Chapter 7
Multi-Fractal Analysis for Feature Extraction from DNA Sequences

Witold Kinsner
University of Manitoba, Canada

Hong Zhang
University of Manitoba, Canada

ABSTRACT

This paper presents estimations of multi-scale (multi-fractal) measures for feature extraction from deoxyribonucleic acid (DNA) sequences, and demonstrates the intriguing possibility of identifying biological functionality using information contained within the DNA sequence. We have developed a technique that seeks patterns or correlations in the DNA sequence at a higher level than the local base-pair structure. The technique has three main steps: (i) transforms the DNA sequence symbols into a modified Lévy walk, (ii) transforms the Lévy walk into a signal spectrum, and (iii) breaks the spectrum into sub-spectra and treats each of these as an attractor from which the multi-fractal dimension spectrum is estimated. An optimal minimum window size and volume element size are found for estimation of the multi-fractal measures. Experimental results show that DNA is multi-fractal, and that the multi-fractality changes depending upon the location (coding or non-coding region) in the sequence.

INTRODUCTION

Deoxyribonucleic acid (DNA) has become one of the most examined molecules on the planet. Scientist around the world have been trying to unravel its secrets for many purposes. Genetic information is currently used to raise better plants and animals, create enhanced pharmaceuticals for humans, and for gene therapy in medicine, as well as for a plethora of other possible applications. Of particular interest is carbon-based DNA computing (e.g., Paum, Rozenberg, & Salomaa, 1998; Lipton & Baum, 1995; Bell & Marr, 1990). Science, as a whole, has benefited from the study

DOI: 10.4018/978-1-4666-0264-9.ch007

of genetics because of the increased understanding of biological processes that all organisms share.

In recent decades, a significant amount of research has been directed towards **sequencing and understanding** the entire human genome through the Human Genome Project (HGP) launched in 1986. The goal of the HGP was to find the location of the approximately 1×10^5 human genes, and to read the entire sequence of the human genome (about 3×10^9 base pairs, bp). An exponential grow rate of that research has resulted in reaching the goal in 2003, 50 years after the formulation of the double helix (e.g., Kieleczawa, 2008; Clayton & Dennis, 2003; França, Carrilho, & Kist, 2002). Many other genomes have also been sequenced since. For example, the rice and the mouse genomes were completed in 2002, followed by the genomes of the rat and chicken in 2004. A year later, the genomes of the chimpanzee and the dog were completed. The Cancer Genome Atlas (TCGA) pilot project started in 2005, and was escalated to a large-scale project in 2009. Modern high-throughput genome analysis techniques have accelerated considerably the rate not only of sequencing, but also of arriving at significant results and insights (e.g., Defense TechBriefs, 2009; TCGARN, 2008; Mitchelson, 2007).

On the other hand, the traditional DNA analysis methods of finding genes and their location at chromatosomes through testing their biological function have been inherently slow. Although numerous faster techniques have been developed, there is still a need to augment them with new approaches. Therefore, robust computational solutions to the gene-finding problem could provide a valuable resource for the HGP and for the molecular-biology community (e.g., Datta & Dougherty, 2006; Nunnally, 2005).

Today, we can already spell out the entire alphabet in the entire genomes of a variety of organisms. We have also learned many individual "words" (genes) in the DNA sequence. To follow the analogy of a language, this could represent a written language. How is this written language expressed (spoken)? We have learned that the three-character codons are like phonemes in the spoken language, with several codons producing the same results (protein). We have also learned a number of regulatory elements in the sequence that could signify the punctuation in the written language, and intonation in the spoken language. Unfortunately, we can comprehend just a few words in that language. We still must learn how to interpret those instructions in a more comprehensive manner. What is the Rosetta stone for DNA? This paper is intended to be a step towards this solution.

Most of the current research in the deciphering the meaning of DNA sequences is approached from the lowest base-pair level. Its main objective is to search for patterns or correlations existing in the DNA sequence related to codons, amino acids, and proteins. A number of gene-finding systems have been developed in recent decades. These systems use a variety of sophisticated computational data-miming techniques, including neural networks (Uberbacher & Mural, 1991), dynamic programming (Snyder & Stormo, 1993), rule-based methods (Solovyev, Salamov, & Lawrence, 1994), decision trees (Hutchinson & Hayden, 1992), probability reasoning (Guigo, Knudsen, Drake, & Smith, 1992), and hidden Markov chains (Henderson, Salzberg, & Fasman, 1997), and other machine-learning schemes such as genetic programming, and support vector machines (e.g., Baldi & Brunak, 1998). Most of the approaches are based on **local** measures only. In addition, many of the techniques rely on the statistical qualities of exons in the gene, thus using only the *known* gene pool as a training set for their classification. Although the model-driven techniques have demonstrated some success, improved techniques should be developed.

An approach to finding such improved techniques is to consider long-range relations (in addition to short-range relations) in the DNA sequence, spanning 10^4 nucleotides, (Voss, 1992; Karlin & Brendel, 1993; Borovik, Grosberg, & Frank-