# Microsatellite's Detection using the S -Transform Analysis based on the Synthetic and Experimental Coding

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

Afef Elloumi Oueslati[3]
University of Carthage
National School of Engineers of Carthage (ENICarthage)
Electrical Engineering Department
Tunisia

Imen Messaoudi[2]
University of Carthage, Higher Institute of Information Technologies and Communications
(ISTIC) Industrial Computing Department
Tunisia

Zied Lachiri[4]
University of Tunis El Manar
SITI Laboratory National School of Engineers of Tunis
BP 37, le Belvédère, 1002, Tunis
Tunisia

***Abstract*—Microsatellite in genomic DNA sequence, or Short tandem repeat (STR). It is a class of tandem repeat that have repeated pattern with size of 2- 6 base-pairs adjacent to each other. The detection of the specific tandem repeat is an important part of genetic diseases identification and it is also used in DNA fingerprinting and in evolutionary studies. Many tools based on string matching have been developed to detect microsatellites. However, these tools are based on prior information about repetitions in the sequence which cannot be always obtainable. For this, the signal processing techniques were suggested to overcome the limitations of the bioinformatic tools. In this paper, we use a new variant of the S-Transform which we apply to short tandem repeats signals. These signals are firstly obtained by applying different coding techniques to the DNA sequences. To further study the performance of the proposed method, we establish a comparison with different bioinformatics approaches (TRF, Mreps, Etandem) and three other methods of signal processing: The Adaptive S-Transform (AST), the Empirical Mode and Wavelet Decomposition (EMWD) and the Parametric Spectral Estimation (PSE) considering the AR model. This study indicates that our approach outperforms the earlier methods in identifying the short tandem repeat, in fact, our method detects the exact number and positions of trinucleotides present in the tested real DNA sequence.**

***Keywords*—DNA sequence; microsatellites; synthetic and experimental coding; s-transform; bioinformatic tools; Empirical Mode and Wavelet Decomposition (EMWD); Parametric Spectral Estimation (PSE)***

## I. Introduction

Computational analysis of the DNA sequences is a fundamental subject, which aims to understand the biological functionality of all living organisms. A particular attention was turned to microsatellites, which ensure many biological functions. In fact, they are implicated in cell metabolism, mismatch repair system [1], regulation of chromatin organization [2], genes activity and in many other functions [3].

A microsatellite sequence, also called Short Tandem Repeat (STR), represents two or more adjacent copies of a short nucleotide pattern unit [4]. The STR is defined by a specific period (pattern unit). The microsatellite's period is typically between 2 and 6 nucleotides per unit appointed di-, tri-, tetra-, penta- and hexa-nucleotides, respectively [5]. These elements (STRs) have a length less than 150 base pair [6]. Microsatellites considerably occur at different locations within the organism genome. They are very redundant, reduced and dispersed therefore, microsatellites are detected through automatic tools due to their importance on the one hand, for the human genome; approximately 10% of the DNA consists of microsatellites [7]. This special repeat can be a direct cause of many human diseases such as Huntington's chorea, spinal and bulbar muscular atrophy [8], myotonic dystrophy [9] and Friedreich's ataxia [10]. On the other hand, for other genomes, microsatellite elements are useful in many research domains such as DNA forensics [11], population genetic analysis [12], conservation biology and phylogenetics [13], [14].

Taking into account the importance of these regions, many researches focused on studying tandem repeats or microsatellites using the bioinformatic tools [15],[16]: the MISA [17], Sputnik [18], Mreps [19], EMBOSS (etandem and equitandem) [20], RepeatMasker [21], and TRF [22] . These tools use repeats candidates and compare them to DNA consensus sequences to detect microsatellites. These algorithms use a regular expression [21], the Hamming distance [15], the recursive match and the penalty scores [22]. They also use *k*-mers with suffix trees [23] and Heuristic alignment procedure [24].

Among these tools, Tandem Repeats Finder (TRF) is the most used one for detecting the short tandem repeats in DNA

sequences [22]. Nevertheless, it is not easy to use due the need to carefully choose settings. This is not only specific to this tool, as most bioinformatics tools need also prior information for the input parameters of the system such as [16]: pattern, pattern size, number of repeats, reference sequence or score [25]. However, sometimes, we do not have this prior information because of lack the short tandem repeats characteristics.

Aiming to overcome these limitations, scientists tried to find effective approach based on the signal processing techniques without prior information on targeted sequence. These approaches mainly use periodicity to detect STRs [26]. In this sense, the spectral analysis based on the exact periodic subspace decomposition and the autoregressive model (AR) were carried out [27]. On the other hand, methods providing a time-frequency representation have been proposed [29], [30]. Thus, the Short Time Fourier transform [28] and the Complex Morlet wavelet transform was used for patterns visualization [31], [32]. In an attempt to detect Microsatellites, the adaptive and modified S-transform has been also used [33], [34].

In this paper, we are interested in the microsatellite's identification in the genomic sequences. As part of the genomic signal processing domain, this work proposes a new method that combines the S-transform and a particular coding technique. Our detection system achieves accurate results without using any prior knowledge about the input data. This paper is organized as follows. Section 2 presents the S-Transform which we will use as a time-frequency representation technique. A coding step is recommended to directly apply the S-transform on the DNA sequence. The different coding techniques used for the genomic sequences coding are described in section 3. In Section 4, the STRs detection algorithm has been detailed and illustrative examples are included. Section 5 provides the experimental results and evaluates the short tandem repeat identification performance by comparison to other methods. Finally, Section 6 concludes this paper.

## II. S-Transform as Analysing Technique

The S-Transform (ST) is a time–frequency distribution which was developed by Stockwel et al. in 1994 for analyzing geophysics data [35]. It is a hybrid technique of the Short Time Fourier Transform (STFT) and the Continuous wavelet Transform (CWT). It retains the phase information as in the STFT and provides a variable resolution similar to CWT. There are several ways to deal with these characteristics. Here, we present three existing variants of the S-Transform: The Standard S-Transform (SST) [35], the Generalized S-Transform (GST) [36] and the Width Window Optimized S-Transform (WWOST) [37]. Finally, we propose our S-Transform modification aiming to enhance the time-frequency resolution of microsatellites representations.

### A. The Standard S-Transform (SST)

The S-Transform, in its standard form, consists in calculating the Fourier Transform of a signal $x(t)$ multiplied by a gaussian window. Therefore, the Standard S-Transform calculation formula is:

$$S(t,f) = \int_{-\infty}^{+\infty} x(\tau) w(\tau,f) e^{-j2\pi\tau}\, \mathrm{d}\tau \quad (1)$$

Where $f$ represents the frequency, $\tau$ represents the time, $w$ is the gaussian window and $t$ controls the position of $w$ on the $\tau$ -axis.

In the time domain, the gaussian window is given by:

$$W(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}} \quad (2)$$

Where $\sigma$ is the gaussian standard deviation. It depends on the frequency as follows:

$$\sigma(f) = \frac{1}{f} \quad (3)$$

This function controls the window's width. The S-Transform can be defined as:

$$S(t,f) = \int_{-\infty}^{+\infty} x(\tau) \frac{|f|}{\sigma\sqrt{2\pi}} e^{\frac{-(\tau-t)^2 f^2}{2}} e^{-j2\pi f\tau}\, d\tau \quad (4)$$

With lowest frequency, S-Transform performs well in the frequency domain. While, with highest frequency, S-Transform gives better resolution in the time domain. The main drawback to the S-Transform is then the time-frequency resolution. More efficient representations were introduced by proposing several modifications [33]. The S-Transform optimization consists in controlling the gaussian window's width by adding new parameters [34].

### B. The Generalized S-Transform (GST)

The Generalized S-Transform is proposed by McFadden [36] as a modified form of the Standard S-Transform. This modification consists in introducing a novel parameter $\alpha$. This parameter controls the gaussian window's width as follows:

$$\sigma(f) = \frac{\alpha}{f} \quad (5)$$

Consequently, the Generalized S-Transform is written as follows:

$$S(t,f) = \int_{-\infty}^{+\infty} x(\tau) \frac{|f|}{\sigma\sqrt{2\pi}\,\alpha} e^{\frac{-(\tau-t)^2 f^2}{2\alpha^2}} e^{-j2\pi f\tau}\, d\tau \quad (6)$$

Width Window Optimized S-Transform (WWOST)

Sejdic and his team [38] have suggested another modification of the gaussian window width by introducing a new parameter $p$ in the expression of $\sigma(f)$.

$$\sigma(f) = \frac{1}{f^p} \quad (7)$$

Thus, the S-Transform becomes as follows:

$$S(t,f) = \int_{-\infty}^{+\infty} x(\tau) \frac{|f|^p}{\sigma\sqrt{2\pi}} e^{\frac{-(\tau-t)^2 f^{2p}}{2}} e^{-j2\pi f\tau}\, d\tau \quad (8)$$

In order to enhance the energy concentration in the time-frequency representation by the S-Transform, we propose another way to control the window width.

### C. Proposed Modification of the S-Transform

In this work, we propose a new variant of the S-Transform by combining the two modified versions of the S-Transform.

The gaussian standard deviation in this case will be defined as:

$$\sigma(f) = \frac{\alpha}{f^p} \quad (9)$$

The S-Transform becomes:

$$S(t,f) = \int_{-\infty}^{+\infty} x(\tau) \frac{|f|^p}{\sigma\sqrt{2\pi}\,\alpha} e^{\frac{-(\tau-t)^2 f^2 p}{2\alpha^2}} e^{-j2\pi f\tau}\, d\tau \quad (10)$$

In Fig. 1, we represent the Gaussian window function around the frequency 0.33 (which is equivalent to periodicity 3 in DNA). We take into account different values of p and α and we provide the temporal and the spectral supports of the correspondent window. When $p = 1$ and $\alpha = 1$, we are in the case of the Standard S-Transform (SST). When $p = 1$ and $\alpha = 2.4$, it is a Generalized S-Transform (GST). For $p = 1.8$ and $\alpha = 1$, it is the Width Window Optimized S-Transform (WWOST). Finally, for $p = 1.2$ and $\alpha = 2.4$, we are in the presence of the proposed S-Transform.

The combination of the parameters $p$ and $\alpha$ in the S-Transform offers more flexibility to the gaussian window to capture periodicity 3 than the anterior versions. It has the characteristic that it minimizes the band in both spatial and frequency domains.

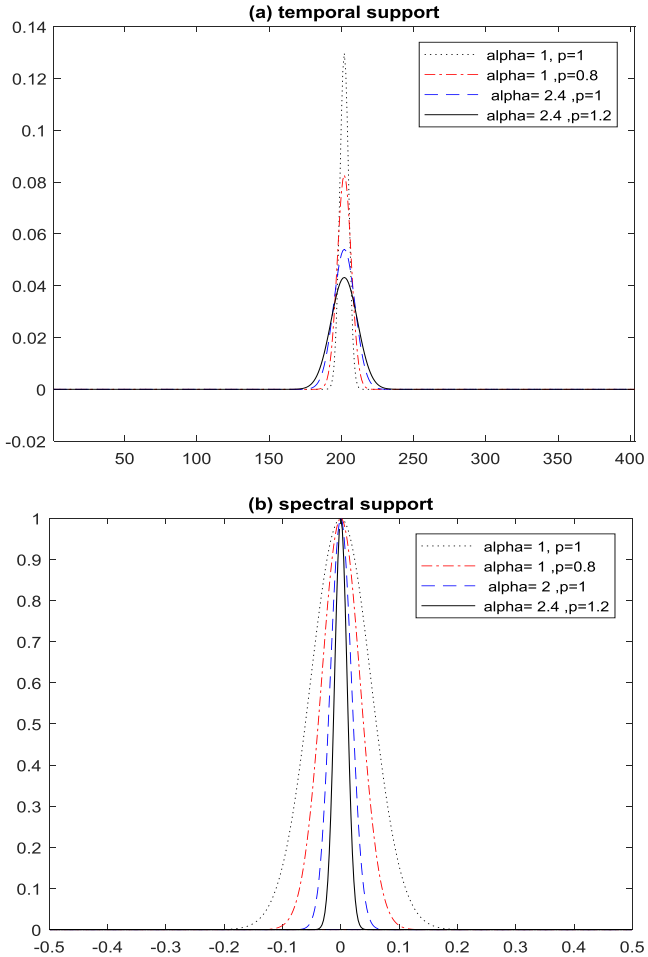


Fig. 1. The Gaussian Window with different Parameters Concentrated Around Periodicity 3; (a) Temporal Support, (b) Spectral Support.

Hence, the importance of this new variant of the S-Transform in terms of detecting the characteristic periodicities is in DNA especially in the microsatellite ones.

## III. DNA Coding Techniques

To be able to apply suitable signal processing methods to the DNA sequence, we must first convert the ATCG string to numeric signal. We would point out that DNA is a character string combining 4 nucleotides: A, C, T and G.

To achieve this conversion operation, different coding techniques have been proposed. The choice of the coding technique is delicate since that each method must be tested to see if it can enhance particular useful information [39]. These techniques can be defined by the substitution of nucleotides by numerical values according to the user's choice. On the other hand, they can be based on statistical or structural properties of DNA; which will reflect interesting specificities of the sequence. Thus, two large DNA coding methods including synthetic and experimental coding.

### A. Synthetic Coding

The synthetic coding principle consists in assigning a real or an imaginary value to a nucleotide base or a group of nucleotides. The most widely used synthetic mapping techniques are the binary coding, the complex binary [40] and the random walk [41].

*1) Binary coding:* The binary coding is based on simply assigning 0 or 1 to indicate the presence or the absence of a nucleotide base in the original sequence. For example, we can apply the following formula to seek the presence of the base A:

$$U(i) = \begin{cases} 1 \; if \; the \; base \; at \; position \; i \; is \; A \\ 0 \qquad\qquad\qquad otherwise \end{cases} \quad (11)$$

*2) Binary complex coding:* The binary complex coding consists in giving an imaginary value to each nucleotide as follows:

$$U(i) = \begin{cases} 1 + j \; if \; the \; base \; at \; position \; i \; is \; A \\ 1 - j \; if \; the \; base \; at \; position \; i \; is \; T \\ -1 - j \; if \; the \; base \; at \; position \; i \; is \; C \\ -1 + j \; if \; the \; base \; at \; position \; i \; is \; G \end{cases} \quad (12)$$

*3) Random walks:* DNA nucleotide can be classified according to their chemical structure [41]. We found in the pyrimidine class the nucleotides (C, T) and (A, G) in the purine one. The random walk is based on assigning the value -1 if the base is C or T and 1 in case if the base A or G.

### B. Exprimental Coding

Experimental coding techniques make use of experimental tables to reflect the chemical and the structural properties of DNA in the produced signal. As examples, we present here the EIIP [42], EIIPc [43] and the PNUC coding [44].

*1) EIIP:* The EIIP mapping is based on the electrons energy measurement which is delocalized in nucleotides [42]. The energy values corresponding to each nucleotide are illustrated in Table I.

TABLE I. THE EXPERIMENTAL VALUES FOR THE EIIP CODING TECHNIQUE

| **Nucleotide** | A | G | C | T |
|---|---|---|---|---|
| **EIIP** | 0.1260 | 0.0806 | 0.1340 | 0.1335 |

*2) EIIPc:* Here, the EIIPc technique [43] uses the electrons energy measurement delocalized in amino acids. The DNA sequence is read by overlapping codon and then, a numerical value according to the following Table II, is allocated to each position.

*3) PNUC:* The PNUC coding [44] is based on the scratches curvatures measurement related to tri-nucleotides during nucleosome positioning. The correspondent experimental values are given in Table III. The principle consists in values assignment after an overlapping reading of the tri-nucleotides in the sequence.

In the next section, we tried these several coding techniques by using the S-Transform to select the most convenient one for tandem repeat detection.

TABLE II. THE EXPERIMENTAL VALUES FOR THE EIIPC CODING

| **Codon** | **EIIPc** | **Codon** | **EIIPc** |
|---|---|---|---|
| CTA/CTC/CTG/CTT /TTA/TTG | 0.0000 | GTA/GTC/GTG/GTT | 0.0058 |
| ATA/ATC/ATT | 0.0000 | CCA/CCC/CCG/CCT | 0.0198 |
| AAC/AAT | 0.0036 | CAC/CAT | 0.0242 |
| GGA/GGC/GGG/GGT | 0.0050 | AAA/AAG | 0.0371 |
| GAA/GAG | 0.0057 | GCA/GCC/GCG/GCT | 0.0373 |
| TAC/TAT | 0.0516 | TGC/TGT | 0.0829 |
| TGG | 0.0548 | ACA/ACC/ACG/ACT | 0.0941 |
| TCA/TCC/TCG/TCT /AGC/AGT | 0.0829 | CGA/CGC/CGG/CGT /AGA/AGG | 0.0959 |
| GAC/GAT | 0.1263 | CAA/CAG | 0.0761 |
| ATG | 0.0823 | TTC/TTT | 0.0946 |

TABLE III. THE EXPERIMENTAL VALUES FOR THE PNUC CODING

| ***Codon*** | ***PNUC*** | ***Codon*** | ***PNUC*** |
|---|---|---|---|
| AAA/ TTT | 0 | AGA/ TCT | 3.3 |
| AAC/ GTT | 3.7 | AGC/ GCT | 7.5 |
| AAG/ CTT | 5.2 | AGG/ CCT | 5.4 |
| AAT/ ATT | 0.7 | ATA/ TAT | 2.8 |
| ACA/ TGT | 5.2 | ATC/ GAT | 5.3 |
| ACC/ GGT | 5.4 | ATG/ CAT | 6.7 |
| ACG/ CGT | 5.4 | CAA/ TTG | 3.3 |
| GAA/ TTC | 3 | GCA/ GGG | 6 |
| GAC/CTG | 5.4 | GCC/ GGC | 10 |
| TTA/ TAA | 2 | TCA/TGA | 5.4 |
| CAG/ CTG | 4.2 | CGA/ TCG | 8.3 |
| CCA/TGG | 5.4 | CGC/ GCG | 7.5 |
| CCC/ TGC | 6 | CTA/ TAG | 2.2 |
| CCG/ CGG | 4.7 | GGA/ TCC | 3.8 |

## IV. MICROSATELLITES DETECTION METHOD

After presenting the mathematical principle of the S-Transform as well as the methods for DNA coding, we present our algorithm for the microsatellite's detection. The model we are providing here is based on the new variant of the S-transform and the PNUC coding. The correspondent flowchart is given in Fig. 2.

The algorithm consists, first, in converting the DNA nucleotides by the PNUC technique. After that, the obtained signal is smoothed using the OTSU method to maximize energy concentration. Afterwards, the new variant of the S-Transform is applied to the preprocessed signal. Then, we apply in a binarization operation of the S-transform time-frequency representation to locate the short tandem repeat. The final step consists to extract motif from microsatellite detected.

### A. Choice of the Coding Technique

In order to choose the appropriate coding ,which allows the best detection of existing periodicities in the genomic sequences with the best concentration of energy, we test different coding techniques. As exprimental coding we consider: PNUC, EIIP and EIIPc. As for synthetic coding, we consider the random walk and the binary and complex indicators. For experiments, we investigate an artificial microsatellite sequence *S* that contains the periodicites: 2, 3, 4, 5 and 6 bp.

A description of the short tandem repeats (STRs) existing in the sequence *S* is illustrated in Table IV.

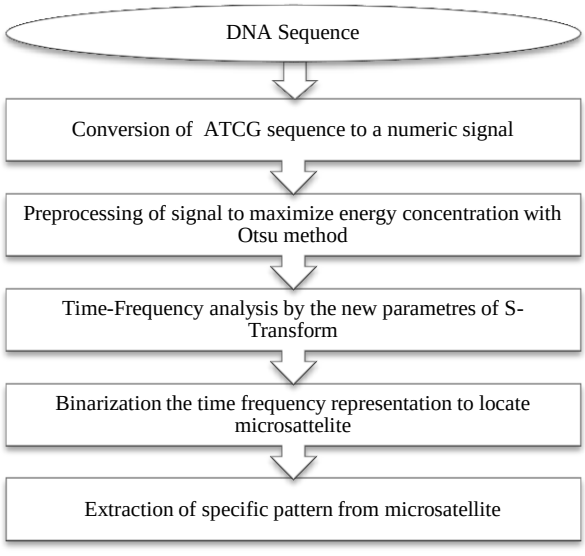


Fig. 2. Flowchart of the Proposed Method for Microsatellites Detection.

*S*=GTGTGTGTGTGTGTGTGTGTGTGTGAAAAAAAAAAAAAAAAA
AAACGTCGTCGTCGTCGTCGTCGTCGTCGTCGTCGTCGTCGTC
GTAAAAAAAAAAAAAAAAAAAAAACTGGCTGGCTGGCTGGCTGGCT
GGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGAAAAAAA
AAAAAAAAAAAAACCGTACCGTACCGTACCGTACCGTACCGTACC
GTACCGTACCGTACCGTACCGTACCGTACCGTACCGTAAAAAAAA
AAAAAAAAAAAACCTACGCCTACGCCTACGCCTACGCCTACGCCT
ACGCCTACGCCTACGCCTACGCCTACG

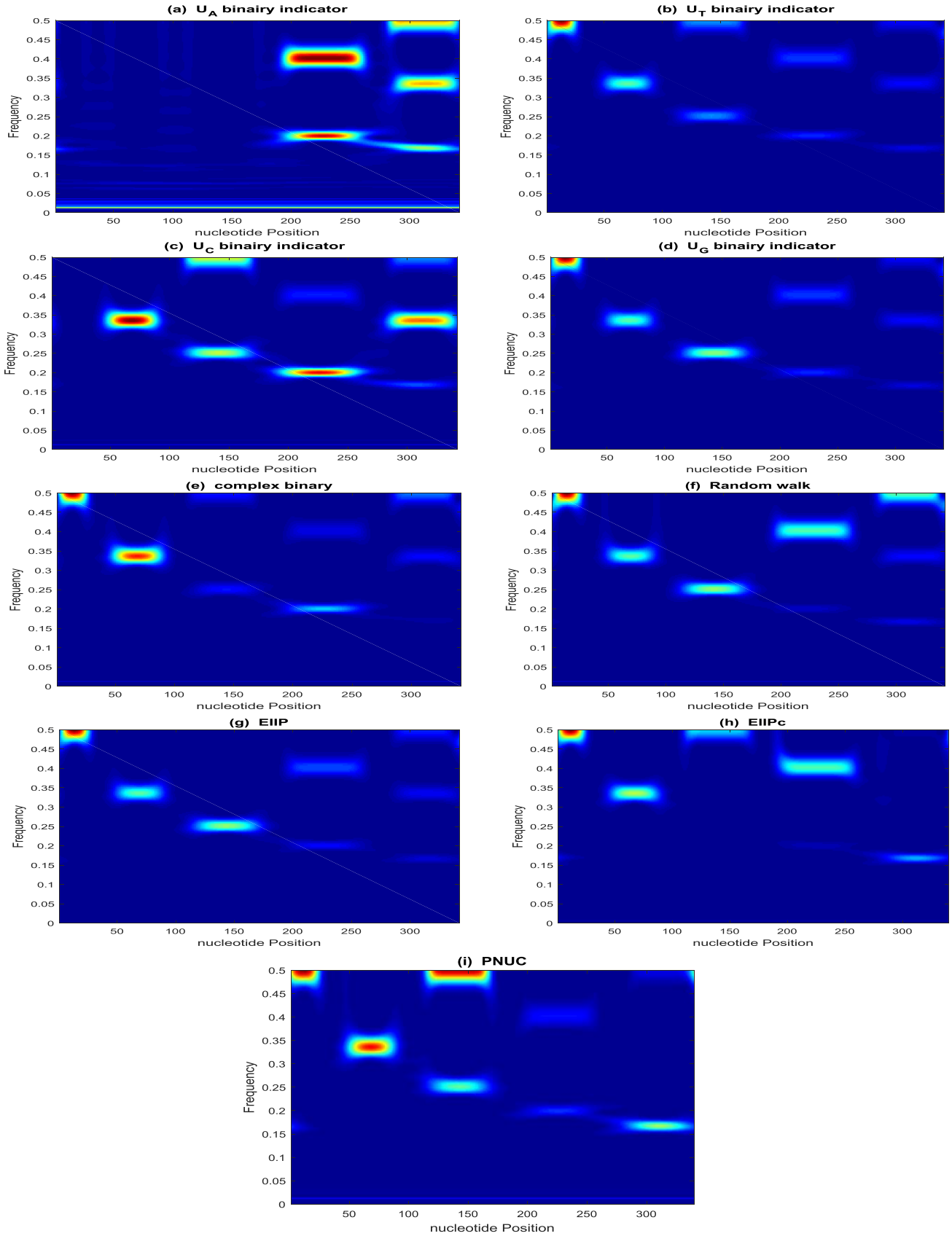


Fig. 3. Time-Frequency Representations of an Artificial Microsatellite Coded with Synthetic and Experimental Technique..

TABLE IV. THE EXISTING STRS IN THE SEQUENCE *S*

| **Periodicity** | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| **Pattern** | GT | CGT | CTGG | CCGTA | CCTACG |

In Fig. 3, we provide the time fequency presentation of this sequence after application of the new S-Transform with $\alpha = 2.4$ and $p = 1.2$.

The evaluation of the coding technique is done based on a visual inspection of the time-frequency representation. The latter, must enhance all the periodicities in the DNA sequence both in the time and the frequency plans.

For the synthetic coding group, we notice that the random walk coding doesn't give a good temporal localization. It is the complex binary technique that gives the best result since it detects the majority of periodicities (all periodicities except periodicity 6). As for the binary coding, we consider the $U_A$, $U_C$, $U_T$ and $U_G$ indicators. Based on the correspondent time-frequency representations, we find that the best result is obtained with $U_G$.

For the experimental tables based-on coding, PNUC is shown to be the best method compared to EIIP and EIIPc.

The binary coding depends on the nucleotide constituting the sequence. In fact, if the microsatellite does not contain the nucleotide G, the binary indicator $U_G$ will not detect it. It is the same case with other indicators as well. Therefore, we will use the PNUC coding.

### B. OTSU Thresholding

In this sub-section, we study the effect of preprocessing on the microsatellite time-frequency representation by the new S-Transform. For this aim, we take an artificial sequence characterized by periodicity 3 which is caused by two different motifs (CGT and TAC). The two microsatellites contained in the sequence *Seq* have almost the same size.

*Seq*=CCCCCCCCCCCGTCGTCGTCGTCGTCGTCGTCGTCGTAAAAAAAA
AAATACTACTACTACTACTACTACTACCCCCCCCCCCGTCGTCGTCGT
CGTCGTCGTCGTCGTCGTAAAAAAAAAAATACTACTACTACTACTACTA
CTACCCCCCCCCCCCCGTCGTCGTCGTCGTCGTCGTCGTCGTAAAAAAA
AAAATACTACTACTACTACTACTACTACTAC

The time-frequency representation of the sequence Seq is given in Fig. 4(a). As we can see, the two STRs are detected but one of them is more expressed than the other. So, we need a method that allows the detection of the two microsatellites (i.e. the two STRs must have same energy in the time-frequency plan). That is why we thought of using the OTSU thresholding technique on the signal before applying the S-Transform.

The principle of this preprocessing technique consists in calculating the histogram shape-based threshold. This method assumes that the data can be divided into two classes, and then determines the optimum threshold that minimizes the combined spread of the two classes [45].

In order to obtain the threshold established by the OTSU's method, we firstly select the location of the minima/maxima of the signal. Then, we prepare the peak strengths histogram in absolute value. Finally, we apply the OTSU's algorithm.

Fig. 4(*b*) shows the time-frequency resolution improvement after preprocessing.

Indeed, we notice that the preprocessing we choose here allowed us to easily identify all the existing periodicities in the artificial sequence. It enhances the energy concentration for the pattern CGT as for the pattern TAC. By contrast, the ST without taking into account the OTSU's method cannot well highlight the pattern TAC.

### C. New Variant of the S-Transform

After preprocessing the DNA sequence, we move to the analysis step. Since the new variant of the S-Transform proposed is a general form of the latter, we test its parameters in such way that SST, GST and WWOST will be also verified. Therefore, we have to modify the frequency-dependent control parameters $\alpha$ and $p$ to get the appropriate window length allowing the microsatellite periodicities (period=1/frequency) detection. Fig. 5 illustrates the time-frequency representations of the DNA sequence *Seq* by testing different values of $\alpha$ and$p$; *Seq* is the same sequence presented in sub-section *B*.

These sub-figures demonstrate that the proposed S-Transform gives the best time-frequency resolution, so it is the most suitable setting to study microsatellites. The choice of these parameter values is done in an empirical way. We retain the values with a narrowest frequency band: $\alpha = 2.4$ and $p = 1.2$.

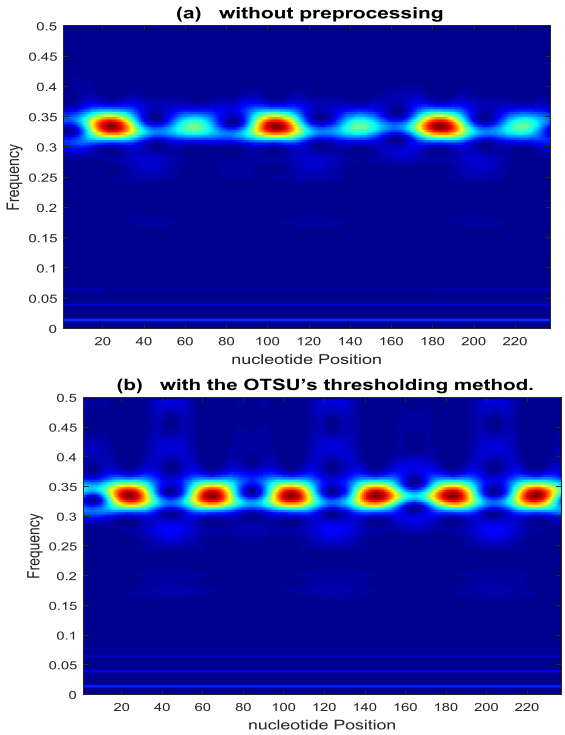


Fig. 4. Time Frequency Representation of Seq: (a) without Preprocessing (b) with the OTSU's Thresholding Method.

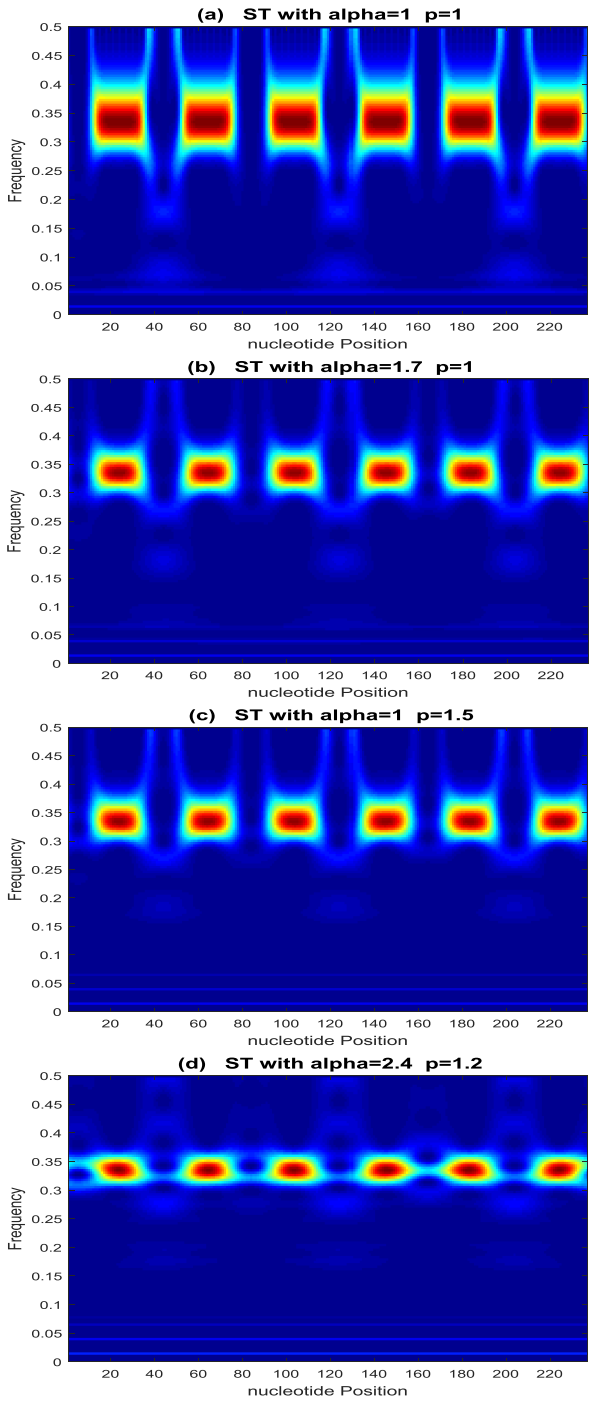


Fig. 5. Time Frequency Representation of Seq by: (a) SST, (b) GST, (c) WWOST and d) the New Variant of S-Transform.

## D. Binarization

After enhancement of the microsatellite representation in the time-frequency plan with the particular values, we go on to the detection step. However, in order to delimit the start and the end of the short tandem repeat from the ST representation, we must eliminate the noise existing in it. For this aim, we thought of transforming the time frequency representation into a binary one using a thresholding operation. The binarization step consists in giving the value of 0 or 1 to a pixel after comparing it to a threshold. In our work, we tested different threshold values. The optimal STRs detection was obtained for a threshold equal to 0.83. In Fig. 6, we present the time-frequency representation of *Seq* after binarization considering the best threshold value.

## E. Extraction Pattern from Short Tandem Repeat

After identifying the microsatellite length and its periodicity. We want now to determine its specific pattern. So, we use an automatic algorithm to capture repetitive pattern.

The extraction pattern consists in comparing a DNA sequence of size p to DNA sequence of length n. The consensus repeat pattern is the most repeated pattern of the sequence.

In the previous example, the algorithm above gave us the results shown in Table V.

This table presents multiple short tandem repeats that exist in seq with its characteristics: (beginning, end, periodicity, pattern). And we notice that the patterns location is made with good precision.

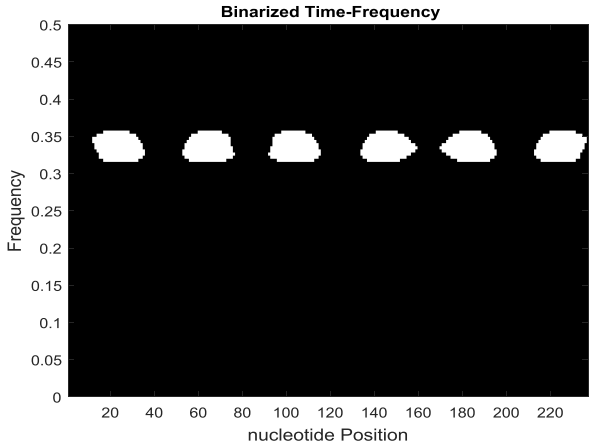


Fig. 1: The Binarized Time-Frequency Representation of Seq.

TABLE V. SHORT TANDEM REPEAT DETECTION IN SEQ

| Start (bp) | End (bp) | Period | Pattern |
|---|---|---|---|
| 12 | 37 | 3 | CGT |
| 53 | 78 | 3 | TAC |
| 92 | 117 | 3 | CGT |
| 133 | 160 | 3 | TAC |
| 169 | 196 | 3 | CGT |
| 212 | 237 | 3 | TAC |

## V. Experimental Results

In this section, we will test the performance of our algorithm in detecting STRs. The sequence X64775 of Oryza sativa Indica Group is selected for our experimentations. This DNA sequence is obtained from the NCBI database [46]. It has a short tandem repeat starting at 142 base-pairs and extending to 186 base-pairs. The repetition of the pattern 'GGC' is the characteristic of this repeat region. We chose this sequence Due to its common use in previous studies [27], [33].

The result obtained after applying each step of our algorithm is illustrated in Fig. 7. We also give the ST presentation of the sequence without applying the OTSU method (Fig. 7(*b*)).

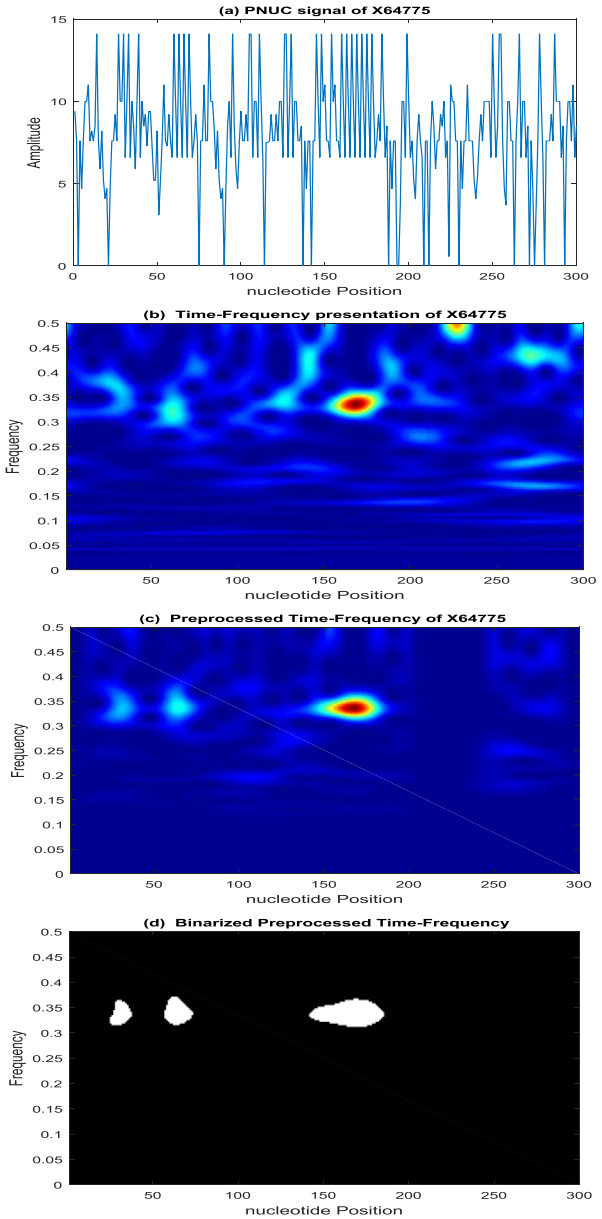


Fig. 6. (a) PNUC Signal Representation, (b) ST Representation of the PNUC Signal, (c) ST Representation after Preprocessing the PNUC Signal with OTSU, (d) Binarized time-Frequency Representation.

The sub-figures (*b*) and (*c*) demonstrate the role played by the OTSU method in smoothing the ST representation; which enhances the three regions with high energy around the frequency 0.33 (i.e. Periodicity 3). In fact, these zones represent repeats of trinucleotides motifs in the X64775 sequence. Periodicity 3 is well localised in both the time and the frequency domains. Hence, the importance of our algorithm is in characterizing microsatellites by the correspondent frequency and locating their position.

To evaluate the efficiency of the proposed method, we compared the obtained results, first, with bioinformatics tools. So, we chose: Mreps, Etandem and TRF.

For Mreps and Etandem, we kept the default parameters. For TRF, in the beginning we also kept the default settings. Then, we changed the settings as follow: match=2, mismatch=5 and indels=5, for the Minimum Alignment Score= 30.

The obtained results are presented in Table VI.

We notice that periodicity 3 is detected one time by Etandem and TRF with the default setting. however, Mreps detects 2 regions of periodicities 3 and one region with periodicity 6. As for our method, it locates 3 regions with periodicity 3. Our results are the nearest to those of TRF considering the adjusted parameters; which are the most suitable.

We compared as well with analysis techniques as the Adaptive S-Transform (AST), the Empirical Mode Wavelet Decomposition (EMWD) and the Parametric Spectral Estimation (PSE) considering the AR model.

The results of the microsatellite detection in X64775 by the methods ATS, EMWD and PSE are detailed in [27] and [33].

From Table VII, Periodicity 3 is detected by EMWD only two times. The remaining techniques identify 3 regions with periodicity 3. These techniques succeed to identify the microsatellites listed in the NCBI database. Whereas, PSE and our method detect another short tandem repeat similar to one detected by TRF in terms of localization. Only our proposed method detected the same three microstatellites detected also by TRF, from the standpoint of periodicity and position.

TABLE VI. Microsatellites Detection in X64775 with Bioinformatic Tools and Our Proposed Method

| **Method** | | **Region** | **Period** | **Pattern** |
|---|---|---|---|---|
| **Etandem** | | 140-202 | 3 | GCG |
| **Mreps** | | 59-73<br>147-163<br>159.-182 | 3<br>6<br>3 | -<br>-<br>- |
| **TRF** | **Default settings** | 145-188 | 3 | GGC |
| | **match=2, mismatch=5, indel=5, score=30** | 29-43<br>59-73<br>145-188 | 3<br>3<br>3 | CGC<br>CGG<br>GGC |
| **Proposed method** | | 27-37<br>59-71<br>146-183 | 3<br>3<br>3 | CGC<br>CGG<br>GGC |

TABLE VII. MICROSATELLITES DETECTION IN X64775 WITH ANALYSIS TECHNIQUES AND PROPOSED METHOD

| Method | Region | Period | Pattern |
|---|---|---|---|
| **AST** | 61-79<br>108-116<br>160-186 | 3<br>3<br>3 | GGC<br>CGG<br>GGC |
| **EMWD** | 57-72<br>140-187 | 3<br>3 | CGG<br>GGC |
| **PSE** | 49-57<br>59-76<br>141-188 | 3<br>3<br>3 | TAC<br>CGG<br>GGC |
| **Proposed method** | 27-37<br>59-71<br>146-183 | 3<br>3<br>3 | CGC<br>CGG<br>GGC |

To conclude, we succeeded in finding an efficient method for STRs detection. Furthermore, the obtained results match those of bioinformatics tools with the advantage of being independent from any prior knowledge about the searched repeat.

## VI. CONCLUSION

This study reveals the advantage of signal and image processing tools in highlighting short tandem repeats in DNA sequences instead of bioinformatics ones. The system, we proposed here, is based upon using a DNA coding technique and the S-transform.

First, we have investigated the role played by the coding technique in enhancing the time-frequency representation of microsatellites. Thus, we have tested six coding techniques which are: PNUC, EIIPc, EIIP, the binary coding, the complex binary coding and the random walk. The best resolution was obtained with the PNUC coding technique.

Next, we have presented our new approach for the microsatellites' detection. The algorithm consists of four steps.

As a first step, we encoded the DNA sequence into a numerical signal using the PNUC technique. Secondly, we preprocessed the obtained signal with the Otsu's method in order to maximize the useful information. Then, we applied a new variant of the S-Transform to get time frequency presentation of the sequence subject of study. The latter representation allowed us to easily localize the microsatellite position and periodicity after proceeding by a binarization step. The final step consists of extracting the pattern from the microsatellites, automatically.

To prove the effectiveness of our method, we have compared results with those of some bioinformatics tools: TRF, Mreps and Etandem. We have also established a comparison with other signal processing tools, which are: AST, Parametric Spectral Estimation and EMWD. In all cases, our approach outperforms these methods in terms of STR detection.

The main advantage of our algorithm consists in being independent from any prior knowledge of the repeat's characteristics. Moreover, it offers the possibility to get a simple graphic visualization of microsatellites.

In the future work, this approach can be extended to identify tandem repeats with higher repetitions unit length (minisatellites and satellites).

## REFERENCES

[1] LI,You-Chun,Korol,Abraham B.,Fahima,Tzion, and al., "Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review." Molecular ecology, vol. 11, no 12, p. 2453-2465,2002.

[2] LI, You-Chun, Korol, Abraham B., Fahima, Tzion, and al."Microsatellites within genes: structure, function, and evolution." Molecular biology and evolution, vol. 21, no 6, p. 991-1007, 2004.

[3] Kimura, Miyako, Sakamuri, Rama Murthy, Groathouse, Nathan A., *and al*. "Rapid variable-number tandem-repeat genotyping for Mycobacterium leprae clinical specimens."Journal of clinical microbiology, vol. 47, no 6, p. 1757-1766, 2009.

[4] Charlesworth, Brian, Sniegowski, Paul, andStephan, Wolfgang. "The evolutionary dynamics of repetitive DNA in eukaryotes." *Nature*, vol. 371, no 6494, p. 215, 1994.

[5] Ellegren, Hans. "Microsatellites: simple sequences with complex evolution." Nature reviews genetics, vol. 5, no 6, p. 435, 2004.

[6] Jarne, Philippe et Lagoda, Pierre JL." Microsatellites, from molecules to populations and back." Trends in ecology & evolution, vol. 11, no 10, p. 424-429, 1996.

[7] Pearson, Christopher E., Edamura, Kerrie Nichol, et Cleary, John D.., "Repeat instability: mechanisms of dynamic mutations." Nature Reviews Genetics, vol. 6, no 10, p. 729, 2005.

[8] P. Leeflang, Esther, Zhang, Lin, Tavare, Simon, et al. "Single sperm analysis of the trinucleotide repeats in the Huntington's disease gene: quantification of the mutation frequency spectrum." Human Molecular Genetics, vol. 4, no 9, p. 1519-1526, 1995.

[9] Hannan, Anthony J. "Tandem repeats mediating genetic plasticity in health and disease." Nature Reviews Genetics, vol. 19, no 5, p. 286, 2018.

[10] Jones, Lesley, Houlden, Henry, et Tabrizi, Sarah J. " DNA repair in the trinucleotide repeat disorders."The Lancet Neurology, vol. 16, no 1, p. 88-96, 2017.

[11] Moretti, Tamyra R., Baumstark, Anne L., Defenbaugh, Debra A., et al. "Validation of short tandem repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples." Journal of Forensic Science, vol. 46, no 3, p. 647-660, 2001.

[12] Kimura, Miyako, Sakamuri, Rama Murthy, Groathouse, Nathan A., et al. " Rapid variable-number tandem-repeat genotyping for Mycobacterium leprae clinical specimens."Journal of clinical microbiology, vol. 47, no 6, p. 1757-1766, 2009.

[13] Jarne, Philippe et Lagoda, Pierre JL." Microsatellites, from molecules to populations and back." Trends in ecology & evolution, vol. 11, no 10, p. 424-429, 1996.

[14] Richard, Guy-Franck, Kerrest, Alix, et Dujon, Bernard,"Comparative genomics and molecular dynamics of DNA repeats in eukaryotes." Microbiology and Molecular Biology Reviews, vol. 72, no 4, p. 686-727, 2008.

[15] Lim, Kian Guan, Kwoh, Chee Keong, Hsu, Li Yang, et al. " Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance." Briefings in bioinformatics, vol. 14, no 1, p. 67-81, 2012.

[16] Zribi, Soumaya, Oueslati, Afef Elloumi, et Lachiri, Zied. "Tandem repeat search tools performance for the Arabidopsis thaliana genome." In Advanced Technologies for Signal and Image Processing (ATSIP), 2016 2nd International Conference on. IEEE.p. 330-335, 2016.

[17] Thiel, Teresa, Michalek, W., Varshney, R., et al. "Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.)."Theoretical and applied genetics, vol. 106, no 3, p. 411-422, 2003.

[18] Abajian, Chris, "Sputnik - DNA microsatellite repeat search utility", 1994.

[19] Kolpakov, Roman, Bana, Ghizlane, et Kucherov, Gregory. " mreps: efficient and flexible detection of tandem repeats in DNA. " Nucleic acids research, vol. 31, no 13, p. 3672-3678, 2003.

[20] Sarachu, Martín et Colet, Marc. "wEMBOSS: a web interface for EMBOSS. " Bioinformatics, vol. 21, no 4, p. 540-541, 2004.

[21] Tarailo‐Graovac, Maja et Chen, Nansheng. "Using RepeatMasker to identify repetitive elements in genomic sequences. "Current protocols in bioinformatics, , vol. 25, no 1, p. 4.10. 1-4.10. 14 2009.

[22] Benson, Gary, et al. "Tandem repeats finder: a program to analyze DNA sequences. " Nucleic acids research, vol. 27, no 2, p. 573-580, 1999.

[23] Saha, Surya, Bridges, Susan, Magbanua, Zenaida V., et al. "Computational approaches and tools used in identification of dispersed repetitive DNA sequences. " Tropical Plant Biology, vol. 1, no 1, p. 85-96, 2008.

[24] RIVALS, Eric."A survey on algorithmic aspects of tandem repeats evolution." International Journal of Foundations of Computer Science, vol. 15, no 02, p. 225-257, 2004.

[25] Cao, Minh Duc, Balasubramanian, Sureshkumar, et Bodén, Mikael. " Sequencing technologies and tools for short tandem repeat variation detection. " Briefings in bioinformatics, vol. 16, no 2, p. 193-204, 2014.

[26] Tran, Thao T., Emanuele II, Vincent A., et Zhou, G. Tong. "Techniques for detecting approximate tandem repeats in DNA." In : Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on. IEEE.p. V-449, 2004.

[27] Zhou, Hongxia, Du, Liping, et Yan, Hong." Detection of tandem repeats in DNA sequences based on parametric spectral estimation." IEEE transactions on information technology in biomedicine, vol. 13, no 5, p. 747-755, 2009

[28] Sharma, Deepak, Issac, Biju, Raghava, G. P. S., et al. "Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation." Bioinformatics, vol. 20, no 9, p. 1405-1412, 2004.

[29] Sussillo, D., Kundaje, A., & Anastassiou, D. "Spectrogram analysis of genomes." EURASIP Journal on Advances in Signal Processing, 2004(1), 790248,2004.

[30] Touati, Rabeb, Messaoudi, Imen, Ouesleti, Afef Elloumi, et al. "Helitron's Periodicities Identification in C. Elegans based on the Smoothed Spectral Analysis and the Frequency Chaos Game Signal Coding." (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 9, No. 4, 2018.

[31] Messaoudi, Imen, Oueslati, Afef Elloumi, et Lachiri, Zied. "Revealing Helitron signatures in Caenorhabditis elegans by the Complex Morlet Analysis based on the Frequency Chaos Game Signals." In : *IWBBIO*. p. 1434-1444, 2014.

[32] Padole, Mamta C. "Recognizing Short Tandem Repeat Regions in Genomic Sequences Using Wavelet." In : Mathematics and Computers in Sciences and in Industry (MCSI), 2014 International Conference on. IEEE. p. 288-294, 2014.

[33] Sharma, Sunil Datt, Saxena, Rajiv, et Sharma, Sanjeev Narayan. "Identification of microsatellites in DNA using adaptive S-transform." IEEE journal of biomedical and health informatics, vol. 19, no 3, p. 1097-1105, 2015.

[34] Sharma, S. D., Saxena, Rajiv, Sharma, S. N., et al. "Short tandem repeats detection in DNA sequences using modified S-transform." International Journal of Advances in Engineering & Technology, vol. 8, no 2, p. 233, 2015.

[35] Stockwell, Robert Glenn, Mansinha, Lalu, and Lowe, R. P. "Localization of the complex spectrum: the S transform. "IEEE transactions on signal processing, vol. 44, no 4, p. 998-1001, 1996.

[36] Mcfadden, P. D卅, Cook, J. G., et Forster, L. M. "Decomposition of gear vibration signals by the generalised S transform." Mechanical systems and signal processing, vol. 13, no 5, p. 691-707, 1999.

[37] Djurović, Igor, Sejdić, Ervin, et Jiang, Jin. "Frequency-based window width optimization for S-transform." AEU-International Journal of Electronics and Communications, vol. 62, no 4, p. 245-250, 2008.

[38] Sejdic, Ervin, Djurovic, Igor, et Jiang, Jin. "A window width optimized S-transform." *EURASIP* Journal on Advances in Signal Processing, vol. 2008, p. 59, 2008.

[39] Zribi, Soumaya, Messaoudi, Imen, Oueslati, Afef Elloumi et Lachiri, Zied. "Tandem repeats detection based on S-transform applied on synthetic and experimental codings." In : Control, Automation and Diagnosis (ICCAD), 2017 International Conference on. IEEE. p. 314-319, 2017.

[40] A .Arneodo, C.Vaillant, B.Audit, F.Argoul, Y.d'Aubenton-Carafa, C.Thermes , Multi-scale coding of genomic information: from DNA sequence to genome structure and function. Phys Rep vol.498,no 2:45–188,2011.

[41] .Bai, Y.Liu, T.Wang "A representation of DNA primary sequences by random walk", Mathematical Biosciences, Elsevier, Vol. 209, n°.9, pp 282–291, 2007.

[42] V.Parisi , V.De Fonzo,F. Aluffi-Pentini : finding tandem repeats in DNA sequences. Bioinformatics 19(14):1733–1738

[43] M. Kobayashi, H. Toyoizumi, "Genomic Sequence Analysis using Electron-Ion Interaction Potential", University of Aizu, Graduation Thesis. March, 2005.

[44] D.S. Goodsell and R.E. Dickerson, "Bending and curvature calculations in B-DNA", Nucleic Acids Research, vol. 22, n° 24, pp: 5497-5503, Oxford University Press, 1994.

[45] H.Chen and R. Gururajan. "Otsu's threshold selection method applied in de-noising heart sound of the digital stethoscope record." In : Advances in Information Technology and Industry Applications. Springer Berlin Heidelberg, p. 239-244,2012.

[46] The NCBI GenBank database, 2014. Available: http://www.ncbi.nlm.nih.gov/Genbank/.