Prediction of coding region and mutations in Human DNA by effective numerical coding and DSP technique

Subhajit Kar
Department of Electronics
West Bengal State University
Kolkata,India
subhajitkar.wbsu@gmail.com

Madhabi Ganguly
Assistant Professor
Department of Electronics
West Bengal State University
Kolkata,India
ray_madhabi@yahoo.co.in

Subhro Ghosal
Assistant Professor
Department of Electronics
APC College
Kolkata,India
subhroapc@gmail.com

2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) | 978-1-7281-8529-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCIS51004.2021.9397102

***Abstract*—Proper numerical conversion of DNA sequences into digital signals unbolts the possibility to employ digital signal processing tools for analyzing genomic data. In this paper, a given DNA sequence is converted into numerical representation based on two bit binary representation of nucleotide bases Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). The four bases A, C, T and G in a DNA primary sequence groupedintopurine{A,G}/pyrimidine{C,T}, amino{A,C}/keto{G,T} and weak-H bond{A,T} /strong- H bond{C,G}.Digital signal processing tools can be applied to the generated numerical sequence to predict the exons or coding regions in a gene, un-translated regions as well as any substitution mutation in coding region. Two staged digital filter comprises of a bandpass filter which essentially extracts the period three components from the sequence and secondly a low pass filter to eliminate high frequency noise present in the output spectra due to long range correlation of nucleotide bases are used. The locations of exons and untranslated regions found in the experiment conform to the original one as obtained from GenBank database.**

Keywords—DNA (Deoxyribonucleic-Acid), Digital Signal Processing (DSP), Mutation, IIR (Infinite impulse response) Filter, Untranslated region, Exon region.

I. INTRODUCTION

Bioinformatics in today's science has provided a major breakthrough in medical research. Use of bioinformatics data such as genomic sequence analysis has great potential to detect any abnormality in the human genome and hence paved the way for their rectification using very recent biological technologies like CRISPER (gene editing technology). Recently, DSP techniques are found to be very useful in every research area of bioinformatics for diagnosis and disease management [1,2,3,4]. It has the advantage that it can analyze a great amount of data within a few milliseconds and compare these data with a given reference set to find abnormalities in a gene. DSP techniques can identify hidden periodicities, nucleotide distribution, and features that cannot be revealed easily by conventional methods such as DNA symbolic and graphical representation otherwise it will require much tedious, expensive, time consuming biological process. DNA is a Deoxyribonucleic acid that contains the genetic instructions used in the development and functioning of all known livingorganisms. The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine(C), and thymine (T).

Mutation in a gene can lead to serious disease because a mutated exon can code for a faulty protein which essentially disturbs the biological process of the body. Various forms of mutation could be found in DNA like base substitutions, indels, and duplication. Frameshift mutation which is responsible for many diseases and disorders in humans is a type of indel where insertion or deletion of multiple bases not divisible by three took place.

To find the protein coding regions and untranslated regions present in the gene we have employed three-base periodicity factor which refers to the sharp peak at frequency f =1/3 in the Power spectrum. The property arises in the exon region due to short-term correlation present in the bases and codon bias [5]. Most of the GSP algorithms are based on the 3 base periodicity property, as this feature is very prominent in most of the genes. Exons are responsible for coding proteins. Figure 1 describes the protein formation from gene with coding and untranslated regions.

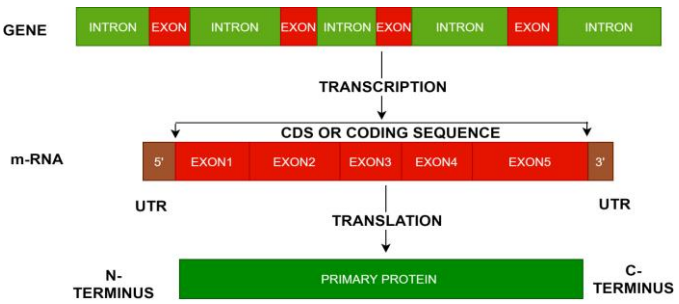

Figure 1:Central dogma of molecular biology. Different arrangement of exons produce different proteins

II. MATERIALS AND METHOD

A. Dataset

Sequences of the Dataset for the study are taken from NCBI website. These sequences are part of various well-known datasets like HMR195, BURSET & GUIGO [6, 7]. The dataset contains various types of sequences like Circular RNA, LncRNA (Long non-coding RNA), mRNA and genes. The characteristic features of the sequences are described in the Table 1.

TABLE I. DEMOGRAPHIC OF THE DATASET USED IN THIS STUDY

Accession Number	Types of Seq	Sequence length	Description	Exons or CDS
NR_130107	Long Non Coding RNA	1913 bp	Homo sapiens gastric carcinoma proliferation enhancing transcript 1	1-1903
NR_004855	Long Non Coding RNA	500 bp	Homo sapiens hepatocellular carcinoma up-regulated long noncodingRNA	1-484
AF044311	Gene	4606 bp	Homo sapiens gamma-synuclein gene	53-173, 988-1029, 1356-1483, 3953-4024, 4314-4334
AF059734	Gene	2401	Homo sapiens homeodomain transcription factor (HESX1) gene	335-491, 1296-1495, 1756-1857, 1953-2051
M10051	mRNA	4723	Human insulin receptor mRNA	139-4287
AB037886	mRNA	2129	Homo sapiens mRNA for NESH	499-1599
NM_133494	Circular RNA	4114	Homo sapiens NIMA related kinase 7	308-1216
NM_000573	Circular RNA	8587	Homo sapiens complement C3b/C4b receptor 1	112-6231

B. Proposed Numerical Representation

In order to analyze different properties by application of digital signal processing, DNA sequence must be converted into numerical sequence. There are different types of numerical representations ─ Electron Ion Interaction scheme (EIIP)[8], Voss representation[9], Complex number representation[10], Paired numeric representation[11], Genetic code context[12], two bit binary representation[13]. Previous studies showed that two bit binary number representation scheme greatly enhances the prediction accuracy of the DNA protein coding region[14-15].

In this study the DNA sequence is converted into two bit binary sequence according to the chemical property ─ Purine{A,G}/Pyrimidine{C,T} [16].Purines and Pyrimidines are the two families of nitrogenous bases that make up nucleic acids.

According to the proposed classification rule, the two bit binary value of the four bases are ─ A = 01, C = 00, G = 10, T = 11 which satisfy the formula A $\oplus$G = 11, C$\oplus$T = 11.

Based on this rule a sample DNA sequence is converted in the following way ─

A	C	T	G	A	T	G	C	A
01	00	11	10	01	11	10	00	01

(1)

The proposed numerical representation could find any substitutions in a coding sequence in the following way.

Transition:

In this type of mutation, a purine base is substituted by another purine or a pyrimidine base is substituted by another pyrimidine base.

If a single mutation took place in 3rd position of the sequence where base A is changed to base G then the two corresponding sequences can be given as-

C	T	A	G	A	T	C	Original seq
C	T	G	G	A	T	C	Mutated seq

(2)

Corresponding to the classification rule above two sequences become:

00	11	01	10	01	11	00	Encoded original seq.
00	11	10	10	01	11	00	Encoded mutated seq.

(3)

To find out the mutation in the given sequence an operating rule based on logical EX-OR operation has been introduced [17].

Using the operating rules in the two sequences obtained from(3) a new sequence is generated. The analysis of this sequence will find out the mutation in the given sequence.

00	00	**11**	00	00	00	00

(4)

Here the subsequent zeros imply no change in the mutated sequence with respect to the original sequence. The bold portion indicates a change in the sequence and it is originated from the same class. As the original sequence contains 'A' in 3rd position so it can be surely said that it is changed to 'G' and vice-versa as both belong to the same class Purine. In

similar way it can be definitely said that if the original sequence contains 'C' in that position then it must have been altered to 'T' and vice versa as both of them belong to the same class Pyrimidine.

Transversion:

In transversion procedure, a purine base is substituted by a pyrimidine base and vice-versa.

Assuming a single mutation took place in 5th position of the sequence where base A is changed to base T the corresponding sequences are obtained as:

A	T	C	A	A	T	G	Original seq
A	T	C	A	T	T	G	Mutated seq

(5)

Corresponding to the classification rule above two sequences become,

01	11	00	01	01	11	10	Encoded original seq
01	11	00	01	11	11	10	Encoded mutated seq

(6)

Applying the operating rules in two sequencesobtained from (6) a new sequence is generated below.

00	00	00	00	**10**	00	00

(7)

The bold portion of the above sequence indicates a mutation between different class i.e., if the original base is 'A' then it may be changed to base 'C' or base 'T'. But the result '10' suggest 'A' must be changed to 'T' as A ⊕T = '10' otherwise it must be '01' as in the case of A ⊕C = '01'. In similar way all the other point mutations could be located very easily.

C. Design of Digital Filter

Choosing an appropriate filter is very much essential because digital filters have the potential to reduce or enhance certain properties of a signal [18]. We have utilized a 3rd order IIR elliptical bandpass filter with passband centered at 2π/3 to extract the three base properties expected in exonic regions of eukaryotic genes. The output PSD contains high frequency noises which are expected due to long range correlation present in the nucleotide bases. To remove noise we have incorporated a moving average filter and allowed the signal to pass through it. Finally we get a smooth PSD with peaks identifying the exons present in the sequence. We have implemented the weighted moving average filter using the Gaussian window of length 30 which gives very good results.

According to the proposed representation a DNA sequence " ACATGAC……" would result in a single sequence representation given as: x[n] = {01000111100100……} corresponding to the Purine/Pyrimidine classification. Here *n* represents the base index. This proposed representation utilizes a useful DNA structural property in representation, in addition to reducing complexity in subsequent processing. The spectral content of the signal can be measured using the given formula: PSD[k] = $|X[k]|^2$ where, $X[k]$ is the filter output of the indicator sequence $x[n]$[19].

D. Evaluation Parameters

The evaluation of the implemented coding measure is done using various parameters at nucleotide and exon level. To discriminate between introns and exons a proper threshold value is calculated using the following formula:

$$T_m = \{(\text{mean } P_{3i} + \text{mean } P_{3e}) + \text{std}\}/2 \quad (8)$$

Where mean P_{3i} indicates mean period three values obtained from introns & mean P_{3e} indicates mean period three values obtained from exons. std stands for standard deviation [20]. In addition, Receiver operating characteristic curve (ROC) at various threshold levels are calculated to reflect the accuracy of the method.

III. Results And Discussions

The converted numerical sequences is applied through IIR band-pass filter and weighted moving average FIR filter respectively to get the resulted power spectral content diagram. The output power spectral density plot for the coding rule is shown in Figure 2 for gene AF044311. The peaks above threshold in the Figure 2 are predicted as exons whereas comparatively low amplitude regions are predicted as introns. Figure 3 depicts the untranslated and coding regions of AF044311 mRNA. All the PSD plots are normalised to get the best output. Various parameters like the signal to noise ratio, standard deviation, mean are derived from the PSD plot.

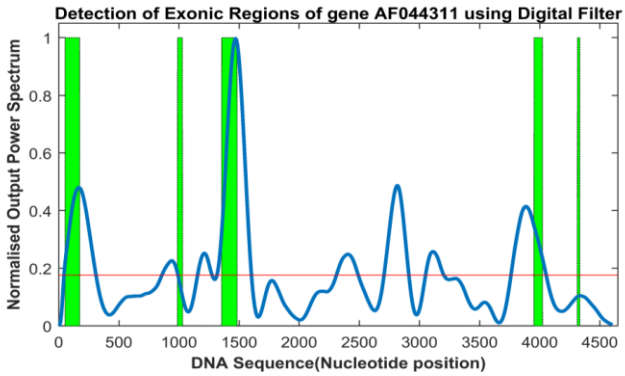

Figure 2: Digital filtering technique to predict the exons of gene AF044311. The shaded regions denote the actual exons as specified by GeneBank.

The coding regions of the selected sequences could be assumed from the power spectrum using an appropriate threshold value. For threshold selection mid-threshold value is considered as discussed in previous section..

Various evaluation parameters are measured to find the accuracy of classification rule to distinguish between introns and exons. Results are listed in Table 2.

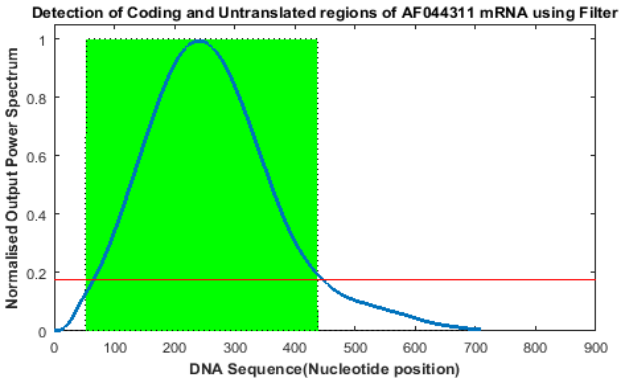

Figure 3: Digital filtering technique to predict CDS and Untranslated regions of AF044311 mRNA. The shaded region denotes the actual CDS while initial and trailing unshaded regions denote the 5' and 3' UTR respectively.

TABLE II. OUTPUT OF THE VARIOUS EVALUATION PARAMETERS

Sequence Accession Number	**Mid Threshold Value**	**Discrimination Measure**	**Miss Rate**	**Wrong Rate**	**Signal Noise Ratio**
AF044311	0.177	1.4	0 .2	0 .5	5.47
AF059734	0.230	1.33	0.25	0.2	3.70
M10051	0.2804	16	0	0	2.98
AB037886	0.232	3.4	0	0	3.65
NR_130107	0.2437	2.5	0	0	2.76
NR_004855	0.4257	7	0	0	1.92
NM_133494	0.34	3	0	0	5.6
NM_000573	0.29	1.8	0	0	3.1

The result shows very good discrimination between coding and non-coding regions and hence proved to be well efficient. Low miss rate and wrong rate implies all the exons are identified correctly with very small error. Accuracy of the method also measured using the ROC curve with the calculation of the area under the curve. The area under the ROC curve (AUC) for the gene AF059734 is calculated as 0.77, for the gene AF044311 it is 0.88. In addition to the coding region, the study also investigated the boundaries of untranslated regions that are part of mRNA but do not take part in translation and hence not a part of the protein. Since LncRNA does not code for protein so, only the genes and mRNAs are considered in this case. For circular RNAs both the untranslated regions are covalently linked together and hence it is difficult to discriminate between 5' and 3' regions. The findings are listed in Table 3. Figures in the bracket are the original boundaries as obtained from NCBI website.

TABLE III. PREDICTED RANGES OF CODING AND UNTRANSLATED REGIONS

Sequence Accession Number	**5' UTR**	**CDS**	**3'UTR**
M10051	1-253 **(1-138)**	254-4252 **(139-4287)**	4253-4722**(4288-4723)**
AB037886	1-519 **(1-498)**	520-1454 **(499-1599)**	1455-2129**(1600-2129)**
AF044311	1-66 relative to mRNA **(1-52)**	67-446 relative to mRNA**(53-436)**	447-708 relative to mRNA**(437-708)**
AF059734	1-165 relative to mRNA**(1-334)**	166-772 relative to mRNA**(335-892)**	773-892 relative to mRNA**(Not Exist)**
NM_133494	NA	397-733 relative to mRNA**(308-1216)**	NA
NM_000573	NA	411-6277 relative to mRNA**(112-6231)**	NA

The above observations clearly suggest that period three based filtering approach is suitable to identify the translated and untranslated regions with great accuracy. The results are tallied with the GenBank data collected from the NCBI website to assess the method. The coding regions in a gene are classified as 'CDS' in the GenBank dataset and can be accessed in the separate FASTA file. Finally, the effect of various types of mutations on filtered spectrum is investigated in this study. The outcomes for mRNA AB037886 and gene AF044311 are listed in tables4 and 5.

TABLE IV. TABLE 4: EFFECTS OF MUTATION ON PSD OBTAINED FROM SEQUENCEAB037886

Type of Mutation	Location	Peak Amplitude of Spectra	Mean	Standard Deviation	Signal To Noise Ratio
Without Mutation	1-2129	0.8344	0.2281	0.2359	3.6580
Frameshift	Single base deletion at 500	0.8553	0.2270	0.24	3.7683
Frameshift	Double base deletion at 500 and 501	0.8481	0.2445	0.2449	3.46
Indels	Triple base deletion at 500,501 and 502	0.8343	0.2290	0.2363	3.64
Frameshift	Single base insertion at 601 (G)	0.8148	0.2075	0.2072	3.9265
Frameshift	Double base insertion after600 (GA)	0.7942	0.2151	0.21	3.6927
Indels	Triple base deletion after 600 (GAG)	0.842	0.2251	0.2361	3.7434

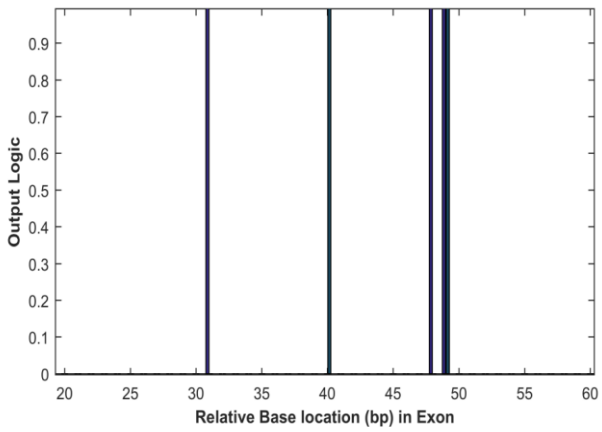

Figure 4: Spotting substitution mutation using the proposed coding scheme in gene AF044311

TABLE V. EFFECTS OF MUTATION ON PSD OBTAINED FROM SEQUENCE AF044311

Type of Mutation	Location	Peak Amplitude Spectra	Mean	Standard Deviation	Signal To Noise Ratio
No Mutation	53-173 988-1029 1356-1483 3953-4024 4314-4334	0.47 0.23 0.98 0.41 0.11	0.1966	0.1718	5.47
Substitution	Base Position 121 'G' replaced by 'A' & Position1441 'G' replaced by 'C'.	0.45 0.22 0.99 0.40 0.10	0.178	0.1693	5.6
Substitution	Base Position 61-62 'CT' replaced by 'AC' & Position 4000 'AG' replaced by 'TA'.	0.43 0.23 0.99 0.39 0.11	0.1762	0.1682	5.66

Summarizing the above tables, it is obvious that frameshift mutation reduces the spectral height significantly although significant changes in spectral height in case of substitution mutation are not seen. In some cases, it is observed that spectral height increases due to insertion or deletion of 3n+1 or 3n+2 bases [21]. Thus, it is very much difficult to find out the substitution mutation using the spectral height method. Wavelet analysis can provide a time-frequency scalogram to visualize thedifference between normal and mutated sequences but it is not effective to locate point mutations [22].The proposed method of numerical representation provided a solution by exactly locating the mutation spot and the type of substitution. To investigate the efficacy of the proposed coding procedures we have manually substituted the base 'A' with 'T', 'C' with 'A', 'C' with 'G' and 'C' with 'T' in the 31st, 40th, 48th and 49th base positions corresponding to 4th exon of gene AF044311. The mutations will generate a coding sequence according to the proposed way such that the mutation in the same class corresponding to 49th position produces '11', the mutation in a different class corresponding to 31st and 48th positions produces '10' and mutation in different class in 40th position will produce '01' while rest of the positions are '00'indicating no substitution took place. The process is demonstrated in the fig 4 where bit '1' is represented by a blue or grey line and bit'0' represented by no bar in respective positions.

IV. CONCLUSIONS

The proposed algorithm found efficient in predicting the exonic regions responsible for coding protein as well as untranslated regions of mRNA. At the same time single or multiple substitution mutations taking place in the exons found effectively. The implemented algorithm will be useful to design effective software that can easily scan a DNA sequence to detect any abnormalities in a coding sequence. The main advantages of the method are it is less time consuming, and cost effective compared to the current wetlab technologies. Cancer is the most common human genetic disease caused by mutations in a number of growth controlling genes. Other prominent diseases and abnormalities like diabetes, Thalassemia, Anemia, mental retardation, bipolar disorder are also associated with genetic mutation. Development in bioinformatics can greatly contribute to cancer treatment by pointing out the exact locations of mutation in a coding region. The next challenge to be addressed in this regard is to reconstruct the original structure of the gene so that the production of abnormal protein is replaced by the original one. Extension of the proposed work is needed to find out insertion or deletion mutations took place in the coding.

[1] S. Barman, M. Roy, S. Biswas and S. Saha, "Prediction of cancer cell using digital signal processing", Ann. Fac. Eng. Hunedoara, vol. 9, no. 3, pp. 91, 2011.

[2] S. Aydin, "Determination of autoregressive model orders for seizure detection", TUBUTAK, Turk J Elec Eng & Comp Sc, 18, (1), pp. 23-29, 2010

[3] K.B. Ramesh, Prabhu, K. S. Shankar, B.P. Mallikarjunaswamy and E.T. Puttaiah, "Genomic Signal Processing (GSP) Of Rheumatic Arthritis (RA) Using Different Indicator Sequences", IJCSMC, vol. 2, no. 5, pp. 332-337, May 2013.

[4] P. P. Vaidyanathan and B.-J. Yoon, "The role of signal processing concepts in genomics and proteomics", Journal of the Franklin Institute, vol. 341, pp. 111-135, 2004.

[5] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences", CABIOS, vol. 113, pp. 263-270, 1997.

[6] S. Rogic, A. K. Mackworth and B. F. Ouellette, "Evaluation of gene-finding programs on mammalian sequences", Genome Res., vol. 11, no. 5, pp. 817-832, 2001.

[7] M. Burset and R. Guigo, "Evaluation of gene structure prediction programs", Genomics, vol. 34, pp. 353-367, 1996.

[8] Achuthsankar S. Nair and Sreenadhan S. Pillai, "A coding measure scheme employing electron-ion interaction pseudo potential (EIIP)," Bioinformation, vol. 1, pp. 197-202, October 2006.

[9] R. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences", Phys. Rev. Lett., vol. 68, pp. 3805-3808, 1992.

[10] P. D. Cristea, "Conversion of nucleotides sequences into genomic signals", J. Cell Mol Med., vol. 6, no. 2, pp. 279-303, 2002.

[11] M. Akhtar, J. Epps and E. Ambikairajah, "On DNA numerical representations for period-3 based exon prediction", IEEE Workshop on Genomic Signal Processing and Statistics, 2007.

[12] C. Yin, "Representation of DNA sequences in genetic codon context with applications in exon and intron prediction", Journal of Bioinformatics and Computational Biology, vol. 13, no. 02, p. 1550004, 2015.

[13] S.Kar., M.Ganguly, & S. Das. "Using DIT-FFT algorithm for identification of Protien coding region in Eukaryotic gene" . Biomedical Engineering: Applications, Basis and Communications, vol. 31,no. 01, pp. 1950002,2019

[14] M.Mabrouk,"Advanced genomic signal processing methods in DNA mapping schemes for gene prediction using digital filters".Am J Signal Process ,7 (1) :12, 2017.

[15] N. Yu, Z. Li and Z. Yu, "Survey on encoding schemes for genomic data representation and feature learning—from signal processing to machine learning", Big Data Mining and Analytics, vol. 1, no. 3, pp. 191-210, September 2018.

[16] Y. H. Yao, X. Y. Nan, T. M. Wang, A new 2D graphical representation - Classification curve and the analysis of similarity/dissimilarity of DNA sequences, J. Mol. Struct. Theochem. , 764: 101-108, 2006.

[17] W. Chen, B. Liao, Y. Liu, W. Zhu and Z. Su, "A numerical representation of DNA sequences and its applications" in , MATCH Commun Math Comput Chem, vol. 60, pp. 291-300, 2008.

[18] R. Kakumani, V. Devabhaktuni and M. Omair Ahmad, "Prediction of protein-coding regions in DNA sequences using a model-based approach", ISCAS 2008, vol. 18, no. 21, pp. 1918-1921, 2008

[19] D. Anastassiou, "Genomic signal processing", IEEE Signal Process. Mag., vol. 18, no. 4, pp. 8-20, Apr. 2001.

[20] H. Kwan, B. Kwan and J. Kwan, "Novel methodologies for spectral classification of exon and intron sequences", EURASIP J. Adv. Signal Process., vol. 2012, no. 1, 2012

[21] L. Wang and L. D. Stein, "Localizing triplet periodicity in DNA and cDNA sequences", BMC BioInf., vol. 11:550, pp. 1–8, 2010.

[22] T. Meng, A. T. Soliman, M.-L. Shyu, Y. Yang, S.-C. Chen, S. Iyengar, et al., "Wavelet analysis in current cancer genome research: A survey", IEEE/ACM Trans. Comput. Biol. Bioinformatics, vol. 10, no. 6, pp. 1442-14359, Dec. 2013.

.