Disponible en ligne sur
ScienceDirect
www.sciencedirect.com

Elsevier Masson France
EM|consulte
www.em-consulte.com

IRBM ●●● (●●●●) ●●●–●●●

IRBM

TAIMA

Inferring Helitron Structures from 1D and 2D Representations Based on the Chaos Game Theory

I. Messaoudi [*], A. Elloumi Oueslati, Z. Lachiri

Université de Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis, LR Signal, Images et Technologies de l'Information, BP 37, le Belvédère, 1002, Tunis, Tunisia

Received 18 October 2015; received in revised form 10 September 2016; accepted 11 January 2017

Abstract

This work introduces a new procedure to highlight helitron structures in the C. elegans DNA sequences. In this sense, the DNA strings are converted to numerical representations either in one or two dimensions. As a one dimensional coding, we choose the Frequency Chaos Game Signal (FCGS) which encodes the DNA based on the Chaos Game theory. Afterwards, a time–frequency representation – generated by the complex Morlet wavelet analysis – transforms the obtained signal into a 2D representation. Through the examples furnished in this paper, we demonstrate the accuracy of the method in emphasizing the helitron borders especially with high orders of FCGS. A comparison with the 2D spectrogram representation confirms the obtained results.

Keywords: Frequency Chaos Game Signal; Complex Morlet wavelet transform; Short-time Fourier transform; Helitrons; C. elegans

1. Introduction

The chromosomal DNA sequences are known by their heterogeneity in terms of length, composition, structure and function; which makes their study a difficult task. Among the complex DNA structures that attracted the scientists in last decades we turn our attention to helitrons. The Helitron is a specific class of Transposable-Elements (TEs) which is present in large portion of eukaryotic genomes [1]. In general, TEs can catch gene fragments and transpose them within the genome through the use of a specific mechanism (from where name "jumping genes"). This operation can lead to changes in nucleic acid sequences like insertion and deletion mutations; which contribute in turn in the genomes evolution [2]. In the case of helitron, the transposition mechanism is known by rolling-circle. The introduction of this structure was done by Vladimir Kapitonov and Jerzy Jurka in 2001 [3]. Because of its high polymorphism (heterogeneity in size and sequence composition) as well as the lack of typical features in its structure, the automated identification of helitron remains particularly recalcitrant. Nevertheless, a large number of helitrons contain TC dinucleotide downstream the 5'-end and palindromic sequences upstream the 3'-end which helped in someway to identify these elements. In this context, some analysis tools have been proposed. For example, Recon [4] and Spectral Repeat Finder [5] are based on the search of the repetitive sequences [2]. In other hand, HelitronFinder both uses the 3'-end and the palindrome sequences to detect helitrons [6]. HelSearch is another program which manually searches for the end structures in helitrons like hairpins and loops. After that, it compares these structures with known consensus sequences using MUSCLE (an alignment algorithm) [2,7]. The main drawback of these methods is the necessity of a prior knowledge about the helitron repetitive sequences and termini. In this respect, results obtained by these techniques are not accurate since full lists of consensus sequences do not exist. The techniques introducing an alignment module, require also an important runtime and most of them are obstructed by the sequences length. In line with this, we furnish two ways to

[*] Corresponding author.
E-mail addresses: imen.messaoudi@enit.rnu.tn (I. Messaoudi), Afef.Elloumi@enit.rnu.tn (A. Elloumi Oueslati), Zied.lachiri@enit.rnu.tn (Z. Lachiri).

http://dx.doi.org/10.1016/j.irbm.2017.01.002
1959-0318/© 2017 AGBM. Published by Elsevier Masson SAS. All rights reserved.

highlight helitron sequences. The first method consists in coding the DNA strings into one dimensional signal on the base of the Chaos Game theory. We give the name of the "Frequency Chaos Game Signal" (FCGS) to this coding. As time series, the FCGS allows following the frequency evolution of nucleotides' occurrence along the genome [8]. With high orders of the FCGS representation we can easily detect the helitron presence without the need of any prior knowledge about the enhanced region. The same thing goes to the second method which is a two dimensional representation of DNA using wavelets. The latter technique consists in applying the complex Morlet wavelet to the FCGS signal. As a result, the time–frequency repartition is shown to provide striking signatures of helitrons. In both the time and the time–frequency representations, the boundaries of helitrons are highlighted when we increase the FCGS level. In the light of this, we divide the paper into four sections. Section 2 gives an overview on our coding technique: the Frequency Chaos Game Signal (FCGS) as well as the complex Morlet wavelet analysis. Section 3 exposes the results. Finally, Section 4 concludes the work.

2. Coding the DNA sequences into one dimensional and two dimensional signals

The DNA sequences are long chainlike molecules composed of four bases 'A' (Adenine), 'T' (Thymine), 'C' (Cytosine) and 'G' (Guanine). To ensure a direct interpretation of DNA, these characters could be transformed into a series of numerical values. In this section, we provide two ways to represent the DNA sequences into explicit signals and images.

2.1. Coding DNA by the Frequency Chaos Game Signal

The Frequency Chaos Game Signal (FCGS) is a new DNA coding which replaces the DNA characters by their frequency of occurrence in the genome. The fundamental concept is inspired from the Chaos Game theory. The procedure consists in calculating the matrix of words occurrence frequency for the whole chromosome using the Frequency Chaos Game Representation (FCGR). The calculus depends on a defined word length (k). Effectively for each k-lengthen word we extract the appropriate frequency of appearance; then we assign this value to the considered position [8]. As illustrative example, we consider the sequence "ACCTAGCTGGA" which we encode by three levels of FCGS: FCGS$_1$, FCGS$_2$ and FCGS$_3$ (see Fig. 1).

In general, the choice of the FCGS level (k) is arbitrary; but for ease of calculation we achieve the eleventh level (FCGS$_{11}$).

2.2. Mapping DNA sequences into images based on the Complex Morlet wavelet analysis

Because the DNA contains a broad scope of regular structures with varying size and composition, one must choose a tool that operates like a microscope and serves therefore to capture all the important trends in these sequences. In this framework, the wavelets offer a powerful way to view the details of time varying and transient phenomena in genomic signals. The wavelets offer a multi-resolution decomposition of a signal by decomposing it into its elementary constituents across scale [9]. According to the following equation, the continuous wavelet transform (CWT) of a signal $X(t)$ is obtained by convolution with a set of compressed and dilated versions of a specific function called mother wavelet $\psi(t)$:

$$T_\psi(X)(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} X(t)\psi^\star(\frac{t-b}{a})\,dt \tag{1}$$

There are a variety of types of wavelet functions. The Complex Morlet wavelet is proven to be the most appropriate wavelet to study transients and non-stationary signals (such is the case of the genomic signals). It is a Gaussian-windowed complex sinusoid defined by:

$$\psi(t) = \pi^{-\frac{1}{4}}\,(e^{i\omega_0 t} - e^{-\frac{1}{2}\omega_0^2})e^{-\frac{1}{2}t^2} \tag{2}$$

Here ω_0 corresponds to the number of oscillations of the wavelet [10].

The absolute value of the obtained coefficients is then plotted in the time-scale plan. It is also possible to use the frequency for representing the wavelet coefficients instead of the scale since the frequency set is proportional to scale one [10,11]. This representation is called scalogram and communicates the time frequency localization property of signals. In this work, we use the complex Morlet wavelet analysis to map DNA as a 2D image (scalogram).

3. Results and discussions

It is well-known that helitrons are complex DNA entities; their identification remains difficult despite the continued efforts to this end. The approaches used up to now have not taken into account the signal processing tools to detect the presence of helitrons. In fact, most of the proposed works within the framework of genomic signal processing were concentrated about segmenting genes into coding (exon) and non-coding regions based on the 3-bp periodicity in exons. For this, different techniques have been used: the smoothing [12], the filtering [13], the Auto-Regressive technique [13,14], the Choi–Williams Distribution [15,16], the Fourier Transform [16,17], the Wavelet Transform [18,19] and the Pitch Synchronous analysis [20]. In this work, we supply new perspectives in terms of DNA visualization and highlighting its hotspots. The first method, the Frequency Chaos Game Signal, is a 1D coding which represents a natural way for segmenting DNA without the need for any preprocessing or analysis technique. In the following we will demonstrate the effectiveness of this method in revealing the helitron entity in the C. elegans genome. For this, we consider an example of Helitron Y1A (which positioned on the chromosome I at [1317449–1319319]). The data is retrieved from the NCBI database http://www.ncbi.nlm.nih.gov/. Since the FCGS method offers a multitude of representative signals (FCGSs) for a same input sequence, we choose to code the chromosome I by the first eleven levels of FCGS. The FCGSs representations of the considered helitron are given in Fig. 2. For this example, amplitudes are multiplied by 100 for clarity purpose. Further,