Biomedical Signal Processing and Control 64 (2021) 102207

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

New methodology for repetitive sequences identification in *human* X and Y chromosomes

Rabeb Touati [a],[b],*, Asma Tajouri [a], Imen Mesaoudi [b], Afef Elloumi Oueslati [b], Zied Lachiri [b], Maher Kharrat [a]

[a] University of Tunis El Manar, LR99ES10 Human Genetics Laboratory, Faculty of Medicine of Tunis (FMT), Tunisia
[b] University of Tunis El Manar, SITI Laboratory, National School of Engineers of Tunis, BP 37, Le Belvédère, 1002, Tunis, Tunisia

ARTICLE INFO

Keywords:
Repetitive sequences
Satellites
Wavelet transform
Canny edge detection
Human genome
New repetitions database
Convolutional neural network (CNN)

ABSTRACT

Repetitive DNA sequences occupy the major proportion of DNA in the human genome and even in the other species' genomes. The importance of each repetitive DNA type depends on many factors: structural and functional roles, positions, lengths and numbers of these repetitions are clear examples. Conserving such DNA sequences or not in different locations in the chromosome remains a challenge for researchers in biology. Detecting their location despite their great variability and finding novel repetitive sequences remains a challenging task. To side-step this problem, we developed a new method based on signal and image processing tools. In fact, using this method we could find repetitive patterns in DNA images regardless of the repetition length. This new technique seems to be more efficient in detecting new repetitive sequences than bioinformatics tools. In fact, the classical tools present limited performances especially in case of mutations (insertion or deletion). However, modifying one or a few numbers of pixels in the image doesn't affect the global form of the repetitive pattern. As a consequence, we generated a new repetitive patterns database which contains tandem and dispersed repeated sequences. The highly repetitive sequences, we have identified in X and Y chromosomes, are shown to be located in other human chromosomes or in other genomes. The data we have generated is then taken as input to a Convolutional neural network classifier in order to classify them. The system we have constructed is efficient and gives an average of 94.4% as recognition score.

1. Introduction

Repetitive DNAs are sequences with multiple copies in the genome. They are rarely associated with clearly defined biological functions. Some of the moderately-repetitive sequences may be involved in gene expression regulation. Other mobile DNA can be constituted by transposable genetic elements (TEs) that are involved in the genome evolution process. The transposition mechanism and the structure of these TEs are the keys to dividing this DNA into classes. Retrotransposons, are an example of TEs class that move via an RNA intermediate. This RNA is transcribed from the DNA and subsequently copied back into DNA. As repetitive DNA we can find tandem repeats or scattered repeated sequences. These repetitive DNA sequences can be classified into two types: highly repetitive or moderately repetitive sequences [1,2].

The major repetitive sequences in all eukaryotic cells are classified into five types according to the sequence's length. In this classification, the microsatellite sequences (Short Tandem Repeat: STR) are the smallest. They are characterized by periodicity between 2 and 4 nucleotides per unit. The second class is constituted by the minisatellites with a length varying between 10 and 60 base pairs (bp). The third class is composed of the satellites which can contain up to 100 nucleotides (100–200 base pairs) [3–5]. The retrotransposons like SINE and LINE are part of the fourth-class which is characterized by a length varying from 50 bp to 6 kb. The final class consists of Ribosomal RNA gene repeat (rDNA) which is the longest with a length between 9 and 45 kb.

In the Human genome, rare fragile sites are chromosomal DNA regions especially characterized by repetitive sequences. In fact, in these regions, DNA damage occurs more frequently than in other locations. Due to chromosome structure, the common fragile sites can be sensitive to replication stress, and they are often rearranged in cancer. In the mammalian centromeres and telomeres, the presence of repetitive sequences is necessary in order to protect chromosomes from damage. For example, alphoid DNA is a kind of DNA satellite having a length of 173 bp. This DNA is located in the middle of a chromosome and makes up the

* Corresponding author at: University of Tunis El Manar, LR99ES10 Human Genetics Laboratory, Faculty of Medicine of Tunis (FMT), Tunisia.
E-mail addresses: Rabeb.touati@enit.utm.tn, Rabeb.touati.enit@gmail.com (R. Touati).

https://doi.org/10.1016/j.bspc.2020.102207
Received 20 December 2019; Received in revised form 23 July 2020; Accepted 1 September 2020
Available online 19 October 2020

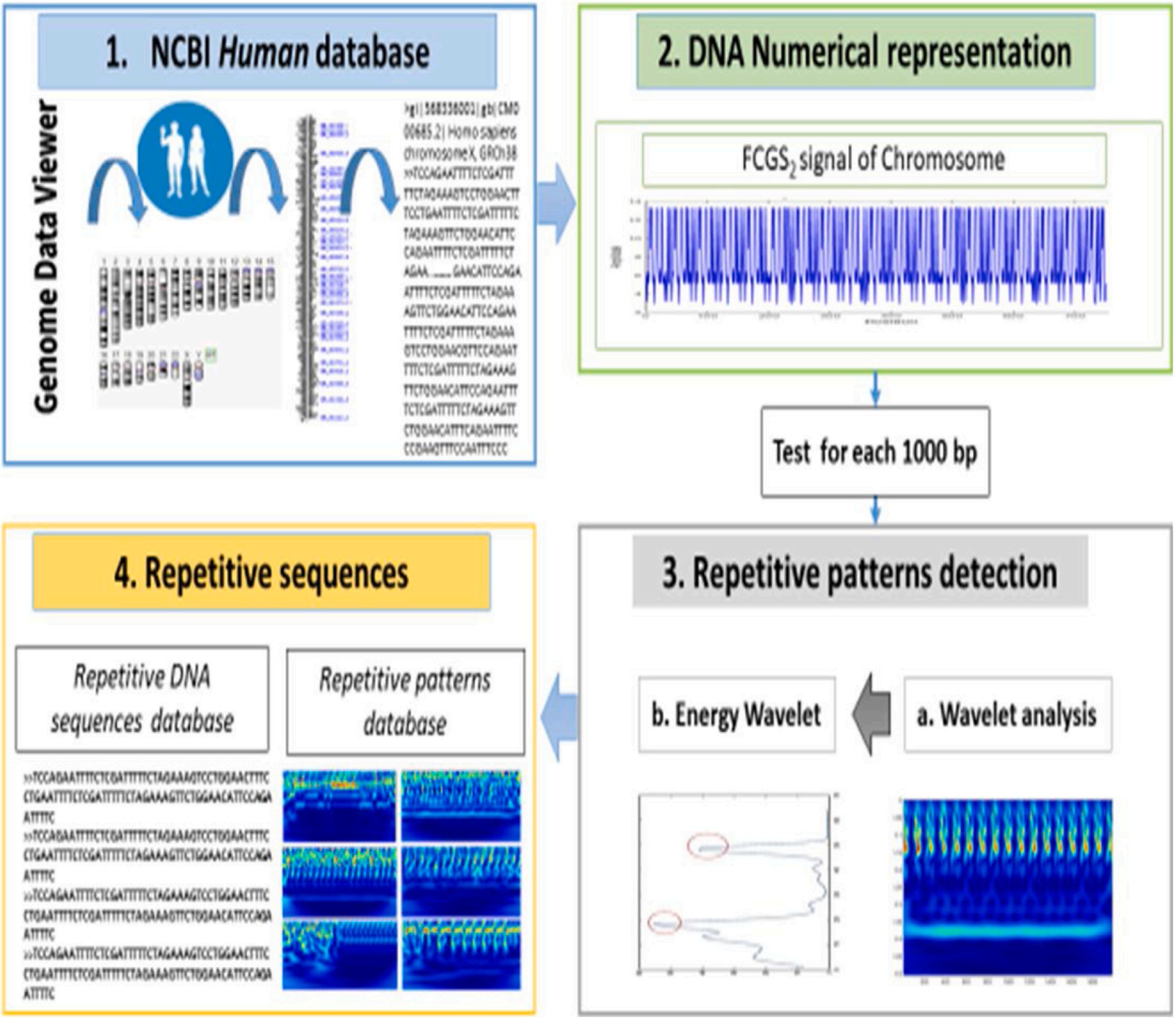

1. NCBI Human database
Genome Data Viewer
2. DNA Numerical representation
FCGS$_2$ signal of Chromosome
Test for each 1000 bp
3. Repetitive patterns detection
b. Energy Wavelet
a. Wavelet analysis
4. Repetitive sequences
Repetitive DNA sequences database
Repetitive patterns database

Fig. 1. Organizational flowchart of the identification of the Repetitive sequences.

larger part of the Human centromeres region [6]. Moreover, telomeres regions located at the chromosome extremities are made up of repeat sequences of 5–7 bp. These elements are called telomere repeats [7]. The repetitive sequence ‘TTAGGG’ is one example. The chromosome integrity is protected by telomere repeats [8,9]. In fact, telomeres hinder the chromosomes’ fusion and protect them against degradation by exonucleases [10].

These repetitive functional elements are not susceptible to become fragile sites because they are hidden in heterochromatin. This heterochromatin prevents unusual DNA structures occurrence leading to recombination by not yet identified mechanisms [11].

Repetitive sequences are abundant in various genomes, from bacteria to mammals, and they cover nearly half of the Human genome [5]. Finding new common repetitive sequences within and between different chromosomes and genomes is an important theme of research in biology. In fact, the detection of all repetitive sequences in DNA could serve in elucidating important biological phenomena. To identify the repetitive sequences, different bioinformatics tools were used [12,13]. Their principle is based on comparison between DNA consensus sequences and repeats candidates. The Mreps [14], MISA [13], Sputnik [15], EMBOSS (etandem and equitandem) [16], TRF [17] and RepeatMasker [18] are obvious examples. In the comparison step, these tools used different approaches such as regular expression [18], Hamming distance [12], recursive match and penalty scores [17]. Localizing new repetitive sequences presents always technical challenges. This is due to the ambiguities that such repeats can create in alignment and assembly programs [19]. In this work, we have developed a new algorithm to detect repetitive patterns that correspond to new repetitive sequences. For this purpose, we used a combination of coding techniques, signals, and image processing techniques. As a result, we have constructed a repetitive sequence database which we subdivided into two sub-databases. The first one contains the existing and validated repetitive sequences. The second DNA repetitive database regroups the newly detected sequences.

In this context, we called "new repetitive sequence", a sequence that was not detected by all current bioinformatics systems as well as alignment programs. In this research, we converted all of the DNA sequences into a synthetic image representation. After that, we extracted all patterns that correspond to the repeat DNA sequences. The second part of this work consists in classifying the obtained data. A deep learning model is chosen for this purpose: Convolutional neural network (CNN).

This paper is divided into four sections. After the introduction, we describe the materials and methods. In Section 2, we first present the biological database subject of this study. We also introduce the coding technique we used to transform the biological data into a numerical one. After that, we describe how we convert the obtained signal into an image based on the wavelet analysis. Further, we introduce the CNN architecture we establish for the repetitive DNA classification. The final parts of this section consist of the employed detection steps and the adopted

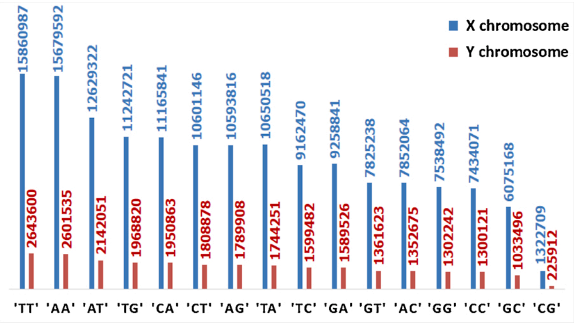


Fig. 2. Dinucleotide occurrence in X and Y chromosomes in the human genome.

evaluation system. In Section 3, we provide and discuss the results in terms of repetitive DNA sequences detection and classification. Finally, Section 4 concludes the paper.

2. Material and methods

Two-thirds of the human genome consists of repetitive DNA sequences [20]; which confers great importance to identification and localization of these elements. In this section, we expose a novel approach for the repetitive DNA sequence identification. This method is effective in detecting dispersed or tandem repeats such as minisatellites and satellites. The detection system is composed of four main blocks. The first one consists in extracting the Human DNA sequences from existing database. The second block is the DNA coding into a numerical representation. The third block consists of "Find Human Repetitive Sequences" (FHRS) method which we propose to the Repetitive DNA sequences detection. It is the application of the wavelet analysis and thus for detecting the repetitive patterns. The last block consists of determining the repetitive sequences and the repetitive DNA sequences database establishment. Fig. 1 shows the corresponding flowchart.

2.1. Human sequences database (DNA library)

The human genome (*Homosapiens*) contains 22 autosomes and two chromosomes that determine human sex: X and Y, with a total number of 46 chromosomes. We find one pair of sex chromosomes in each human cell. In females, the cell contains two X chromosomes, while in males we have one X and one Y chromosome. A detailed description of the human DNA material is available in the NCBI database (National Center for Biotechnology Information) [21]. From the human DNA data, we count 2.91-billion base pairs (bp) consensus sequence in the euchromatic portion [22]. Given that this is a huge amount of data, we based our work only on X and Y chromosomes. Even, at the level of these two chromosomes, we have an important mass of data. As an example, we give in Fig. 2 the number of apparition of dinucleotides in both X and Y chromosomes.

Our goal is to find repetitive DNA on these chromosomes. It is important to mention that the more complex the genome is, the more difficult is to find new repetitive sequences within. Therefore, the challenge presented in this work is identifying new repetitive DNA sequences in human X and Y chromosomes.

2.2. DNA coding for numerical representation

Aiming to visualize repetitive patterns in the human genome, the DNA sequences have to be transformed into numerical data. This transformation is called “DNA coding”. In this work, we opted for a special coding technique called “Order 2 Frequency Chaos Game Signal” ($FCGS_2$) [23,24]. The $FCGS_2$ coding is a statistical representation of DNA. In the proposed method, chromosomes are transformed based on the occurrence probability of the successive dinucleotides groups. This technique represents the time-frequency evolution of the dinucleotides in the chromosome. In the following, we give the transformation equation (eq. 1).

$$\begin{cases} FCGS_2(x) = \sum_x \sum_i P_{2_{nucléotide}}(i,x), \\ P_{2_{nucléotide}} = N_{2_{nucleotide}} / Length_{Chr}. \end{cases} \quad (1)$$

where $N_{2_{nucleotide}}$ is the occurrence number of dinucleotides group in the whole chromosome and $Length_{Chr}$ is the chromosome’s length.

In this work, we coded the entire human chromosomes X and Y. The sequence that represents chromosome X is a signal with a length of 156,040,895 bp. As for chromosome Y, it is a signal of size equal to 57227415bp.

2.3. Find human repetitive sequences (FHRS) approach

The identification of repetitive DNA sequences is taking greater and greater importance these days. Many algorithms, using various knowledge fields, have been implemented for repetitive sequences localization. In this context, signal processing approaches were used to detect repetitive sequences, according to the correspondent periodicity [25–29]. In this paper, we propose an efficient algorithm based on the signal and image processing tools to localize repetitive DNA sequences. This method has the advantage of being independent from prior knowledge about the repeated sequences. This section presents the new algorithm we designed to detect the repetitive DNA-sequences after transforming them into numerical signals. This algorithm is called Find Human Repetitive Sequences (FHRS). It contains three steps:

- DNA signals to DNA images transformation: the scalogram representation;
- Energy calculation of each scalogram image which is obtained by the wavelet analysis. After that, retaining the image whose energy amplitude exceeds a chosen threshold (equal to 10 here);
- Finding the reference repetitive sequence in the retained image. It is the longest repeated unit in the considered DNA sequence.

2.3.1. The DNA time frequency representation by the complex Morlet analysis

The scalogram representation of a DNA sequence is an image that we obtain by wavelet analysis and encode in the RGB space (three color channels: Red, Green, and Blue). This time-frequency representation is shown to be efficient in terms of visualizing and detecting repetitive patterns. Here, the idea is to use this type of DNA image to find repetitive patterns that correspond to periodic sequences.

The motivation behind this choice is that changing a pixel in the image has no influence on the overall shape of the repetitive pattern. Indeed even if the repetition pattern contains variations in nucleotide composition, this does not greatly impact the overall shape of the repetitive pattern at the level of DNA image. Furthermore, our choice for this method is reinforced by its performance in characterizing different classes of transposable elements [30,31]. For the wavelet analysis, we use the complex Morlet wavelet which is best suited to localize repetitive DNA in the time-frequency domain. The principle consists of applying the wavelet analysis to the signal obtained by the $FCGS_2$ coding. This analysis is done by decomposing a given DNA signal into a sum of basic functions called wavelets. The latter wavelets are issued from the mother wavelet by two operations: expansion and translation. These wavelets take into account both time and frequency variations, which allow them to easily capture all the different hidden frequencies in the signal [32–34]. Unlike the mother wavelet, which only has a time-varying parameter expressed by the function ψ(t), the daughter wavelet expression depends on time and scale parameters (a and b

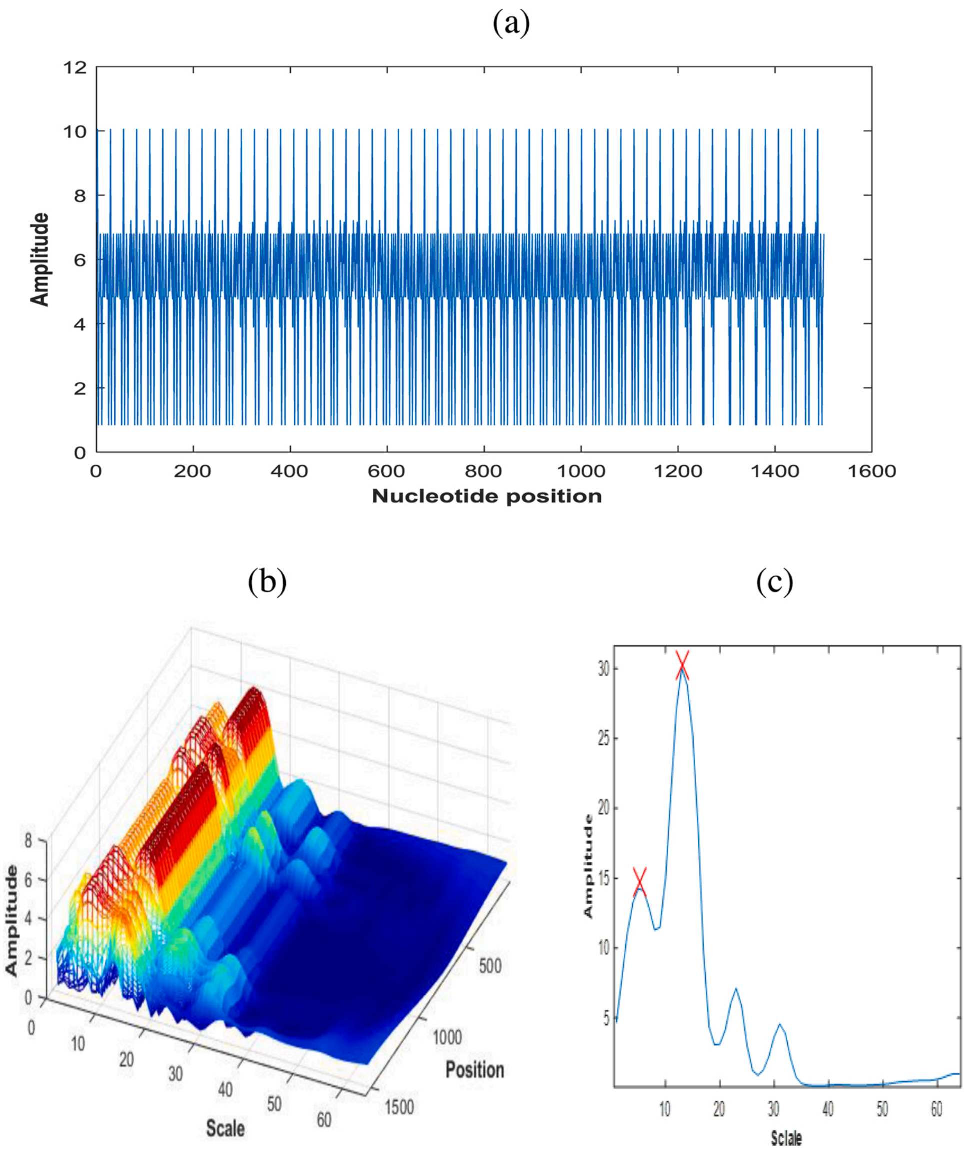


Fig. 3. Illustration of the repetitive DNA detection steps based on CWT analysis: a) DNA coding with FCGS$_2$ b) 3D scalogram c) Energy peaks greater than 10 indicates the existence of repetitive sequences.

respectively). It is generated following this equation:

$$\psi_{a,\,b}(t) = \frac{1}{\sqrt{a}}\psi^{*}\left(\frac{t-b}{a}\right),\ \ \mathrm{a} > b \in \mathbb{R} \tag{2}$$

where* indicates the conjugate complex. As we have chosen a Gaussian-windowed complex sinusoid (complex Morlet) to be applied as analysis window, the Continuous Wavelet Transform (CWT) will be written as:

$$\psi_{cmor}(t) = \Pi^{-\frac{1}{4}}\left(e^{i\omega_0 t} - e^{-\frac{1}{2}i\omega_0{}^2}\right)e^{-\frac{t^2}{2}} \tag{3}$$

Here the oscillation's number (ω_0) must be greater than 5 (admissibility condition). The continuous wavelet coefficients of a DNA signal $x(t)$ is a matrix which elements are calculated by the following formula:

$$W_{(a,b)}[x(t)] = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} x(t)\psi^{*}\left(\frac{t-b}{a}\right)dt \tag{4}$$

The modulus of these coefficients $|W_{(a,b)}|$ provides the scalogram representation of the DNA sequence.

2.3.2. *The energy calculation of the DNA scalograms*

Since chromosomes X and Y are too long, we decompose $x(t)$, which is the correspondent FCGS$_2$ signal, in a set of segments. Each segment $x_i(t)$ has a size of 1000 bp. After segment cut, we apply the CWT wavelet and calculate the correspondent energies. As a result, we obtain a new database of the human DNA representations. In total, we count 156,041 images of the X chromosome and 57,228 images of the Y chromosome. The wavelet coefficients matrix contains the time-frequency information about a signal. To further explore this information, we calculate the scale-energy (E) of each nucleotide position, according to following equation:

$$\begin{cases} E_i(a) = \sum_{b=1}^{1000}\left|W_{(a,b)}[x_i(t)]\right|^2, \\ for\ \ each\ \ i = 1 : Length_{Chr}/1000. \end{cases} \tag{5}$$

Here, the parameter a represents the scale in the wavelet analysis; it varies from 1 to 64. As for the indicator i, it represents the image number.

By applying (eq.5), we obtain a vector that contains the energy of the DNA scalogram. Peak values higher than 10 in the vector indicate the existence of repetitive patterns in the DNA image. Fig. 3 shows an

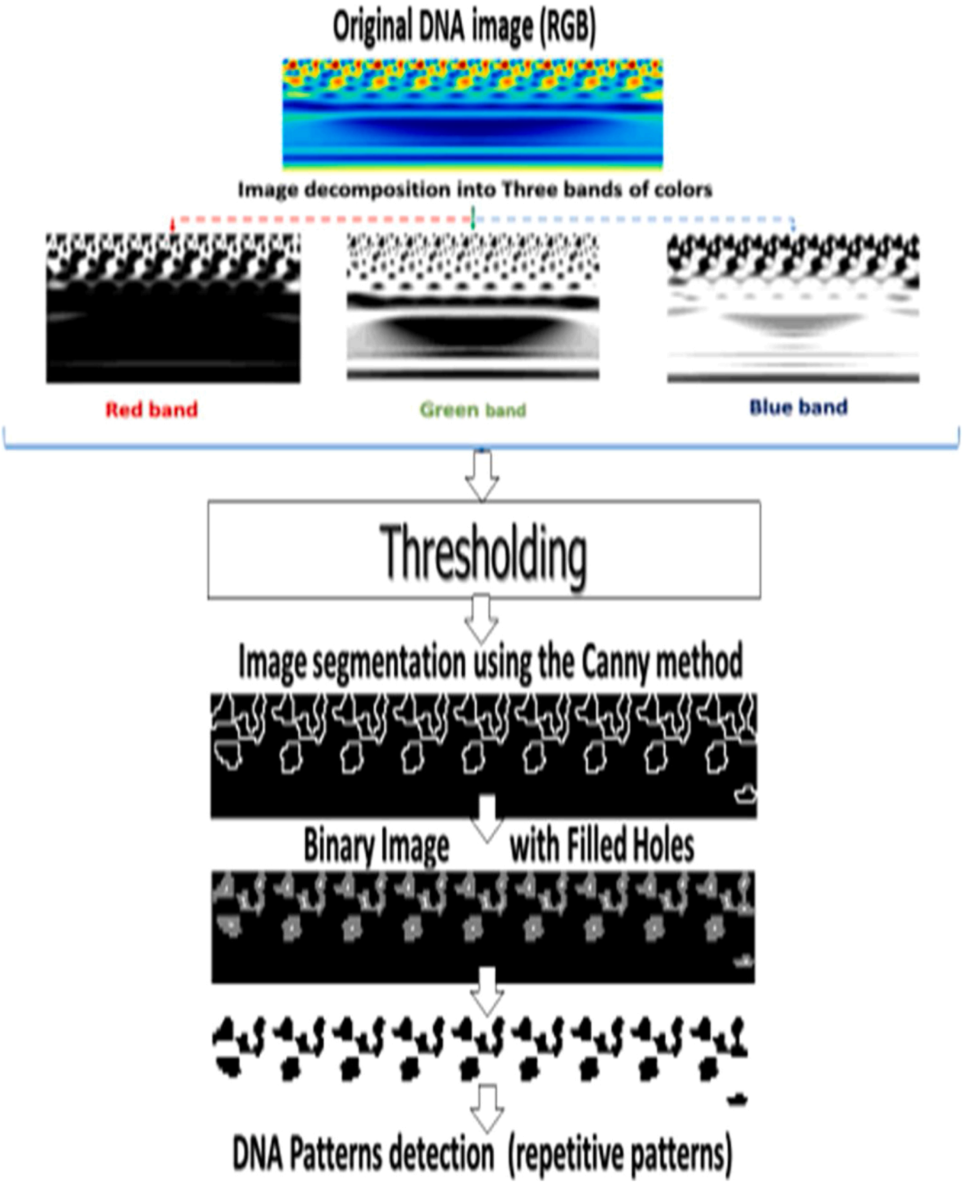

Original DNA image (RGB)
Image decomposition into Three bands of colors
Red band
Green band
Blue band
Thresholding
Image segmentation using the Canny method
Binary Image with Filled Holes
DNA Patterns detection (repetitive patterns)

Fig. 4. Flowchart diagram of the adopted segmentation methodology to extract the repetitive patterns.

example of the $FCGS_2$ signal, the correspondent scalogram in a 3D representation and the energy wavelet of a sequence located in chromosome X of the human genome. This sequence corresponds to the portion [342,500 bp: 344,000 bp] in the *PPP2R3B* gene.

As we can see, magnitude of the energy wavelet indicates the presence of periodicities in the sequence. If we consider the frequency content, we can note that the repetitive sequence is characterized by a specific frequency band. The limits of this frequency band correspond to the repetitive DNA portion in the analyzed sequence. As for the 3D representation, it contains repetitive patterns of particular shape that are related to the DA repetitions. Following this method, we have constructed our database of the repetitive DNA images. The patterned images were selected according to the energy-wavelet peaks. The generated database was named "*repeat-Data*".

2.3.3. The reference repetitive sequence search

For each DNA image into the *repeat-Data* database, we aim to identify a DNA-reference sequence, to which corresponds the existing repetitive pattern in the scalogram. This DNA-reference sequence is the longest subsequence in terms of size and repetition numbers. After this step, we have built a database that contains the location and the repetition number of all the localized sequences of reference. As we focus on detecting new repetitive sequences in the human genome, we verified the availability of the reference repetitive sequence in the public databases. For this, we checked if this sequence is annotated or not in both *DFAM* and *NCBI* databases. Hence, if our new repetitive sequence is not listed in these public databases, we added it to our new database. This new repetitive sequence is called "*New-repeat-Data*".

2.4. Patterns extraction based on adaptive local thresholding and morphological processing

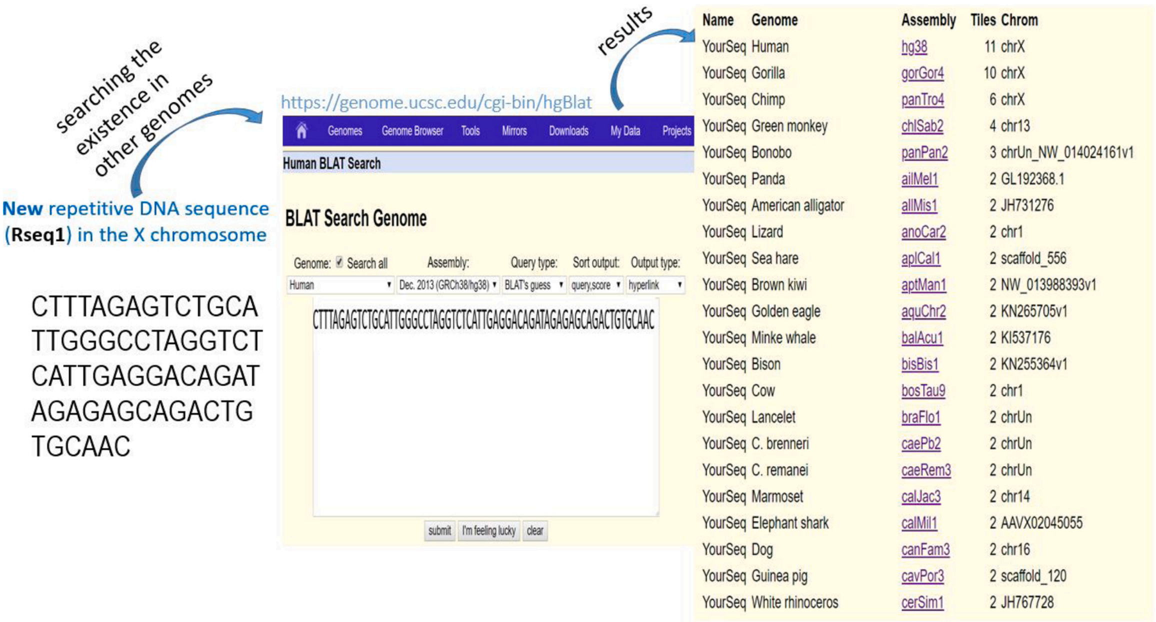
After collecting the new repetitive sequences using the FHRS algorithm, we move on to the step of extracting the repeat patterns using image processing tools. The Fig. 4 summarizes the proposed methodology of extracting tandem repeat patterns in the DNA images. It illustrates the results obtained when we considered the “TRseq1” sequence. The sequence is 261 base pairs lengthen; its position is 28,076,765 bp to 28,077,025 bp along the human X chromosome.

As in this example, the data we are treating here is the set of

Table 1
Position of “Rseq1” on both X and Y chromosomes of the human genome.

Start (bp)	End (bp)	Start (bp)	End (bp)
26,609	26,669	41,241	41,301
26,792	26,852	42,400	42,460
28,317	28,377	243,312	243,372
34,474	34,534	244,958	245,018
34,657	34,717	246,787	246,847
41,058	41,118	248,556	248,616

scalogram images that we stored before in the database "New-*repeat-Data*". The main goal of this part of work is to detect and localize the repetitive patterns in the scalogram representations. That’s why we based our work on a segmentation algorithm. Our method consists first in decomposing the DNA image into three color channels (red, green and blue) and choosing the blue one. This choice is justified after testing all the color bands. The best segmentation result corresponds to the bleu channel since it is best contrasted compared to the others. Then for a binarization purpose, a simple thresholding is applied to keep only the pixels having an intensity value less than or equal to 26. Then, to keep only the region of interest, we have used an edge detection technique. The Canny edge detector provides good detection and localization relatively to other operators [35,36]. The algorithm detects brightness discontinuities in the image using a Canny filter. It is a multi-stage algorithm used to detect a wide range of edges in images [37,38]. The Canny operator uses double thresholds: high and low thresholds. The high threshold algorithm detects important and significant information like lines and contours in the image. The low threshold algorithm ensures that no details are missing. The Canny edge detector is widely used to locate sharp intensity changes and to find object boundaries in an image, especially in computer vision domains. The classification of one pixel as an edge, using the Canny edge detector, is achieved by gradient


searching the existence in other genomes
New repetitive DNA sequence (Rseq1) in the X chromosome
results
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
Projects
submit
I'm feeling lucky
clear

CTTTAGAGTCTGCA
TTGGGCCTAGGTCT
CATTGAGGACAGAT
AGAGAGCAGACTG
TGCAAC

https://genome.ucsc.edu/cgi-bin/hgBlat

Human BLAT Search

BLAT Search Genome

Genome: ☑ Search all: Human
Assembly: Dec. 2013 (GRCh38/hg38)
Query type: BLAT's guess
Sort output: query,score
Output type: hyperlink

CTTTAGAGTCTGCATTGGGCCTAGGTCTCATTGAGGACAGATAGAGAGCAGACTGTGCAAC

Name	Genome	Assembly	Tiles	Chrom
YourSeq	Human	hg38	11	chrX
YourSeq	Gorilla	gorGor4	10	chrX
YourSeq	Chimp	panTro4	6	chrX
YourSeq	Green monkey	chlSab2	4	chr13
YourSeq	Bonobo	panPan2	3	chrUn_NW_014024161v1
YourSeq	Panda	ailMel1	2	GL192368.1
YourSeq	American alligator	allMis1	2	JH731276
YourSeq	Lizard	anoCar2	2	chr1
YourSeq	Sea hare	aplCal1	2	scaffold_556
YourSeq	Brown kiwi	aptMan1	2	NW_013988393v1
YourSeq	Golden eagle	aquChr2	2	KN265705v1
YourSeq	Minke whale	balAcu1	2	KI537176
YourSeq	Bison	bisBis1	2	KN255364v1
YourSeq	Cow	bosTau9	2	chr1
YourSeq	Lancelet	braFlo1	2	chrUn
YourSeq	C. brenneri	caePb2	2	chrUn
YourSeq	C. remanei	caeRem3	2	chrUn
YourSeq	Marmoset	calJac3	2	chr14
YourSeq	Elephant shark	calMil1	2	AAVX02045055
YourSeq	Dog	canFam3	2	chr16
YourSeq	Guinea pig	cavPor3	2	scaffold_120
YourSeq	White rhinoceros	cerSim1	2	JH767728

Fig. 5. Illustrative example of using BLAT algorithm to search a new repetitive DNA sequence in the whole human genome and in other genomes.

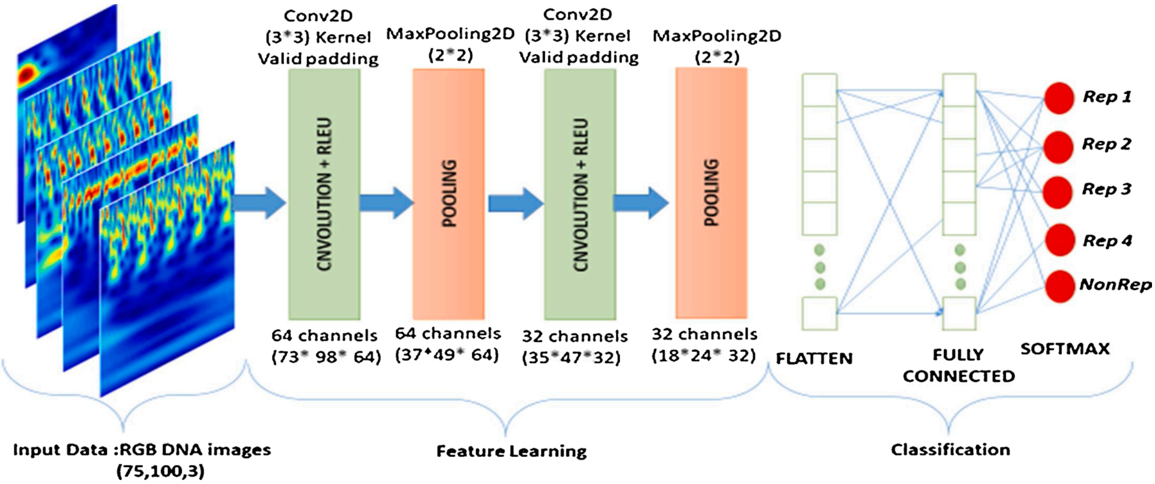


Fig. 6. Architecture of convolutional neural network for DNA images recognition. First layer is convolutional layer. It consists of 64 channels with kernel size of 3*3 voxels. The second is the maxpooling layer. Output of maxpooling layer is the input of the third layer: convolutional layer with 32 channels. Each convolutional layer is the input of the fourth layer: maxpooling layer. Then, the output of maxpooling layer is concatenated, a vector is formed and then inputted to the fully connected layer. The images from the dataset (N = 980) were splitted into 80% for training (780 images) and 20% for testing (200 images). Multiple epochs were used in the training procsess, where the epoch’s number used is equal to 100.

magnitude computation of this pixel. The result is then compared with one of its neighbors, where the maximum intensity varies the most. Finally, we fill the holes in areas of interest based on morphological operators [39]. The result is an image that only contains repetitive patterns. Based on this method, we can then extract and isolate the particular regions of repetitive DNA patterns.

2.5. DNA-reference sequences location in other human chromosomes and other species

After finding the DNA repetitive sequences in the human X and Y chromosomes (which can be tandem or scattered repeated sequences), we verified their existence in other chromosomes or even in other genomes. To achieve this goal, we have used two public bioinformatics algorithms: BLAT [40] and DFAM [41]. For each new repetitive sequence we detected, we searched it in the whole human genome and in all other genomes using the BLAT platform. As an example, we consider the new scattered repeated sequence “Rseq1”.

Rseq1="CTTTAGAGTCTGCATTGGGCCTAGGTCTCATTGAGGACA-GATAGAGAGCAGACTGTGCAAC".

It is a 61 base pair (bp) lengthen sequence with a repetition number equal to12 in the whole human genome. The corresponding positions on both X and Y chromosomes are given in the following table (Table 1).

After localizing “Rseq1” in X and Y chromosomes, we searched for the existence of this sequence in other regions. Fig. 5 shows the result of the checking of the “Rseq1” existence in other species. As we can see, “Rseq1” exists in several genomes such as; *Human*, *Gorilla*, *Chimpanzee*, *Greenmonkey*, *Bonobo*, etc.

After proving the existence of the newly discovered repetitive sequence in all genomes, we tried to find whether this sequence is located in genes. We, especially, searched for its existence in exonic regions or in other families of DNA. If this sequence exists nowhere in these DNA types, we classified it as a new repetitive DNA sequence type. On the other hand, we verified the uniqueness of these new sequences using our approach FHRS, and thus by comparing the repetitive patterns in the scalogram representations.

In order to ensure that our work is as meaningful and effective as possible, we thought of establishing a classification system to classify these new datasets (new repetitive DNA sequences). For this reason, we considered the scalogram representation (2D image) as input data to the system. As for the classifier, we have chosen CNNs as they are efficient in terms of images classification Fig. 6.

2.6. Convolutional neural network: CNN

CNN is a special neural networks type which works using data having a grid topology [42]. CNNs classification technique were developed by LeCun et al. (in 1998) in the aim to recognize handwritten characters from bank checks. CNNs is a deep learning model inspired by the visual mechanism of living organisms. It uses convolutional layers to the features extraction from input data. In the CNN model, convolutional layer neurons are able to extract higher-level abstraction features from features extracted at the previous layer. CNN was applied with success in DNA studies [43–46], Breast Cancer Cell Segmentation [47,48], medical diagnosis [49,50], character recognition [51] and in other areas of application.

In this work, we used CNN to establish a system of new repetitive DNA sequences recognition in human X and Y chromosomes. For this, we took the RGB scalogram representations of DNA as the input of the classification system with a size of 75 × 100.

The DNA images are passed, then, through a stack of convolutional layers, where we used filters with a very small receptive field (3 × 3). These filters act in the role of a scanner as they capture motifs in different orientations (up/down, center, left/right). Each neuron output on a convolutional layer is the result of a convolution operation between the kernel matrix and the neuron input. As for Max-pooling, it is performed over a 2 × 2 pixel window. For each convolutional layer, the second layer is a global max-pooling layer. Each one of max-pooling layers only outputs the maximum value of all of its respective convolutional layers outputs. The second layer is considered as a sample-based discretization process. This process has a goal to down the sample of input and to reduce its dimensionality.

After transforming the image into a suitable form for the Multi-Level Perceptron, the image must be flattened into a column vector. The result is a flattened output that is fed to a feed-forward neural network.

A back-propagation was applied to every iteration of training. A Fully-Connected layer was added to ensure a non-linear combination learning of the high-level features (which are represented by the output of the flatten layer). The Fully-Connected layer is learning a possibly non-linear function in that space. Over an epoch’s series, using the Softmax Classification technique our model is eligible to distinguish between dominating and certain low-level features in images and it can classify repetitive DNA classes.

After transforming the image into a suitable form for the Multi-Level Perceptron, the image must be flattened into a column vector. The result is a flattened output that is fed to a feed-forward neural network. A back-propagation was applied to every iteration of training. A Fully-Connected layer was added to ensure a non-linear combination learning of the high-level features (which are represented by the output of the flatten layer). The Fully-Connected layer is learning a possibly non-linear function in that space. Over an epoch’s series, using the Softmax Classification technique our model is eligible to distinguish between dominating and certain low-level features in images and it can classify repetitive DNA classes.

3. Results

Only sexual chromosomes provide opportunities to know the evolution mechanisms from one specie to another. These mechanisms can depend on the accumulation of repetitive sequences [2]. In this work, we first applied the FHRS technique to detect new repetitive sequences within human sexual chromosomes (X and Y). After that, we entered these sequences to a CNN based on classification system aiming at recognizing them.

3.1. New repetitive DNA detection results

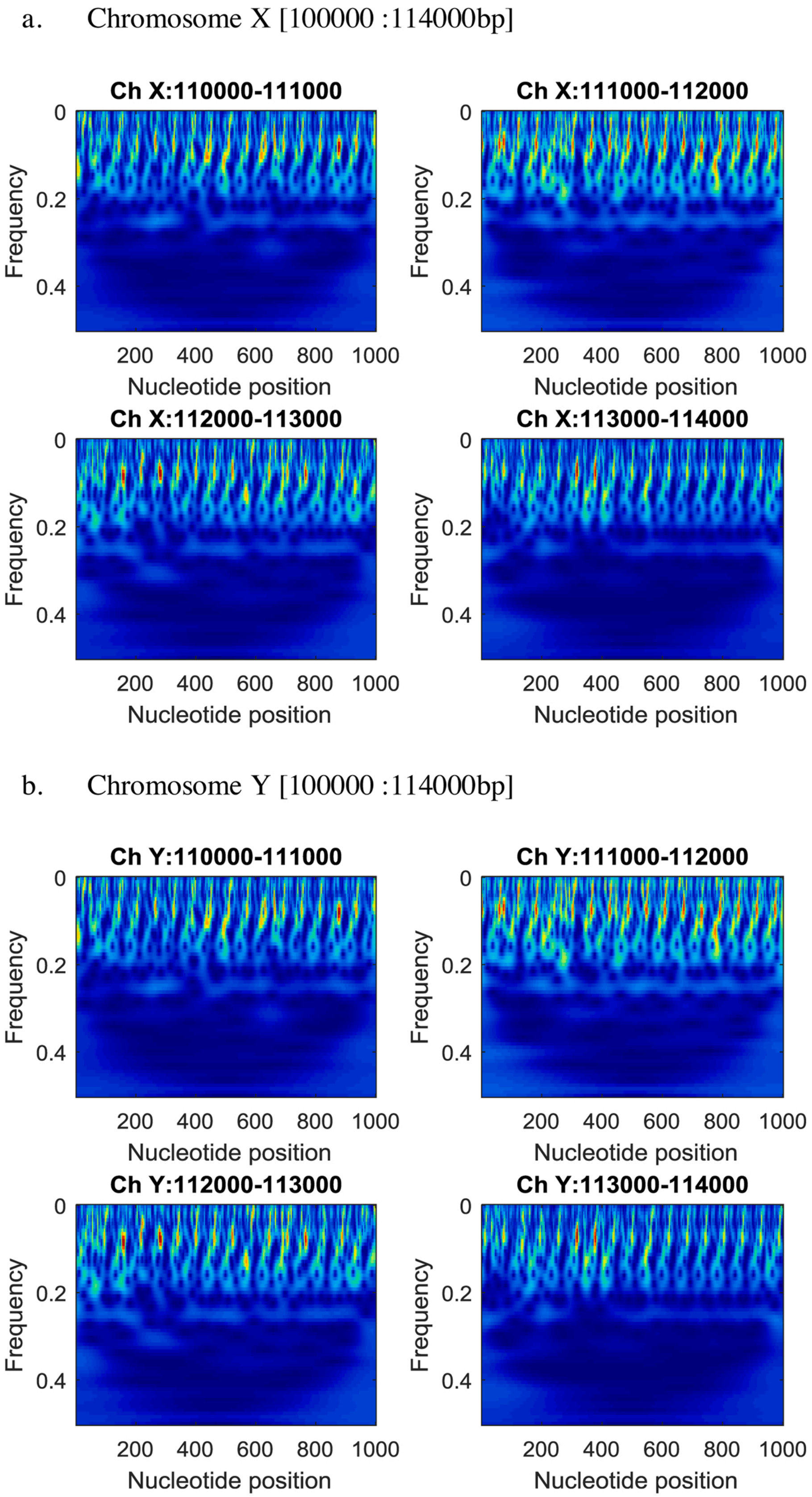
In this work, we used the FHRS approach (Find Human Repetitive Sequences), which combines wavelet analysis and a specific coding technique, to represent repetitive patterns in the form of an image. This method has the advantage of identifying new repetitive sequences without using any prior knowledge about the input DNA sequence. Based on this, we have discovered various new repetitive DNA sequences within sexual chromosomes, be they tandem or interspersed. After that, we have looked for the existence of these sequences in the whole human chromosomes or in other genomes. Afterward, we checked if these sequences exist or not in genes. Finally, we classed these repetitive sequences in terms of their relative location to heterochromatin, telomere, and centromere.

As a result, we have constructed a database comprising two sub-databases. The first one contains newly discovered repetitive sequences of type satellites and minisatellites. The second one encloses existing repetitive sequences.

Here, the new repetitive sequences database provides the composition of the new highly repetitive DNA sequences and the correspondent locations. The repetitive sequences are of different sizes and are classified into two types: tandem repeat sequences or interspersed repeat sequences. We called this new database "*New-repeat-Data*".

With our approach, highly conserved repetitive DNA sequences, having no annotations in the DNA library (NCBI or DFAM), have been found in the human genome.

In the telomere of X and Y chromosomes, we have found highly short



a. Chromosome X [100000 :114000bp]

b. Chromosome Y [100000 :114000bp]

Fig. 7. Telomere image signature of homologue regions corresponding to the minisatellite “Rseq2” (*CTTTAGAGTCTG*)$_n$ within X and Y chromosomes.

Table 2
Positions of the new discovered repeat sequence “Rseq3” in 12 chromosomes of the human genome.

Chromosome	Repetition number	Chromosome	Repetition number
Chr X	2302	Chr 6	3016
Chr 1	3447	Chr 7	2562
Chr 2	4193	Chr 8	2405
Chr 3	3450	Chr 9	1916
Chr 4	3647	Chr 10	837
Chr 5	3263	Chr 11	2021

or long repetitive sequences. The sequence “Rseq2” (Rseq2=CTTTAGAGTCTG) is an example of short Minisatellite of 21 base pairs. Its repetition number is 312 extending from 26,304 bp to 249,544 bp. In addition, the sequence $(CCCTAA)_n$, which is annotated in NCBI database, has been well localized using our algorithm.

As long repetitive Minisatellite sequences, we have discovered a new sequence “Rseq1” of 61 base pairs and a repetition number of 12. These repetitive sequences exist in the same location within great portions of chromosome Y. Fig. 7 shows an example of the global signature of a new telomeric repetitive sequence with a 71000bp of size.

On the other hand, a high repetitive sequence “Rseq3” (Rseq3=‘TTTAAAGAT’ of size equal to 9 bp) has shown as a new repetitive sequence in the human genome. This short repetitive DNA sequence was found also in many species such as chimpanzees, bonobo, and even in SARS−COV2 (COVID-19) coronavirus genome with a repetition number of 2. Table 2 shows the location of this microsatellite in some chromosomes of the human genome.

Other sequences are found to be very high repetitive in the human genome, like the sequence “Rseq4” (Rseq4= ‘GTATACA’) which appears in the X chromosome 1375 times. This sequence exists also in the COVID-19 coronavirus.

Furthermore, we have found a new minisatellite with a size of 61bp in human. Using the BLAT algorithm, this sequence was also found in the X chromosome of *Gorilla* (*gorGor4*) with a position of 15499bp to 15,559 bp. Fig. 8 shows the method adopted to localize this repetitive sequence in other regions.

Fig. 8 is divided into two result blocks. In the first one, we expose the scalogram corresponding to the new repetitive DNA sequence. The second one contains the sequence location result in all the other genomes using the BLAT algorithm.

In the first result block, we provide the scalogram representation of the DNA sequence we have located at the X chromosome of the human genome (Xp22.33, position: [321001:322000bp]). The scalogram representation makes possible to see all the specific repetitive patterns. After that we extracted the reference sequence which is the maximum repetitive sequence having a maximum size in the DNA sequence. Then, we have found two new repetitive sequences that were not referenced by the current bioinformatic systems or sequence alignment programs. Locations of these two new repetitive sequences in both X and Y chromosomes are given by Table 3. The repetitive patterns in the scalograms prove the presence of two microsatellites: Rseq5 whose size is 61bp and reRseq6 size is 28 bp.

These sequences are:

- Rseq5=‘AAAAAAAAAAAAAAAAGAAAAGCCGGGCGTGGTGGTGGGTGCCTGTGGTCCCAGCTGCTCGGGACGCTGAGGTGGGAGGATTGCTTGAGCCCAGGAGTTTGACACCAGCATGGGCAATATGGTAAGACCC’.
- Rseq6=‘CCCAGGAGTTTGACACCAGCATGGGCAA’.

Table 3
Positions of the two new discovered repeat sequences Rseq5 and Rseq 6 in the X and Y chromosomes of the human genome.

Start	End	Size (bp)	sequence
321,472	321,602	130	
321,612	321,742	130	“Rseq5”
321,752	321,882	130	
321,267	321,295	28	
321,419	321,447	28	
321,561	321,589	28	
321,701	321,729	28	“Rseq6”
321,841	321,869	28	
322,148	322,176	28	
323,095	323,123	28	

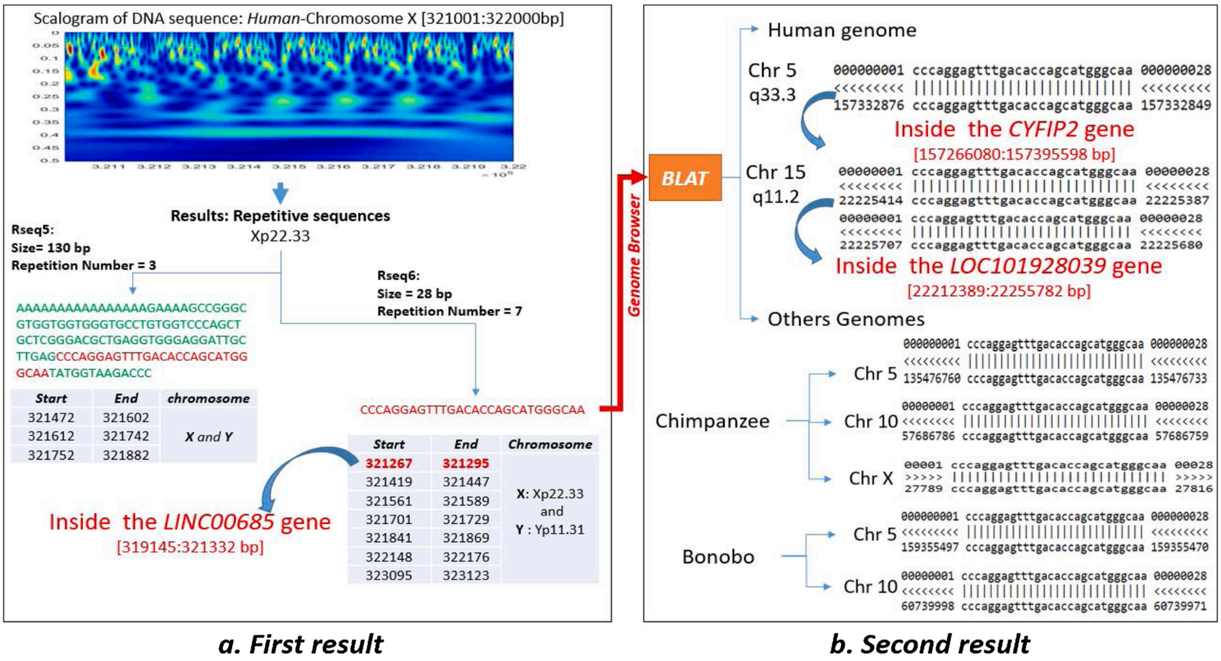

Scalogram of DNA sequence: *Human*-Chromosome X [321001:322000bp]
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
3.211
3.212
3.213
3.214
3.215
3.216
3.217
3.218
3.219
3.22
×10⁵
Results: Repetitive sequences
Xp22.33
Rseq5:
Size= 130 bp
Repetition Number = 3
AAAAAAAAAAAAAAAAGAAAAGCCGGGC
GTGGTGGTGGGTGCCTGTGGTCCCAGCT
GCTCGGGACGCTGAGGTGGGAGGATTGC
TTGAGCCCAGGAGTTTGACACCAGCATGG
GCAATATGGTAAGACCC
Start End chromosome
321472 321602
321612 321742 X and Y
321752 321882
Inside the LINC00685 gene
[319145:321332 bp]
Rseq6:
Size = 28 bp
Repetition Number = 7
CCCAGGAGTTTGACACCAGCATGGGCAA
Start End Chromosome
321267 321295
321419 321447
321561 321589 X: Xp22.33
321701 321729 and
321841 321869 Y : Yp11.31
322148 322176
323095 323123
Genome Browser
BLAT
Human genome
Chr 5
q33.3
000000001 cccaggagtttgacaccagcatgggcaa 000000028
<<<<<<<<< |||||||||||||||||||||||||||| <<<<<<<<<
157332876 cccaggagtttgacaccagcatgggcaa 157332849
Inside the CYFIP2 gene
[157266080:157395598 bp]
Chr 15
q11.2
00000001 cccaggagtttgacaccagcatgggcaa 00000028
<<<<<<<< |||||||||||||||||||||||||||| <<<<<<<<
22225414 cccaggagtttgacaccagcatgggcaa 22225387
00000001 cccaggagtttgacaccagcatgggcaa 00000028
<<<<<<<< |||||||||||||||||||||||||||| <<<<<<<<
22225707 cccaggagtttgacaccagcatgggcaa 22225680
Inside the LOC101928039 gene
[22212389:22255782 bp]
Others Genomes
Chimpanzee
Chr 5
000000001 cccaggagtttgacaccagcatgggcaa 000000028
<<<<<<<<< |||||||||||||||||||||||||||| <<<<<<<<<
135476760 cccaggagtttgacaccagcatgggcaa 135476733
Chr 10
00000001 cccaggagtttgacaccagcatgggcaa 00000028
<<<<<<<< |||||||||||||||||||||||||||| <<<<<<<<
57686786 cccaggagtttgacaccagcatgggcaa 57686759
Chr X
00001 cccaggagtttgacaccagcatgggcaa 00028
>>>>> |||||||||||||||||||||||||||| >>>>>
27789 cccaggagtttgacaccagcatgggcaa 27816
Bonobo
Chr 5
000000001 cccaggagtttgacaccagcatgggcaa 000000028
<<<<<<<<< |||||||||||||||||||||||||||| <<<<<<<<<
159355497 cccaggagtttgacaccagcatgggcaa 159355470
Chr 10
00000001 cccaggagtttgacaccagcatgggcaa 00000028
<<<<<<<< |||||||||||||||||||||||||||| <<<<<<<<
60739998 cccaggagtttgacaccagcatgggcaa 60739971
a. First result
b. Second result

Fig. 8. Repetitive DNA sequence detection in X chromosome. (a): The 2-D representation provides a visual way to see three characterized long repetitive sequences; (b): Location of these repetitive sequences in other regions.

Table 4
Position corresponding to the new discovered scattered repeat sequence Rseq6 (28bp) in different chromosomes in the human genome.

Start	end	Chromosome	Gene	location in gene	location in genome
321,267	321,295	X and Y	LINC00685	intron 1/1	Xp22.33 and Yp11.31
157,332,849	157,332,876	5	CYFIP2	intron 21/31	5q33.3
22,225,387	22,225,414	15	LOC101928039	uncharacterized	15q11.2
22,225,680	22,225,707				
237,615,395	237,615,421	1	RYR2	intron 37/104	1q43

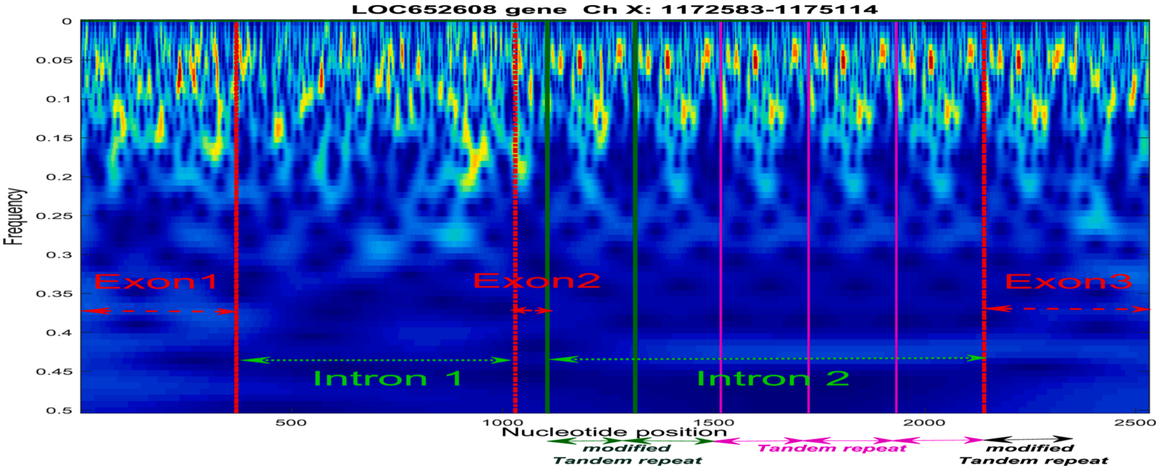


Fig. 9. LOC652608 Gene in the X chromosome contains a tandem repeat sequence: Rseq7 started in Intronic region (Intron 2) until Exonic region (Exon 3).

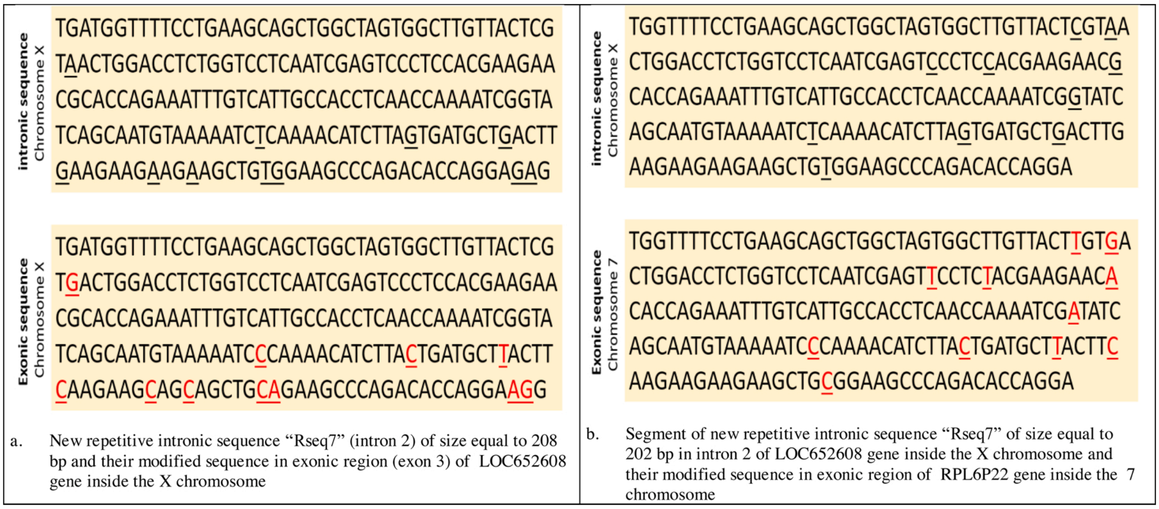
After the localization of these two repetitive DNA sequences (Rseq5 and Rseq6), we have chosen to use the BLAT alignment tool in order to see if these sequences have other locations in the other human chromosomes or in other genomes. Indeed, the repetitive sequences that migrate to different regions of the genome have a great importance and they have been classified as conservative mobile DNA sequences. Their importance will be higher if these conservative regions are localized in genes.

As a result, we have found the Rep2 sequence at the position 321,267 bp to 321447 bp in the intronic region of a non-protein coding RNA 685 (*LINC00685*) gene, and thus in both X and Y chromosomes [52].

In the sub-figure b of Fig. 8 (second result), we show that the new repetitive sequence Rep2 is located, not only within other chromosomes (1, 5, 15, X and Y) of the human genome, but also in other genomes like *chimpanzee* and *bonobo*. Results shown in Table 4 prove that Rep2 has been located in intronic region of different chromosomes of the human genome: 1, 5, 15, X and Y.

In fact, the sequence "Rseq6" presents a special intronic conservative region located, not only in different chromosomes but also in different genomes. Rseq6 sequence that have a size of 29 bp has been localized in two genes corresponding to *chimpanzee* genome. It is located at the position: 135476733–135476760 in the *DDX46* gene ([135444165:135519361 bp]) of the chromosome 5. It is also localized at the position: 86759–57686796 bp positions in the *FAM13C* gene ([57587233:57707637bp]) of the chromosome 10.

In addition, we present another example of a special new repetitive sequence "Rseq7" which has been found using our approach. The Fig. 9 shows the time-frequency representation of the *LOC*652,608 gene which has a size of 2532 bp. The gene is found at the position: 1172583–1175114 bp in the X chromosome of the human genome. This


intronic sequence
Chromosome X
Exonic sequence
Chromosome X
intronic sequence
Chromosome X
Exonic sequence
Chromosome 7

TGATGGTTTTCCTGAAGCAGCTGGCTAGTGGCTTGTTACTCG
TAACTGGACCTCTGGTCCTCAATCGAGTCCCTCCACGAAGAA
CGCACCAGAAATTTGTCATTGCCACCTCAACCAAAATCGGTA
TCAGCAATGTAAAAATCTCAAAACATCTTAGTGATGCTGACTT
GAAGAAGAAGAAGCTGTGGAAGCCCAGACACCAGGAGAG

TGATGGTTTTCCTGAAGCAGCTGGCTAGTGGCTTGTTACTCG
TGACTGGACCTCTGGTCCTCAATCGAGTCCCTCCACGAAGAA
CGCACCAGAAATTTGTCATTGCCACCTCAACCAAAATCGGTA
TCAGCAATGTAAAAATCCCAAAACATCTTACTGATGCTTACTT
CAAGAAGCAGCAGCTGCAGAAGCCCAGACACCAGGAAGG

a. New repetitive intronic sequence "Rseq7" (intron 2) of size equal to 208 bp and their modified sequence in exonic region (exon 3) of LOC652608 gene inside the X chromosome

TGGTTTTCCTGAAGCAGCTGGCTAGTGGCTTGTTACTCGTAA
CTGGACCTCTGGTCCTCAATCGAGTCCCTCCACGAAGAACG
CACCAGAAATTTGTCATTGCCACCTCAACCAAAATCGGTATC
AGCAATGTAAAAATCTCAAAACATCTTAGTGATGCTGACTTG
AAGAAGAAGAAGCTGTGGAAGCCCAGACACCAGGA

TGGTTTTCCTGAAGCAGCTGGCTAGTGGCTTGTTACTTGTGA
CTGGACCTCTGGTCCTCAATCGAGTTCCTCTACGAAGAACA
CACCAGAAATTTGTCATTGCCACCTCAACCAAAATCGATATC
AGCAATGTAAAAATCCCAAAACATCTTACTGATGCTTACTTC
AAGAAGAAGAAGCTGCGGAAGCCCAGACACCAGGA

b. Segment of new repetitive intronic sequence "Rseq7" of size equal to 202 bp in intron 2 of LOC652608 gene inside the X chromosome and their modified sequence in exonic region of RPL6P22 gene inside the 7 chromosome

Fig. 10. Two examples of conserved intronic repetitive sequences (satellites) and noncoding sequence located in coding region such as senescence [53].

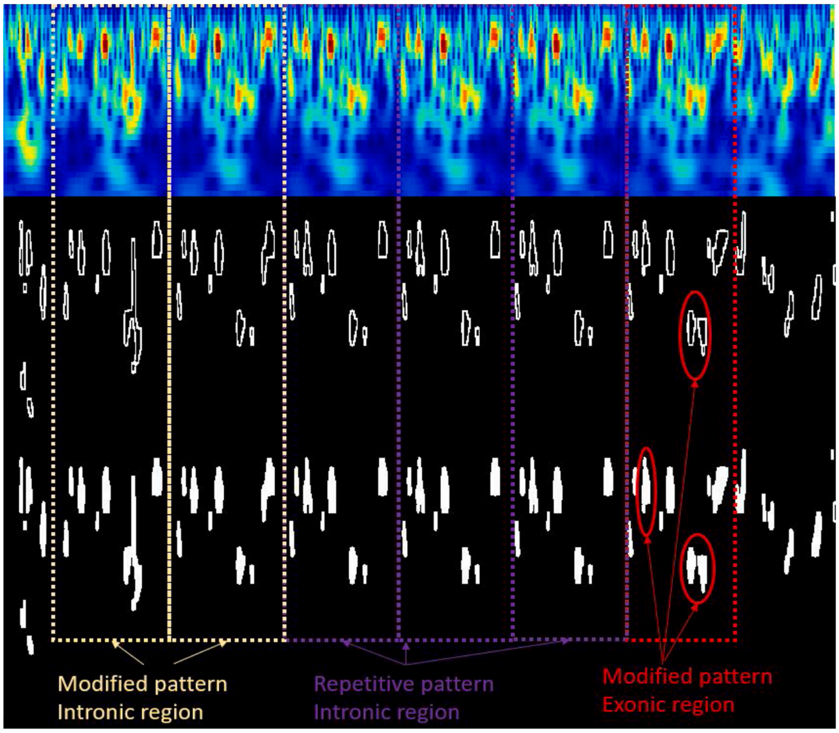


Fig. 11. Example of DNA image segmentation by which we can obtain the begining and the end of the repetitive patterns located in intronic region (Intron 2), and the corresponding modified sequences (especially in exonic region) with the modification region.

pseudo-gene is a 60S ribosomal protein L6-like. The DNA image shown in Fig. 9 demonstrates three exonic regions and two intronic regions.

We can clearly see that the second intronic region is composed by a specific tandemic sequence which we called "Rseq7". The correspondent modified version has the same size as "Rseq7" which is equal to 208 bp.

This particular repetitive sequence starts in the intronic zone: Intron2 until reaching and exceeding the exonic zone: Exon3; with a modification of 11 nucleotides.

Intron 2 is a noncoding sequence (208 bp) which is composed of multiple repetitions of "Rseq7".

Rseq7='TGATGGTTTTCCTGAAGCAGCTGGCTAGTGGCTTGT-TACTCGTAACTGGACCTCTGGTCCTCAATCGAGTCCCTCCACGAA-GAACGCACCA-GAAATTTGTCATTGCCACCTCAACCAAAATCGGTATCAGCAATG-TAAAAATCTCAAAACATCTTAGTGATGCTGACTTGAAGAAGAA-GAAGCTGTGGAAGCCCAGACACCAGGAGAG'.

Then, we searched this new tandem repeat "Rseq7" in the other chromosomes. As a result, we found that this sequence exists in 7 chromosomes with some nucleotides modifications. Moreover, we have located this modified intronic sequence in genes regions of other chromosomes of the human genome.

Fig. 10 shows two reference sequences and the modified version. The first exonic sequence example corresponds to the *LOC*652608 gene in located in the X chromosome (Fig. 10a). The second exonic sequence corresponds to the *RPL6P*22 gene in which is located in the chromosome 7 (Fig. 10b).

For these two examples the nucleotides variation number between the intronic sequence "Rseq7" and the exonic sequence is equal to 11 base pairs but with different locations.

On the other hand, we have chosen to use image processing techniques to extract the repetitive sequences. The idea consists in segmenting the scalogram image in order to extract the repetitive patterns. For this purpose, we developed a new segmentation algorithm applied to the DNA scalograms. Fig. 11 illustrates the obtained results by our segmentation algorithm with a thresholding value equal to 26. It shows the location of the "Rseq7" repetitive sequences and the correspondent

Table 5
Location of repetitive intronic satellites sequence "Rseq7" and the corresponding exonic modified sequences in different chromosomes of the human genome.

Start	end	Chromosome	Gene	location in gene	location in genome	sequence	Description
1,174,105	1,174,312	X and Y	LOC652608	Intron 2	Xp22.33 and Yp11.2	Rseq7(208 bp)	60S *ribosomal protein* L6-like
1,174,313	1,174,520						
1,174,521	1,174,728						
1,174,729	1,174,936			Exon 3		modified Rseq7	
45,781,761	45,781,958	1	RPL6P1	Exon 1	1p34.1	modified Rseq7	*ribosomal protein* L6 pseudogene 1
65,573,946	65,574,144	4	EPHA5	Exon 1	4q13.1-q13.2	modified Rseq7	EPH receptor A5
			RPL6P10	Exon 1	4q13.2	modified Rseq7	*ribosomal protein* L6 pseudogene 10
137,722,471	137,722,672	7	DGKI	Intron 2	7q33	modified Rseq7	OTTHUMP00000208597
14,070,714	14,070,911		RPL6P21	Exon 2	7p21.3	modified Rseq7	*ribosomal protein* L6 pseudogene 21
64,141,719	64,141,920		AC091685.2	Exon 2	7q11.21	modified Rseq7	*ribosomal protein* L6 pseudogene 11
33,859,542	33,859,740	8	LOC105379364	uncharacterized	8p12	modified Rseq7	uncharacterized LOC105379364
			RPL6P22	Exon 1	8p12	modified Rseq7	*ribosomal protein* L6 pseudogene 22
83,151,812	83,152,006	12	RPL6P25	Exon 1	12q21.31	modified Rseq7	*ribosomal protein* L6 pseudogene 25
112,405,884	112,406,338		RPL6	Exon 6	12q24.13	modified Rseq7	*ribosomal protein* L6
6,462,328	6,462,526	18	RPL6P27	Exon 1	18p11.31	modified Rseq7	*ribosomal protein* L6 pseudogene 27

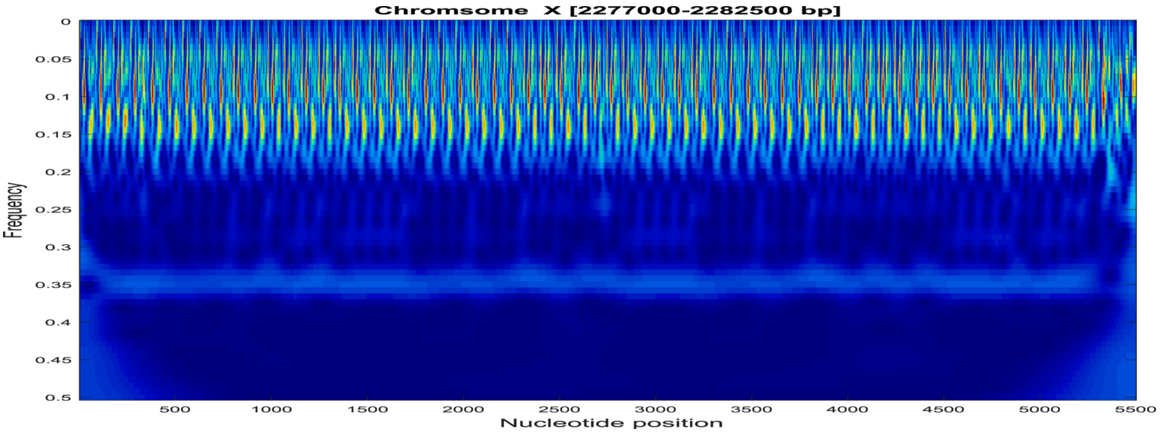


Fig. 12. Scalogram corresponding to a DNA sequence in X chromosome that contains repetitive sequences in intronic region.

modified versions. Here, we can see in the first subfigure (scalogram) that the repetitive pattern is located at: 1173583bp-1175114 bp in the X chromosome of the human genome. The second subfigure presents the segmented image. The repetitive patterns correspond to the repetitive sequences which start in intronic sequences and end in exonic region with some nucleotides modification (11 nucleotides) in the beginning and in the end (Fig. 11).

After the repetitive sequences localization, we checked if these sequences are located in other regions in the human genome and even in the genomes of other species. Table 5 shows the location of the repetitive sequence “Rseq7” and its modified repetitive sequences in different gene regions of different chromosomes in the human genome. We can note that this new repetitive sequence characterizes a ribosomal protein (RPs) region in the human genome. The ribosomal RNA gene repeat (rDNA) is the largest repetitive region in the eukaryotic genome. The genome stability depends on the stability of the rDNA, the latter affects cellular functions

The next example in Fig. 12 shows highly repetitive patterns in the X chromosome at position: 2277000–2282500 bp (Xp22.33 region) in the human genome. This region contains tandem repeat sequences and interspersed repeat sequences. In addition, the localization results have shown that these specific patterns are localized in the intronic region of the *DHRSX* gene ([2,219,506 bp: 2,500,974 bp]) in the X chromosome and even in other genes located in other chromosomes.

DHRSX gene is a new gene discovered in 2014 at the Xp22.33 and Yp11.2 in the human genome. It has been shown that the protein encoded by this gene is implicated in the positive regulation of starvation induced autophagy [54].

The scalogram represented in Fig. 12 indicates the presence of repetitive patterns in intronic regions. The reference sequence corresponding to tandem repeat sequence “Rseq8” has a size equal to 89bp and 14 as a repetition number. Other repetitive sequences are localized in these intronic regions which are:

- “Rseq9” with a size of 42 bp and 26 as repetition number
- “Rseq10” with a size of 19 bp and 63 as repetition number
- “Rseq11” with a size of 6 bp and 123 as repetition number.

All these repetitive sequences are minisatellite type. In the NCBI database, these regions are defined as a low complexity G-rich repetition and there is no further given information.

- Rseq8="AGGGAGAGAGAGGGAGGGCAAACGAGAGGGAGAGAGAA-GGAGGAGGAGGAAATGGGGGAAAGAGAGAGAAAGAGAGATGGA-GAGGGAAC"
- Rseq9="AGAGAGATGGAGAGGGAACAGGGAGAGAGAGAGGGAGGGC-AAAC"

Table 6
Location of the intronic repetitive sequence “Rseq8” in the X and Y chromosomes of the human genome.

Start	End	Repetition types	Gene	Location in gene	location in genome
2,277,547	2,277,635				
2,277,636	2,277,724	Tandem			
2,277,725	2,277,813				
2,277,903	2,277,991	Dispersed			
2,278,791	2,278,879				
2,278,880	2,278,968	Tandem			
2,278,969	2,279,057		DHRSX	Intron 4/6	Xp22.33 and Yp11.2
2,279,147	2,279,235	Tandem			
2,279,236	2,279,324				
2,280,290	2,280,378				
2,280,379	2,280,467	Tandem			
2,280,468	2,280,556				
2,280,646	2,280,734	Tandem			
2,280,735	2,280,823				

The Table7 provides the locations of “Rseq9” in the X chromosome of other genomes.

Table 7
Position of “Rseq9” in X chromosome of other genomes.

Start	End	chromosome	genome
1,833,384	1,833,425	X	*Gorilla*
1,927,705	1,927,746		
1,927,883	1,927,924		
1,928,326	1,928,367		
1,928,414	1,928,455	X	*Chimpanzee*
1,928,503	1,928,544		
1,928,592	1,928,633		
1,928,771	1,928,812		
2,220,841	2,220,882	X	*Bonobo*
1,800,783	1,800,824	X	*Rhesus*

- Rseq10="AGAGAGATGGAGAGGGAAC"
- Rseq11= "AGAGAGAA"

These repetitive sequences are also located at the same position in intronic region within the *DHRSX* gene in the Y chromosome of the human genome.

Table 6 details the location of the new repetitive sequence “Rseq8” inside the X and Y chromosomes.

Furthermore, this repetitive sequence is located inside the intronic region of the *DHRSX* gene with tandem repeat and dispersed repeat forms.

Fig. 13 shows an example of another repeat tandem pattern found in the X chromosome at position 27210460−27211308bp in the human

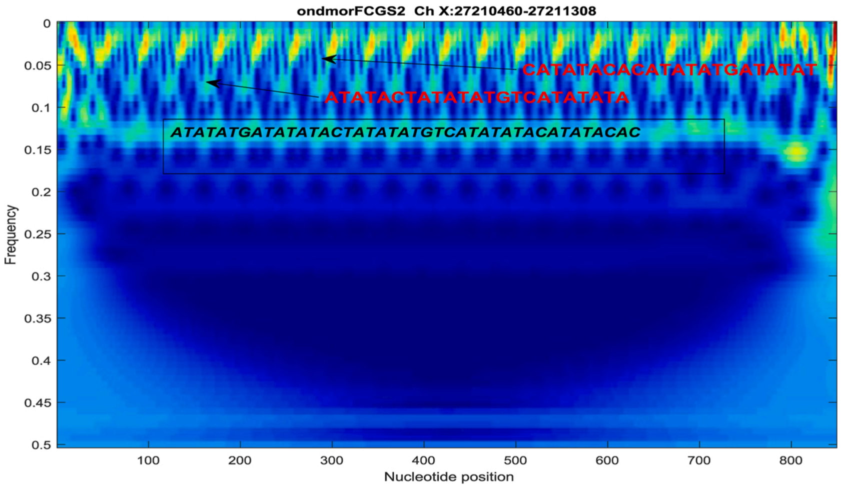
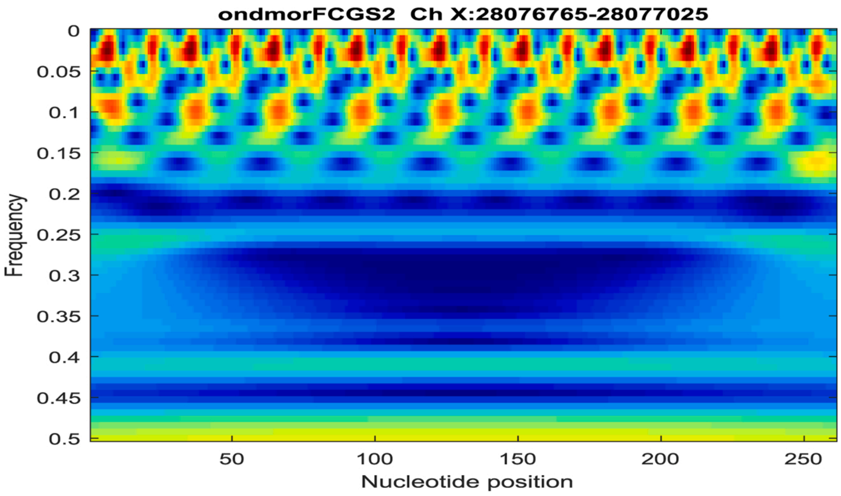


Fig. 13. Scalogram representation of a new discovered tandem repeat sequence “Rseq12”:(ATATATGATATATACTATATATGTCATATATACATATACAC)$_n$.



CTGTATAACCTAAATAATATAGGTTATAT
CTGTATAACCTAAATAATATAGGTTATAT
CTGTATAACCTAAATAATATAGGTTATAT
CTGTATAACCTAAATAATATAGGTTATAT
CTGTATAACCTAAATAATATAGGTTATAT
CTGTATAACCTAAATAATATAGGTTATAT
CTGTATAACCTAAATAATATAGGTTATAT
CTGTATAACCTAAATAATATAGGTTATAT
CTGTATAACCTAAATAATATAGGTTATAT

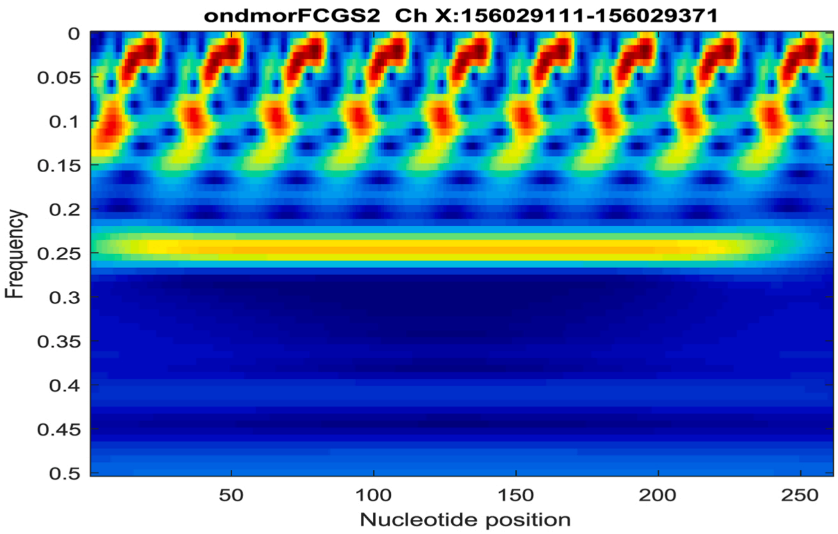
Fig. 14. Scalogram image corresponding to DNA sequence “TRseq1” (with size equal to 261) containing the tandem repeat sequence “Rseq13” with a repetition number equal to 9.

genome. The annotation of this sequence in the NCBI database indicates the presence of the simple repeats classes (TA)n ([27,210,497 : 27,210,679]), (CATATA)n ([27,210,682 : 27,210,757]) and (TA)n ([27,210,758 :27211323]). These confirmed repetitive sequences have been also located with our approach. In addition we discovered the new repetitive sequences:‘ATATATGATATATACTATATATGTCATATATA-CATATACAC’, ‘ATATATGATATATAC’, ‘TGATAT’, ‘TACATA’ and ‘GATATA’ These sequences have been localized inside the LOC105373150 gene ([27153368:27399005]) within the Xp21.3 region of the human genome.

- Rseq12=“ATATATGATATATACTATATATGTCATATATACATATACAC”

The short repetitive sequence "TACATA" (6 bp) appears 22 times in this DNA sequence and has 69,710 as a repetition number in the X chromosome.

After searching for the existence of this tandem repeat sequence



TCTCTGCGCCTGCGCCGGCGCGGCGCGCC
TCTCTGCGCCTGCGCCGGCGCGGCGCGCC
TCTCTGCGCCTGCGCCGGCGCGGCGCGCC
TCTCTGCGCCTGCGCCGGCGCGGCGCGCC
TCTCTGCGCCTGCGCCGGCGCGGCGCGCC
TCTCTGCGCCTGCGCCGGCGCGGCGCGCC
TCTCTGCGCCTGCGCCGGCGCGGCGCGCC
TCTCTGCGCCTGCGCCGGCGCGGCGCGCC
TCTCTGCGCCTGCGCCGGCGCGGCGCGCC

Fig. 15. Scalogram image corresponding to DNA sequence “TRseq2” confirm the existing of the “Rseq14” tandem repeat sequence $(\text{TCTCTGCGCCTGCGCCGGCGCGGCGCGCC})_n$ annotated in [45] as a minisatellites sequence which their repetition number equal to 9.

“Rseq12” in other locations, we have searched it in the X chromosome of other genomes: *Bonobo* genome [27,166,538 bp: 27,166,578 bp]; *Chimpanzee* genome [27158028:27158068 bp].

Using our algorithm, we have successfully found 9 repetitions of another new short repetitive sequence as a tandem repeat sequence (TRs). We called this sequence of 29 base pairs “Rseq13”.

- Rseq13=“CTGTATAACCTAAATAATATAGGTTATAT”

Fig. 14 shows the scalogram of a new repetitive DNA sequence that we called “Rseq13”. The sequence has a size of 261 bp and it is localized at 28076765–28077025 bp in the X chromosome. It is a tandem repeat sequence, with patterns of 29 bp length: “Rseq13”. The NCBI and the Dfam databases don’t indicate the existence of such repetitive sequence (“Rseq13”). With our approach we succeeded to detect this tandem repeat without any prior knowledge about its existence.

The repetitive sequence “Rseq13” is located not only in the X chromosome of human genome but also in other genomes like in the X chromosomes of *Bonobo* (at [28,032,917 bp-28,033,158 bp]), *Chimpanzee* ([28,028,604 bp-28,028,816 bp]) and *Gorilla* ([28,333,971 bp-28,334,231 bp]) with two nucleotides modification.

Fig. 15 shows the scalogram of a new DNA sequence “TRseq2” with a size of 261 bp. The sequence is positioned at 156029111–156029371 bp in the X chromosome. As we can see, the scalogram contains a repetitive pattern corresponding to a tandem repeat sequence: “Rseq14”. This subsequence ("TCTCTGCGCCTGCGCCGGCGCGGCGCGCC") has a size of 29 bp and 9 as a repetition number.

Rseq14 is not annotated as a tandem repeat in the NCBI or the Dfam databases but it is defined as a TAR1of the telomeric satellite family [55].

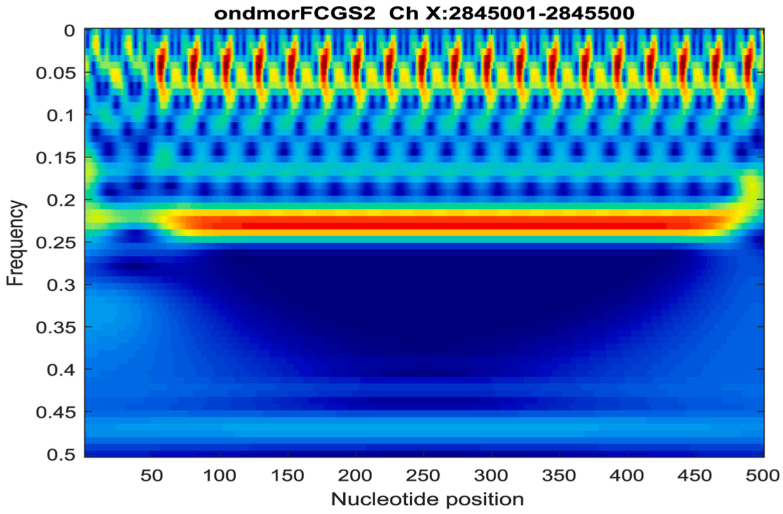
Table 8
Repetitive sequence location corresponding to “Rseq14” in the X human chromosome and in other genomes.

Start	End	Repetition number	Repetition type	chromosome
12,491	12,751	9		5
12,520	12,780	9		5
57,215,631	57,215,891	9		Y
156,029,111	156,029,371	9		X
10,629	10,950	2	TAR1 : Satellitetelomeric	1
181,167	181,311	2		
10,754	11,079	2		12
10,601	10,629	1		16
101,980,093	101,980,000	1		15
135,076,184	135,076,000	1		11

In Table 8, we provide the localization results of “Rseq14” in the whole human genome and in other genomes.

Fig. 16 shows the scalogram of another new DNA sequence: “TRseq3” with a size of 500 bp and extending from 2845001bp to 2845500bp in the X chromosome of human genome. The sequence contains a tandem repeat sequence: “Rseq15” (CGTGTGTATGTATATTTATATACA), which size is a 24 bp and its repetition number is equal to 18. This sequence is not annotated as a tandem repeat sequence in the NCBI database nor in the Dfam database.

Our "New-repeat-Data" database of all new discovered repetitive



ATATTTATATACACATGTGTATGTATATTTATATACACATGTGTATG
TATATATACACGTGTGTATGTATATTTATATACACGTGTGTATGTAT
ATTTATATACACGTGTGTATGTATATTTATATACACGTGTGTATGTA
TATTTATATACACGTGTGTATGTATATTTATATACACGTGTGTATGT
ATATTTATATACACGTGTGTATGTATATTTATATACACGTGTGTATG
TATATTTATATACACGTGTGTATGTATATTTATATACACGTGTGTAT
GTATATTTATATACACGTGTGTATGTATATTTATATACACGTGTGTA
TGTATATTTATATACACGTGTGTATGTATATTTATATACACGTGTGT
ATGTATATTTATATACACGTGTGTATGTATATTTATATACACGTGTG
TATGTATATTTATATACACGTGTGTATGTATATTTATATACACGTGT
GTATGTATATTTATATACACGTGTGTATGT

Fig. 16. Scalogram image corresponding to the DNA sequence "TRseq3" that contains "Rseq15" as tandem repeat motif.

Table 9
Description of the input data to the CNN classification system.

CLASS	Repetitive pattern with size X	NUMBER
Rep1	X>100	180
Rep2	60<X<100	150
Rep3	30<X<60	200
Rep4	X<30	250
NonRep	NONE	200

sequences are presented in "Supplementary Material" file. To conclude, we succeeded to implement an efficient algorithm for repetitive sequences detection. The sequences we detected are of two types: satellites and minisatellites. On the other hand, we have obtained better results than those of the bioinformatics tools. The main advantage presented by this work is being independent of any prior knowledge about the searched repeat.

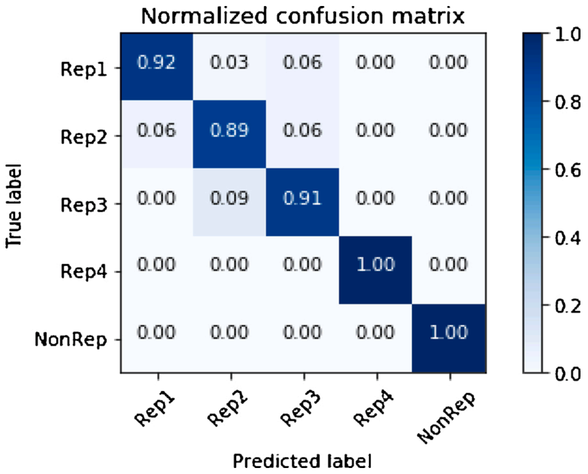


Fig. 17. Confusion Matrix result obtained by our classification system.

3.2. CNN classification results

In this section, we present the results of using CNN model to classify DNA scalograms obtained in the first part of this work. Our goal is to identify the different classes of the new repetitive sequences we discovered and stocked in the "New-repeat-Data" database. As a data, we randomly took 200 non-repetitive sequences (NonRep) and 780 repetitive sequences (Rep). Repetitive sequences data consists of 780 sequences divided into 4 classes depending on their repetitive pattern length (Table 9). These classes are: Rep1 (with a size >100), Rep2 (with a size between 60 and 100), Rep3 (with a size between 30 and 60) and Rep4 (with a size <30). In globally, our constructed database contains five classes that four contain scalograms of repetitive sequences and one contains scalograms without repetitive sequences. For the classification purpose, all the dataset (980 scalogram images) was splitted into 80% for training (784 images) and 20% for testing (196 images). Thus, by such classification system we can discover images that contain similar repetitive patterns. We can also differentiate these images from others that don't contain repetitions.

The Fig. 17 represents the classification results of the four repetitive DNA classes (images with repetitive patterns) against one class of non-repetitive DNA (images with no repetitive patterns).

Table 10
Evaluation measurements of our classification system.

Class	Precision	Recall	F1-score
Rep1	94	92	93
Rep2	89	89	89
Rep3	88	91	90
Rep4	100	100	100
NonRep	100	100	100
avg/ total	94.2	94.4	94.4

With the CNN model, we distinguished different specific types of DNA images. The score ranges from 89% to 100%. The obtained results yield an average score of 94.4%.

The confusion matrix of the classification rates confirms that our system is efficient in distinguishing between small repetitive patterns (Rep4) and non-repetitive DNA sequences (NonRep) with score equal to 100%. This result is quite clear, since the scalogram images of these two classes are very different.

The following Table 10 contains three evaluation measurements: precision, recall and F1-score which we used to evaluate our classification system.

Overall, our system gives good results in recognizing the four new repetitive DNA sequences with an average of 95% in precision, recall and F1-score.

4. Conclusion

Genetic knowledge improvement of the human genome is a complex and a continuous research process. To contribute to this process, bioinformatics and signal and images processing tools have been applied to reveal hidden spectral features of DNA sequences. Although the repetitive DNA sequences occupy 40% of the *Human* genome, the localization of these sequences remains insufficient as it is a very difficult task.

In this paper, we proposed a new algorithm based on the signal and image processing tools to extract the repetitive patterns from DNA images that correspond to the repetitive DNA sequences. The main goal of this is to create a new database that contains locations of all the new discovered repetitive sequences. As an example of the obtained results, we found a new modified repetitive sequence that can characterize 60S ribosomal protein: “Rseq7”. Therefore, deeper studies that may give a biological interpretation of these results will be welcome.

In this article, we proposed a novel and highly-effective method for DNA images prediction based on CNN model. In our prediction system, the obtained accuracy scores over 100 fold cross validation ranged from 89% to 100% with an overall score of 94.4%.

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Declaration of Competing Interest

The authors declare that there are no conflict of interest exists and no competing interests regarding the publication of this paper.

CRediT authorship contribution statement

Rabeb Touati: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing. **Asma Tajouri:** Writing - review & editing, Validation. **Imen Mesaoudi:** Writing - review & editing. **Afef Elloumi Oueslati:** Validation, Formal analysis. **Zied Lachiri:** Validation. **Maher Kharrat:** Conceptualization, Supervision.

Acknowledgment

This study was founded by the Ministry of Higher Education and Research, LR99ES10 Human Genetics Laboratory.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at https://doi.org/10.1016/j.bspc.2020.102207.

References

[1] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, J. D. Gocayne, The sequence of the human genome, Science 291 (5507) (2001) 1304–1351.
[2] N.L. de Freitas, A.B. Al-Rikabi, L.A.C. Bertollo, T. Ezaz, C.F. Yano, E.A. de Oliveira, M. de Bello Cioffi, Early stages of XY sex chromosomes differentiation in the fish Hoplias malabaricus (Characiformes, Erythrinidae) revealed by DNA repeats accumulation, Curr. Genomics 19 (3) (2018) 216–226.
[3] C. Ramel, Mini-and microsatellites, Environ. Health Perspect. 105 (suppl 4) (1997) 781–789.
[4] M.A. Biscotti, E. Olmo, J.P. Heslop-Harrison, Repetitive DNA in Eukaryotic Genomes, 2015.
[5] T.J. Treangen, S.L. Salzberg, Repetitive DNA and next-generation sequencing: computational challenges and solutions, Nat. Rev. Genet. 13 (1) (2012) 36.
[6] E.W. Jabs, M.G. Persico, Characterization of human centromeric regions of specific chromosomes by means of alphoid DNA sequences, Am. J. Hum. Genet. 41 (1987) 374–390.
[7] E.H. Blackburn, J.G. Gall, A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in tetrahymena, J. Mol. Biol. 120 (1978) 33–53.
[8] J.A. Stewart, M.F. Chaiken, F. Wang, C.M. Price, Maintaining the end: roles of telomere proteins in end-protection, telomere replication and length regulation, Mutat. Res. Mol. Mech. Mutagen. 730 (1-2) (2012) 12–19.
[9] R.K. Moyzis, J.M. Buckingham, L.S. Cram, M. Dani, L.L. Deaven, M.D. Jones, J. R. Wu, A highly conserved repetitive DNA sequence,(TTAGGG) n, present at the telomeres of human chromosomes, Proc. Natl. Acad. Sci. 85 (18) (1988) 6622–6626.
[10] V.A. Zakian, Structure and function of telomeres, Ann Rev Genet 23 (1989) 579–604.
[11] J.C. Peng, G.H. Karpen, Epigenetic regulation of heterochromatic DNA stability, Curr. Opin. Genet. Dev. 18 (2) (2008) 204–211.
[12] Kian Guan Lim, Chee Keong Kwoh, Li Yang Hsu, et al., Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance, Brief. Bioinformatics 14 (1) (2012) 67–81.
[13] Teresa Thiel, W. Michalek, R. Varshney, et al., Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.), Theor. Appl. Genet. 106 (3) (2003) 411–422.
[14] Roman Kolpakov, Bana Ghizlane, Kucherov, et al., Gregory. “mreps: efficient and flexible detection of tandem repeats in DNA, Nucleic Acids Res. 31 (13) (2003) 3672–3678.
[15] Chris Abajian, Sputnik - DNA Microsatellite Repeat Search Utility, 1994.
[16] Sarachu, Martín et Colet, Marc, wEMBOSS: a web interface for EMBOSS, Bioinformatics 21 (4) (2004) 540–541.
[17] Gary Benson, et al., Tandem repeats finder: a program to analyze DNA sequences, Nucleic Acids Res. 27 (no 2) (1999) 573–580.
[18] Tarailo-Graovac, Maja et Chen, Nansheng, Using RepeatMasker to identify repetitive elements in genomic sequences, Curr. Protoc. Bioinformatics 25 (no 1) (2009) 14, p. 4.10. 1-4.10.
[19] P. Flicek, E. Birney, Sense from sequence reads: methods for alignment and assembly, Nat. Methods 6 (11s) (2009) S6.
[20] A.J. de Koning, W. Gu, T.A. Castoe, M.A. Batzer, D.D. Pollock, Repetitive elements may comprise over two-thirds of the human genome, PLoS Genet. 7 (12) (2011), e1002384.
[21] The NCBI, GenBank Database, 2019. Available: http://www.ncbi.nlm.nih.gov/Genbank/ (Accessed 1 September 2019).
[22] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, J. D. Gocayne, The sequence of the human genome, Science 291 (5507) (2001) 1304–1351.
[23] R. Touati, S. Haddad-Boubaker, I. Ferchichi, I. Messaoudi, A.E. Ouesleti, H. Triki, M. Kharrat, Comparative genomic signature representations of the emerging COVID-19 coronavirus and other coronaviruses: high identity and possible recombination between Bat and Pangolin coronaviruses, Genomics 112 (6) (2020) 4189–4202.
[24] R. Touati, A.E. Oueslati, I. Messaoudi, Z. Lachiri, The Helitron family classification using SVM based on Fourier transform features applied on an unbalanced dataset, Med. Biol. Eng. Comput. 57 (10) (2019) 2289–2304.
[25] M. Buchner, S. Janjarasjitt, Detection and visualization of tandem repeats in DNA sequences, Ieee Trans. Signal Process. 51 (9) (2003) 2280–2287.
[26] S.D. Sharma, S.N. Sharma, R. Saxena, Identification of short exons disunited by a short intron in eukaryotic DNA regions, IEEEACM Trans. Comput. Biol. Bioinform. (2019).
[27] V.R. Chechetkin, A.Y. Turygin, Search of hidden periodicities in DNA sequences, J. Theor. Biol. 175 (4) (1995) 477–494.

[28] D. Sharma, B. Issac, G.P.S. Raghava, R. Ramaswamy, Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation, Bioinformatics 20 (9) (2004) 1405–1412.

[29] R. Touati, I. Messaoudi, A.E. Oueslati, Z. Lachiri, 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) (2018).

[30] R. Touati, I. Messaoudi, A.E. Oueslati, Z. Lachiri, A combined support vector machine-FCGS classification based on the wavelet transform for Helitrons recognition in C. elegans, Multimed. Tools Appl. 78 (10) (2019) 13047–13066.

[31] R. Touati, I. Messaoudi, A.E. Oueslati, Z. Lachiri, Distinguishing between intra-genomic helitron families using time-frequency features and random forest approaches, Biomed. Signal Process. Control 54 (2019), 101579.

[32] A. Grossmann, J. Morlet, Decomposition of Hardy functions into square integrable wavelets of constant shape, Siam J. Math. Anal. 15 (4) (1984) 723–736.

[33] R.J.E. Merry, Wavelet theory and applications: a literature study, DCT rapporten 2005 (2005).

[34] A.H. Najmi, J. Sadowsky, The continuous wavelet transform and variable resolution time-frequency analysis, Johns Hopkins APL Tech. Dig. 18 (1) (1997) 134–140.

[35] M. Kumar, R. Saxena, Algorithm and technique on various edge detection: a survey, Signal & Image Processing 4 (3) (2013) 65.

[36] P. Sahni, N. Mittal, Breast cancer detection using image processing techniques. Advances in Interdisciplinary Engineering, Springer, Singapore, 2019, pp. 813–823.

[37] J. Canny, A computational approach to edge detection, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1986) 679–698.

[38] P. Bao, L. Zhang, X. Wu, Canny edge detection enhancement by scale multiplication, IEEE Trans. Pattern Anal. Mach. Intell. 27 (9) (2005) 1485–1490.

[39] P. Soille, Morphological Image Analysis: Principles and Applications, Springer Science & Business Media, 2013.

[40] W.J. Kent, BLAT—the BLAST-like alignment tool, Genome Res. 12 (4) (2002) 656–664. Available 2019, https://genome.ucsc.edu/ (Accessed 2019).

[41] T.J. Wheeler, J. Clements, S.R. Eddy, R. Hubley, T.A. Jones, J. Jurka, R.D. Finn, Dfam: a database of repetitive DNA based on profile hidden Markov models, Nucleic Acids Res. 41 (D1) (2012) D70–D82 (Accessed 2019), http://www.dfam.org/home.

[42] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. Ieee 86 (11) (1998) 2278–2324.

[43] S.M. Abd–Alhalem, N.F. Soliman, S. Eldin, S.E. Abd Elrahman, N.A. Ismail, E.S. M. El-Rabaie, F.E.A. El-Samie, Bacterial classification with convolutional neural networks based on different data reduction layers, Nucleosides Nucleotides Nucleic Acids (2019) 1–11.

[44] H. Zeng, M.D. Edwards, G. Liu, D.K. Gifford, Convolutional neural network architectures for predicting DNA–protein binding, Bioinformatics 32 (12) (2016) i121–i127.

[45] A. Al-Ajlan, A. El Allali, CNN-MGP: convolutional neural networks for metagenomics gene prediction, Interdiscip. Sci. 11 (4) (2019) 628–635.

[46] M.K. Elbashir, M. Ezz, M. Mohammed, S.S. Saloum, Lightweight convolutional neural network for breast Cancer classification using RNA-Seq gene expression data, IEEE Access 7 (2019) 185338–185348.

[47] J. Zhou, L.Y. Luo, Q. Dou, H. Chen, C. Chen, G.J. Li, P.A. Heng, Weakly supervised 3D deep learning for breast cancer classification and localization of the lesions in MR images, J. Magn. Reson. Imaging 50 (4) (2019) 1144–1151.

[48] A. Ghoneim, G. Muhammad, M.S. Hossain, Cervical cancer classification using convolutional neural networks and extreme learning machines, Future Gener. Comput. Syst. 102 (2020) 643–649.

[49] M. Porumb, E. Iadanza, S. Massaro, L. Pecchia, A convolutional neural network approach to detect congestive heart failure, Biomed. Signal Process. Control 55 (2020), 101597.

[50] A.K. Mukhopadhyay, S. Samui, An experimental study on upper limb position invariant EMG signal classification based on deep neural network, Biomed. Signal Process. Control 55 (2020), 101669.

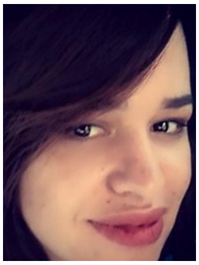
[51] S. Kundu, S. Ari, P300 based character recognition using convolutional neural network and support vector machine, Biomed. Signal Process. Control 55 (2020), 101645.

[52] J. Zhang, D. Fan, Z. Jian, G.G. Chen, P.B. Lai, Cancer specific long noncoding RNAs show differential expression patterns and competing endogenous RNA potential in hepatocellular carcinoma, PLoS One 10 (10) (2015), e0141042.

[53] T. Kobayashi, Genome instability of repetitive sequence: lesson from the ribosomal RNA gene repeat. In DNA Replication, Recombination, and Repair, Springer, Tokyo, 2016, pp. 235–247.

[54] G. Zhang, Y. Luo, G. Li, L. Wang, D. Na, X. Wu, L. Wang, DHRSX, a novel non-classical secretory protein associated with starvation induced autophagy, Int. J. Med. Sci. 11 (9) (2014) 962.

[55] W.R. Brown, P.J. MacKinnon, A. Villasanté, N. Spurr, V.J. Buckle, M.J. Dobson, Structure and polymorphism of human telomere-associated DNA, Cell 63 (1) (1990) 119–132.



Rabeb. Touati: PhD, master and engineer in electrical engineering from the National Engineering School of Tunisia (ENIT). Currently, she has a Post-Doctoral position at the Laboratory of Human Genetics (LR99ES10) at the Faculty of Medicine of Tunis. Her research interest includes biomedical, genomic signal and image processing, bioinformatics, pattern recognition and machine learning.

Asma Tajouri: PhD in Human Genetics from the Faculty of Medicine of Tunis. She has a Post-Doctoral position at the Laboratory of Human Genetics (LR99ES10) at the Faculty of Medicine of Tunis. Her research interests include Human Genetics.

Imen. Messaoudi: Received her PhD degree in electrical engineering from the National Engineering School of Tunisia. She is Assistant professor at the Higher Institute of Information Technologies and Communications from Carthage University. Her research interest includes biomedical and genomic signal processing.

Afef. Elloumi Oueslati: PhD in electrical engineering from the National Engineering School of Tunisia (ENIT). She is Associate Professor at the National School of Engineers of Carthage (ENICarthage). Her research interest includes issues related to signal and image processing applied in the biomedical and genomic fields.

Zied. Lachiri: PhD in electrical engineering from the National Engineering School of Tunisia (ENIT).He is Professor and Research Director in the Signal, Image and Information Technology laboratory (LR-SITI, ENIT). His research interests include pattern recognition, and signal and image processing in biomedical, multimedia, and man-machine communication

Maher. Kharrat: PhD in Human Genetics from the Faculty of Medicine of Tunis (FMT). He is Associate Professor and Research Director in the Genetic Human laboratory (LR99ES10) at the Faculty of Medicine of Tunis (FMT). He currently works at the Faculty of Medicine, University of Tunis El Manar. Dr. Maher does research in the field of Human Genetics.