Protein Coding Regions Identification through the Modified Morlet Transform

Jesús Pascual Mena-Chalco
Institute of Mathematics and Statistics - University of Sao Paulo - Brazil
Roberto Marcondes Cesar-Jr
Institute of Mathematics and Statistics - University of Sao Paulo - Brazil

Abstract

An important topic in biological sequences analysis area is the protein coding regions identification. This identification allows the posterior research for meaning, description or biological categorization of the analyzed organism [1]. Currently, several methods combine pattern recognition with knowledge collected from training datasets of known genes or from comparison with genomic databases. Nonetheless, the accuracy of these methods is still far from satisfactory. New methods of DNA sequences processing and genes identification can be created through *search-by-content* such sequences [2]. The periodic pattern of DNA in protein coding regions, called three-base periodicity (TBP), has been considered proper of coding regions. This phenomenon was not observed for nonprotein coding. The digital signal processing techniques supply a strong basis for regions identification with TBP [2,3].

In this work we introduce a new method for protein coding regions identification with TBP, based on a wavelet transform, called Modified Morlet Transform (MMT), which does not need to be trained on sequences databases. We use a fixed binary mapping rules to create four binary sequences. Where each one represents the positions of each nitrogenate base in DNA sequence. Next the MMT, with different scales is applied to all binary sequences. The module of each normalized coefficient is projected onto the position axis. Projection onto the scale axis reveal which scale carry more signal energy throughout the positions. The result of the projection position axis represents the protein coding region identificator. These projection coefficients correspond to regions with TBP. Thus, we use thresholding coefficients, based on both shrinking values and inflection points, to exclude positions where the associated energy is lower. At the moment, we consider arbitrary length region criterions for discarding possible very short protein coding regions identification. The performance of the proposed transform was examined by analyzing synthetic and real DNA sequences (RGRC2 and F56F11.4 genes of *O. sativa* and *C. elegans* organism, respectively). Preliminary results show that MMT is better than traditional methods by presenting greater sensitivity to TBP and discriminatory capability between protein coding regions.

References

[1] A. W.-C. Liew, H. Yan, and M. Yang, Pattern recognition techniques for the emerging field of bioinformatics: A review, Pattern Recognition 38 (2005), no. 11, 2055-2073.

[2] X. Zhang, F. Chen, Y. Zhang, S. C. Agner, M. Akay, Z. Lu, M. M. Y. Waye, and S. K. Tsui, Signal processing techniques in genomic engineering, Proceedings of the IEEE 90 (2002), no. 12, 1822-1833.

[3] J. P. Mena-Chalco, H. S. Alves, H. Carrer and R. M. Cesar-Jr, Bioinformatics tools for assembling and analysis of chloroplast genomes, International Conference on Bioinformatics and Computational Biology (2004), Rio de Janeiro.