Journal of Computational Intelligence in Bioinformatics
ISSN 0973-385X Volume 5, Number 2 (2012) pp. 149-159

http://www.ripublication.com/jcib.htm

Wavelet Analysis of Coding and Noncoding Regions of DNA Sequences Using Integer Representation

En-Bing Lin and Mojisola Oyapero

Department of Mathematics
Central Michigan University, Mt. Pleasant, MI, 48859 USA
E-mail: lin1e@cmich.edu

Abstract

A very challenging and important task in the field of bioinformatics is being able to differentiate coding regions from noncoding regions in newly sequenced genomes. Fourier analysis can be used to differentiate these regions by using the periodic organization of three bases in coding regions. This method however has limitations because of the imperfect manifestation of the period three periodicity in many cases. It has also been proven that when wavelet is used for analyses it gives better results and is less tedious. Most other work that has been done looks majorly at the coding regions without saying much about the noncoding regions. In this work we will be using the values of wavelet coefficients to analyze these regions so that trends that are common to both coding and noncoding regions can be observed. The three types of wavelets that are employed for this analysis are coiflets, Daubechies and biorthogonal. Denoising will be done using four different threshold selection methods (Rigorous SURE, Heuristic SURE, Minimax and sqtwolog). The signal to noise ratio (SNR) and the percentage root mean square difference (PRD) will be used to evaluate the denoising methods. Finally compression of the signals will be carried out.

Keywords: Wavelets, Coding region, Non-coding region, DNA sequence, threshold, SNR, PRD.

Introduction

Every Species carries a genetic code which plays a vital role in its replication and thus each species possesses its own distinct biological features. There have been several studies on DNA sequences because of its relevance and importance to the genetic makeup of different species. A DNA sequence is a finite string from the alphabet N =

{A, C, G, T} of nucleotides. [1] Knowledge of DNA sequences has become very important for basic biological research, other research areas that use DNA sequencing are fields such as diagnosis, biotechnology, forensic biology and biological systematics. The DNA material in chromosomes is composed of coding and noncoding regions. The coding regions are known as genes and contain the information necessary for a cell to make proteins. [2]

Different methods have been used to identify protein coding regions over the years which include Genscan algorithm [3] and MZFF method. [4] Exploring the measure of spectral content in DNA sequences based on the fact that coding regions show a periodic organization of three bases, which is not found in noncoding regions. A method based on a modified Gabor-Wavelet transform (MGWT) has also been used. This novel transform is tuned to analyze periodic signal components. [5]

Another method computes the 3-base periodicity and the background noise of the stepwise DNA segments of the target DNA sequences using nucleotide distributions in the three codon positions of the DNA sequences. [6] Vaidyanathan, P.P and Byang-Jun Yoon used the digital filters to extract the period-3 component to predict gene locations. [7]

This periodicity shown in coding regions is observed to be imperfect since it is absent and non-uniform in some coding regions. [5] It has been observed that wavelet transforms are less tedious and can be used to probe localized structure of DNA sequences. [8] Wavelets can be used to analyze coding and noncoding regions since the wavelet coefficients of both the coding and non-coding regions exhibit different traits hence we can take the wavelet transforms and be able to differentiate both the coding and noncoding regions.

It has been established in literature that wavelets are localized in both time and frequency while Fourier transform is only localized in time. Wavelets thus give better results when used in multiresolution analysis. Its properties results in some useful applications like compression and denoising which cannot be done by using Fourier transforms. [13] The objectives of this paper include doing wavelet analysis and carrying out denoising and compression. The continuous wavelet transform is taken and the behaviors of both the coding and noncoding regions are studied and then compared using the three dimensional plot (3Dplot), scalograms, the distribution of its means and standard deviations. It is observed that that there is a consistent pattern for the coding regions and also for the noncoding regions.

Materials and Method

Tools and Materials

Wavelet coefficients of coding and noncoding regions are distributed differently. DNA data of three different Eukaryotic organisms; Ateles geoffroy, Bos Taurus and Anolis Carolinensis with NCBI accession numbers U04852, L09603 and L31503 respectively are analyzed. The DNA sequences that are used for analysis in this paper were used by Jishnu S. and Deepa P. Gopinath in their paper. [9] Three types of wavelets will be used for analysis (coiflets, Daubechies and biorthogonal) and their performances will be compared. This work will be using integer representations for

the DNA sequences in order to carry out analysis using wavelets. The choice of representation is based on the fact that both the coding and noncoding regions will be differentiated. The DNA sequence used here is downloaded from the NCBI website. The work is done using MATLAB.

Method

The DNA sequence is downloaded from the NCBI website and it is in character form. It is converted into numerical form using the integer representation. [10] The scheme used here is as follows: A=1, G=2, T=3, C= 4. Once the DNA sequence has been converted using the integer mapping scheme, wavelet transform is applied for analysis.

There are different types of wavelets; however for the purpose of this work three different types of wavelets are applied. They are coiflets, Daubechies and biorthogonal. The results from this analysis are compared for the three wavelets. Furthermore denoising and compression are carried out for these sequences.

The continuous wavelet transform is defined as:

$$[w_\psi x(t)](a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi^*\left(\frac{t-b}{a}\right)dt \quad a > 0,\ b \in R. \qquad (1)$$

Where the symbol * represents the complex conjugate, x(t) is the given signal (DNA sequence) and ψ is a wavelet.

The discrete wavelet transform is defined as:

$$[Dw_\psi x(n)](a,b) = \sum_{n \in Z} x(n) g_{j,k}(n), \quad a = 2^j, b = k2^j, j \in N, k \in Z. \qquad (2)$$

Where g's are the coefficients of the wavelet equation associated with ψ. [13]

Denoising is carried out using four different threshold selection methods which are Rigorous SURE, Heuristic SURE, Minimax and Sqtwolog. The signal to noise ratio (SNR) and the percentage root mean square difference (PRD) are used to evaluate the denoising methods. Analysis is carried out separately for the wavelet transforms of the coding and noncoding regions and comparisons are made. Global comparisons are made by normalizing since the coding and noncoding regions have different lengths. This is done by dividing the wavelet coefficients by the maximum values.

Results and Discussion

Coding and Noncoding Regions

The DNA sequences taken for analysis are in character form; hence they are converted by using the integer representation. First, the coding and noncoding regions of the Anolis Carolinensis genomic sequence are analyzed. The wavelet transform of coding and noncoding regions are similar when the coiflets and biorthogonal wavelets are used, while it looks slightly different for Daubechies. See figures 1 and 2

respectively for scalograms of a coding region and a non coding region using the three different types of wavelets.

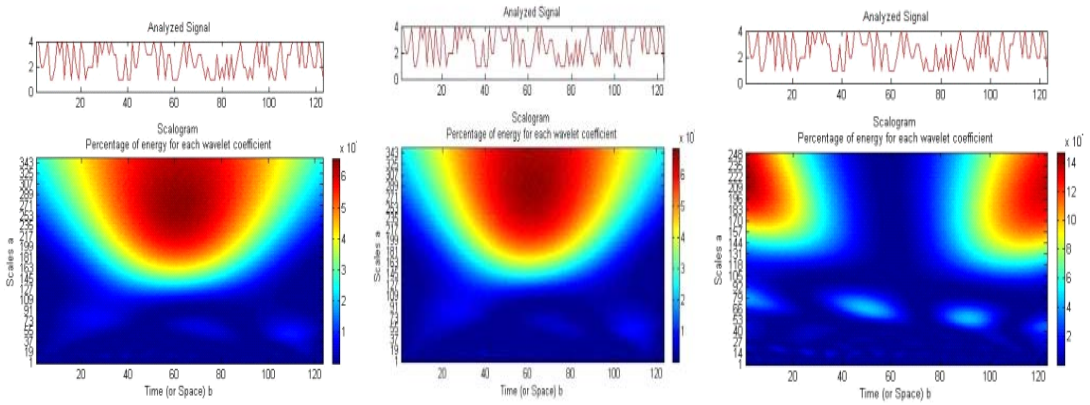

Figure 1: Wavelet transform of a coding region using coiflet, biorthogonal and Daubechies wavelets respectively.

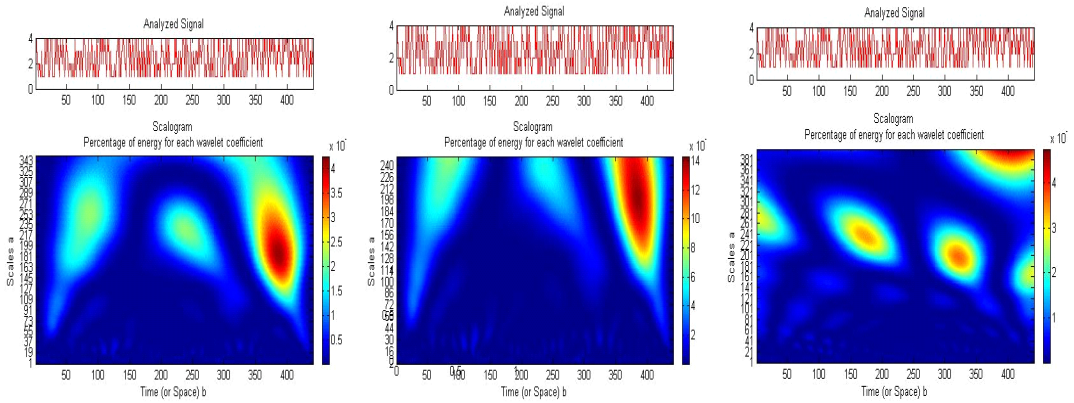

Figure 2: Wavelet transform of a noncoding region using coiflet, biorthogonal and Daubechies wavelets respectively.

It is observed from figures 1 and 2 that the coding and noncoding regions have different structures using the three different types of wavelets. Next, the distribution of the mean and standard deviation for the coding and noncoding regions using the coiflet, biorthogonal and Daubechies wavelets respectively are observed. The results are shown in figures 3 and 4, which indicate that the distributions of the mean and standard deviations for the coding regions are almost symmetrical while for the noncoding regions they are far from been normally distributed. We also see here that the results for the coiflets and biorthogonal are similar while that of Daubechies looks different.

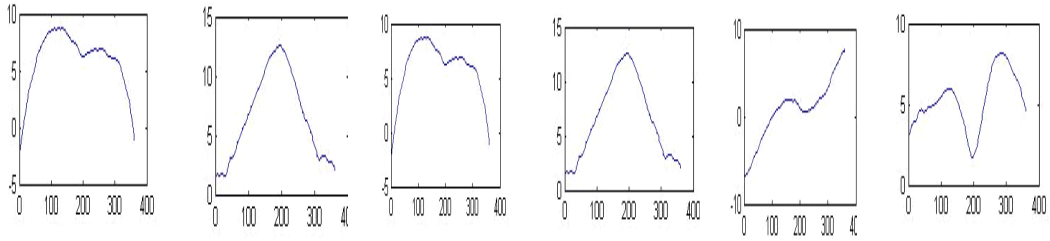

Figure 3: The distribution of the mean and standard deviation of a coding region using coiflet, biorthogonal and Daubechies
wavelets respectively.

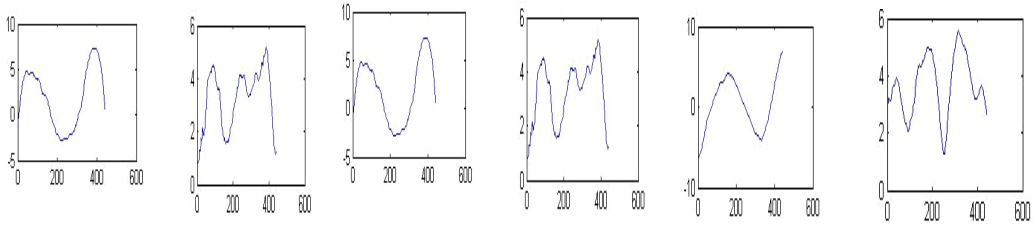

Figure 4: The distribution of the mean and standard deviation of a noncoding region using coiflet, biorthogonal and Daubechies wavelets respectively.

The 3Dplot was also observed for the coding and noncoding regions using the coiflet, biorthogonal and Daubechies wavelets respectively. The results are shown in figures 5 and 6 respectively, from which, it is observed that the 3Dplots for both the coding and noncoding regions exhibit different and distinct traits. The coding regions have a distinct peak while there is more than one distinct peak for the noncoding regions. The plots for the coiflets and the biorthogonal wavelets are similar while the pattern for the Daubechies wavelets is slightly different.

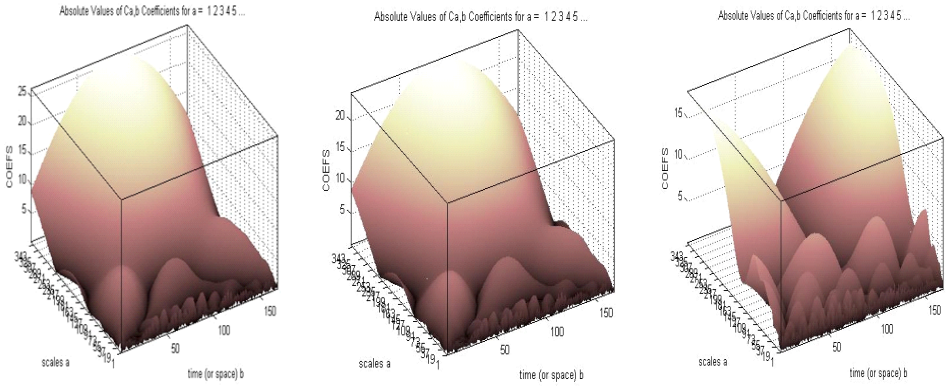

Figure 5: The 3Dplot of the wavelet coefficients of a coding region using coiflet, biorthogonal and Daubechies wavelets respectively.

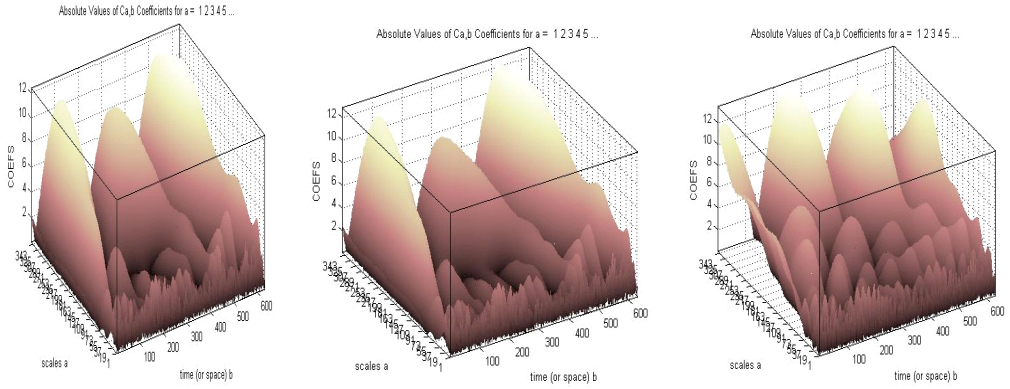

Figure 6: The 3Dplot of the wavelet coefficients of a noncoding region using coiflet, biorthogonal and Daubechies wavelets respectively.

Next a global comparison is done on the wavelet coefficients by normalizing since each of the coding and noncoding regions have different lengths. This is done by defining the function N(a) as in (3) below:

$$N(a) = \frac{w(a,b)}{\max(abs(w(a,b)))} \tag{3}$$

where w(a,b) are the wavelet coefficients is defined in (1) above. A plot of N(a) and the results are shown in figures 7 and 8 below.

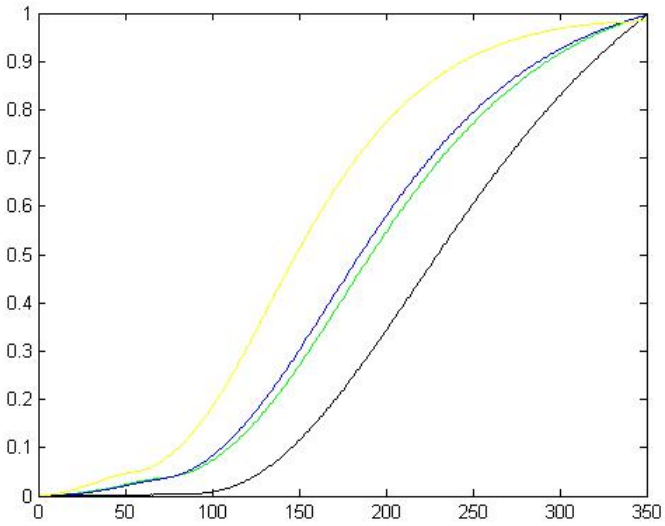

Figure 7: Plot of N(a)for all the coding regions of the Anolis Carolinensis.

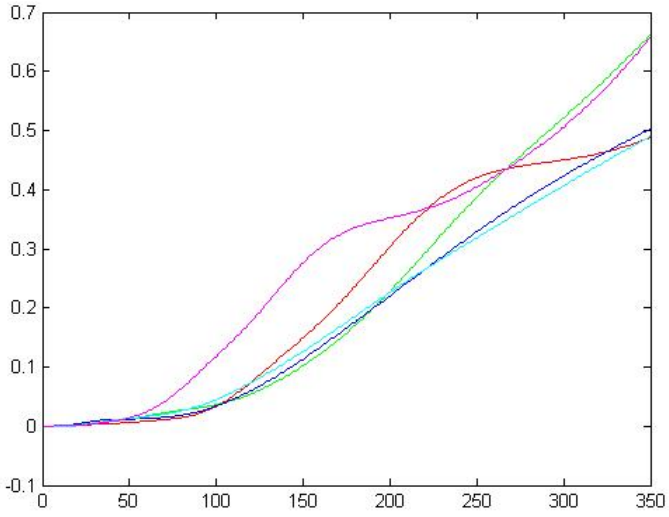

Figure 8: Plot of N(a) for all the noncoding regions of the Anolis Carolinensis.

The plots of N(a) and the results are shown in figures 7 and 8. It can be seen from figures 7 and 8 that for the coding regions, the values of N(a) run from 0 to 1, while for the noncoding regions they run from 0 to 0.7. The pattern of the plots also looks similar in these regions.

The boxplot is based on the quartiles of a data set. Quartiles are values that partition the data set into four groups, each containing 25% of the measurements. The Lower quartile is the 25th percentile, the middle quartile is the median and the upper quartile is the 75th percentile. The interquartile range is the distance between the lower and upper quartiles. The box encloses the interquartile range of the data in the box. [14] The boxplot of N(a) is given in figure 9. The boxplots show that values of N(a) for the coding region is distributed between 0-0.95 while that of the noncoding region is between 0-0.45.

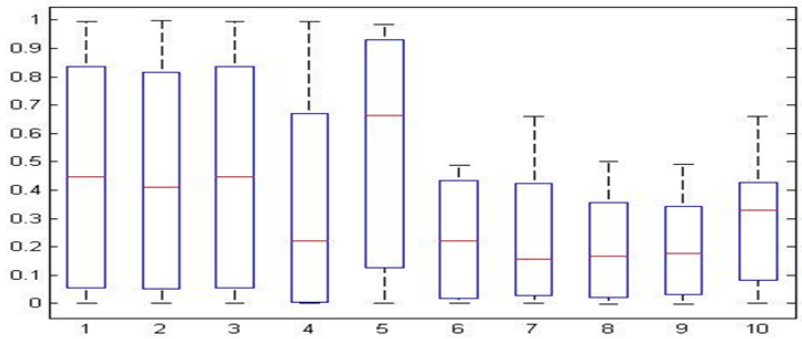

Figure 9: Boxplot of N(a) for both the coding (1-5) and noncoding (6-10) regions of the Anolis Carolinensis.

Coding regions: 1: 1341-1701 2: 3898-4066 3: 5207-5372 4: 6003-6242 5: 9361:9482
Noncoding regions: 6: 1702-3897 7: 4067-5206 8: 5373-6002 9: 6243-9360 10: 9483-9923

Compression

One of the most important applications of wavelet transforms is the compression of the signals. The main objective for signal compression is to compress the data for a specified channel bandwidth or storage requirement while maintaining the highest possible signal quality. [15] The DNA sequences analyzed in this paper are done by representations using integers hence can be represented in form of signals.

The original form of these sequences could be huge and voluminous but when the resulting signal is compressed, it will be possible to store large sequences in a small space and can be decompressed later for future use. The DNA sequences generate a growth exponentially daily and hence there is need for storage space.

Table 1: Compression of three DNA Sequences

DNA	Wavelet	Threshold Values	Retained Energy(%)	Number of zeros(%)
Anolis Carolinensis	coif	0.71	99.91	48.80
	bior	0.71	99.84	45.94
	db	0.71	99.87	46.21
Ateles geoffroy	coif	0.81	99.60	50.20
	bior	0.81	99.60	47.90
	db	0.82	99.40	46.50
Bos Taurus	coif	0.79	99.74	49.36
	bior	0.79	99.72	48.35
	db	0.80	99.68	47.60

The compression procedure contains three steps: Decomposition, Thresholding detail coefficients and Reconstruction. The higher the retained energy and the number of zeros the better performance. Compression is done using discrete wavelet transforms for the three different forms of wavelet. Global thresholding is used and the thresholding method used is removing values near zero. The threshold values, Retained Energy and the Number of Zeros are recorded. The result is given in table 1, from which, it is observed that Coiflets performed best with the highest retained energy for each sequence and the highest number of zeros.

Denoising

Denoising is another important application of wavelet transforms. It consists of restoring a useful signal from observation corrupted by additive noise. [16] Denoising is carried out on the signals from the three DNA sequences by looking at each of the coding and noncoding regions as subsignals (using integer representations) and then

taking an overall average. The four threshold selection methods that are used are Rigorous SURE, Heuristic SURE, minimax and sqtwolog. [11]These methods were compared using the three wavelets. The Signal to noise ratio (SNR) and the percentage root mean square difference (PRD) were used to evaluate our denoising methods.

The formula for the SNR and PRD are given as follows:

$$PRD = \sqrt{\frac{\sum_{n=0}^{N}(V(n)-V_R(n))^2}{\sum_{n=0}^{N}V^2(n)}} \tag{4}$$

where V(n) is original DNA sequences and V_R (n) is denoised version of the DNA sequence.

$$SNR = \log_{10}\frac{\sum_{n=0}^{N}{V_R}^2(n)}{\sum_{n=0}^{N}{S_R}^2(n)} \tag{5}$$

where $S_R(n)$ is the deformation in reconstructed DNA sequence. [12]

The result shows that the method that gave the best thresholding method is minimax with the highest SNR values and the lowest PRD values for the three wavelets. The results for SNR and PRD are shown in tables 2 and 3 respectively.

Table 2: SNR values for the three DNA sequences

DNA	Wavelet	rigrsure	sqtwolog	heursure	minimaxi
Anolis Carolinensis	coif	3.5383	3.5382	3.5382	3.5421
	bior	3.5382	3.5382	3.5382	3.5422
	db	3.5395	3.5388	3.5388	3.5429
Ateles geoffroy	coif	3.2606	3.2606	3.2606	3.2641
	bior	3.2627	3.2611	3.2611	3.2739
	db	3.2638	3.2604	3.2634	3.2735
Bos taurus	coif	3.1881	3.1881	3.1881	3.1967
	bior	3.1886	3.1878	3.1878	3.1939
	db	3.1923	3.1901	3.1913	3.1963

Table 3: PRD values for the three DNA sequences

DNA	Wavelet	rigrsure	sqtwolog	heursure	minimaxi
Anolis Carolinensis	coif	0.2806	0.2813	0.2813	0.2649
	bior	0.2812	0.2813	0.2812	0.2648
	db	0.2764	0.2791	0.2791	0.2622
Ateles geoffroy	coif	0.2965	0.2965	0.2965	0.282
	bior	0.2885	0.2974	0.2974	0.2319
	db	0.278	0.2823	0.2799	0.2312
Bos taurus	coif	0.2949	0.2824	0.2949	0.261
	bior	0.292	0.2951	0.2951	0.2705
	db	0.2789	0.2887	0.2842	0.2617

Conclusion

It has been mentioned in the introduction that the wavelet transform is an improvement over the Fourier transform since it can be used to probe localized structure of the DNA sequences. Wavelets can be used for signal noise reduction and compression. Compression was done using coiflets, biorthogonal and Daubechies; coiflets provided the best results. Furthermore, amongst the different thresholding methods used, the minimax method provides the best result. The three types of wavelets are equally good for visual representations when differentiating coding and noncoding regions. However, the coiflets and biorthogonal give very similar results. It was also observed when carrying out denoising that lower levels give better results. Finally from the global comparison, it indicates that there is a consistent trend for the three types of wavelets. More work will be done in the future to determine what happens if there are different types of DNA sequences or DNA sequences with mutations.

References

[1] Nello Cristianini, M. W., Introduction to Computational Genomics. Cambridge University Press, 2006.

[2] John M. Butler, Forensic DNA Typing. Biology, Technology, and Genetics of STR Marker. Elsevier Academic Press, 2005.

[3] Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. Journal of molecular biology, 26(1), pp: 78-94, 1997.

[4] Zhang, M.Q., Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc. Natl Acad. Sci. USA, pp: 565–568, 1994.

[5] J. P. Mena-Chalco, H. Carrer, Y. Zana, R. M. Ceaser Jr, Identification of Protein Coding regions using modified Gabor Wavelet Transform, IEEE/ACM Transaction on Computational Biology and Bioinformatics, 5(2), pp:198-207,2008.

[6] Changchuan Yin, Stephen S.-T. Yau, Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. Journal of theoretical biology, Journal of Theoretical Biology 247, pp: 687–694, 2007.
[7] Vaidyanathan, P.P., Byang-Jun Yoon, Digital filters for gene prediction applications. Signals, Systems and Computers, Conference Record of the Thirty-Sixth Asilomar Conference, vol.1, pp: 306-310, 2002.
[8] A. A. Tsonis, P. Kumar, J. B. Elsner and P. A. Tsonis, Wavelet Analysis of DNA sequences, Physical Review Letters, The American Physical Society, 53(2), pp:1828-1834, 1996.
[9] Jishnu S. and Deepa P. Gopinath, Wavelet Analysis of Coding and Noncoding Regions of DNA Sequences.10th National Conference on Technological Trends (NCTT09) 6-7 Nov 2009.
[10] Hon Keung Kwan, Arniker, S.B., Numerical representation of DNA Sequences. IEEE International Conference, pp: 307-310, 2009.
[11] Gao Chao, Zhou Shanxue, Wavelet transform threshold noise reduction methods in the oil pipeline leakage monitoring and positioning system. Journal of Electronics (CHINA), 27(3), pp: 405-411, 2010.
[12] Mikhled Alfaouri, Khaled Daqrouq, ECG Signal Denoising By Wavelet Transform Thresholding. American Journal of Applied Sciences, 5(3), pp: 276-281, 2008.
[13] Hans-Georg Stark, Wavelet and Signal Processing: An Application-Based Introduction, Springer, 2005.
[14] McClave, Dietrich, Sincich, Statistics. seventh edition, Prentice Hall, 1997.
[15] N. Jayant, Signal compression: Coding of Speech, Audio, Text, Image and Video. World scientific Publishing Co. Pte, Ltd, 1997.
[16] Michel Misiti. et. al, Wavelets and their applications. ISTE ltd, 2007.