

[1] Department of Mathematics, Aristotle University of Thessaloniki, Thessaloniki, Greece, pakolias@math.auth.gr, apapado@math.auth.gr

Abstract - Periodicity is a structural property of DNA sequences. It is expressed as either nucleotides or words of nucleotides that appear with specific fixed distances in-between. Mainly, there have been observed two types of periodic behaviours in DNA. The first one has been observed in chromatin, which is a basic element of the cell nucleus. The researchers observed that certain di-nucleotides in the DNA of chromatin tends to appear at approximately every 10 to 11 bases. Subsequent studies suggested that the period of chromatin sequences converges to 10.4 bases. Also, more recent studies, which investigated the genome of three organisms, A. thaliana, C.elegans and H.sapiens, suggested that the di-nucleotide AA has almost perfect 10.5-base periodic behaviour in those organisms. One explanation about this type of periodicity is that the distance of 10.5 bases is exactly the "step" of the double strand, which curves the DNA chain and allows these long sequences to suppress into the small area of the nucleus. The second type of periodicity has been observed in areas of the genome that are transcribed and later translated into proteins, also called coding regions. Previous studies have used methods from mathematical analysis, such as the spectral density, and they have shown that in coding regions, there is a tendency of certain nucleotides to reappear every 3-bases. Also, this type of periodicity has only been observed in coding regions, while for non-coding regions similar periodic behaviour has not been observed. As each of the amino acids is encoded with a triplet of nucleotides (codons) and some specific amino acids are more abundant than others, authors concluded that the periodic behaviour, in fact exists, due to this higher frequency of certain amino acids and the period of 3-bases is sue to the triplet nature of the DNA. As the whole genome of each organism is frequently of several billions bases, the information about the periodic behaviour of the coding regions of the DNA would be really helpful into detecting those regions and distinguish between protein encoding regions and non-coding regions. Some algorithmic techniques have already been implemented using this information and they have used mathematical methods, such as the Fourier transformation. Also, some other well-known algorithms use hidden-Markov models, in order to classify between different regions of DNA. In this paper, a semi-Markov model is applied to 3-base periodic sequences, which characterize the protein-coding regions of the gene. Analytic forms of the related probabilities and the corresponding indexes are provided, which yield a description of the underlying periodic pattern. Last, the previous theoretical results are illustrated with synthesized and real data from different organisms.

Keywords - DNA sequences, Periodicity, Semi-Markov chain.