

Haar wavelet based approach for Short Tandem Repeats(STR) Detection

Tirthankar Paul
Information Technology and Electrical
Engineering
University of Oulu
Oulu, Finland
tirthankar.paul@oulu.fi

Seppo Vainio
Biochemistry and Molecular Medicine
Laboratory of Development Biology
University of Oulu
Oulu, Finland
seppo.vainio@oulu.fi

Juha Roning
Information Technology and Electrical
Engineering
University of Oulu
Oulu, Finland
juha.roning@oulu.fi

Abstract— Short Tandem Repeats (STRs)/ Microsatellites are the key factors to find the individuality on a DNA. These are playing an essential role in forensic science. Existing tools for identifying the whole series of STRs/microsatellites necessitate a complex computational method such as fast fourier transform, pattern recognition etc. The study reflects, identifying the STR regions, involving signal processing most preferably Haar wavelet and covering the complete range of STRs in the long chain of DNA sequence. The process demands less computational complexity and time compared to the other tools.

Keywords— Short Tandem Repeat, Signal Processing, Haar Wavelet, Wavelet Transform.

I. INTRODUCTION

Use of short tandem repeats (STR) analysis is increasingly common in forensic applications and has been of a recent research focus. A short tandem repeat is a microsatellite, which is a repeated sequence of nucleotides at specific location of nuclear DNA[1]. STR regions are polymorphic, and the exact number of repeated regions uniquely characterizes of an individual. The goal of STR analysis is to measure the exact number of repeating units. The most common method for STR analysis begins with polymerase chain reaction (PCR) followed by counting the exact number of repeats. Most common STRs consists of sequences of length 4 bases in the 13 standard forensic locations of the USA, but STRs of length 3 and 5 bases are also not uncommon.

Essentially, STR analysis involves counting the repeated number of patterns. In spite of having a finite length of the DNA sequence, this may be computationally challenging to find STR. In this paper, a faster method is proposed based on wavelets that not only count the number of STR units but also provides their locations on the DNA sequence.

STRs have much impact on genetic analysis, paternity decision, forensic science etc. [2]. There are about 40 genetic disorders related to STR mutations' disorders [3,4]. Neurological disorder such as Huntington's disease caused by extension of a massive number of repetition part. About 3% of the entire human genome consists of STRs. The scattering of STR in a chromosome is not uniform, and maximum STRs are originated in the noncoding region. As an example, an STR location is named as D3S1358, where 'D' is DNA, 3 is the chromosome number, 'S' means STR and the digits 1358 is the identifier[5]. Based on repetition unit size (quantity of base-pair(bp)), the tandem repeats are

divided three classes, 1) Satellite repeats, 2) Minisatellite repeats and 3) Microsatellite or Short tandem repeats(STR). Satellite repeats have base-pair from 5 to 171 bp as a repetition unit. For the minisatellite, the range of base-pair is 6 to 64, and finally, the shortest one STRs have 1-6bp long repeat sequence, which can be appeared up to 50 times [6-8].

The Haar wavelet was applied by Padole [9]. The author described a new way to identify the short tandem repeats region using Haar wavelets. The technique is very efficient to detect STR regions. In the study, many different lengths of STR repeats in a different form was shown and detected very effectively. But Haar wavelet is defined as having 2^n sampling points that can only identify 2^n length ($n=0,1,2,\dots$) STR sequence. Therefore, STR region of having 1,2,4,8... bp length-STR, can be distinguished perfectly. On the other way, those STR or microsatellite of length 3 or 5 bp cannot be identified with the conventional Haar wavelet method, as indicated by Padole[9]. So, in the present work, a new wavelet function to detect the three, five and six sequenced length-STR, and together with an algorithm, can identify the complete range of STR lengths.

While a conventional pattern-finding algorithm on a DNA string may work well for STR analysis, our method based on Haar wavelets offers better computational efficiency. The time complexity of Wavelet transforms is $O(\log n)$, where n is the length of the given DNA sequence. In addition to a lower computational cost, our method can be implemented as a tool for detecting the STR region in forensic studies.

The rest of the paper is described as follows. In section II, some related works based on genome signal processing has been discussed. The methodology is shown in section III. In part IV & V, conventional and the new approach that can deal with the problem have been shown consecutively. There are experiments and results discussed in section VI. Finally, in section VII, the work has been concluded.

II. RELATED WORK

Signal processing theory has been playing a vital role in analyzing the genome data. The study with genome data is known as Genome signal processing (GSP). Here in the discussion, most of the earlier studies identified sequences such as protein, tandem repeats and coding region of genome through GSP. FFT, period-3 property, pattern matching were the techniques, which were applied as a part of GSP in genome data. Translation of DNA message is not limited to a character by symbol scheme; mathematical and logical

algorithms can be a follower for the purpose. Many algorithms or mathematical theories are capable of defining the hidden science in natural creations. Some previous works identified protein region through GSP [10-14]. In most of these works, the nucleotides are mapped into a numerical array. Mapping with numerical values is considered single dimension representations, and the numerical values can be any numbers such as atomic weight values, random code of protein or nucleotides. On the other hand, in multi-dimensional/multi-layer representation more than one dimension vectors replace a nucleotide, for example, Voss representation [15]. In many studies, periodic nature was found in the coding region. The periodicity can be captured through some digital filters in GSP that help to predict the gene location on the DNA string [16]. Deepak Sharma, 2004 has proposed spectral repeat finder (SRF) to find discrete repeat sequence [8]. Most of the studies the coding region were transformed to the frequency domain after converting into a numerical array. Therefore, in the power spectrum, the coding region can be viewed by applying period-3 property [17] Tiwari, 1997). The peak of the power signal can be found at the frequency 1/3, due to the recursive nature of the coding region [8]. Translation of DNA message is not limited to a character by symbol scheme; mathematical and logical algorithms can be a follower for the purpose.

III. METHODOLOGY

Wavelet transforms was applied into the genome in some early studies [5, 9, 18-21]. Wavelet is the rapidly decaying wave-like oscillation that has zero mean infinite duration. Wavelet can be different shape and sizes; some well-known wavelets are morlet, daubechies and biorthogonal etc. The strength of wavelets is its wide range of variety of various forms of wavelet analysis. Wavelets are the class of function, which are well localized in time and frequency, analyze the abrupt changes of a signal and an image, and very fast in transformation. A wavelet is a mathematical function, which split a continuous-time signal into different scale modules or several groups of coefficients. These module or coefficient vectors hold information about features of the data in the different scale module. Generally, a frequency range is assigned to each of the module/ components. The discrete wavelet transform (DWT) performs a windowing technique with different sized window. It filters the long interval signal and short interval signal for getting low and high-frequency information respectively [22]. A signal 'x' pass through the low pass filter (scaling function) and high pass filter (wavelet function) and impulse response 'g' convolute with each of the filtered signal [20]. A wavelet transforms data or signals transform into coefficients based signal on wavelet transforms; it is a lossless linear transform. The wavelet function can be represented mathematically as

$$y(n) = \sum_{k=-\infty}^{\infty} x[k].g[n-k] \quad (1)$$

The signal after passed through low and high pass filter simultaneously and down sampled, the coefficient number become half of its original input sample for each filter. The

DWT generates low-pass filtered and high-pass filtered signals which are known as the approximation coefficient and detail coefficient, respectively. The coefficients can be represented through equation 2 and 3 and Fig. 1.

$$y(n)_{low} = \sum_{k=-\infty}^{\infty} x[k].g[2.n-k] \quad (2)$$

$$y(n)_{high} = \sum_{k=-\infty}^{\infty} x[k].h[2.n-k] \quad (3)$$

Wavelet transform matrix form can be described as $W_t = W.X$, Where $W = [L; H]$, X is the input signal. L and H are the low-passes and high-pass filter impulse response. Wavelet transform is useful to find the position of a character in a string array of a time series data, and it can be applied to any data sequence.

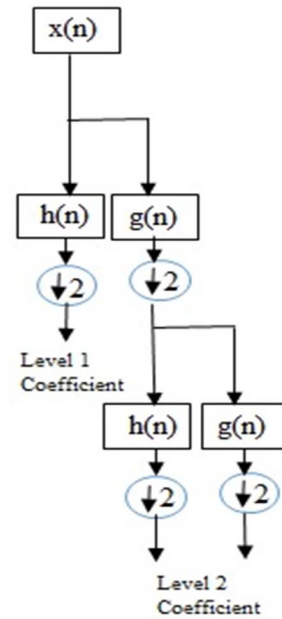


Fig 1. Signal flow diagram of Wavelet transform

Haar Wavelet: Haar wavelet is the basic form of all wavelets, which transforms a discrete signal into two sub-signals of half equal lengths. Haar wavelets are very fast in computation, and it is memory efficient and easily reversible without any data loss [23]. Haar functions are orthogonal functions in the unit interval $[0, 1]$, that can analyze the frequency component of the input signal. It is determined by the sampling point of the power of 2, and its input signal length be 2^n [24]. Fig. 2 is a Haar function (equation 4) wave of its simplest form.

$$h(x) = \begin{cases} 1 & \text{for } x \in [0,1] \\ 0 & \text{elsewhere.} \end{cases} \quad (4)$$

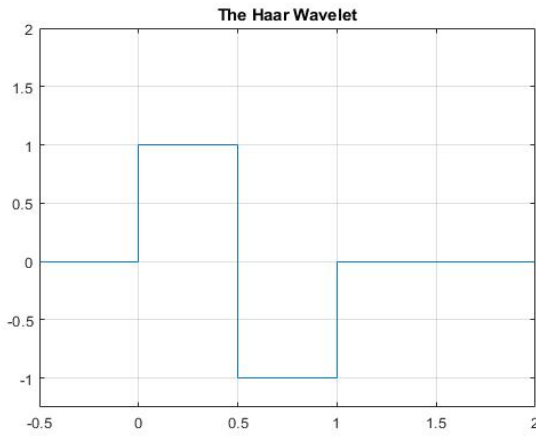


Fig. 2. The Haar Wavelet function.

IV. APPROACH BASED ON HAAR WAVELETS

The solution is proposed to overcome the problems of Haar wavelets for detecting repeating sequences that are not of length 2^n , is following. Computation of Haar wavelet was discussed and explained in earlier works [9, 25]. Considering 1-dimensional vector ‘S’ i.e, [4 2 4 2 1 1 4 2], the length of the signal is 8 ($N = 2^3$). This means it needs three levels of operations. In the process of Haar wavelet computational operation, two coefficients appear a) pairwise averages and b) detailed coefficient. The pairwise averages are for to get a new set of data with “lower-resolution”. Detailed coefficient holds the missing information, and it restores the original data. So, the pairwise average data on the level one is $[(4+2)/2, (4+2)/2, (1+1)/2, (4+2)/2] = [3, 3, 1, 3]$. In this, the new averaging data set has an average loss of information. The detailed coefficient is necessary to get back the original data and the first set of the detail coefficients are $[(4-2)/2, (4-2)/2, (1-1)/2, (4-2)/2] = [-1, -1, 0, -1]$. The original data set of eight values can be reconstructed with the pairwise average data and detailed coefficients without any data loss. This process will be applied for further levels. On the second level, the output of level one will act as an input signal, and the coefficients are as follows.

Pairwise average data on level two: $[(3+3)/2, (1+3)/2] = [3, 2]$.

Detailed coefficient: $[(3-3)/2, (1-3)/2] = [0, -1]$.

and on the level three,

the pairwise average data $[(3+2)/2]=[5/2]$.

Detailed coefficient: $[(3-2)/2]=[1/2]$.

TABLE I. CONVENTIONAL HAAR WAVELET COMPUTATION.

Decomposition Level (n)	Length of Signal	Pairwise Average $S_a = (S_i + S_{i+1})/2$	Detailed coefficient. $S_d = (S_i - S_{i+1})/2$
0	8	4 2 4 2 1 1 4 2	-
1	4	3 3 1 3	-1, -1, 0, -1
2	2	3 2	0, -1
3	1	5/2	1/2

Therefore, the wavelet transform of the signal ‘S’ is $[5/2 \ 1/2 \ 0 \ -1 \ -1 \ 0 \ -1]$. Haar wavelet transform process alternatively

performs averaging and differencing. Process of similar consecutive data points results in zero value because of differencing operation. Therefore, a zero or consecutive zero in detailed coefficient in any level implies that a) there is a sequence repetition in the input signal and b) the level suggests the repetition sequence length. For example, from the signal ‘S’ two zero has been found in the detailed coefficient, one in level 1 and other in level 2. In level one ($n=1$), a zero means one digit repetition. In the same way, two-digit repetition can be found in level 2. The signal ‘S’ has two ‘1’ at 5th and 6th position, and three times ‘4 2’, of which two of them are consecutive. A wavelet gives zero mean value if the wavelet coefficients of the signal come from constant value. This principle is useful to mask the repeat region of a numerical signal with constant intensity[28]. Fig. 3 shows the approach of 2^n length STR detection using Haar wavelet, where DNA sequence is mapped into numerical data then the data is filtered with Haar wavelet. The filtered results indicate the STR regions on the graph.

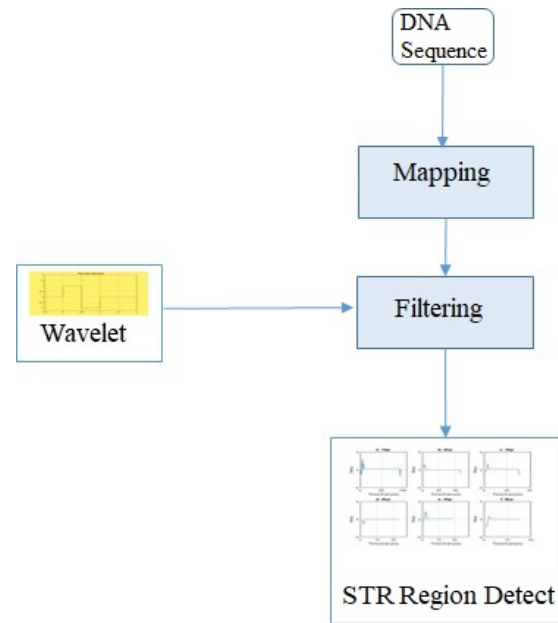


Fig. 3: Block diagram for STR detection using Haar Wavelet.

V. PROPOSED APPROACH

Conventional Haar wavelet is zero mean value wavelet, and masks two or 2^n sequence repeat data. In Fig. 2, the wavelet function values are 1 and -1. The summation of these two values of the wavelet function is zero. Based on the Haar wavelet concept to mask the three and five sequence repetition data, wavelet function should be zero mean value. The three sequence mask wavelet can take any of these sets of values, 1 1 -2, 1 -2 1, -2 1 1 and so on. Any combination of these sequences which results in summation value of zero, can be used for masking three sequence data. Here in Fig. 4, [1 1 -2] is taken to mask the three sequence data. Similarly, for five sequence wavelet signal is $y[n]=[1 \ -1 \ 1 \ -2 \ 1]$ and six bp length-STR can be found simultaneously both 2bp and 3bp length-STR finding operation. Before applying the Haar

CTATCTATCTATCTATCTATCTATCTATC
 TATGAGACAGGGTCTTGCTCTGTCACCCAG
 ATTGGACTGCAGT

- 4) ATGAAATCAACAGAGGCTTGCATGTATCTA
 TCTGTCTAGTCTAGTCTAGTCTAGTCTAG
 TCTAGTCTAGTCTAGTCTAGTCTAGTCTA
 GTCTAGTCTAGTCTATGAGACAGGGTCTT
 GCTCTGTCACCCAGATTGGACTGCAGT

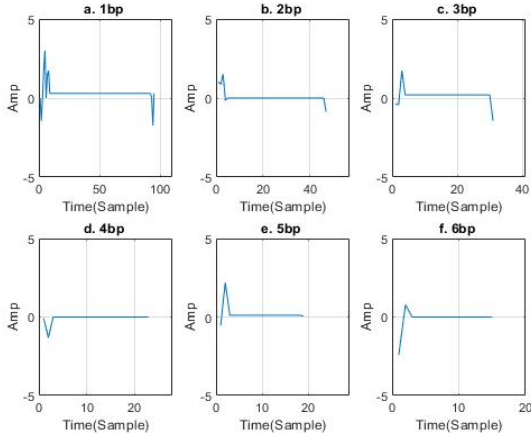


Fig. 5a. Visual representation of STR regions (Example 1).

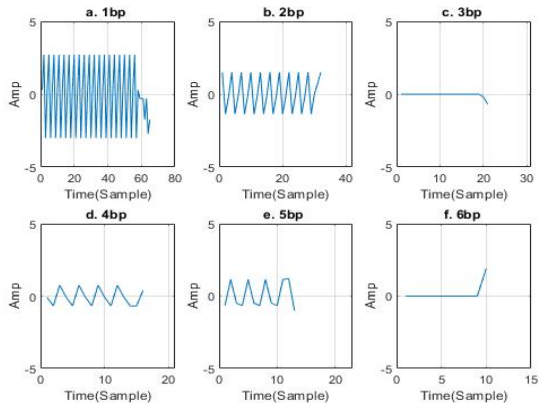


Fig. 5b. Visual representation of STR regions (Example 2).

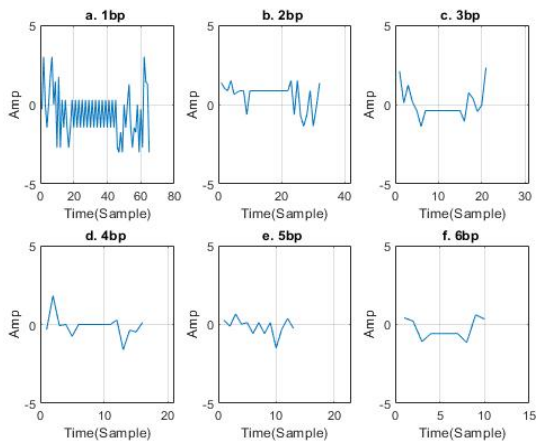


Fig. 5c. Visual representation of STR regions (Example 3).

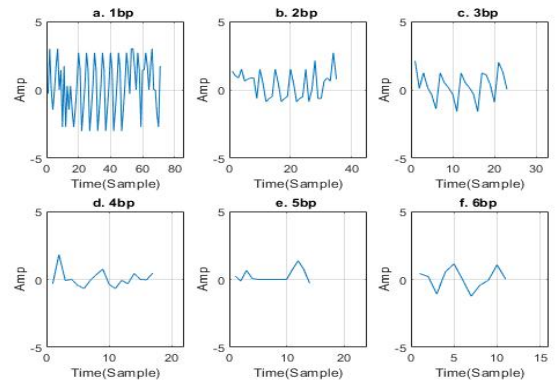


Fig. 5d. Visual representation of STR regions (Example 4).

```
Command Window
The Sequence has 2bp Tandem repeats, & repeat region is

ans =

3x70 char array

' 1 TATATATATA TATATATATA TATATATATA TATATATATA TATATATATA TATATATATA'
' 61 TATATATATA TATATATATA TATATATATA TATATATATA TATATATATA TATATATATA'
' 121 TATATATATA TATATATATA TATATATATA TATATATATA TATATATATA TATATATA '
```

```
Command Window
The Sequence has 3bp Tandem repeats, & repeat region is

ans =

2x70 char array

' 1 TAGTAGTAGT AGTAGTAGTA GTAGTAGTAG TAGTAGTAGT AGTAGTAGTA GTAGTAGTAG'
' 61 TAGTAGTAGT AGTAGTAGTA GTAGTAGTAG TAGTAGTAGT AGTAGTAGTA GTAGT '
```

```
Command Window
The Sequence has 4bp Tandem repeats, & repeat region is

ans =

' 1 TAICTAICTA TCTAICTATC TAICTAICTA TCTAICTAIC TAICTAICTA '
```

```
Command Window
The Sequence has 5bp Tandem repeats, & repeat region is

ans =

2x69 char array

' 1 CTAGTCTAGT CTAGTCTAGT CTAGTCTAGT CTAGTCTAGT CTAGTCTAGT CTAGTCTAGT'
' 61 C '
```

Fig. 6. MATLAB results for STR region and lengths.

Fig. 5a, 5b, 5c and 5d are the analysis of the different sequences. The straight line on $y=0$ axis is the main search finding from the graph. The consecutive points on $y=0$ axis indicate the repeat region for the sequence. The graph of the first example shows in Fig. 5a, which has ' $y=0$ ' line on 2bp, 4bp and 6bp subgraphs (The straight line of 3bp and 5bp graphs have very small amplitude; the figures are very close but not on the ' $y=0$ ' line). 2bp('CA' 'CA' 'CA' 'CA' 'CA' 'CA') sequence repeats appear as 4bp ('CACA' 'CACA'

‘CACA’) repeats and also 6bp(‘CACACA’ ‘CACA,CA’) repeats. But the STR is considered 2bp because it is the shortest tandem repeats for the sequence. Also it is noticeable that there are straight lines on 3bp and 5bp subgraphs, but they are on lies on ‘y=0’ line.

Similarly, 3bp and 6bp repetition can be found in Fig. 5b. Fig. 5c and Fig. 5d show the graph where repeat regions have been found in 4bp and 5bp subgraphs for 4bp and 5bp length STR sequences consecutively.

The STR region can be found by having zeros in the detailed coefficients. The number of zero indicates the number of repetitions of microsatellites. No ‘zero’, means no repetition on the experimented signal. If a zero is found in detailed coefficients that can be tracked until it reaches level 0, to get the microsatellite sequence in the DNA, all these experiments have been done in MATLAB. Fig. 6 shows the MATLAB results of STR regions and STR lengths.

VII. CONCLUSION.

STR analysis plays a key role not only in forensic science, but also in genetic mapping, phylogenetic, and population studies. While the conventional methods rely on string pattern finding, our method proposes a novel approach to perform STR analysis based on Haar wavelets. The time complexity of Wavelet transforms is $O(\log n)$ of n length DNA sequences. Our method is computationally efficient and is readily implementable tool facilitating STR detection for forensic analysis. We show that our approach can not only detect STRs of length 2^n , but also of lengths 3, 5, or any other number. In this paper, we proposed a theoretical analysis and practical implementation in MATLAB to validate our method.

ACKNOWLEDGMENT

InfoTech Oulu Doctoral Program has supported the work. We thank our colleagues from the University of Oulu, who provided insight and expertise that greatly assisted the research.

REFERENCES

- [1] G Benson, “Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids*” Res. 1999;27(2):573–580.
- [2] D. Goldstein and C. Schlotterer, “Microsatellites: Evolution and Applications,” Oxford University Press, 1999
- [3] Pearson CE, Edamura KN, Cleary JD. “Repeat instability: mechanisms of dynamic mutations,” 2005. *Nat Rev Genet* 6: 729–742.
- [4] S. Mirkin, “Expandable DNA repeats and human disease.” *Nature* 447, 932–940 (2007).
- [5] H. Fan, & J. Y. Chu, “A Brief Review of Short Tandem Repeat Mutation,” *Genomics, Proteomics and Bioinformatics*, 5(1), 2007, 7–14.
- [6] McNamara-Schroeder K, Olonan C, Chu S, Montoya MC, Alvir M, Ginty S, and Love JJ.: “DNA fingerprint analysis of three short tandem repeat (STR) loci for biochemistry and forensic science laboratory courses,” *Biochem Mol Biol Educ*. 2006 Sep; 34(5):378-83.
- [7] Serkan Dogan, Adna Asic, Imer Muhovic, Larisa Besic & Damir Marjanovic “Overview of the Genetic STR Clustering among Worldwide Human Populations, *International Journal of Human Genetics*,” 14:3-4, 131-142, 2014,
- [8] Deepak Sharma, Biju Issac, G. P. S. Raghava, R. Ramaswamy; “Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation,” *Bioinformatics*, Volume 20, Issue 9, 12 June 2004, Pages 1405–1412,
- [9] M. C. Padole, “Recognizing Short Tandem Repeat Regions in Genomic Sequences Using Wavelet,” *Proceedings - 2014 International Conference on Mathematics and Computers in Sciences and in Industry, MCSI 2014*, 288–294.
- [10] M Akhtar, J Epps, E Ambikairajah, “On DNA Numerical Representations for Period-3 Based Exon Prediction,” *IEEE International Workshop on Genomic Signal Processing and Statistics*. 2, 2007, pp. 1–4.
- [11] TM Inbamalar, R Sivakumar, “Filtering Approach to DNA Signal Processing,” In: *International Proceedings of Computer Science and Information Tech.* volume 28, 2012, pp. 1–5.
- [12] S Marhon, SC Kremer, “Gene prediction based on DNA spectral analysis: a literature review,” *Journal of computational biology* 18, 2011, 639–76.
- [13] M Akhtar, J Epps, E Ambikairajah, “Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction,” *Journal of Selected Topics in Signal Processing* 2: 2008, 310–321.
- [14] A Rushdi, J Tuqan, “Gene Identification Using the Z-Curve Representation,” In: *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*. volume 2, 2006, pp. 1024–1027.
- [15] Changchuan Yin, Ying Chen, Stephen S-T Yau, “A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering,” *Journal of theoretical biology*, ISSN: 1095-8541, Vol: 359, Page: 18-28, Publication Year: 2014
- [16] B. Ma, D. Qu and Y. Zhu, “A novel adaptive filtering approach for genomic signal processing,” *IEEE 10th INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING PROCEEDINGS*, Beijing, 2010, pp. 1805-1808.
- [17] Shrish Tiwari, S. Ramachandran, Alok Bhattacharya, Sudha Bhattacharya, Ramakrishna Ramaswamy, “Prediction of probable genes by Fourier analysis of genomic sequences,” *Bioinformatics*, Volume 13, Issue 3, June 1997, Pages 263–270.
- [18] J. A. T. Machado, A. C. Costa, & M. D. Quelhas, “Wavelet analysis of human DNA,” *Genomics*, 98(3), 155–163.
- [19] K. K. Hon, R. Atwal, & B. Y. M. Kwan, “Wavelet analysis of DNA sequences,” *2008 International Conference on Communications, Circuits and Systems Proceedings, ICCAS 2008*, 816–820.
- [20] J. K. Meher, M. R. Panigrahi, G. N. Dash, & P. K Meher, “Wavelet Based Lossless DNA Sequence Compression for Faster Detection of Eukaryotic Protein Coding Regions,” *International Journal of Image, Graphics and Signal Processing*, 2012, 4(7), 47–53.
- [21] K. B. Murray, D. Gorse, & J. M. Thornton, “Wavelet transforms for the characterization and detection of repeating motifs,” *Journal of Molecular Biology*, 2002, 316(2), 341–363.
- [22] M. S. Islam, R. Pears, & B. Bacic, “A wavelet approach for precursor pattern detection in time series,” *Journal of Electrical Systems and Information Technology*, 2018, 5(3), 337–348.
- [23] M. I. Mahmoud, M. I. M. Dessouky, S. Deyab, & F. H. Elfouly, “Comparison between Haar and Daubechies Wavelet Transformions on FPGA Technology,” *World Academy of Science, Engineering and Technology*, 2007, 6(3), 23–29.
- [24] R. S. Stankovir, & B. J. Falkowski, “The Haar wavelet transform: Its status and achievements,” *Computers and Electrical Engineering*, 2003, 29(1), 25–44.
- [25] D. Sacharidis, “Constructing Optimal Wavelet Synopses,” *Proceedings of the 2006 International Conference on Current Trends in Database Technology EBDT '06*, Pg 97-104, 2006 10.1007/11896548_10.
- [26] T. Paul, S. Vainio, J. Roning, “Towards Personalised, DNA Signature Derived Music via the Short Tandem Repeats (STR) ,” In: Arai K., Kapoor S., Bhatia R. (eds) *Intelligent Computing. SAI 2018. Advances in Intelligent Systems and Computing*, 2019, vol 857. Springer, Cham.
- [27] S. H. El-Alfy, & A. F. Abd El-Hafez, “Paternity testing and forensic DNA typing by multiplex STR analysis using ABI PRISM 310 Genetic Analyzer,” *Journal of Genetic Engineering and Biotechnology*, 2012, 10(1), 101–112.
- [28] B. Zhang, M. J. Fadili, J. L. Starck, & S. W. Digel, “Fast Poisson noise removal by biorthogonal Haar domain hypothesis testing,” *Statistical Methodology*, 2008, 5(4), 387–396.