Multimedia Tools and Applications (2019) 78:13047–13066
https://doi.org/10.1007/s11042-018-6455-x

A combined support vector machine-FCGS classification based on the wavelet transform for Helitrons recognition in C.elegans

Rabeb Touati [1] · Imen Messaoudi [1,2] · Afef Elloumi Oueslati [1,3] · Zied Lachiri [1]

Received: 15 September 2017 / Revised: 18 July 2018 / Accepted: 23 July 2018 /
Published online: 29 August 2018

Abstract

The Helitrons, an important sub-class of the transposable elements (TEs) class II, have been revealed in diverse eukaryotic genomes. They are mobile elements with great impact on genomic evolution. Till today, there is no systematic classification model of helitrons; that's why we thought of creating an efficient automatic model to identify these sequences. This paper focuses on the discrimination between helitrons and non-helitrons using the Support Vector Machine (SVM). In this study, we use all the SVM kernels and the higher accuracy rates are obtained by reaching the optimal kernels-parameters (d, c and σ). Further, we introduce two methods to represent the genomic sequences in the form of features to be considered later for the classification task: (i) the temporal and the spectral features extracted from the Frequency Chaos Game Signals order 2 ($FCGS_2$) (ii) the features extracted from the Continuous Wavelet Transform (CWT) applied to the $FCGS_2$ signals. The dataset we used regards two types DNA classes in C.elegans: the helitrons and the repetitive DNA sequences that contain microsatellites and do not form helitrons. The classification results prove that the wavelet energy feature is more effective than the

✉ Rabeb Touati
Rabeb.touati.enit@gmail.com

Imen Messaoudi
imen.messaoudi@enit.rnu.tn

Afef Elloumi Oueslati
Afef.Elloumi@enit.rnu.tn

Zied Lachiri
Zied.lachiri@enit.rnu.tn

[1] SITI Laboratory, National School of Engineers of Tunis (ENIT), University Tunis El Manar, BP 37, le Belvédère, 1002 Tunis, Tunisia

[2] Higher Institute of Information Technologies and Communications, Industrial Computing Department, University of Carthage, Carthage, Tunisia

[3] National School of Engineers of Cartage (ENICarthage), Electrical Engineering Department, University of Carthage, Carthage, Tunisia

$FCGS_2$ features in the helitron's recognition system. The performance of our system achieves a high recognition rate (Globally accuracy rate) reaching the value of 92.27%.

Keywords Helitrons · Repetitive DNA · Microsatellites · C.Elegans · $FCGS_2$coding · SVM · Features · Continuous wavelet transform · Kernel tricks

1 Introduction

Helitrons form an important part of Transposable Elements (TEs) DNA class II in eukaryotic genomes [16, 45]. This specific DNA type transposes by a rolling circle replication mechanism via a single-stranded DNA intermediate. During transposition, helitrons are the only TEs DNA that does not create target site duplications [16]. Helitrons frequently capture miscellaneous host genes, some of which can evolve into novel host genes. In the evolution of host genomes, the helitrons seem to play a major role because some of the host genes become essential for the helitrons transposition [34, 45]. They are also involved in a number of biological processes such as heterochromatin development, gene expression regulation [54] and genome rearrangements [51]. In addition, helitrons have the ability to synthesize new genes by nearby exons capture, also by transcription readthrough and unrelated exons placement into common transcripts [2]. Helitrons were firstly discovered in the nematode *Caenorhabditis elegans* [15] then in plants (Arabidopsis thaliana and *Oryza sativa*) [33] and in fruit fly (Drosophila) [4]. At present, they are identified in a diverse range of species like: fungus [11], lucifugus [35], animals [55], vertebrates, specifically in the fish Daniorerio and Sphoeroidesnephelus genomes [34]. In this paper, we focus on the helitrons characterization and classification from a different point of view. Indeed, we harness the power of the signal processing tools to identify these interesting elements in a visual way and we use the support vector machine (SVM) as a classification technique. The SVM solves recognition problems of the two classes and multi-class. It has been widely used in numerous domains due to its inherent discriminating learning and generalization capabilities; it is often applied to solve statistical learning problems [50]. The SVM classification was applied with success in bioinformatics studies [22], DNA [32], molecular genetics [7] and for the identification and characterization of microRNAs and target prediction [12, 28]. In this framework, we propose using this classification strategy to help non-specialists to easily annotate Helitrons. This work was driven by the fact that helitron recognition using the signal processing methods has not been yet addressed.

The paper includes an introduction, a state of the art of the related works, two sections describing the work and a conclusion. In the section proposed method, we provide a description of the DNA coding technique and the analysis methods used to extract features as well as the classification technique. Finally, we give the conclusion of this work. In the section experimental results, we describe how we partitioned the parameters extracted for the C.elegans genome into helitrons and non-helitrons databases. Then, we expose the classification results endorsed by a comparison between results we obtained for two groups of parameters.

2 Related works

Since their discovery, helitrons have attracted widespread attention. To identify and analyze these elements many computational tools were developed: HelitronFinder [5], HelSearch [54],

a combination of BLAST search and the hidden Markov models [41] and HelitronScanner [53]. These tools are typically based on the search for canonical helitrons which begin with a 5′ T (C/T) and terminate with a CTRR 3′as well as the existence of a hairpin structure (~ 11 base pairs). In this sense, a comparison of the searched area with the reference helitron is performed by different alignment algorithms [6, 43].

These methods are generally hindered by the lack of information about the reference sequences as well as the need of an enormous memory space [37, 42]. Besides the asymmetric structure of helitrons, their abundance and diversity in genomes, present an enormous identification and annotation problem. Taking into account all these factors, we can comprehend why the structure and the dynamic of the helitron sequences have not been yet well studied. In fact, the available bioinformatics tools aim to identify the presence of helitron on the basis of previous knowledge of its biological features [14] and do not provide a visual tool to detect the presence of this element in a given sequence. Further, despite the availability of several methods of helitron identification, a systematic classification method based on the information (features) revealed by the sequence itself has not been yet taken forward.

As a solution, we propose in this work the combination of the signal processing tools with the SVM-approach (a supervised learning algorithm) to identify helitronic sequences. The main steps of the systematic recognition of helitrons are based on the choice of the classifier, the features extraction methods and the choice of the non-helitrons databases.

3 FCGS Coding technique and Wavelet transform

To present a DNA sequence (a chainlike molecule composed of four bases: A, T, C, and G) into a numerical form we need to transform these bases into a series of numerical values. For this aim, we provide two ways to represent the DNA sequences into:

- explicit signals when we applied the FCGS coding technique
- explicit images when we applied the CWT to these signals.

3.1 $FCGS_2$ coding technique

Thanks to the development of DNA coding techniques, different methods have emerged like the binary coding [30], the structural bending trinucleotide coding (PNUC) [31], the Frequency Chaos Game Signal (FCGS) [25, 26, 46], etc. The latter technique is a new one dimensional coding. It is based on assigning the frequency value of each sub-pattern in the chromosome to the same group of nucleotides existing in the sequence. In this work, we consider the Frequency Chaos Game Signal of dinucleotides: $FCGS_2$. It is based on the apparition's probability of all two successive nucleotides for an entry DNA sequence [25, 26, 46]. The probability of a given dinucleotide ($P_{2nucleotide}$) is calculated following this equation:

$$P_{2nucleotide} = \frac{N_{2nucleotide}}{LN_{Chr}} \tag{1}$$

$N_{2nucleotide}$ represents the apparition number in the whole sequence of a given dinucleotide. LN_{Chr} represents the length of the chromosome in base pairs.

Our goal is to establish a signals database for both helitrons and the repetitive sequences which are considered as non-helitrons. For this, a dinucleotide (i), at a position (k) is replaced by its occurrence's probability:

$$S_{2nucleotide}(k) = \sum_i P_{2nucleotide}(i,k) \tag{2}$$

The $FCGS_2$ consists in calculating the sum of all dinucleotide indicators ($S_{2nucleotide}$):

$$\mathrm{FCGS}_2(k) = \sum_k S_{2nucleotide} \tag{3}$$

Therefore, the $FCGS_2$ technique represents the temporal evolution of the dinucleotides frequency along the chromosome which means that it reflects the statistical features of the chromosome itself. To enhance such characteristics, we propose to apply a time- frequency method, which is the wavelet transform.

3.2 Wavelet transform

The Wavelet Transform is widely used in the time-frequency analysis of biomedical and biological signals [1, 13, 23, 29, 47]. In this work, we use the wavelet energy features that we extract from the wavelet coefficients' matrix to classify helitrons. The coefficients matrix is obtained by applying the continuous wavelet transform to the $FCGS_2$ signal with the Complex Morlet Wavelet is taken as analysis window. After that, we prepare the features database (energy dataset) of all genomic sequences to be passed to the SVM-classifier for the helitrons classification purpose.

The CWT decomposes a given signal into a sum of windows called wavelets. The latters are obtained by shifting and expanding a mother wavelet ψ(t) [9, 24, 27]. The set of wavelet windows is obtained by:

$$\psi_{s,u}(t) = \frac{1}{\sqrt{s}}\psi^*\left(\frac{t-u}{s}\right), \mathrm{s} > 0, \mathrm{u} \in \mathbb{R} \tag{4}$$

Here * is the complex conjugate.

The Complex Morlet function is expressed by:

$$\psi_{cmor}(t) = \pi^{-\frac{1}{4}}\left(e^{i\omega_0 t} - e^{i\omega_0 2}\right)e^{-\frac{t^2}{2}} \tag{5}$$

Here ω_0 is the oscillation's number.

The continuous wavelet transform is performed by applying this formula:

$$W_{(s,u)}[\mathrm{FCGS}_2(t)] = \frac{1}{\sqrt{s}}\int_{-\infty}^{+\infty}\mathrm{FCGS}_2(t)\psi^*\left(\frac{t-u}{s}\right)dt \tag{6}$$

In the following, we consider the complex Morlet wavelet transform (CWT) along 64 scales with the parameter ω_0 fixed at 5.4285.

The final result is a matrix of coefficients which we use to generate the scalogram representation by calculating the absolute value of these coefficients. Here, we use the scalogram presentation as a new way to visualize the DNA sequences. The time-frequency plan allows us to distinguish a DNA class by a specific signature (specific motifs with different

periodicities). Thanks to this property, we are able to visually recognize a given helitron class. In the following (Fig. 1), we provide the representation that characterizes each helitron class. Since we will classify helitrons and non helitrons (repetitive DNA), we also give an example of the time-frequency signature of a repetitive DNA sequence containing the microsatellite (CAACG)n. The horizontal axis indicates the DNA position in base pairs while the vertical axis indicates the frequency.

Note that these scalograms are the result of the wavelet analysis applied to the $FCGS_2$ of some concatenated helitrons belonging to chromosome I of C.elegans. The scalogram of the microsatellite type in Fig. 1 corresponds however to two concatenated sequences of repetitive (CAACG)n DNA in chromosome I. These scalograms examples present the overall behavior of the considered DNA types. For example, helitrons of type Helitron2_CE, HelitronY2_CE, HelitronY3_CE and Ndnax2_CE have specific signature presented by small repetitive motifs spread over a large frequency band. On the other hand, helitrons of type Helitron1_CE, HelitronY1_CE, HelitronY4_CE and Helitron1A_CE present similar signatures characterized by a pronounced energy around the frequency 0.15 (which is equivalent to periodicity 6). Other similar repetitive motifs are noticed for Helitron1_CE and HelitronY1A_CE around the frequency 0.1 (which corresponds to the periodicity 10). These helitrons present various motifs around different periodicities compared to other helitron families. As for helitrons NDNAX1 and NDNAX3, they have specific repetitive patterns compared to the other helitron classes which renders them very distinctive.

Finally, the microsatellite adopt a different time-frequency signature which facilitates the distinction between this DNA class and helitrons.

From these figures, we can see that the repetitive patterns and the energy concentration around the frequency bands allow to visually differentiate between the helitron classes and the non helitron ones. In the following, we will exploit these results to automatically classify helitrons based on SVM.

4 SVM classification

Our aim is to recognize helitrons from the non-helitron sequences. For this goal, we use the supervised learning model: the Support Vector Machines: SVMs. The SVM classifiers are based on the VapnicChervonenkis (VC) theory and the principle of the structural risk minimization (SRM) [38]. The SVM model was developed by Vapinik in 1998 [50]. The main principle that encompasses this technique, is the structural risk minimization (SRM) [10, 38, 49]. Using the function of Kernel, the original input is set and remodeled to a high-dimensional feature space to achieve optimal classification performance in the new feature space [50]. Maximizing the error margin could give effective discriminate SVMs classifiers. They have also the advantage of being able to deal with samples of a very higher dimensionality. They have been successfully used in different pattern recognition applications like face, EEG signals, speaker and DNA [8, 17, 18, 40, 48, 52]. The SVMs are particularly attractive to the biological sequences analysis due to their ability to handle large dataset and large input spaces [20, 39].

The SVM have better generalization abilities which are achieved through the minimization of the upper bound of the generalization error. It aims to maximize the margin; distance from a separating hyperplan to the closest negative (−1) or positive (+1) sample between classes. One or several hyperplans are constructed in

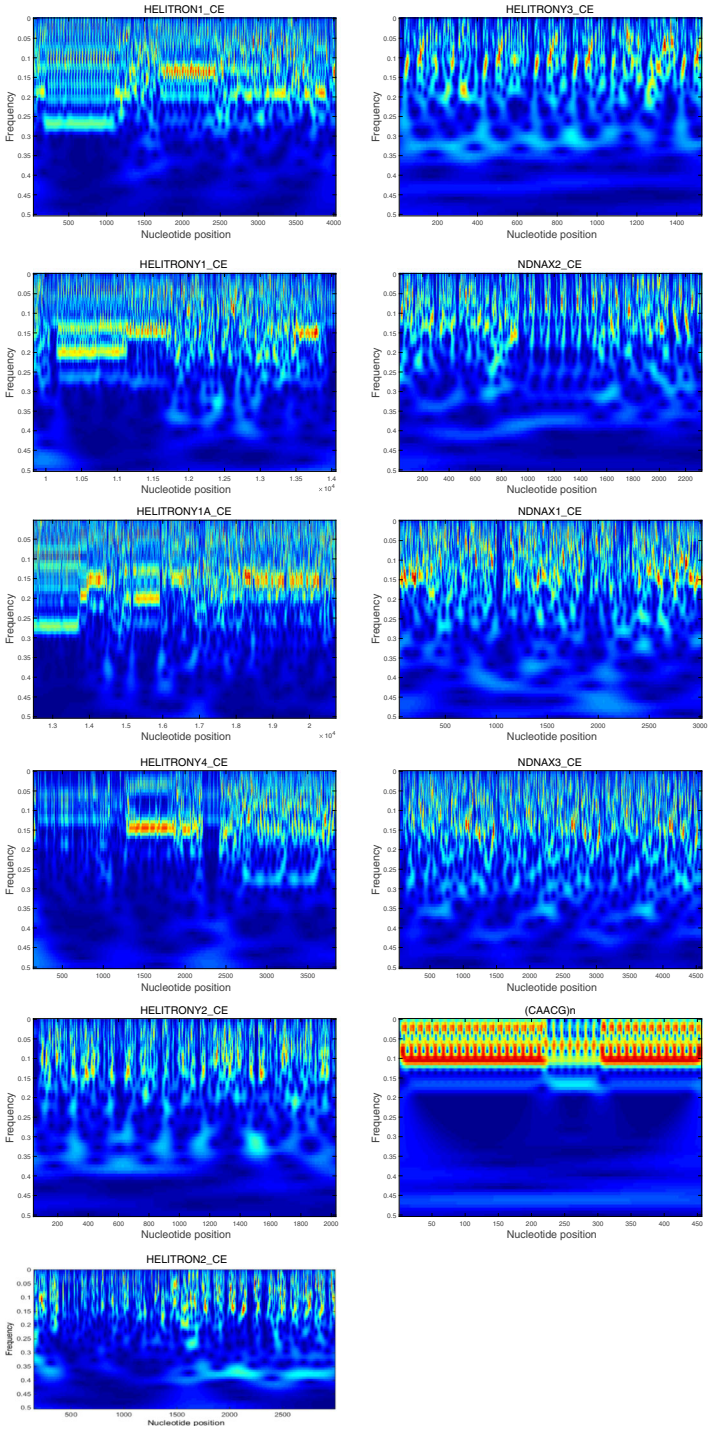


Fig. 1 Examples of Helitron and repetitive DNA signatures

order to separate different classes. Elsewhere, an optimal hyperplan must be found. This optimal hyperplan [3], a linear decision function, has to be with the maximal margin. However, this margin should be between the vectors of the two classes. The hyperplan can be described as:

$$f(x) = w^t x + b \tag{7}$$

Were w is the weight vector, x is the input vector, and b is the bias.

4.1 Kernel functions

The major tricks of SVM are the kernel functions. For the case in which no linear separation is possible, these functions are used. In the case when data are not linearly separable the kernel tricks extend to the class of decision functions. In addition, the kernel function can be explained as a measure of similarity between the input samples x_i and x_j [18], which allows SVM classifiers to meet the separation rule even with highly divergent and complex boundaries.

In what follows, we focus on finding the best kernel and the kernel function parameters to classify helitrons. Several choices for the kernel function are available, including: linear, polynomial, sigmoid, RBF. In the next paragraphs, we present a quick overview of the most frequently used kernel functions.

1) Polynomial Kernel

The Polynomial kernel, a non-stationary kernel, is well adapted for problems where all the training samples are normalized. Its parameters should be carefully tuned; which are the slope σ, the polynomial degree d and the constant term r.

$$K(x_i, x_j) = \left(\sigma x_i^T x_j + r\right)^d, \quad \sigma > 0 \tag{8}$$

Here, we consider $d = 3$ and $r = 0$.

2) RBF kernel

The Gaussian functions (Radial basis functions: RBF) are a family of kernels that measures the distance between feature vectors smoothed by an exponential function [36]. The carefully chosen parameters (c, σ) can play a major role in the performance of the kernel.

Below, we present the equation of RBF (radial basis function) kernel [21].

$$K(x_i, x_j) = \left(\sigma \left\| x_i - x_j \right\|\right)^2, \quad \sigma > 0 \tag{9}$$

The accuracy of the classifier is highly sensitive to the choice of the parameter σ. The latter must be tuned to control the amount of smoothing. In fact, the behaviors of SVM change when σ becomes too small or too large.

In this work, we use the following couples (c, σ):

$$\sigma = c = \left[2^{-6}, 2^{-5}, \ldots 2^{9}, 2^{10}\right] \tag{10}$$

3) Sigmoid kernel

Typically, the kernel must satisfy Mercer's theorem (the kernel is a positive definite function). Despite its widespread use, sigmoid kernel is not positive semi-definite for certain values of parameters.

$$K(x_i, x_j) = tanh(\sigma x_i^T x_j + r) \tag{11}$$

Here, gamma is the scale parameter of the input samples and r is the shift parameter that controls the threshold of mapping ($r = 0$). Hence, the parameters σ and r have to be properly chosen. If this choice is not properly done, it yields to wrong results.

5 Material and Method

This section is devoted to the helitron recognition in C.elegans genome based on the $FCGS_2$ coding technique and the SVM classification.

5.1 Material

For this work, the C.elegans sequences were retrieved from the NCBI public database [44]. Two genomic datasets were then used for the current study: one is the "helitrons" dataset and the other represents the "non-helitrons" dataset. In this context, the non-helitronic sequences we have chosen are also consisting of small repetitive motifs (one or more nucleotides) that frequently appear in the genome. The basic repetitive patterns we used here are of a microsatellite's type with a ranging length from 2 to 5 base pairs (pb) [7, 28]. Our choice goes to the following patterns: (A)n, (AATAG)n, (ATG)n, (ATGGTG)n, (ATTG)n, (CA)n, (CAA)n, (CAGG)n, (CAGA)n,(CAACG)n, (CAAT)n and (CAATA)n. This choice is justified by the fact that helitrons contain themselves microsatellites sequences; which misleads the helitron recognition rates in most of bioinformatics tools.

The helitron classes, we have investigated here, are of the number of ten: Helitron1_CE (H1), Helitron2_CE (H2), HelitronY1_CE (Y1), HelitronY1A_CE (Y1), HelitronY2_CE (Y2), HelitronY3_CE (Y3), and HelitronY4_CE (Y4), NDNAX1_CE (N1), NDNAX2_CE (N2) and NDNAX3_CE (N3). These families possess complex and variable structures. More of this, the size of the helitronic sequences varies according to the family. Globally, the apparition number of helitrons in the C.elegans genome varies from 77 to 1093 according to their family (Table 1 in the next sub-section). The variability in terms of length, composition and structure makes difficult the identification of these elements.

5.2 Method

In the following Fig. 2, we provide the flowchart describing this work.

The adopted methodology is composed by five steps.

– The first step consists in extracting all chromosomes data for the C.elegans model [44] and generating the corresponding $FCGS_2$ sequences. In this way, the DNA database will be converted to a 1-D signals database.

Table 1 The Helitron occurence number in training and testing databases in C.elegans

Helitron type	Helitron number		
	Total	training	testing
H1	197	132	65
H2	469	313	156
Y1	483	322	161
Y1A	**1093**	729	364
Y2	337	225	112
Y3	117	24	93
Y4	532	415	117
N1	77	52	25
N2	188	126	62
N3	134	94	40

Bold entries provide an idea about the best result that we obtained.

- The second step consists in extracting the helitron and the non-helitron sequences from the NCBI database. Here the non helitronic DNA consists in repetitive sequences that contain microsatellites and do not form helitrons.
- In the third step, helitron and non-helitron signals database regarding dimmers (sequences of 2 bp length) is prepared. Then, a signal database that corresponds to all helitron types is established as well as another database that do not contain helitrons but repetitive DNA of microsatellite type [44].
- In the fourth step, we prepared the database which contains the corresponding temporal, spectral and time frequency features of each of these helitronic and non-helitronic sequences. In this way two types of features were extracted: the combination of the spectral and the temporal features and the energy calculated based on the wavelet transform. The ultimate database was splitted into two sub-databases: 70% for training and 30% for testing. For SVM-classification system, the Table 1 specifies the helitrons

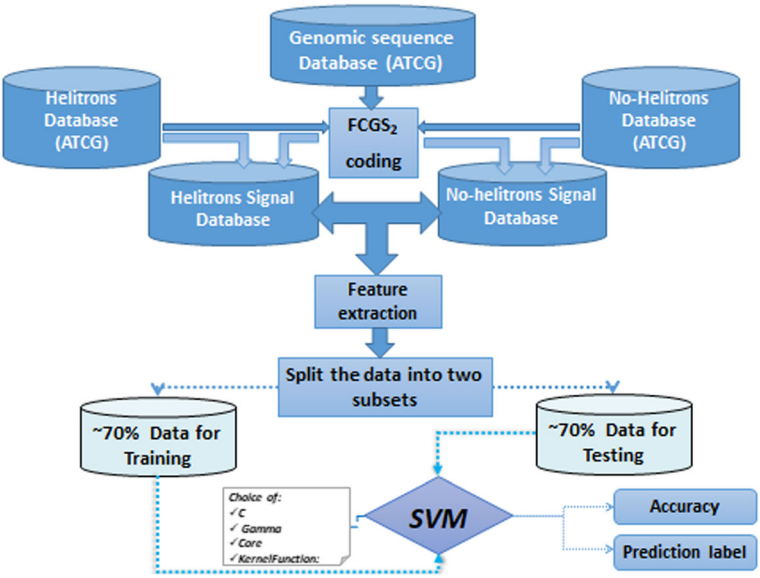


Fig. 2 SVM -Helitron recognizer flowchart

number which we considered for training and test and thus for each helitron class. In the other hand, we took the same number for the sequences which contain microsatellites.

From Table 1, we can see that the helitron of type helitronY1A_CE has the highest occurrence number in the genome.

– Finally, the SVM technique was used for classification. The classification step involves separating data into testing and training sets. In order to have accurate results (when the system tends to give better recognition rates), all kernel functions were tested.

5.2.1 Features extraction

For the features extraction task, several methods have been reported in the signal processing field including: the time domain, the frequency domain (like the Fourier transform) and the Time-frequency domain (such as the short-time Fourier and the wavelet transforms). Here, we propose using these methods to extract features from $FCGS_2$ signals to be considered later in the classification step. In the features extraction step, various independent variable values are prepared as input of the classifier to predict the corresponding class to which belongs the independent variable.

a) Temporal features

For the temporal features, we use statistical measures including: maximum number of peaks ($Pics_{occurence}$), average (μ), standard deviation (Std), Mahalanobis distance (X), variance (Var), median(median), energy (E) and Root Mean Square (RMS).

$$Pics_{occurence} = \frac{number(\max FCGS2_{Chri})}{LN_{Chri}} \tag{12}$$

Where i is the number of the chromosome (Chr) and LN is the length of the chromosome i. For the chromosomes I, IV and V the $Maxpic$ feature presents the number of apparition of the dinucleotide 'TT' and for the other chromosomes it presents the number of the dinucleotide 'AA'.

b) Spectral features

As spectral features, we use the following parameters: mean power spectral density (Smean), Power spectral density (PSD) and the Power root mean square (Prms).

Here, the Fourier transform is used to convert the time-based signal into the frequency domain. The features we extract are:

- Mean power spectral density (Smean): it is the average Power spectral density. It measures the energy of a signal when distributed in the frequency domain. Its mathematical expression is given by:

$$Smean = mean\left(\frac{2*\left\|\frac{X(f)}{N}\right\|}{N}\right) \tag{13}$$

Here N is the signal length and $X(f)$ is the Fourier transform of the signal x.

- Power spectral density (PSD): the computation of PSD is done by applying the Fast Fourier Transform on the autocorrelation function ($r_{xx}(\tau)$).
- Power root mean square (Prms): the Prms measures the power of the signal's magnitude. It is calculated from the following equation:

$$Prms_{freq} = \sqrt{\sum \mathrm{FCGS2}(f)^2} = \sqrt{\sum_{i=1}^{N-1} FFT\left(\mathrm{FCGS2}_i^2\right)} \tag{14}$$

We further use the absolute value of Prms_freq whose expression is as follows:

$$Prms\_abs = |Prms\_freq| \tag{15}$$

c) Time-frequency features

For the time-frequency features, the genomic sequences are firstly transformed into signals using the $FCGS_2$ technique. The signals are secondly transformed into the time–frequency domain based on the continuous wavelet transform (CWT). Finally, a vector containing the wavelet coefficients and their relative energies are used as features for helitrons and non-helitron classification (Fig. 3).

The procedure consists of calculating the energy at each scale of the wavelet decomposition using the following formula:

$$E_{wavelet} = \sum_{s=1}^{L} \left| W_{(s,u)}[FCGS_2(t)] \right|^2 \tag{16}$$

Here, L is the length of the $FCGS_2$ signal. Since we have considered 64 scales for the CWT computation (which means that the wavelet coefficients matrix contains 64 power spectra), we have calculated 64 relative energy values.

The sub-figures (a) and (c) in Fig. 4 provide the scalogram representation (absolute value of the wavelet coefficients) of two helitron examples; While the correspondent energy (energy concentration around frequencies) are illustrated in the sub-figures (b) and (d). For the scalograms, the horizontal axis indicates the helitron's position in base pairs. As for the energy representation, the horizontal axis gives the energy amplitude.

The first example is an HelitronY4_CE which is positioned at [354,726 bp–355,316 bp] in the C.elegans chromosome I. The second example is an NDNAX2_CE which is positioned at [5,640,357 bp- 5,640,991 bp] in the same chromosome.

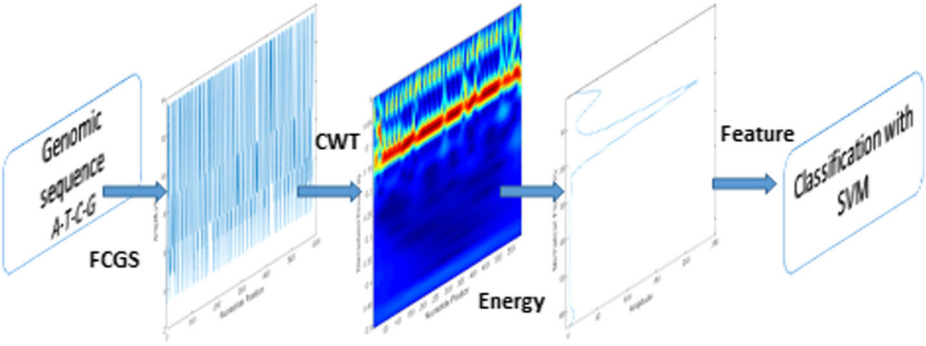


Fig. 3 SVM -Helitron recognizer flowchart based on features extracted from CWT

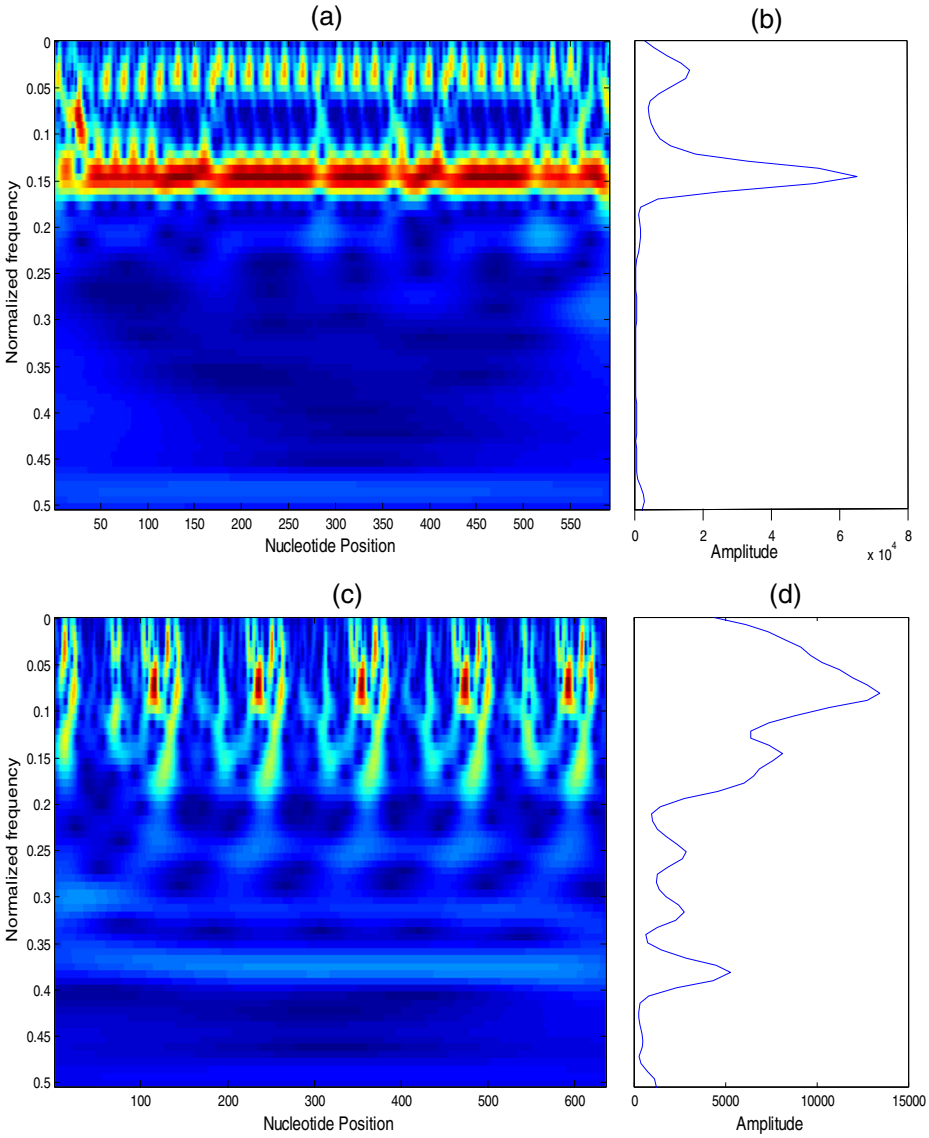


Fig. 4 The scalogram and the relative energy representations of two helitrons

As it can be seen, the scalogram serves to visualize the signature of helitrons [13]. Regarding the energy plot, the pronounced peaks indicate the presence of repetitive motifs in the correspondent scalogram. For example, in the case of HelitronY4_CE, the highest energy is concentrated around the frequencies 0.05 and 0.15. This translates into a special repetitive motif within the frequencies band [0–0.05] in the scalogram representation. We also note that the most pronounced peak in the spectrum corresponds to the frequency 0.15, which is equivalent to the periodicity 6. In addition, the frequency 0.027 indicates the existence of the periodicity 35 in the DNA sequence.

As for the helitron NDNAX2, particular repetitive motifs with a special energy pattern mark their presence in the frequency band [0–0.2]. The energy plot proves that the highest energy is

included into this sub-band. The large energy band has high amplitudes that correspond to the main periodicities: 6, 10 and 20. Finally, Fig. 5 presents the scalogram representation (sub-figure (e)) and the energy-spectrum (sub-figure (f)) of a repetitive DNA sequence of type (CAGA)n. The sequence has the size of 376 bases pair and the position [13,273,486 bp–13,273,861 bp] at the C.elegans chromosome I. This sequence consists of the repetition of the microsatellite 'CAGA', which size is obviously of 4 bp. The repetition of this subsequence in the DNA sequence creates a periodicity 4. This periodicity is well translated in the scalogram representation in the form of high energy frequency band around the frequency 0.25. Likely, we can also see the existence of the periodicity 5 in the scalogram. These two periodicities correspond to the peaks with the highest amplitudes in the sub-figure (f).

Based on this, we can see that the time-frequency is an effective way to represent DNA in such way all periodicities that characterize the considered sequence can be detected.

Next, we will exploit this manner to extract information about DNA to be taken into account by the SVM classification.

I. Experimental results

In this work, we mixed helitron sequences with the genomic sequences that do not contain helitrons and whose number of training and testing are equal to the helitrons ones. One third of data is taken then for testing and two-thirds for training.

We used, further, all the temporal, the spectral and the time-frequency features for all the sequences. As for the kernel tricks, we used the Linear, Polynomial, RBF and Sigmoid functions. For the SVM-kernels parameters, we found that using the Cross-Validation function gives the optimal value of two parameters: the kernel width (σ) and the regularization parameter (c) [19]. It is known that the recognition task is roughly divided into two stages: the feature extraction and the recognition. The performance of the recognition system strongly depends on the choice of the feature extraction method. Thus, we tried different methods to extract features from the genomic data. In fact, we based the helitron prediction on two features databases: the first database contains the temporal and the spectral features extracted from the $FCGS_2$ signals; while the second database contains the relative energy obtained by the

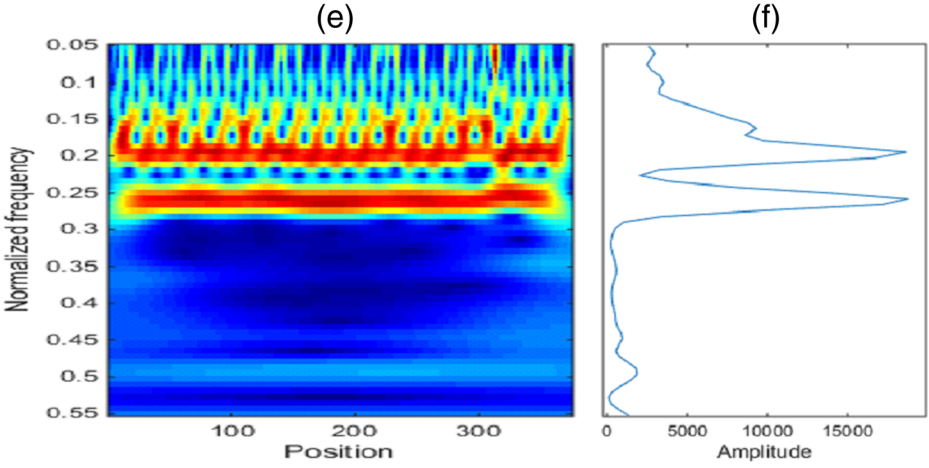


Fig. 5 The scalogram and the relative energy representations of repetitive DNA sequence of type (CAGA)n

continuous wavelet transform of the $FCGS_2$ signals. Consequently, a comparison between the accuracy rates of the helitron prediction based on of the two methods was conducted. The first database consists of the combination of 8 temporal features and 3 spectral features. The parameters we considered are as follows:

- The temporal features are: max, μ, Median, Std, X, Var, E and RMS
- The spectral features are: Smean, PSD and Prms

As for the kernel tricks, we used Linear, Polynomial, RBF and Sigmoid functions which parameters are varied according to the cross-validation function. For each kernel function we defined and calculated the best parameters (d, c and σ) which give the best accuracy rates for all helitron types' recognition. In the following table we provide the best recognition accuracy rate for all helitron classes in the C.elegans. The results of the classification are based on the combination of the temporal (8 parameters) and the spectral features (3 parameters).

The prediction performances, illustrated in Table 2, show very good results. In fact, the accuracy values are between 75.88% and 95.65, depending on the helitron family. The best accurate prediction (which is 95.65) is obtained for the helitron NDNAX1 class. In addition, six helitron types were highly predicted (with an accuracy rate greater than 91%) which are: HelitronY1_CE, helitronY1A_CE, HelitronY3_CE, NDNAX1_CE, NDNAX2_CE and NDNAX3_CE. On the other hand, the SVM-Linear performs the better results for HelitronY1_CE (ACC = 92.43%), Helitron1_CE (ACC = 83.5%), Helitron2_CE (ACC = 75.88%) and HelitronY2_CE (ACC = 83.66%). When we used the cross validation, we found that with the Polynomial kernel $d = 2$ we get the best values of kernel parameter. As for the RBF kernel, different values of c and σ have given best accuracy of some helitrons classes.

Table 2 The best SVM results of helitrons identification with different SVM-kernels and using the group of temporal and spectral feautures extracted from $FGCS_2$

Helitron type	Helitron Rate (%)	Non-Helitron Rate (%)	KERNEL	SVM parameters	SVM ACC %
H1	81.53	77.96	RBF	c = 65,536 and σ = 0.000000015625	79.66
	83.05	77.98	Linear		**80.5**
H2	77.30	66.66	RBF	c = 65,536 and σ = 0.000000015625	66.31
	76.59	75.17	polynomial	d = 2	**75.88**
Y1	89.07	84.03	RBF	c = 60 and σ = 0.0000015625	86.55
	94.95	89.91	polynomial	d = 2	**92.43**
Y1A	95.16	94.08	RBF	c = 60 and σ = 0.01	**94.62**
	95.16	88.70	polynomial	d = 2	91,93
Y2	90.09	74.25	RBF	c = 60 and g = 0.01	82.17
	88.11	79.20	Linear		**83.66**
Y3	94.11	91.17	RBF	c = 600 and σ = 0.05625	**92.64**
	94.11	82.35	polynomial	d = 2	88.23
Y4	80.64	84.51	RBF	c = 600 and σ = 0.015625	**82.58**
	83.22	78.70	polynomial	d = 2	80.96
N1	95.65	95.65	RBF	c = 600 and σ = 0.015625	**95.65**
	95,65	91.30	Linear		93.47
N2	94.73	87.7	RBF	c = 600 and σ = 0.015625	**91.22**
	94.73	85.96	polynomial	d = 2	90.35
N3	95	92.5	RBF	c = 600 and σ = 0.015625	**93.75**
	97.5	77.5	polynomial	d = 2	87.5

Bold entries provide an idea about the best result that we obtained.

Given the variability of the helitron's sequences in composition and size, the wavelet coefficients matrix leads to a set of features which are not balanced in length. However, the SVM method is limited when it is applied at imbalanced datasets. For this reason, we need to apply a reduction method to obtain balanced datasets for the wavelet analysis while conserving the useful information. Therefore, we have chosen the energy measure to balance these features. Based on the energy vector calculated from the wavelet coefficients matrix, the features database was established in a second step for both helitrons and non-helitronic sequences. The experimental results are illustrated in Table 3 which represents the best SVM results.

This table shows that the high prediction accuracy (average of all accuracy which is equal to 92.27%) of our method is due to the ability of the time-frequency features to capture helitron/non-helitron attributes.

We notice that the best rate for the NDNAX3 class, which is 96.25%, was obtained using the RBF-Kernel with $c = 65{,}536$ and $\sigma = 0.000000015625$. Also, the HelitronY3_CE class was recognized with an accuracy rate reaching 92.64% with these parameters.

Overall, the kernel width $\sigma = 0.000000015625$, the penalty $c = 60$ and the SVM-RBF have given best performance in terms of recognition rates for the Helitron1_CE class with a global accuracy: ACC = 95.76%. Six other notable helitron classes were showing high accuracy rates: HelitronY1_CE, HelitronY1A_CE, NDNAX1_CE, NDNAX2_CE, HelitronY4_CE, and Helitron2_CE with the respective values of 95.43%, 95.38%, 93.48%, 92.98%, 91.24 and 85.46%. As it can be noted, with the polynomial kernel width parameter $d = 2$, we obtained the best global accuracy rate reaching the value of 88.11% for the HelitronY2_CE class.

Based on these results, we can see that introducing a set of time-frequency features reveal interesting results that can categorize the helitrons sequences.

Table 3 The best SVM results of helitron identification with different SVM-kernels using the feautures extracted from the continuous Wavelet Transform

Helitron type	Helitron Rate (%)	Non-Helitron Rate (%)	KERNEL	SVM parameters	SVM ACC %
H1	98.30	93.22	RBF	σ = 0.000000015625 and c = 60	**95.76**
	91.5	96.6	polynomial	d = 2	94.07
H2	84.39	86.52	RBF	σ = 0.000000015625 and c = 60	**85.46**
	82.27	81.56	polynomial	d = 2	81.91
Y1	97.48	93.27	RBF	σ = 0.000000015625 and c = 60	**95.38**
	87.40	89.91	polynomial	d = 2	88.65
Y1A	95.69	95.16	RBF	σ = 0.000000015625 and c = 60	**95.43**
	87.09	86.02	polynomial	d = 2	86.56
Y2	91.08	84.15	RBF	σ = 0.000000015625 and c = 65,536	82.67
	88.11	88.11	polynomial	d = 2	**88.11**
Y3	94.11	91.17	RBF	σ = 0.000000015625 and c = 65,536	**92.64**
	91.18	85.29	polynomial	d = 2	88.23
Y4	93.43	89.05	RBF	σ = 0.000000015625 and c = 60	**91.24**
	87.59	90.51	polynomial	d = 1 AND c = 0.01	89.05
N1	91.30	95.65	RBF	σ = 0.000000015625 and c = 60	**93.48**
	95.65	78.26	Linear		86.95
N2	98.24	87.71	RBF	σ = 0.000000015625 and c = 60	**92.98**
	100	82.45	polynomial	d = 2	91.23
N3	100	92.5	RBF	σ = 0.000000015625 and c = 65,536	**96.25**
	97.5	87.5	polynomial	d = 2	92.5

Bold entries provide an idea about the best result that we obtained.

Now, if we compare these results (Table 3) with those obtained when we used the first group of parameters (Table 2), we can clearly see that the SVM classifier better recognizes helitrons when it is based on the features extracted from the wavelet transform. Indeed, the best global accuracy rate using the temporal and the spectral features extracted from the $FCGS_2$ signals reached 88.29%. However, the best global accuracy rate we obtained using the energy wavelet features attained 92.27%.

From the present work, we can conclude that the choice of the features we have to extract from the helitronic signals play a major role in their recognition. It turns that using a time-frequency analysis gives better results than using temporal and spectral analysis in terms of the helitron classification. In addition, the SVM-classifier parameters have shown a great influence on the classification results.

6 Conclusion

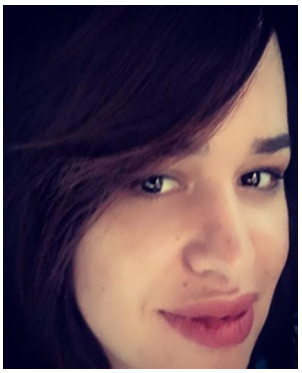
In this work, we developed a highly accurate method for predicting the Helitron sequences. A support vector machine classification approach based on all SVM-kernels has been adopted for the helitron recognition in C.elegans. The obtained results revealed very encouraging classification accuracies. The detailed experimental results presented here have shown the great effect of the feature extraction step on the helitrons classification rates. In fact, for the feature extraction, we have proposed two methods. The first method consists in extracting temporal and spectral features from the $FCGS_2$ of helitrons and non-helitrons sequences (a repetitive DNA which contains microsatellites). However, the second method relies on the continuous wavelet transform of the $FCGS_2$ signals of helitrons and non-helitron sequences. To make a balanced features database, we extracted the relative energy from the wavelet coefficients matrix. The classification results have shown the superiority of the time-frequency analysis compared to the temporal and spectral analysis in terms of the helitron classification. Furthermore, we demonstrated that choosing the optimal parameters for the SVM-kernels (d, c, and σ) would greatly help improve the accuracy rates of the helitron prediction.

References

1. Amin HU, Malik AS, Ahmad RF (2015) Feature extraction and classification for EEG signals using wavelet transform and machine learning techniques. Australas Phys Eng Sci Med 38:139–149. https://doi.org/10.1007/s1324
2. Barbaglia AM, Klusman KM, Higgins J, Shaw JR, Hannah LC, Lal SK (2012) Gene capture by Helitron transposons reshuffles the transcriptome of maize. Genetics 190:965–975. https://doi.org/10.1534/genetics.111.136176
3. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 2:273–297
4. Dias GB, Heringer P, Kuhn GC (2016) Helitrons in Drosophila: chromatin modulation and tandem insertions. Mob Genet Elements 62:e1154638
5. Du C, Caronna J, He L, Dooner HK (2008) Computational prediction and molecular confirmation of Helitron transposons in the maize genome. BMC Genomics 9:51. https://doi.org/10.1186/1471-2164-9-51
6. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340

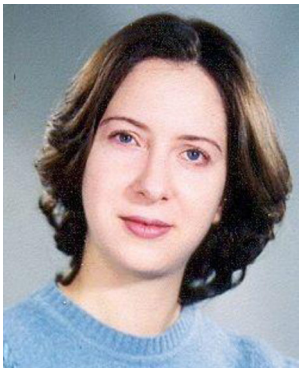
7. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16: 906–914. https://doi.org/10.1093/bioinformatics/16.10.906
8. Ghimire D, Jeong S, Lee J, Park SH (2017) Facial expression recognition based on local region specific features and support vector machines. MTAP 76:7803–7821. https://doi.org/10.1007/s11042-016-3418-y
9. Grossmann A, Morlet J (1984) Decomposition of hardy functions into square integrable wavelets of constant shape. SIAM J Math Anal 15:723–736. https://doi.org/10.1137/0515056
10. Gutschoven B, Verlinde P (2000) Multi-modal identity verification using support vector machines (SVM). In: Information Fusion. FUSION 2000. Proceedings of the Third International Conference on IEEE, Vol. 2, pp. THB3–3, July. 2000
11. Hood ME (2005) Repetitive DNA in the automictic fungus Microbotryumviolaceum. Genetica 124:1–10. https://doi.org/10.1007/s10709-004-6615-y
12. Huang Y, Yang YB, Gao XC et al (2017) Genome-wide identification and characterization of microRNAs and target prediction by computational approaches in common carp. Gene Reports 8:30–36
13. Jahankhani P, Kodogiannis V, Revett K (2006) EEG signal classification using wavelet feature extraction and neural networks. In: Modern Computing IEEE John Vincent Atanasoff 2006 International Symposium 120–124
14. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenetic Genome Res 110:462–467. https://doi.org/10.1159/000084979
15. Kapitonov VV, Jurka J (2001) Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci 98:8714–8719. https://doi.org/10.1073/pnas.151269298
16. Kapitonov VV, Jurka J (2007) Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet 23: 521–529. https://doi.org/10.1016/j.tig.2007.08.004
17. Kaur B, Singh D, Roy PP (2017) A novel framework of eeg-based user identification by analyzing music-listening behavior. MTAP 76(24):25581–25602. https://doi.org/10.1007/s11042-016-4232-2
18. Kumar M, Gromiha MM, Raghava GP (2011) SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. J Mol Recognit 24:303–313. https://doi.org/10.1002/jmr.1061
19. Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley
20. Li L, Luo Q, Xiao W et al (2017) A machine-learning approach for predicting palmitoylation sites from integrated sequence-based features. J Bioinforma Comput Biol 15:01: 1650025. https://doi.org/10.1142/S0219720016500256
21. Lin HT, Lin CJ (2003) A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. Neural Comput 3:1–32
22. Mateos A, Dopazo J, Jansen R, Tu Y, Gerstein M, Stolovitzky G (2002) Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. Genome Res 12:1703–1715 http://www.genome.org/cgi/doi/10.1101/gr.192502
23. Mena-Chalco J, Carrer H, Zana Y, Cesar RM (2008) Identification of protein coding regions using the modified Gabor-wavelet transform. IEEE/ACM TCBB 5:198–207
24. Merry RJE, Steinbuch M (2005) Wavelet theory and applications. Literature Study, Eindhoven University of Technology, Department of Mechanical Engineering, Control Systems Technology Group
25. Messaoudi I, Oueslati AE, Lachiri Z (2014) Building specific signals from frequency chaos game and revealing periodicities using a smoothed Fourier analysis. IEEE/ACM Trans Comput Biol Bioinform 11: 863–877. https://doi.org/10.1109/TCBB.2014.2315991
26. Messaoudi I, Oueslati AE, Lachiri Z (2015) 2D DNA representations generated using a new coding and the time-frequency analysis. JMIHI 5:1035–1044. https://doi.org/10.1166/jmihi.2015.1498
27. NAJMI AH, SADOWSKY J (1997) The continuous wavelet transform and variable resolution time-frequency analysis. Johns Hopkins APL Tech Dig 18:134–140
28. Nigatu D, Sobetzko P, Yousef M et al (2017) Sequence-based information-theoretic features for gene essentiality prediction. BMC Bioinformatics 18:1: 473. https://doi.org/10.1186/s12859-017-1884-5
29. Orhan U, Hekim M, Ozer M (2011) EEG signals classification using the K-means clustering and a multilayer perceptron neural network model. Expert Syst Appl 38:13475–13481. https://doi.org/10.1016/j.eswa.2011.04.149
30. Oueslati AE, Ellouze N, Lachiri Z (2007) 3D spectrum analysis of DNA sequence: application to *Caenorhabditis elegans* genome. In: Bioinformatics and Bioengineering (BIBE 2007) 864–871
31. Oueslati AE, Messaoudi I, Lachiri Z, Ellouze N (2015) A new way to visualize DNA's base succession: the Caenorhabditis elegans chromosome landscapes. Med Biol Eng Comput 53:1165–1176. https://doi.org/10.1007/s11517-015-1304-9

32. Öz E, Kaya H (2013) Support vector machines for quality control of DNA sequencing. JIAP 2013:85. https://doi.org/10.1186/1029-242X-2013-85
33. Poulter RTM, Goodwin TJD (2005) DIRS-1 and the other tyrosine recombinase retrotransposons. Cytogenet Genome Res 110:575–588. https://doi.org/10.1159/000084991
34. Poulter RT, Goodwin TJ, Butler MI (2003) Vertebrate helentrons and othernovel Helitrons. Gene 313:201–212. https://doi.org/10.1016/S0378-1119(03)00679-6
35. Pritham EJ, Feschotte C (2007) Massive amplification of rolling-circle transposons in the lineage of the bat Myotislucifugus. Proc Natl Acad Sci 104:1895–1900. https://doi.org/10.1073/pnas.0609601104
36. Schiilkopf B (2001) The kernel trick for distances. Adv Neural Inf Proces Syst 13:301–307
37. Schlötterer C (2000) Evolutionary dynamics of microsatellite DNA. Chromosoma 109:365–371. https://doi.org/10.1007/s004120000089
38. Shawe-Taylor J et al (1998) Structural risk minimization over data-dependent hierarchies. IEEE Trans Inf Theory 44:1926–1940. https://doi.org/10.1109/18.705570
39. Song J, Li F, Takemoto K et all (2018) PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. J Theor Biol 443:125–137 https://doi.org/10.1016/j.jtbi.2018.01.023
40. Suo H, Li M, Lu P, Yan Y (2008) Using SVM as back-end classifier for language identification. EURASIP ASMP 2008:674859. https://doi.org/10.1155/2008/674859
41. Sweredoski M, DeRose-Wilson L, Gaut BSA (2008) Comparative computational analysis of nonautonomous helitron elements between maize and rice. BMC Genomics 9:467. https://doi.org/10.1186/1471-2164-9-467
42. Takezaki N, Nei M (1996) Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics 144:389–399
43. Tempel S (2007) Dynamique des hélitronsdans le génomed'arabidopsisthaliana: développement de nouvellesstratégiesd'analyse des élémentstransposables. PHD Thesis, IRISA, Université de Rennes I. https://tel.archives-ouvertes.fr/tel-00185256
44. The NCBI GenBank database. [Online]. Available: http://www.ncbi.nlm.nih.gov/Genbank/. Accessed 15 Sept 2005
45. Thomas J, Pritham EJ (2015) Helitrons, the eukaryotic rolling-circle transposable elements. Mobile DNAIII ASMscience 3:893–926. https://doi.org/10.1128/microbiolspec.MDNA3-0049-2014
46. Touati R, Messaoudi I, Oueslati AE, Lachiri Z (2018) Helitron's periodicities identification in C. Elegans based on the smoothed spectral analysis and the frequency Chaos game signal coding. Int J Adv Comput Sci Appl 9(4). https://doi.org/10.14569/IJACSA.2018.090438
47. Touati R, Messaoudi I, Oueslati AE, Lachiri, Z (2018) Classification of Helitron's Types in the *C. elegans* Genome based on Features Extracted from Wavelet Transform and SVM Methods. Bioinformatics 127–134. https://doi.org/10.5220/0006631001270134
48. Valli I, Marquand AF, Mechelli A et al (2016) Identifying individuals at high risk of psychosis: predictive utility of support vector machine using structural and functional Mri data. Front Psychiatry 7:52. https://doi.org/10.3389/fpsyt.2016.00052
49. Vapnik V (2013) The nature of statistical learning theory. Springer Science & Business Media
50. Vapnik VN, Vapnik V (1998) Statistical learning theory. Wiley, New York
51. Wicker T, Sabot F, Hua-Van A et al (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982. https://doi.org/10.1038/nrg2165
52. Xie D, Li A, Wang M, Fan Z, Feng H (2005) LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. Nucleic Acids Res 33:W105–W110. https://doi.org/10.1093/nar/gki359
53. Xiong W, He L, Lai J, Dooner HK, Du C (2014) HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proc Natl Acad Sci 111:10263–10268. https://doi.org/10.1073/pnas.1410068111
54. Yang L, Bennetzen JL (2009) Structure-based discovery and description of plant and animal Helitrons. Proc Natl Acad Sci 106:12832–12837. https://doi.org/10.1073/pnas.0905563106
55. Zhou Q et al (2006) Helitron transposons on the sex chromosomes of the Platyfish Xiphophorus maculatus and their evolution in animal genomes. Zebrafish 3:39–52. https://doi.org/10.1089/zeb.2006.3.39



Rabeb Touati ENIT, Tunisia -received her engineering degree in telecommunication from and PhD student in Signal Processing from the National Engineering School of Tunisia (ENIT). She was Assistant at The National Institute of Applied Science and Technology (INSAT). Her research inte-ests include genomic signal processing.

Imen Messaoudi ENIT, Tunisia- received her bachelor's degree in Electronic from PHD in Signal Processing from the National Engineering School of Tunisia (ENIT) and her PHD in Signal Processing from the National Engineering School of Tunisia (ENIT). She was Assistant at The Informatic Department of the Higher Institute of Accountancy and Business Administration (ISCAE). She is, currently, Assistant at the Higher Institute of Biotechnology of Sidi Thabet (ISBST).Her research interests include biomedical signals particularly ECG and genomic signal processing.



Afef Elloumi Oueslati ENIT, Tunisia -received her Diploma in Electrical Engineering and her PhD in Signal Processing from the National Engineering School of Tunisia (ENIT). She is Assistant Professor at the Electrical Department of the Higher School of Technologies and Informatics (ESTI). Her research interests include issues related to biomedical signals especially ECG and genomic signal processing. She has published research papers at international journal, and national and international conference proceedings.

Zied. Lachiri ENIT, Tunisia – was born in Tunis, Tunisia. He received the B.S degree in Electrical Engineering, the M.S. degree in automatic and signal processing and the PhD. degree in Electrical Engineering from the National School of Engineering of Tunis (ENIT-Tunisia), in 1993, 1997 and 2002, respectively. In 2002 he joined the Physic and Instrumentation Department, INSAT, as Assistant Professor and became Associate Professor and Professor in 2007 and 2012. He is currently a Professor at the Department of Electrical Engineering, ENIT and Research Director with the Signal, Image and Information Technology laboratory (LR-SITI, ENIT). His research interests include pattern recognition, signal processing and image processing applied in biomedical, multimedia and man machine communication. He is member of the European Association for Signal, Speech and Image Processing.