Implementation and Development of Signal Processing Tools for Genomic data

Dissertation submitted in partial fulfillment of the requirements for the Degree of

MASTERS OF TECHNOLOGY

IN

ELECTRONICS & COMMUNICATION ENGINEERING

By

Sanyogita Sharma

Enrollment No.: 152005

UNDER THE GUIDANCE OF

Mr. Pardeep Garg

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
WAKNAGHAT, SOLAN - 173234, INDIA
May 2017

TABLE OF CONTENTS

DECLARATION BY THE SCHOLAR

I hereby declare that the work reported in the M-Tech dissertation entitled **“ (IMPLEMENTATION AND DEVELOPMENT OF SIGNAL PROCESSING TOOLS FOR GEONOMIC DATA)”** submitted at **Jaypee University of Information Technology, Waknaghat India,** is an authentic record of my work carried out under the supervision of **(Mr. PARDEEP GARG)**. I have not submitted this work elsewhere for any other degree or diploma.

( )

SANYOGITA SHARMA

Enrollment no. 152005

Department of Electronics and Communication Engineering

Jaypee University of Information Technology, Waknaghat, India

Date ( )

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

(Established by H.P. State Legislative vide Act No. 14 of 2002)
P.O. Waknaghat, Teh. Kandaghat, Distt. Solan - 173234 (H.P.) INDIA
Website: www.juit.ac.in
Phone No. (91) 01792-257999
Fax: +91-01792-245362

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech dissertation entitled **"Implementation and Development of Signal Processing Tools for Genomic data"** which is being submitted by Sanyogita sharma in fulfillment for the award of Master of Technology in Electronics and Communication Engineering by the Jaypee University of Information Technology, is the record of candidate's own work carried out by her under my supervision. This work is original and has not been submitted partially or fully anywhere else for any other degree or diploma.

Mr. Pardeep Garg

Assistant Professor
Department of Electronics & Communication Engineering
Jaypee University of Information Technology, Waknaghat.

ACKNOWLEDGEMENT

Firstly, I would like to express my sincere gratitude to my guide Mr. Pardeep Garg for his constant encouragement, patience, motivation, valuable suggestions and constant support throughout my thesis work. His guidance has helped me throughout this study and also motivated me for writing this report. I would also like to thank him for providing me his precious time whenever I needed his support and supervision. His discussion on current research issues, valuable advice and suggestions encouraged me in innumerable ways and helped me to improve my intellectual maturity.

I give my special thanks to Prof. S.V. Bhooshan, Head of the Electronics and Communication Engineering Department, for all the facilities provided. I am also very thankful to all the faculty members of the department, for their constant encouragement during the project. I also take the opportunity to thank all my friends who have directly or indirectly helped me in my thesis work. Last but not the least I am very much thankful to God for showering warm blessings and my parents for their moral support and continuous encouragement while carrying out this study.

Date: 01-05-2017 **Sanyogita sharma**

ABSTRACT

In literature review we have studied number of algorithms which are used to measure the power spectrum of the DNA sequences with N/3 period property and are helpful in distinguishing the protein coding region and the non-coding region in the DNA sequences. A large number of algorithms have been advanced in last few years for the gene prediction. In my work I have firstly done the analysis on various window functions, digital signal processing algorithm and the mapping scheme. The hanning window function is being used as it provide the best performance for big data sets then we make use of short time Fourier transform as a DSP algorithm which provide better performance with the hanning window. After this analysis I have used another digital signal processing algorithm maximum entropy method which provides much better performance as compared with the short time Fourier transform. The mapping schemes paired- numeric, EIIP is used for numerical conversion of data. The performance depends on the spectral analysis of the windowed data. At each position of the DNA sequence the Fourier transform is calculated of the sequence.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVATIONS

DNA	Deoxyribonucleic Acid
CT-STFT	Continuous Time Short Time Fourier Transform
DT-STFT	Discrete Time Short Time Fourier Transform
DSP	Digital Signal Processing
FFT	Fast Fourier Transform
DFT	Discrete Fourier Transform
STFT	Short Time Fourier Transform
MEM	Maximum Entropy Method
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
DSP	Digital Signal Processing

CHAPTER 1

1 INTRODUCTION

The Cell is the structural and functional unit of all living organism. Living organism cab be classified into two parts [1]

1) Prokaryote – A single cell microscopic organism which has no nucleus with a membrane also no other specialized organs, includes the bacteria [1].

2) Eukaryote-A eukaryote is also a organism whose cells contain a nucleus and other organelles enclosed within membranes [1].

1.1 DNA

Deoxyribonucleic Acid (DNA) includes all the necessary information to run a cell. DNA can be considered as the blue print for cell machinery [1]-[2]. DNA is made up of linear chains of subunits called nucleotides [1].The four possible nucleotide bases are A (Adenine), C (Cytosine), G (Guanine) and T (Thymine). DNA encodes information for building proteins. Genes are the contiguous stretch which is along DNA. A gene has two regions called exons and introns [3]. Introns and exons are considered to be parts of genes. Exons are said to be the protein coding region and introns are known as non coding region [1]-[3]. Introns are parts of genes that do not relate with the coding for proteins directly [2]. The DNA generally contains chromosomes which are combination of genic and intergenic regions as shown in figure 1.1 [4]. The region which includes exons has the period 3-property that is generally not found in region which includes the introns [1]-[4]. The region that actually encodes the gene product is smaller than the introns are known as exons.

Deoxyribonucleic acid a self-replicating material.DNA is present in all living organisms as the main constituent of chromosomes [1]-[5]. The nucleotide carries the cell's genetic information and hereditary characteristics via its nucleotides. In DNA each individual nucleotide of the DNA sequence is converted to numerical values through a mapping scheme [3]. A number of techniques have been presented to map the DNA nucleotides into the numerical values [1]. The different types of mapping schemes are (1) compact representation (2) minimum redundancy (3) compatibility with different mathematical tools (4) distance between all nucleotides pairs are equal [3]-[4]. DNA mapping can be completed with various window functions and techniques like DSP Techniques which are very important in genomics DNA research to disclose genome features to recognize the periodicities which may be hidden [1]-[2]-[5]. After the mapping is being done, the signal processing techniques can be implemented to classify period-3 region in the given DNA sequence [6]. Genomic information is said to be digital in a every sense.

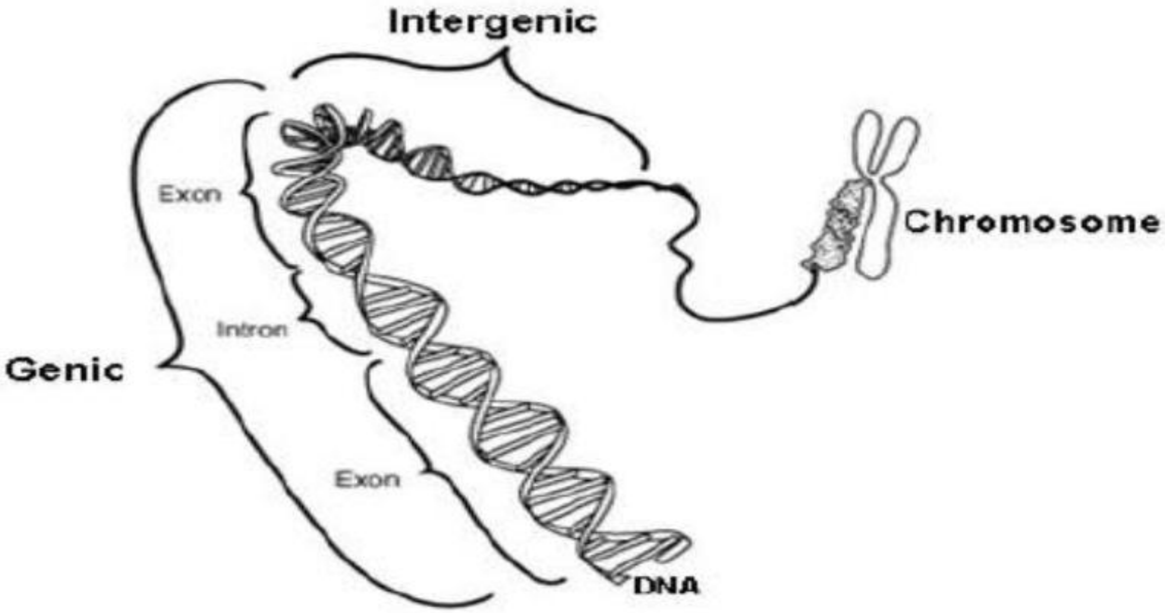


Fig 1.1 -Eukaryote DNA structure which is made up of genic and intergenic region [4]

1.1.1Numerical Representations of DNA Sequences

There are many types of methods which are being used to convert the data sequence into numerical form such techniques are known as mapping [1]-[6]. DNA molecules are made up of four nucleotides, adenine (A), thymine (T), cytosine(C), and guanine (G) [7].

The number of character strings of DNA molecules which are mapped into numerical Sequences [4]. In literature many approaches have been used which concentrate on gene prediction [3] .Voss representation technique is known to be an efficient method in finding the coding and non-coding regions in a DNA sequences [2]-[4]. It offers a graphical and numerical representation, base distribution has efficient spectral detector [7]. The Voss and tetrahedron techniques are equivalent to the techniques used for the formation of power spectrum but [1]-[2]-[4]

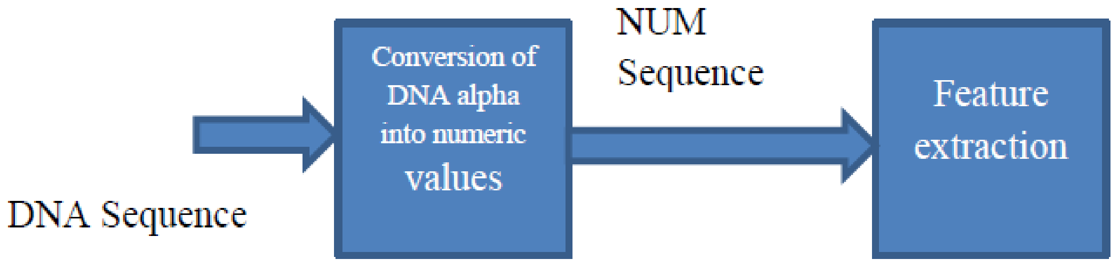


Fig 1.2- Process of mapping [5]

Firstly the conversion of numerical signals are made to extort features after that the window function is applied to the dataset afterwards the DSP algorithm is applied to the same dataset. The results provide the accurate detection of the end-points of exons. The paired numeric or Voss mapping technique are very familiarly used in estimation now days. They are used to map the given nucleotides and into the four binary sequences.

Table1.1- Comparison of mapping schemes [7]

METHODS	MERITS	DEMERITS
Voss	Efficient spectral detector of base distribution and periodicity features offering numerical and graphical visualization [4]-[7].	Linearly dependent set of representation redundancy [4]-[7].

Paired Numeric	Locate pattern and sequences in genomes [5].	
EIIP	Periodicity detection [6]	Reduced redundancy[6]

1.2USE OF WINDOW

A transform basically helps to construct signal which is in time domain to the signal which is in frequency domain so that we can analyze the number of frequencies present in a signal [4]-[8]. The spectral leakage is caused by discontinuities in the original number of periods in a signal and can be improved using window functions [3]-[4]-[7]. In digital signal processing a window function is commonly known as tapered function which is a mathematical function that is zero outside the chosen interval 2]-[6]-[7]. For explaining in detail we can consider a function that will be constant inside the interval and zero outside that interval is known to be rectangular window function [8]. When a function or a waveform sequence is multiplied by a window function, the product is also zero-valued outside the interval [3]-[6]-[9]. There are many window functions in digital signal processing which are being discussed as follows [4]:

1.2.1RECTANGULAR WINDOW

The rectangular window is said to be the simplest the window [10]

.

$$\mathrm{w(n)} = \begin{cases} 1\,, & 0 \text{ to } \mathrm{N}-1 \\ 0, & \text{elsewhere} \end{cases}$$

The sudden changes in rectangular window reduce losses and improve the dynamic range. Therefore other windows are designed to overcome this limitation [5]-[10].

1.2.2HANN WINDOW

The Hann window was named after the researcher Julius von Hann and also it is also known as the Hanning window [7]-[10]. The window function is considered that minimize the level of side lobe [10].

$$w(n) = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right)$$

1.2.3BLACK MAN HARRIS WINDOW

A Black man Harris window is a part Hanning family, it can be implemented by summing more shifted version of sinc function, which is used to minimize the level of side-lobe in the function [4]-[10].

$$w(n) = a_0 - a_1\cos(\frac{2\pi n}{N-1}) + a_2\cos(\frac{4\pi n}{N-1}) - a_3\cos(\frac{6\pi n}{N-1})$$

1.2.4GAUSSIAN WINDOW

When we take Fourier transform of a Gaussian function it is also considered to be Gaussian [10].

$$\mathrm{w(n)} = e^{-\frac{1}{2}(\frac{n-(N-1)/2}{\sigma(N-1)/2})^2}$$

1.2.5 TRIANGULAR WINDOW

The triangular window is said to be the 2nd order window. It is the product between the two rectangular windows which has width N/2 [6]-[10].

$$w(n) = 1 - \left|\frac{n-\frac{N-1}{2}}{\frac{L}{2}}\right|^2$$

1.2.6 KAISER WINDOW

The Kaiser, or Kaiser-Bessel, window that was discovered by Jim Kaiser and it is a approximation of the DPSS window using Bessel functions [3]-[7].

$$w(n) = \frac{\mathrm{I}_0(\pi\alpha\sqrt{1-(\frac{2\mathrm{n}}{\mathrm{N}-1}-1)^2}}{\mathrm{I}_0(\pi\alpha)}$$

1.2.7 BARTLETT WINDOW

In Bartlett window the ending part of the window is always zero while the ending part of the triangular window is non zero [2]-[3]. The coefficients of a Bartlett window can be computed as follows [10]:

$$\mathrm{w(k)}=\begin{cases} \frac{2k}{K}, \ 0 \le k \le \frac{K}{2} \\ 2-\frac{2k}{K}, \ \frac{K}{2} \le k \le K \end{cases}$$

1.2.8 NUTTALL WINDOW

The coefficients of nuttall window generally differ from the Blackman-Harris window but if we analyze the coefficients of blackman harris then we can see that it produces slightly lower side lobes [6]-[7]. If the nuttall window produces maximum side lobes then they can be minimized to get better output [10].

$$\mathrm{w(n)} = a_0 - a_1\cos(\frac{2\pi n}{N-1}) + a_2\cos(\frac{4\pi n}{N-1}) - a_3\ \cos(\frac{6\pi n}{N-1})$$

1.2.9 PARZEN WINDOW

The Parzen window is considered to be the approximation of Gaussian window [10].

$$\mathrm{w(n)} = \begin{cases} 1-6(\frac{\mathrm{n}}{\mathrm{N}/2})^2 \ , & 0 \ll |\mathrm{n}| \ll {}^{\mathrm{N}}\!/_4 \\ 2(1-\frac{|\mathrm{n}|}{\mathrm{N}/2})^3 \ , & {}^{\mathrm{N}}\!/_4 < |\mathrm{n}| \ll {}^{\mathrm{N}}\!/_2 \end{cases}$$

1.2.10 BOHMAN WINDOW

The Bohman window is basically the convolution or a product between the two half-duration cosine pulses [9]-[10].Also the product of a triangular window with a single cycle of a cosine in time domain [10].

$$w(n) = \left(1 - \frac{|n|}{N/2}\right)\cos\left(\pi\frac{|n|}{N/2}\right) + \frac{1}{\pi}\sin\left(\pi\frac{|n|}{N/2}\right), 0 \le |n| \le N/2$$

1.2.11 TUKEY WINDOW

The Tukey window is a form of rectangular window in which the first and last samples are equal to the parts of a cosine function. If input is r ≥1, then hann window can be obtained from tukey window [10-[11].

$$w(n) = \begin{cases} \frac{1}{2}\left[1 + \cos\left(\pi\left(\frac{2n}{\alpha(N-1)} - 1\right)\right)\right], & 0 \le n \le \frac{\alpha(N-1)}{2} \\ \frac{1}{2}\left[1 + \cos\left(\pi\left(\frac{2n}{\alpha(N-1)} - \frac{2}{\alpha+1}\right)\right)\right], & (N-1)\left(1 - \frac{\alpha}{2}\right) < n \le (N-1) \end{cases}$$

1.2.12 CHEBYSHEV WINDOW

The Chebyshev window for a specified side lobe level has the narrowest main lobe [10].

$$w(n) = w_0(n - \frac{N-1}{2})$$

1.2.13 HAMMING WINDOW

The Hamming window has the maximum side lobe level and its height is considered to be one –fifth of the hann window [12]. Hann and hamming window are related to each other. Its function is as follows [10]

$$\mathrm{w\ (n)} = \alpha - \beta \cos\left(\frac{2\pi n}{N-1}\right)$$

1.2.14 FLAT TOP WINDOW

The flat-top window is considered to be a negative-valued window function that has negligible losses in frequency domain [12].

$$\mathrm{w(n)} = a_0 - a_1 \cos(\frac{2\pi n}{N-1}) + a_2 \cos(\frac{4\pi n}{N-1}) - a_3 \ \cos(\frac{6\pi n}{N-1}) + a_4 \cos(\frac{8\pi n}{N-1})$$

CHAPTER 2

DSP TOOLS

2.1 DFT (Discrete Fourier Transform)

The discrete Fourier transform (DFT) generally changes the form of a sequence with finite length in which the samples are equally spaced [3]-[4]-[7]. This process is known to be a complex function of frequency [9]. The interval of the DTFT in which it is sampled is the reciprocal of the duration of input sequence [7]-[9]. If the particular sequence eliminates all the non-zero values present in the function, after estimating its DTFT we come to know that it is a continuous DFT [6]. The DFT generally provides the discrete samples of the original data [9]. In digital signal processing a may be defined as the quantity or signal that generally varies with time, the elements like pressure of a sound wave, radio signal and temperature readings all these elements function over a finite interval of time [12]. The DFT of a signal of length N, at frequency k is defined as follows [4]

$$F(k_f) = \sum_{n=0}^{N-1} f(n)e^{-i2\pi k_f n/N}$$

Where $F[k_f]$ is the DFT coefficient [11]. The power spectrum analysis for a DFT is basically used on the components of frequency for a signal that is suppressed by the noise [5]-[11]. For DFT power spectrum analysis of a DNA sequence the four binary sequences indicators are required [10]-[12].

2.2FFT (Fast Fourier Transform)

A fast Fourier transform (FFT) algorithm generally computes the discrete Fourier transform (DFT) of a sequence which is much faster in speed [8]-[13]. The basic idea behind Fourier analysis is that it firstly converts original signal to a frequency domain from the time domain and sometimes the opposite [12]-[13]. An FFT is a fast process it can compute by dividing the DFT matrix into small factors. It is helpful in reducing the complexity of the

DFT of which is be computed [3]-[10]-[12]. The DFT can be obtained by splitting a sequence of values [7]. An FFT gives us a new and faster way to compute the same results as of DFT. The difference in speed can be advantageous, especially for large set of data and that could be in thousands or millions [2]. The enormous speed helps in making the calculation of the DFT accurate [5]. In the existence of error, many of the FFT algorithms are more accurate than evaluating the DFT. The DFT is defined by the formula[8]

$$X_{k_f} = \sum_{n=0}^{N-1} x\,(n)e^{-i2\pi k_f n/N} \qquad k_f = 0, \dots\dots, N-1$$

2.3 STFT (Short Time Fourier Transform)

The short-time Fourier transform (STFT), or also known as short-term Fourier transform, is a type of Fourier transform which is used to determine the frequency and phase component of a signal as it changes with time[2]-[3]-[4]. The basic process for computing the STFT is to split a long time signal into shorter segments which should be of equal length [10]-[13]. Then we compute the Fourier transform of each segment separately. The result is for Fourier spectrum of each of the shorter segment [5]-[7]-[12].

2.3.1 Continous –Time STFT

In the continuous-time Transform the function which is to be transformed is multiplied by a window function that will be a non-zero function for a short period of time. Mathematically, this is written as [3]-[7]-[8]:

$$\{X(t)\}(\tau, \omega) = \int_{-\infty}^{\infty} x(t)\omega(t-\tau)e^{-j\omega t}\,dt$$

2.3.2 Discrete-Time STFT

In the discrete time Transform the data which is to be transformed are broken up into frames which could overlap each other [2]. This can be expressed as:

$$\{x[n]\}(k, \omega) = \sum_{-\infty}^{\infty} x[n]\omega[n-k]e^{-j\omega n}$$

Where $x[n]$ denotes signal $w[n]$ denotes window function. In this k is discrete and ω is continuous [14],

$$\{\mathrm{x(t)}(\tau,\omega) = [\mathrm{X}(\tau,\omega)]^2$$

2.4 MAXIMUM ENTROPY METHOD

Estimating the power spectrum of the sequence is equal to estimating a autocorrelation of the sequence [4]-[15]. If we consider the data, the autocorrelation can only be estimated for specific periods that are the small sections of the initial data $|K| < N$ [2]-[6]-[15]. As a result the autocorrelation will be set to zero for$|K| \geq N$.There are many signals which are non-zero in the lag $|K| \geq N$,these signal require the windowing function which can increase the resolution and accuracy of the estimated spectrum [4]-[7] .If we see the case of narrowband processes the autocorrelation of the processes decay slowly with K [15]. In the classical methods the autocorrelation is extrapolate (extending some method to an unknown situation by assuming that the results will have no effect with it) with zeros [11]-[15]. With the help of extrapolation the effect of window will be mitigated and we can get a more accurate estimate of power spectrum [9]. It is difficult to perform the extrapolation so the maximum entropy method is very helpful method in examining the power spectrum [4]-[7]. If we consider a autocorrelation $r_x(k)$ for lag$|k| \geq n$,then we will extrapolate the $r_x(k)$ for $|k| > n$.The extrapolated value is denoted by $r_e(k)$ which is as follows [6]-[15]:

$$\mathrm{P_x}\left(\mathrm{e^{j\omega}}\right) = \sum_{-\mathrm{n}}^{\mathrm{n}} \mathrm{r_x(k)e^{-jk\omega}} + \sum_{|\mathrm{k}|>n} \mathrm{r_e\ (k)e^{-jk\omega}}$$

The $P_x(e^{j\omega})$ is considered to be a valid power spectrum [7]-[8]. $P_x(e^{j\omega})$ should be real valued and non negative for all values of ω. This value to be non negative does not guarantee the unique extrapolation [15]

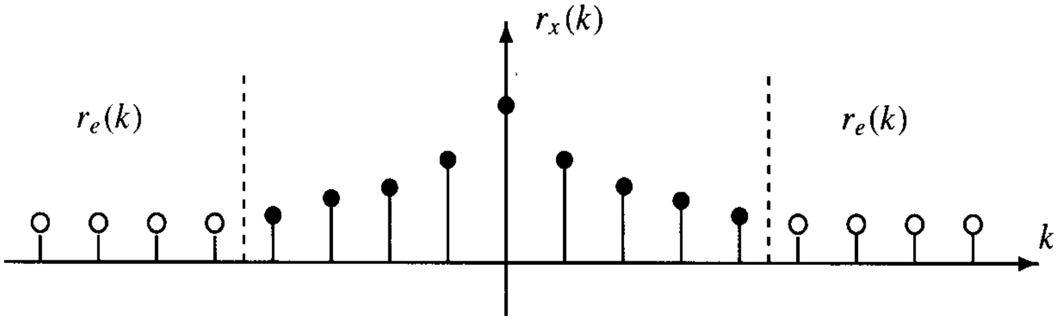


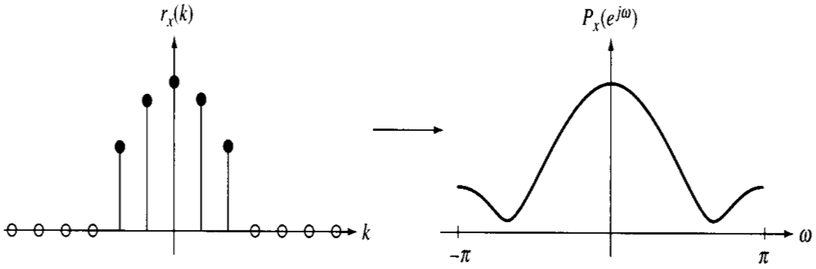
Fig2.1- Extrapolating the autocorrelation sequence[15]

The $P_x(e^{j\omega})$ is considered to be a valid power spectrum [3]. $P_x(e^{j\omega})$ should be real valued and non negative for all values of ω. This value to be non negative does not guarantee the unique extrapolation [12]-[15]. Therefore some additional constraint needs to be imposed to the allowable extrapolation [15]. One such idea was suggested by the BURG to maximize the entropy. Basically entropy is a measure of randomness and uncertainty [3]-[5]-[13]. A maximum entropy extrapolation is equivalent to finding the sequences with autocorrelation that make x(n) as random as possible [15].

For a Gaussian random process the power spectrum will be as follows [3]

$$H(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln P_x(e^{j\omega})d\omega$$

The maximum entropy power spectrum is the one that maximises the above equation



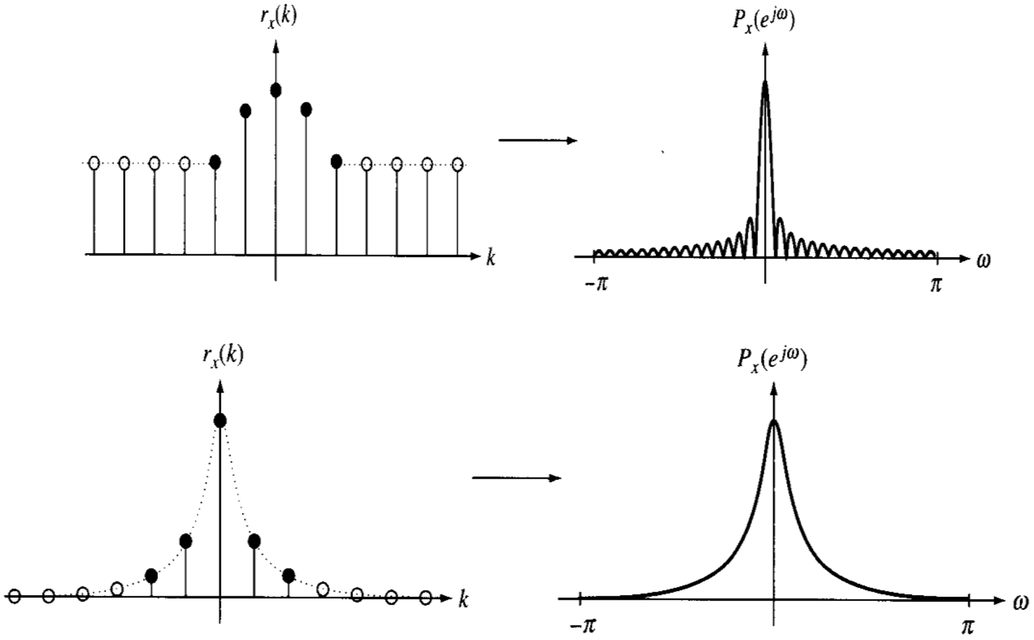


Fig 2.2- Different extrapolations of partial autocorrelation sequences and their corresponding power spectral densities [15]

The MEM spectrum can be computed as follows:

(1) Firstly the autocorrelation equations are solved for the all pole cofficients [5]- [15].

(2) Then the MEM spectrum is produced by incorporating these parameters[15]

(3) If the P_{mem} ($e^{j\omega}$) is considered to be an all pole power spectrum then $r_x(k)$ it satisfies the Yule Walker equation that is[3]-[15]

$$r_x(l) = -\sum_{k=1}^{p} a_p(k_f) r_x\,(k_f - l) \qquad \text{for } l > 0$$

The maximum entropy method basically extrapolates the autocorrelation sequence with then noise suppression [7]. The MEM has been studied as a spectrum analysis tool. In case we have any information or constraint in a process where we have set of some autocorrelation values the best way to estimate the power spectrum is to find the Fourier transform of the given autocorrelation sequences along with the extrapolation which performs the maximum entropy extrapolation [6]-[9]-[15].

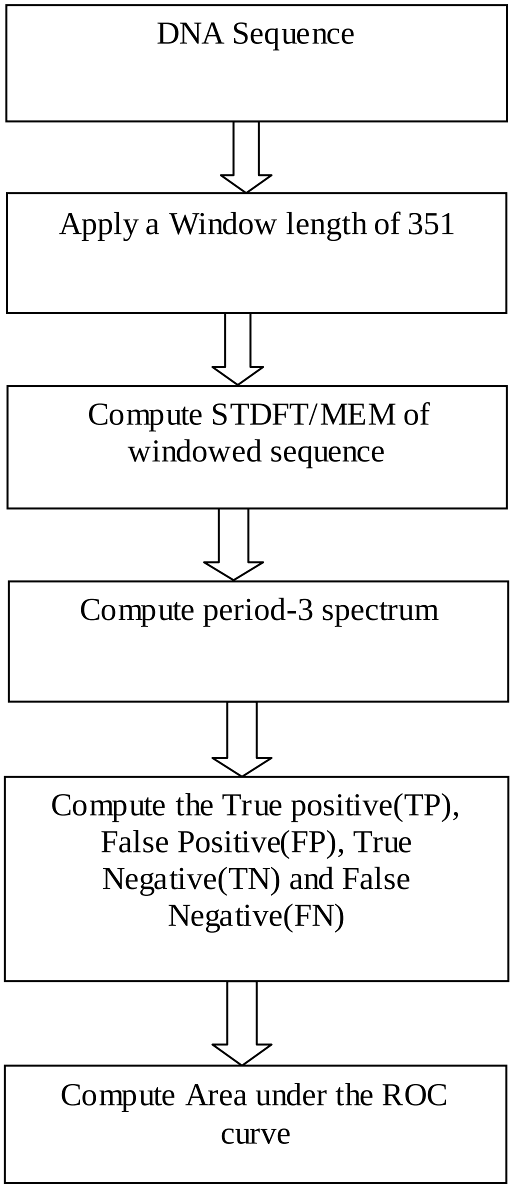
The MEM estimate is better than the classical methods used in the power spectrum estimation which depends on what type of process is being analyzed [3]. By this method we can improve the spectral density. The autocorrelation function matches with the known values. It is the basic application for maximum entropy modeling and is used where data is in spectral form [4]-[6]. The use of this technique is based on the sources of the spectral data and also is dependent on the amount of information known about the spectrum that will be applied to the model [15].

CHAPTER 3

OBJECTIVE

After the study of different window functions and Digital signal processing algorithm, it is analyzed that there is a large difference in the performance of the different window functions [11]. The 3-base periodicity is used as a basic element [4]-[11]. The basic algorithm for the identification of protein coding region is as follows

Table 3.1-The flow chart for protein coding region identification [16]


DNA Sequence
Apply a Window length of 351
Compute STDFT/MEM of windowed sequence
Compute period-3 spectrum
Compute the True positive(TP), False Positive(FP), True Negative(TN) and False Negative(FN)
Compute Area under the ROC curve

(1) The data is collected from the different sites for instance NCBI (National center for biotechnology information)[12]

And worm base where the DNA of different species are being preserved and used by the researchers for their research [6]-[11].

(2) The collected data is then converted into numerical form so that it becomes easy to access and a proper window function can be applied to it to analyze the performance of the evaluated data [11]-[12].

(3)When the window function is applied the shape of the spectrum varies for different window functions which will be used further in the performance analysis [4]-[6].

(4) The window function try to read all the data provided and if some left then the use of window is helpful in that case. That is why the window size of 351 is used [11].

(5) The digital signal processing algorithm is used to get the periodicity-3 property. The DSP algorithm is applied to the windowed sequence to get the spectral analysis [11]-[16].

(6) After applying the algorithm the values of true positive, true negative, false positive and false negative is computed which provide us the sensitivity and specificity [12]-[13].

(7)The ROC curve is being formed from the values of sensitivity and specificity [4]-[5].

CHAPTER 4

LITERATURE REVIEW

In [4] it can be seen that there is a relationship between the lengths of a DNA sequence, now a day's bioinformatics and genomic signal processing are having greater advancement. Various techniques are developed for gene prediction over the past many years. In order to apply DSP tools to DNA sequence, symbolic nucleotides of DNA must be transformed into a numerical sequence and these are affecting the performance of the algorithm.

The paper [15] involves Genomic signal processing (GSP)[7]. It is the engineering research area which is concerned with genomic data analysis. There are number of techniques used for numerical analysis of DNA sequences which are studded in various papers [46] which have biological property and also preserve its biological meaning which is very important..

In [13] this paper various advanced methods of DNA numerical representation for DNA sequence analysis has been presented with their merits and demerits. In DNA sequence analysis if we make use digital signal processing algorithms then it requires the basic conversion of original sequence into the numerical sequence. The choice of numerical representation generally depends on the properties that can be seen in the numerical domain used for the preliminary detection and identification for the protein coding region.

In [14] it is found that, using Digital Signal Processing is possible only if the sequences are converted into numbers. Using digital signal processing in genomics is the basic element for solving most of the problems. In this area prediction of gene position in a genomic sequence and identifying the shortcoming regions in DNA sequence. This paper explains this answer and introduces comparison between these techniques in terms of their precision in exon and introns classification.

In [16] a new Digital Signal Processing based method is introduced to identify the protein coding regions in DNA sequences. Here, the provided DNA sequences are converted into numeric sequences with the help of electron ion interaction potential (EIIP) representation. Then discrete wavelet transformation algorithm is used. Absolute value of the energy is found and also the value of threshold using the data bases available in the National Centre for Biotechnology Information (NCBI) site. The comparative analysis is being done and analysis ensures the efficiency of the proposed system. Existing digital signal processing (DSP) methods provide less accurate results and computationally complex solution with higher background noise. Hence, improvements in accuracy, computational complexity, and reduction in background noise are done in this paper for identification of the protein coding regions in the DNA sequences.

In [10] the proposed algorithm for gene prediction in this paper compares the N/3 spectral components of DNA signal with the corresponding spectrum of period-3 DNA signal. In the DNA N/3 spectrum, for the bases in which the difference between these two spectrums is already defined with threshold level, the signal values are replaced by the different signal of the

CHAPTER 5

OUTLINE OF THE WORK

In the Genomic signal processing there are many elements which are primarily used in the identification of protein coding region [16]-[17]. The elements which are used window functions, digital signal processing algorithm [17].

Many digital signal processing based algorithms have been useful for finding the periodicities in DNA sequence [17]-[18]. Initial focus is on the mapping based schemes. The DSP algorithms such as the short-time discrete Fourier transform (ST-DFT) and Maximum entropy method (MEM) are used for spectral analysis. Firstly ST-DFT is discussed in detail, the basic process for the algorithm [19].

To derive a new matrix-based expression of the DNA spectrum that comprises most of the widely used DSP algorithms in the literature [19].

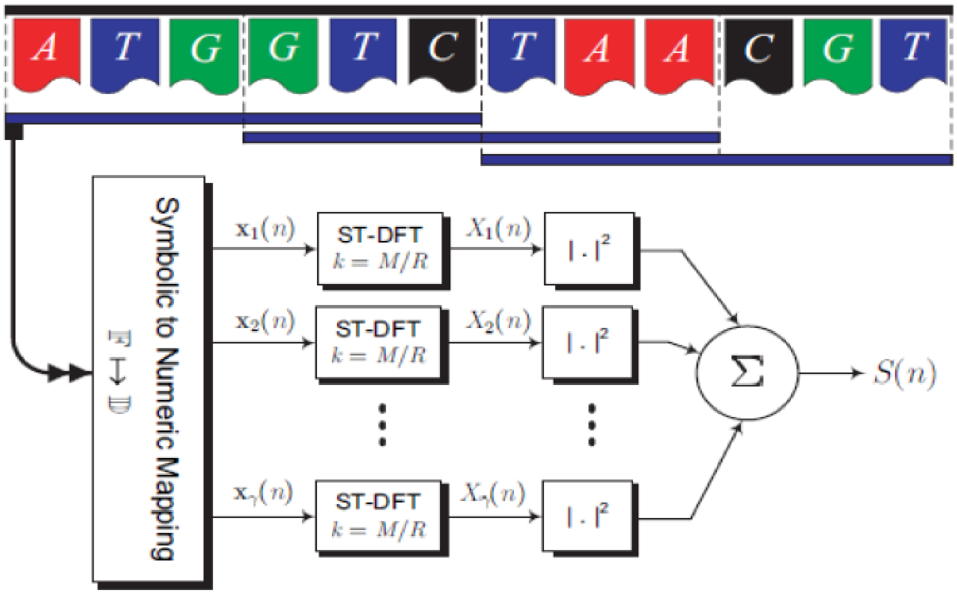


Fig 5.1- System structure to find the DFT of a function [20]

5.1MATRIX FORMULATION OF THE ST-DFT

By using the definition the matrix notation is as follows, we can restate the equation

$$X\left(R_n - \frac{M}{R}\right) = \left[1\mathrm{e}^{-\mathrm{j}2\pi/\mathrm{R}} \ldots \ldots \ldots \mathrm{e}^{-\mathrm{j}2\pi\left(\mathrm{R}-\frac{1}{\mathrm{R}}\right)}\right] = \mathrm{C}^{\mathrm{T}}\tau(\mathrm{n})^{\mathrm{I}}$$

In the analysis of signals [20], the FFT is a frequently used tool for frequency estimation and detection. The FFT provides good frequency resolution but provide no information about the time instant at the result [20]-[21]. This is considered to be the major shortcoming of a simple FFT approach [20]. Therefore some better techniques are used in place of FFT which provide the information of both frequency as well as the time [22]. So if we are analyzing non-stationary signals, then we can use one tool that is Short Time Fourier Transform (or STFT) [21]. In STFT we consider multiple FFTs on same set of data and they are operating on different time instant of input data [20]-[21].

The resulting matrix is made up of many FFTs, each FFT representing the frequency content during a particular window in time where the window can be determined by the FFT chosen length [18]. The resultant matrix helps to get the proper information about the strong signals positions and time of occurrence. Therefore the ST-FT is used in most of the cases [17]-[21].

Another algorithm is being used which provide a better exon prediction as compared to the short time Fourier transform [18]-[22]. The method is known as maximum entropy method. This method is a non parametric approach of spectrum estimation [22]-[23].

5.2 SPECTRAL ANALYSIS

Power spectrum is said to be the Fourier transform of the autocorrelation sequence. Estimation of power spectrum is estimation of autocorrelation [17]. Power spectrum also has certain limitations which are as follows:

(1) The amount of data on which the work has to be done is always limited and in some cases it may be very small [15]. This limitation may be inherent characteristics of data collection. If we consider a example in which data persist for short period of time [19]-[24]. It might be possible that the limited data set remain constant over the short duration [22].

(2) The original data is often ruined by noise or infected by interfering signals [12]-[25]. Thus it is a problem to estimate finite number of noisy measurements.

In signal detection and tracking power estimation plays an important role [19]. Many other applications of the spectrum estimation are harmonic analysis and prediction, time series extrapolation and interpolation, spectral smoothing, bandwidth compression, beamforming and direction finding [26].

The spectrum estimation is generally categorized into two parts [22]. The first part is the classical approach or the non-parametric approach in which we estimate the autocorrelation for the original set of data [22]-[26]. Then the power spectrum is estimated by Fourier transforming of the auto correlated sequences [24]. The second part is the non-classical approach or the parametric approach in which we make use of a model to estimate the power spectrum of the process [22]-[15].

The maximum entropy method is included in the classical or non-parametric approach. Estimating the power spectrum of the sequence is equal to estimating a autocorrelation of the sequence [13]. If we consider the data, the autocorrelation can only be estimated for short period of time that is the small sections of the initial data $|K| < N$ [11]-22]. As a result the autocorrelation will be set to zero for$|K| \geq N$.There are many signals which are non-zero in the lag $|K| \geq N$,these signal require the windowing function which can increase the resolution and accuracy of the estimated spectrum [24] .If we see the case of narrowband processes the autocorrelation of the processes decay slowly with K [17]. In the classical methods the autocorrelation is extrapolate (extending some method to an unknown situation by assuming that the results will have no effect with it) with zeros [15]-[23]. With the help of extrapolation the effect of window will be mitigated and we can get a more accurate estimate of power spectrum [18]. It is difficult to perform the extrapolation so the maximum entropy method is very helpful method in examining the power spectrum [19]-[25].

The $P_x(e^{j\omega})$ is considered to be a valid power spectrum [11]. $P_x(e^{j\omega})$ should be real valued and non negative for all values of ω. This value to be non negative does not guarantee the unique extrapolation [23]-[24]. Therefore some additional constraint needs to be imposed to the allowable extrapolation [19]. One such idea was suggested by the BURG to maximize the entropy. Basically entropy is a measure of randomness and uncertainty [22]. A maximum entropy extrapolation is equivalent to finding the sequences with autocorrelation that make x(n) as random as possible [19].

There are some basic steps which help in estimating the spectrum in the maximum entropy method

(1) Firstly the autocorrelation equations are solved for the all pole cofficients [27].

(2) Then the MEM spectrum is produced by incorporating these parameters[15]

(3) If the P_{mem} $(e^{j\omega})$ is considered to be an all pole power spectrum then $r_x(k)$ satisfies the Yule Walker equation that is [28]

$$r_x(l) = -\sum_{k=1}^{p} a_p(k) r_x(k-l) \qquad \text{for } l > 0$$

The maximum entropy method extrapolates the autocorrelation sequence with the noise suppression [25]. The MEM has been studied as a spectrum analysis tool. In case we have any information or constraint in a process where we have set of some autocorrelation values the best way to estimate the power spectrum is to estimate the Fourier transform of the autocorrelation sequences along with the extrapolation which performs the maximum entropy extrapolation [24]-[25].

5.3 EXON PREDICTION ALGORITHM

If we compare the period-3 DNA component of DNA signal with corresponding spectrum then the algorithm will let us know about the noise which is present in the non coding region. When we plot ROC (Receiver operating characteristics) the optimum threshold can be determined from set of sequences collected from the dataset [10]. We will see if the noise can be suppressed, so that we can get a improved detection [10]-[15].

Initialization: First we generally set the base location with the mapping schemes [15]. **Step 1-**Then we use a window with length N = 351 so that we can select all the nucleotide from the base location of the DNA sequence [11].

Step 2- After that the four binary sequences $x_A[n]$, $x_T[n]$, $x_G[n]$ and $x_C[n]$ are obtained for the windowed sequence of Step 1.

Step 3: Each of the binary sequences obtained in Step 2 is provided as input to an digital signal processing algorithm. The corresponding outputs $y_A[n]$, $y_T[n]$, $y_G[n]$ and $y_C[n]$ are estimated using the algorithm [11].

Step4: Compute the values for sensitivity and specificity of the provided data. Plot the graph between sensitivity and specificity [11].

Step 5: plot the roc curve and afterwards compute the AUC.

Using the digital signal processing tool that is maximum entropy method we have estimated the power spectrum [11]-[15]. It follows certain steps for performance analysis of different DNA for Identification of Protein Coding Regions which are as follows:

(a)The data set with different accession numbers are collected from the NCBI site or some other websites where the gene information of different species are kept in the form of A, G, C, T nucleotides [11]. The data sets which are being used for analysis are AF007189, AF058762, AF0282233, AF059734 and F56F11.4 [16]. The data sets will then be converted into numerical form with different mapping techniques which are commonly used [22].

(1)The data set AF007189 contains single exon with 1601 base pairs which will lie in the region as follows:

Table 5.2-Number of nucleotides (AF007189)

Total base pairs	Start	End
1601	477	1139

(2)The data set AF058762 contains exon on two positions with 3036 base pairs as follows:

Table 5.3- Number of nucleotides (AF058762)

Total base pairs	Start	End
3036	115	482
	1867	2662

(3)The data set AF0282233 contains exons at three positions with 4547 base pairs as follows

Table 5.4- Number of nucleotides (AF0282233)

Total base pairs	Start	End
4575	68	392
	1483	1673
	3211	3558

(4)The data set AF059734 contains exons at four different positions with 2401 base pairs as follows:

Table 5.5- Number of nucleotides (AF059734)

Total base pairs	Start	End
2401	335	491
	1296	1495
	1756	1857
	1953	2051

(5)The data set F56F11.4 contains exons at five different positions with 8100 base pairs as follows :

Table 5.6- Number of nucleotides (F56F11.4)

Total base pairs	Start	Start
8100	929	1137
	2528	2857
	4114	4377
	5465	5644
	6342	7605

(b) The mapped sequence is applied the window size of 351 to include all the nucleotides that may be very large in size [28]. 351 is set as standard size in most of the implementations.

(c) The digital signal processing algorithm is applied then to the windowed sequence which will provide the power spectrum of the positions of the exons [28]-[29].

(d) After implementation of DSP algorithm we calculate the four parameters True positive, True negative, False positive and false negative [28].

Table 5.7- statistical procedure to obtain the sensibility and the specificity in coding region prediction

	Coding region	Non Coding region
Positive Prediction	True Positive(TP)	False Positive(FP)
Negative Prediction	False Negative(FN)	True Negative(TN)

(e) We can easily predict the true positive rate and false positive rate in other words which is known as sensitivity and specificity.

The sensitivity gives the measure of the proportion of coding nucleotides that have been correctly predicted as coding [17].

$$\text{Sensitivity (Sn)} = TP/(TP+FN) = TPR$$

The specificity is the proportion of predicted coding nucleotides that are actually from the coding region [19].

$$\text{Specificity (Sp)} = TN/(TN+FP) = \text{True Negative Rate} = 1 – FPR$$

(f) To draw a ROC curve, only the true positive rate (TPR) and false positive rate (FPR) are needed [22]. Since TPR is equivalent to sensitivity and FPR is equal to 1 − specificity, the ROC graph is sometimes called the sensitivity vs (1 − specificity) plot [17]-[18].

(g) Then we compute the value of area under the curve [22].

CHAPTER 6

RESULTS AND DISCUSSIONS

In this chapter we will discuss the different analysis which is being done on window functions and the digital signal processing algorithms. As we have discussed in chapter 2 in detail about the different window functions which can be used in exon prediction because each window has different behavior in context of performance. Five DNA sequences having accession numbers AF007189, AF058762, AF0282233, AF059734 and F56F11.4 with window size 351 is taken for performance comparison

Now we will study the performance of various window functions for five different data sets and paired numeric mapping scheme.

(1)The data set with accession number AF007189 with hanning window of size 351

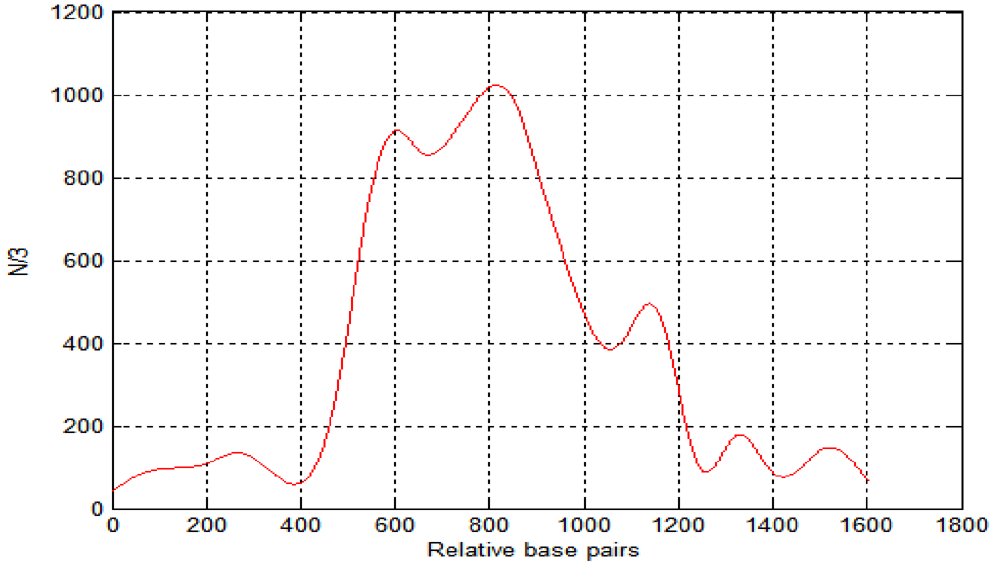


Fig 6.1-Hanning window

(2)The data set with accession number AF007189 with Rectangular window of size 351

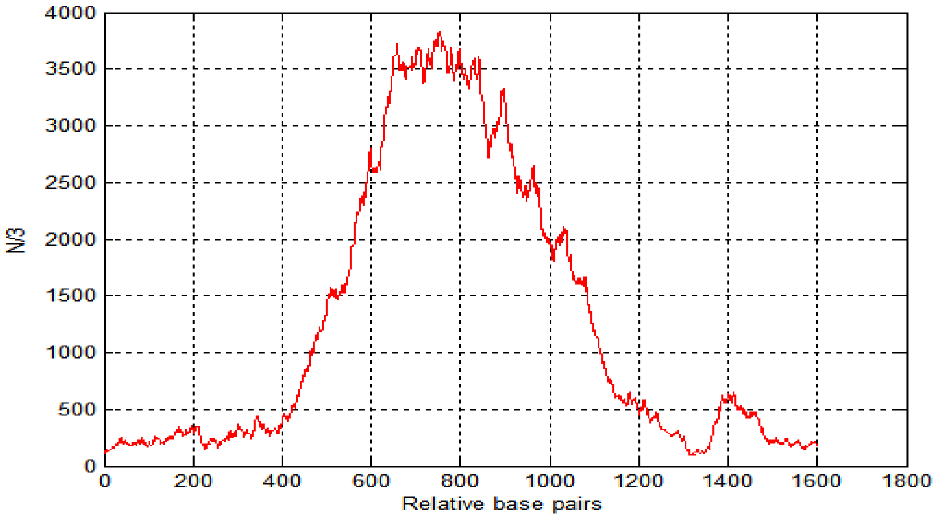


Fig 6.2-Rectangular window

(3)The data set with accession number AF007189 with Blackman window of size 351

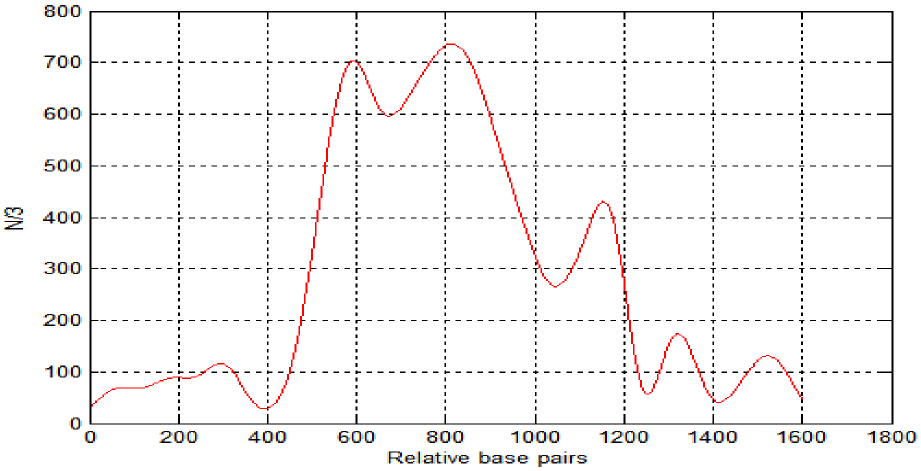


Fig 6.3-Blackman window

(4)The data set with accession number AF007189 with Kaiser Window of size 351

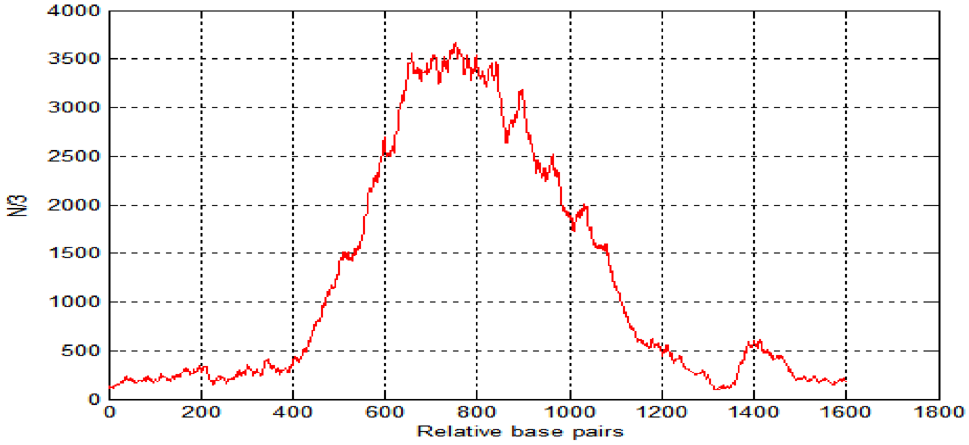


Fig 6.4-Kaiser Window

(5)The data set with accession number AF007189 with Gaussian Window of size 351

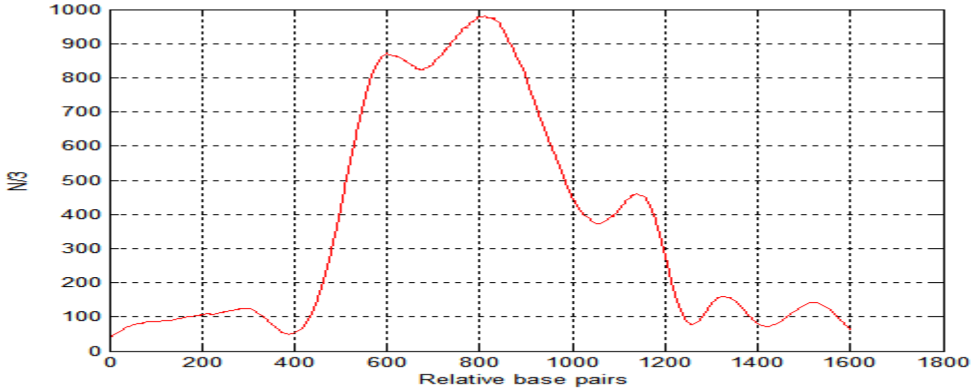


Fig 6.5-Gaussian Window

(6) The data set with accession number AF007189 with Parzen Window of size 351

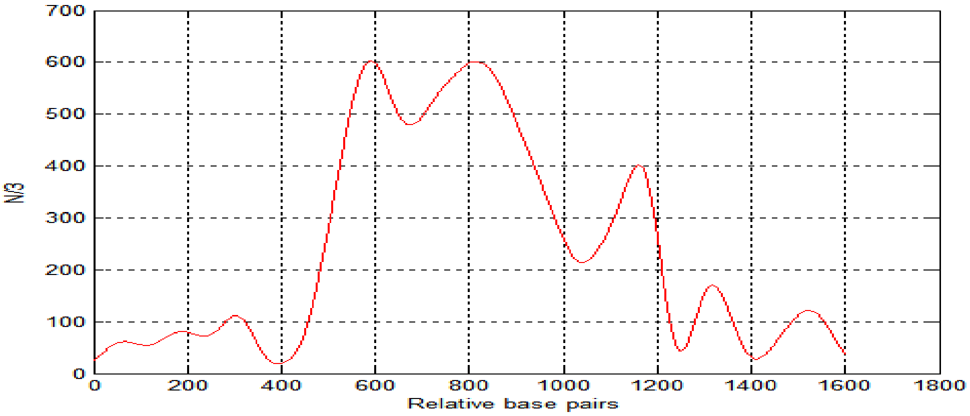


Fig 6.6-Parzen Window

(7) The data set with accession number AF007189 with Barthann Window of size 351

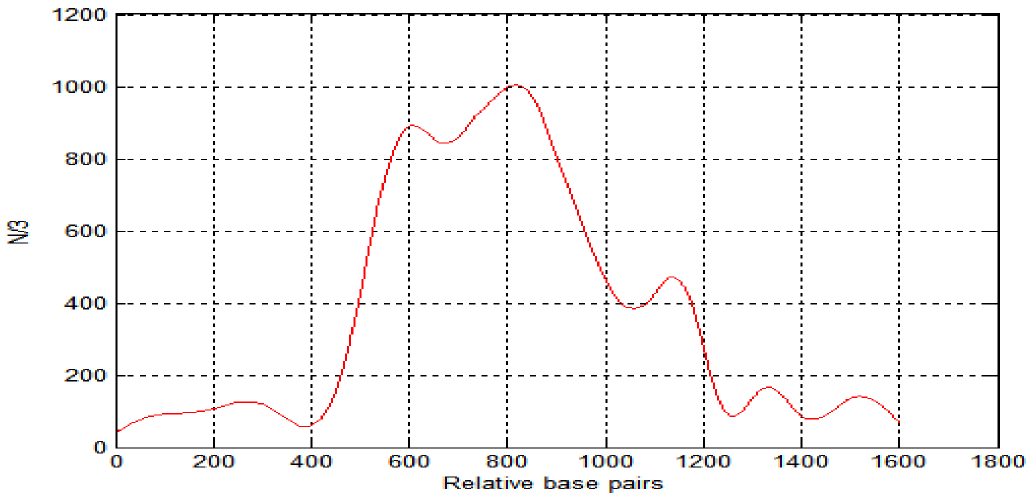


Fig 6.7-Barthann Window

(8) The data set with accession number AF007189 with Bohman Window of size 351

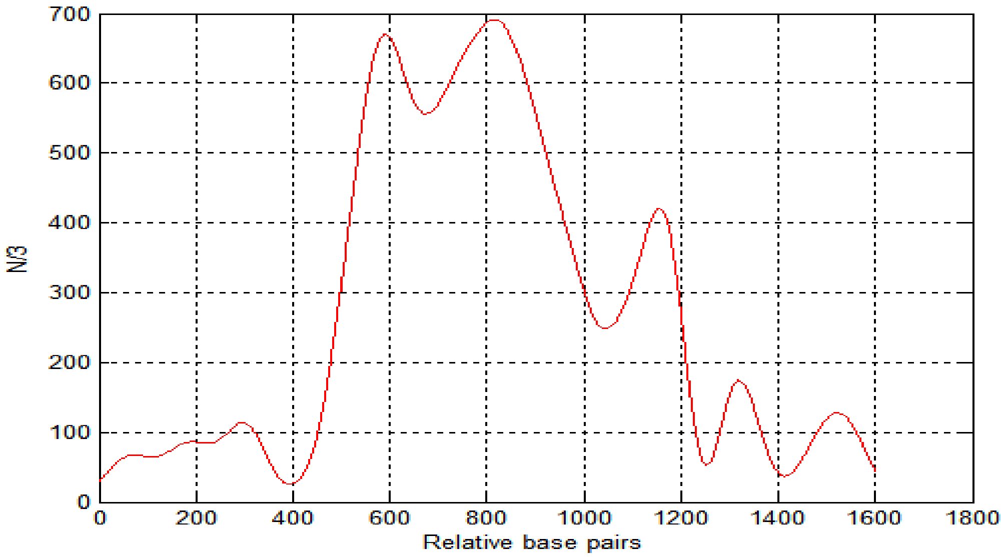


Fig 6.8-Bohman Window

(9) The data set with accession number AF007189 with Chebyshev Window of size 351

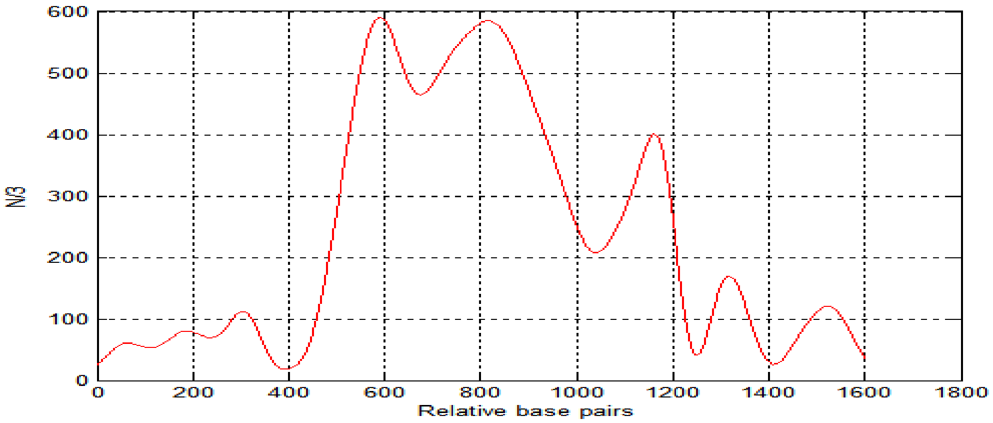


Fig 6.9-Chebyshev Window

(10) The data set with accession number AF007189 with Taylor Window of size 351

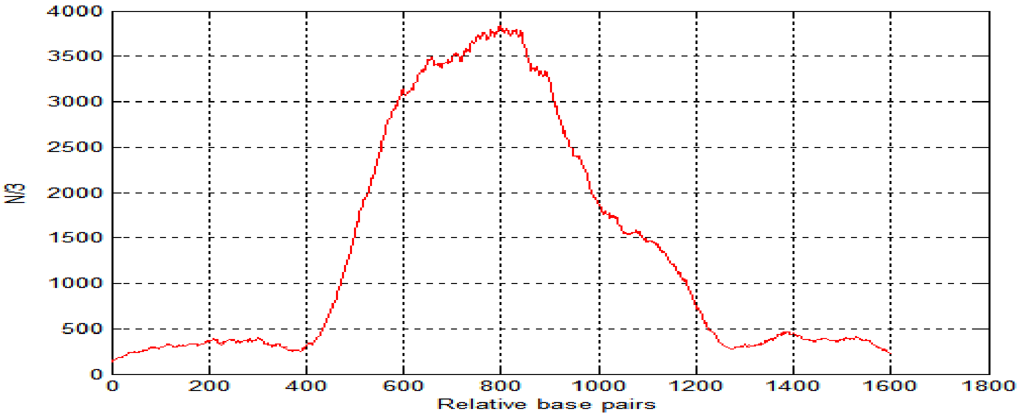


Fig 6.10-Taylor Window

(11) The data set with accession number AF007189 with Triangular Window of size 351

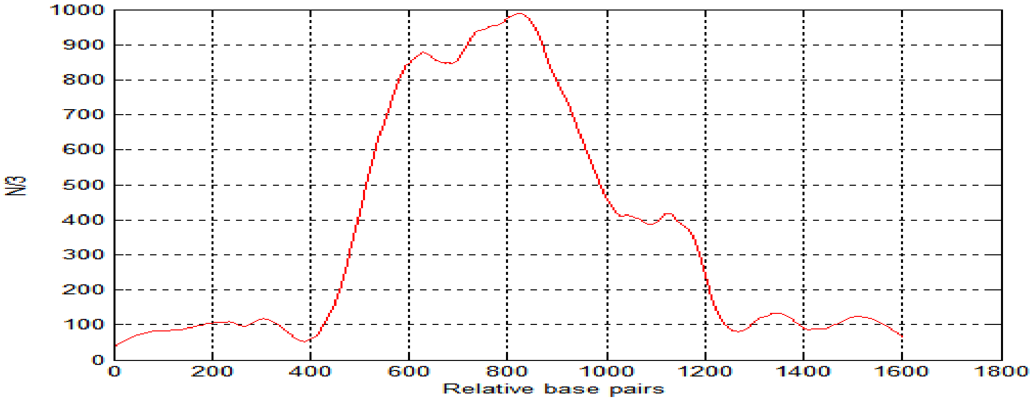


Fig 6.11-Triangular Window

(12) The data set with accession number AF007189 with Tukey Window of size 351

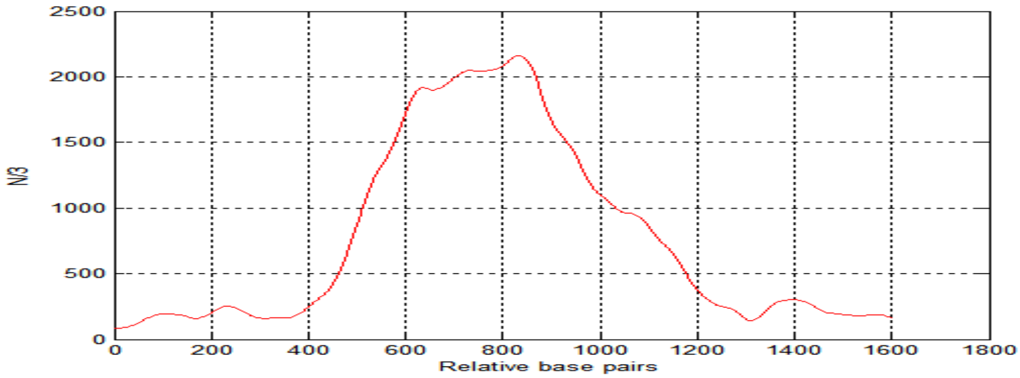


Fig 6.12-Tukey Window

(13) The data set with accession number AF007189 with Blackmanharris Window of size 351

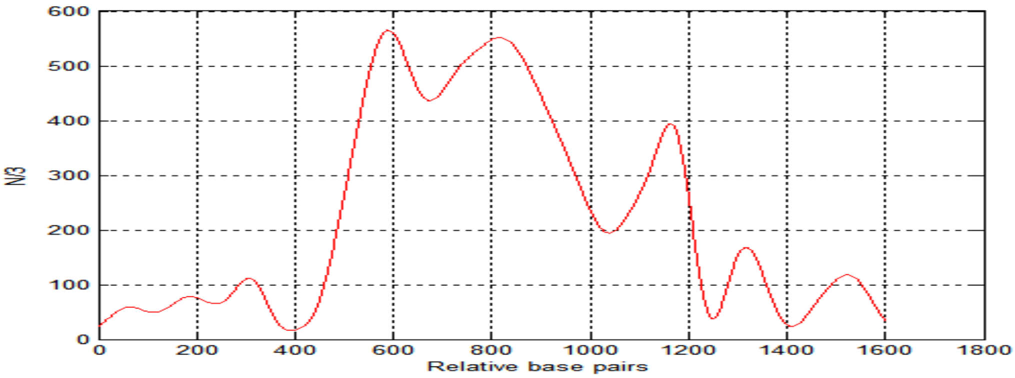


Fig 6.13-Blackmanharris Window

(14) The data set with accession number AF007189 with Bartlett Window of size 351

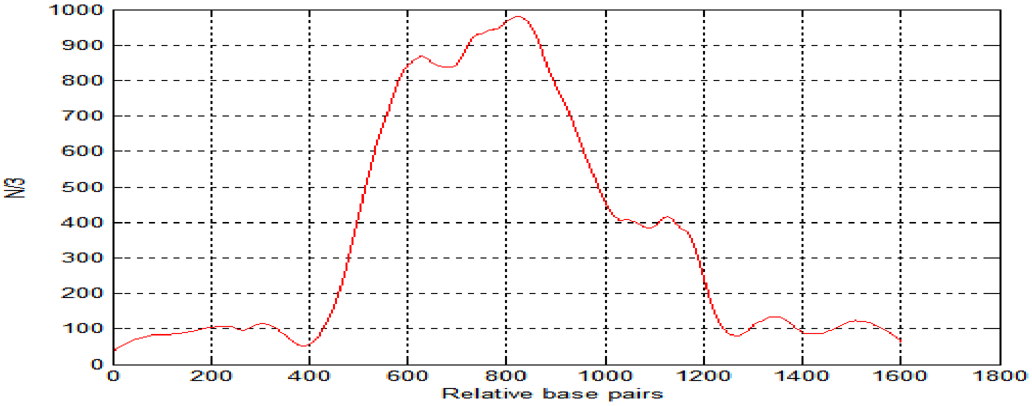


Fig 6.14-Bartlett Window

(15) The data set with accession number AF007189 with Flattop Window of size 351

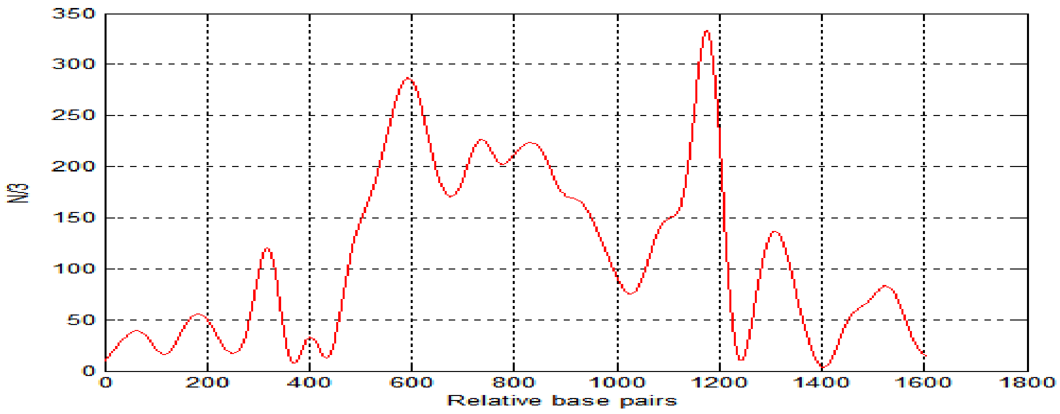


Fig 6.15-Flattop Window

We have also studied the comparison of the window function corresponding to two other mapping schemes voss representation and EIIP. Now we will study the comparison of the window function with a different accession number AF059734 which has four nucleotides positions in it and have Eiip mapping scheme.

(1)The data set with accession number AF059734 with Hanning Window of size 351

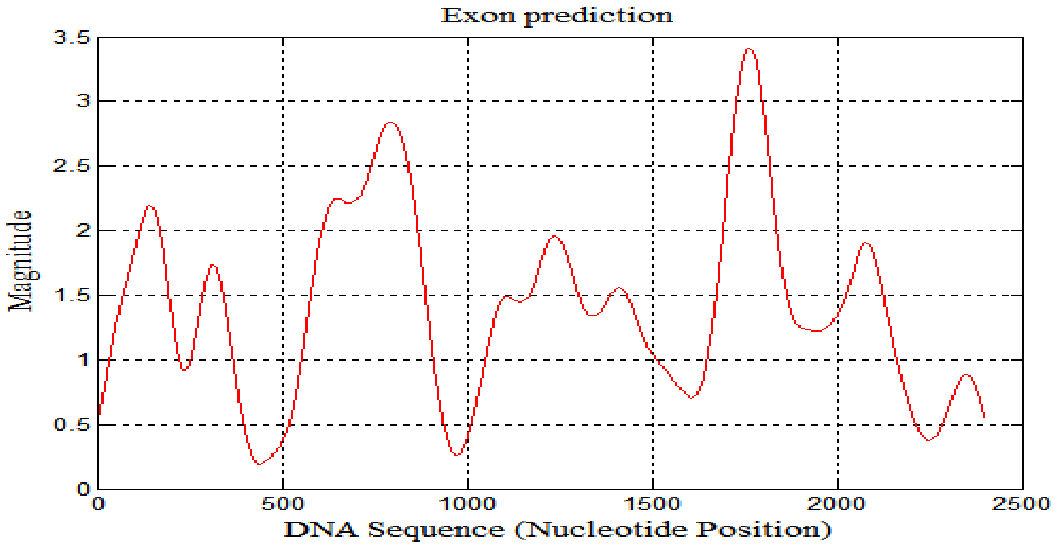


Fig 6.16-Hanning Window

(2)The data set with accession number AF059734 with Rectangular Window of size 351

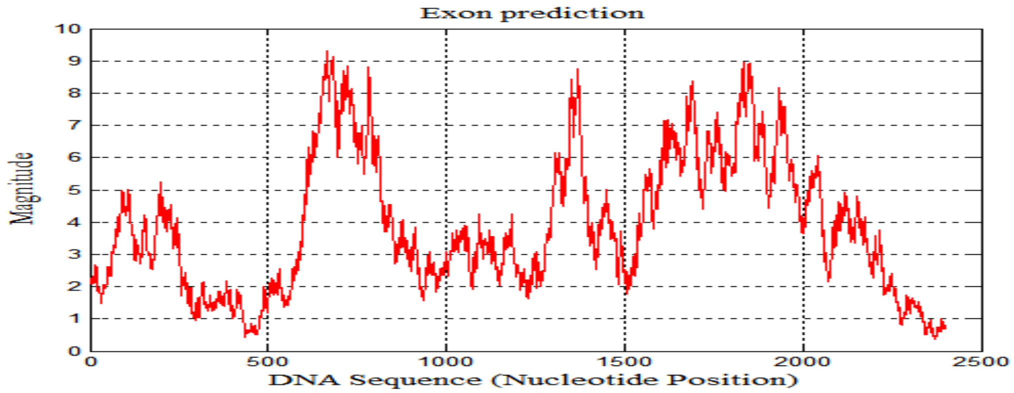


Fig 6.17-Rectangular Window

(3)The data set with accession number AF059734 with Blackman Window of size 351

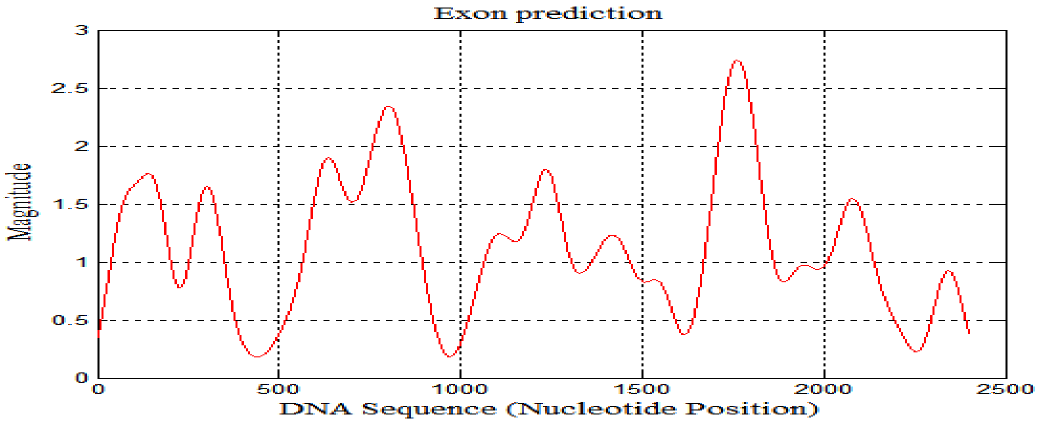


Fig 6.18-Blackman Window

(4)The data set with accession number AF059734 with Kaiser Window of size 351

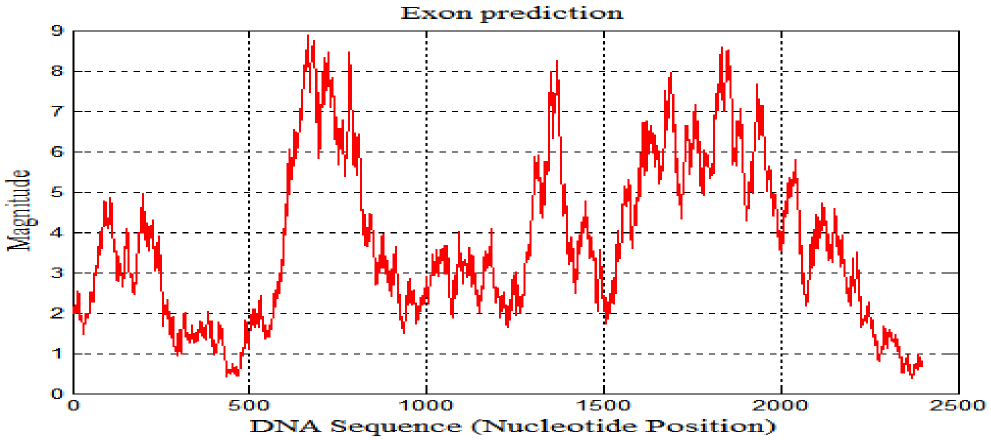


Fig 6.19-Kaiser Window

(5)The data set with accession number AF059734 with Gaussian Window of size 351

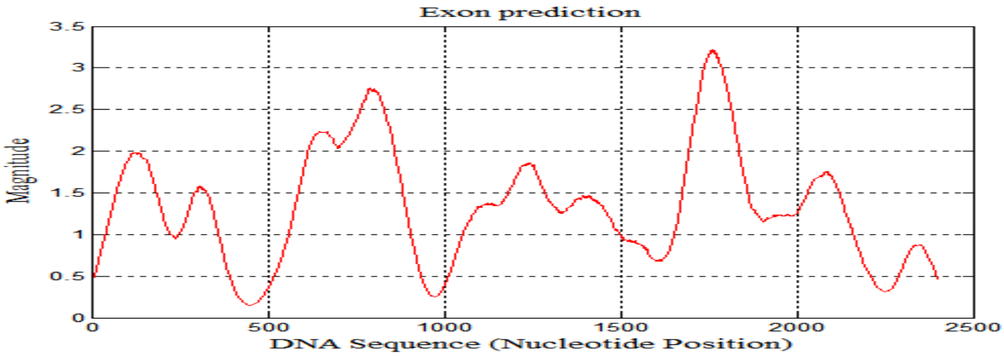


Fig 6.20-Gaussian Window

(6)The data set with accession number AF059734 with Parzen Window of size 351

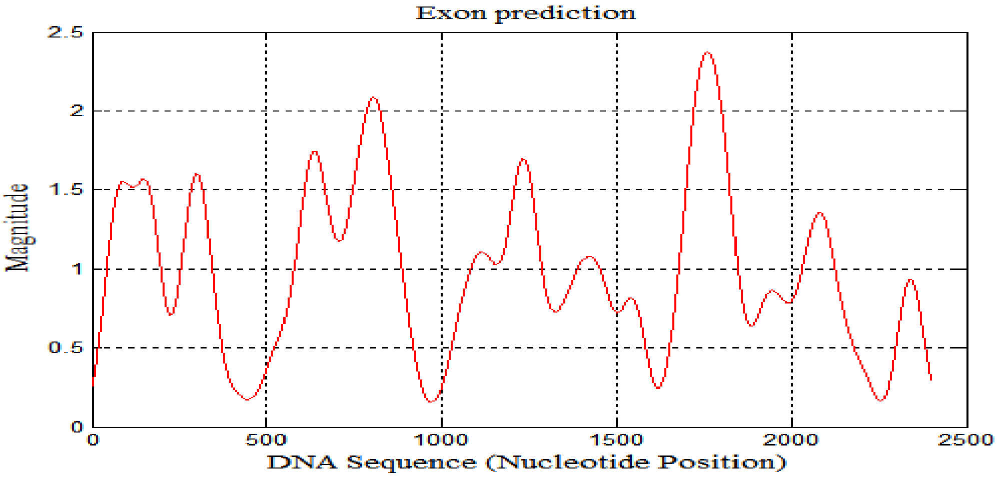


Fig 21-Parzen Window

(7)The data set with accession number AF059734 with Barthann Window of size 351

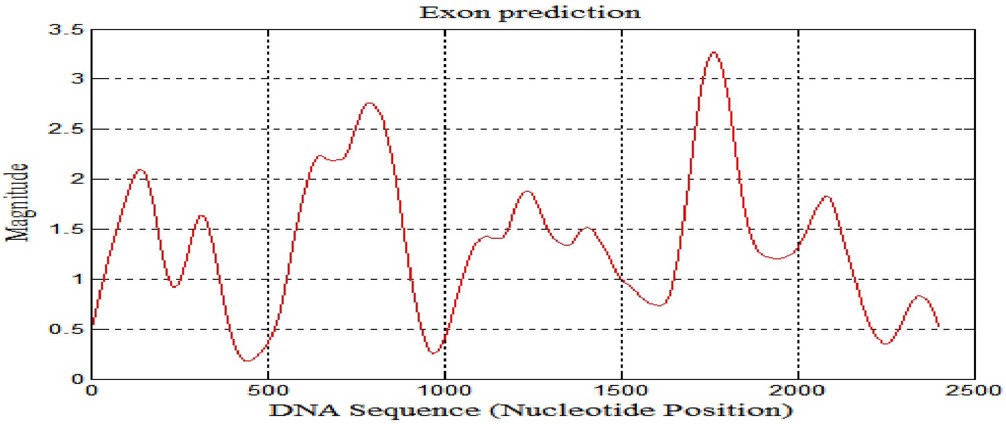


Fig 6.22-Barthann Window

(8) The data set with accession number AF059734 with Flattop Window of size 351

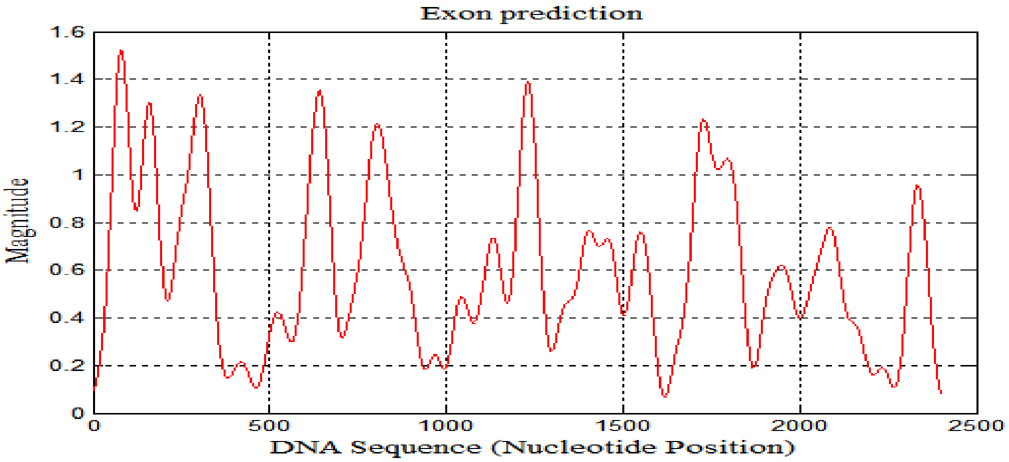


Fig 6.23-Flattop Window

(9) The data set with accession number AF059734 with Bartlett Window of size 351

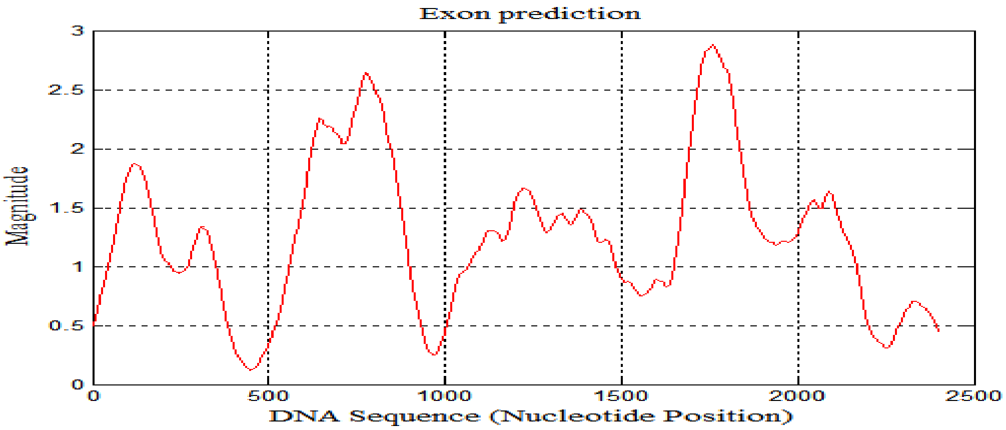


Fig 6.24-Bartlett Window

(10) The data set with accession number AF059734 with Tukey Window of size 351

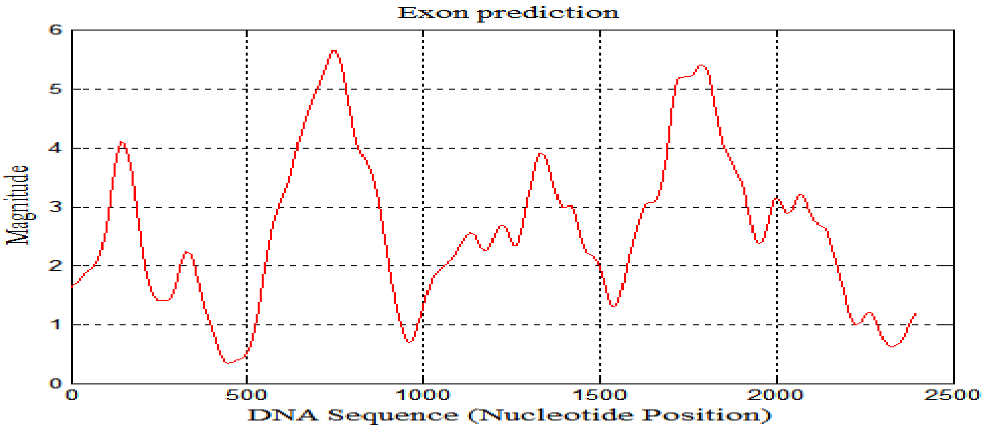


Fig6.25 -Tukey Window

(11) The data set with accession number AF059734 with Triangular Window of size 351

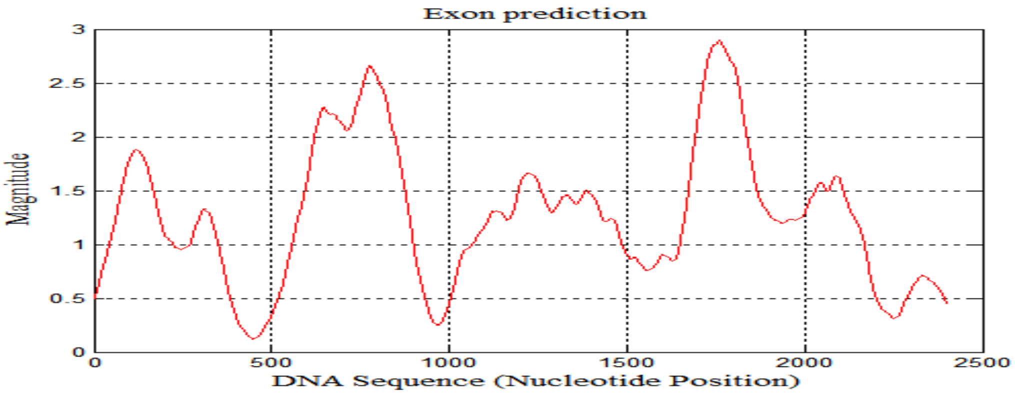


Fig 6.26-Triangular Window

(12) The data set with accession number AF059734 with Taylor Window of size 351

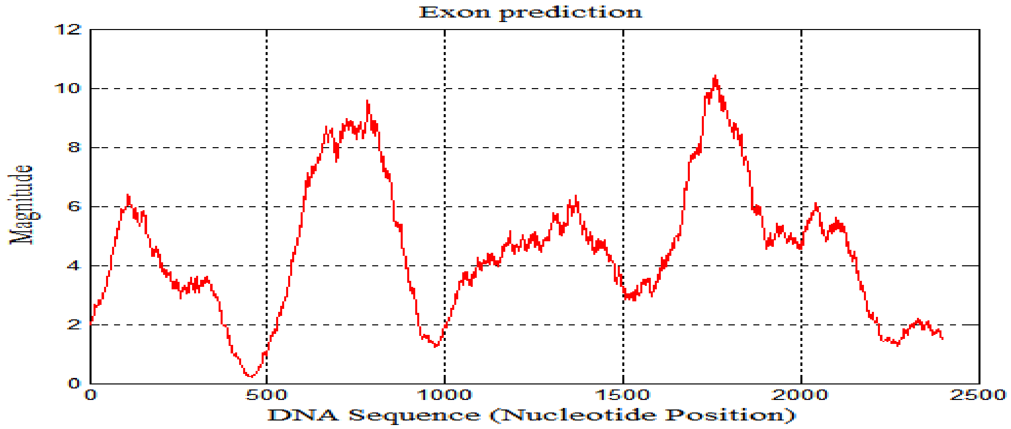


Fig 6.27-Taylor Window

(13) The data set with accession number AF059734 with Nuttall Window of size 351

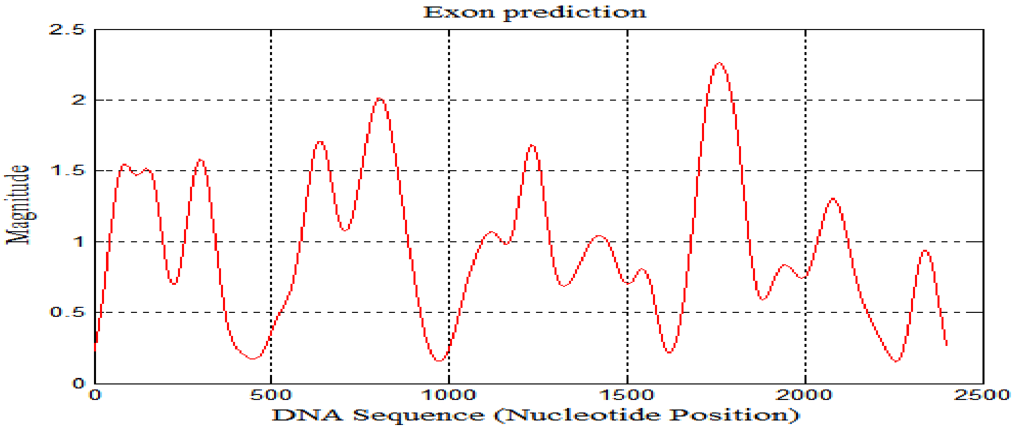


Fig6.28-Nuttall Window

We have the power spectrum of the different window function which is compared with the help of two different mapping schemes. The experimental results have been carried out for the DNA sequences. The performance evaluation has been done by calculating the area under the curve (AUC). Accuracy of the window function is higher if the value of Area under curve (AUC) is more. The results carried out using different window functions are presented in the above figures.

For the value of area under the curve we make use of digital signal processing algorithm. Firstly we make use of short time discrete Fourier transform. Therefore we have computed the values of AUC's as follows:

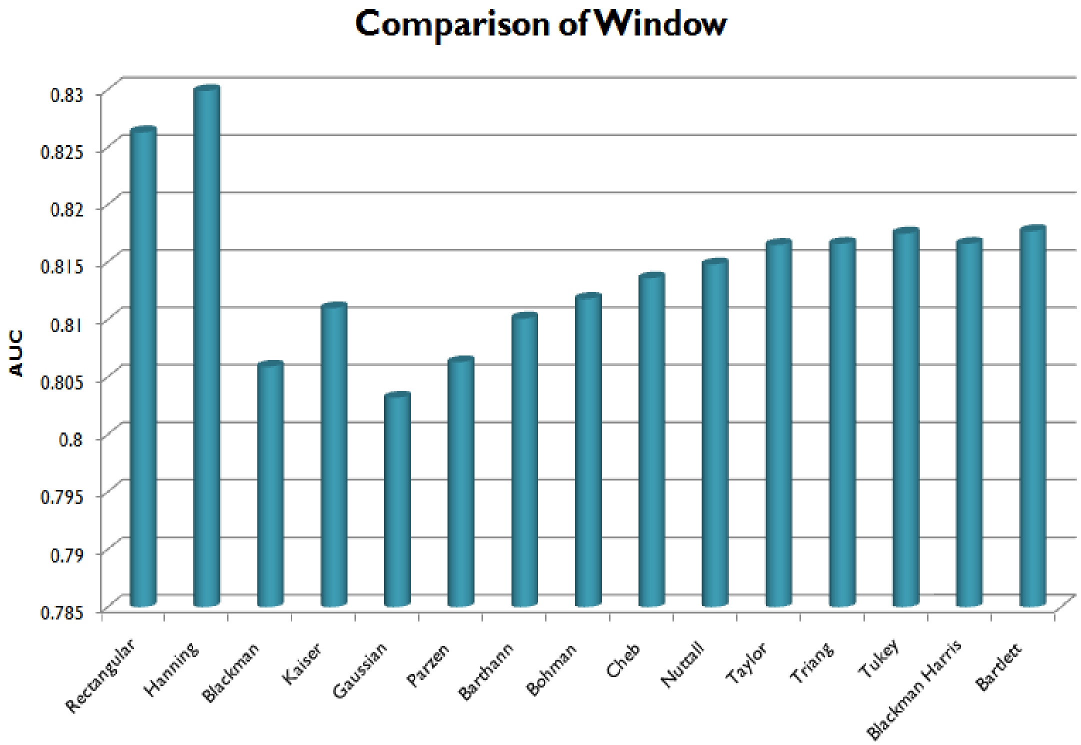
Table 6.1-Compared AUC's of different window functions

GENE ID	**AF007189**	**AF028233**	**AF058762**	**AF059734**	**F56F11.4**
Rectangular Window	0.5136	0.8144	0.8476	0.4674	0.8263
Hanning Window	0.4962	0.8221	0.8429	0.4728	0.8299
Blackman Window	0.5285	0.5285	0.8480	0.4573	0.8059
Kaiser Window	0.5432	0.7876	0.8516	0.4547	0.8110
Gaussian Window	0.5563	0.7721	0.8523	0.4457	0.8032
Parzen Window	0.5678	0.7557	0.8536	0.4407	0.8063
Barthann Window	0.5898	0.7655	0.8542	0.4405	0.8101

Bohman Window	0.5940	0.7643	0.8543	0.4407	0.8118
Chebyshev Window	0.5979	0.7632	0.8541	0.4408	0.8136
Nuttall Window	0.6016	0.7622	0.8538	0.4408	0.8148
Taylor Window	0.6062	0.7612	0.8547	0.4397	0.8165
Triangular Window	0.6105	0.7605	0.8553	0.4391	0.8166
Tukey Window	0.6147	0.7592	0.8561	0.4379	0.8175
Blackman Harris Window	0.6180	0.7584	0.8558	0.4381	0.8166

On the basis of experimental results, it has been concluded that the Hanning window gives more accurate results for the two mentioned sequences. In future the performance of ST-FT using various window functions can be tested for big data sets.

Fig 6.29-Comaprison of window



In STDFT, the window function plays an important role and performance of STDFT differs for every window functions. The ROC curve is computed for all five accession numbers AF007189, AF058762, AF0282233, AF059734 and F56F11.4 with the hanning window function.

(1)The ROC of the data set AF007189

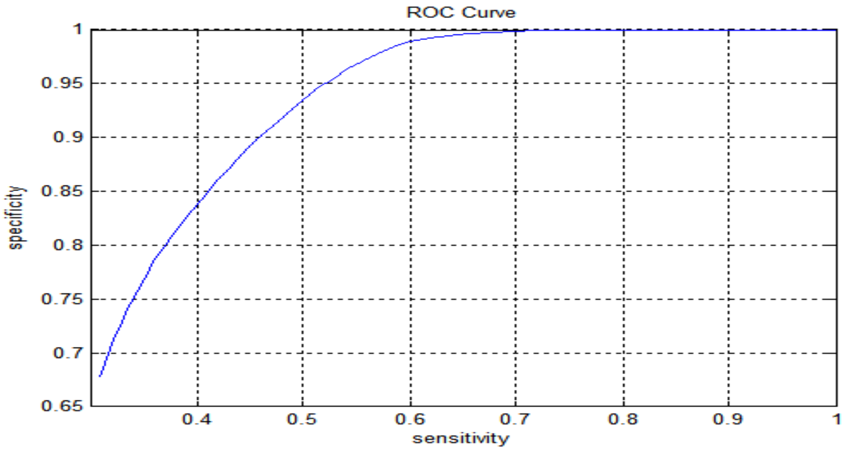


Fig 6.30-ROC for AF007189

(2)The ROC of the data set AF058762

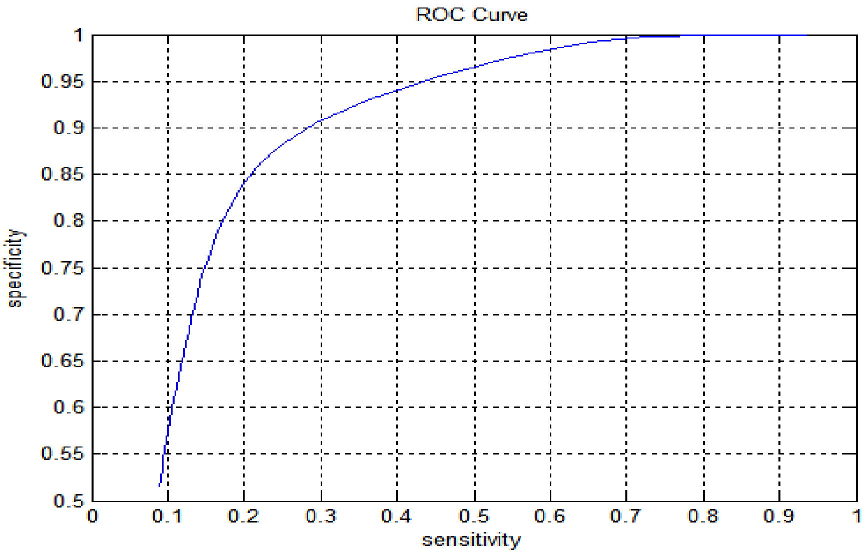


Fig 6.31-ROC for AF058762

(3)The ROC of the data set AF0282233

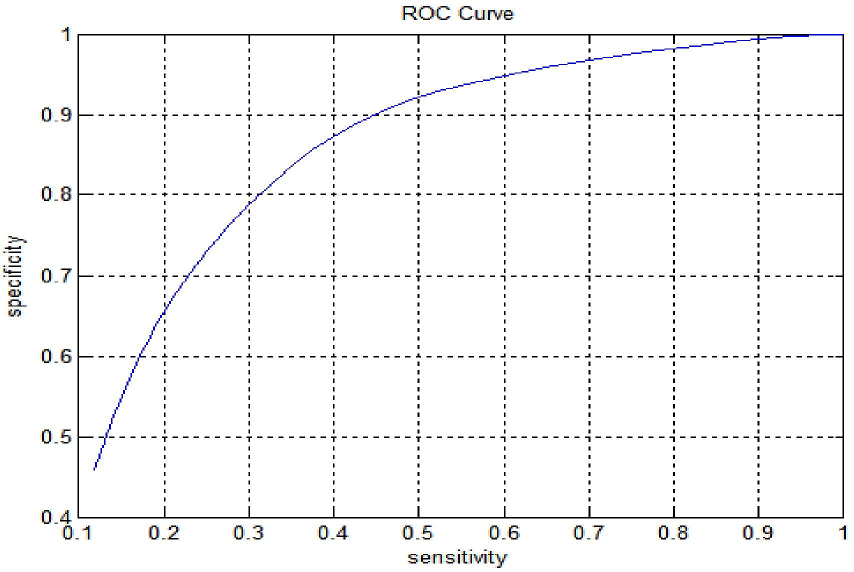


Fig 6.32-ROC for AF0282233

(4)The ROC of the data set AF059734

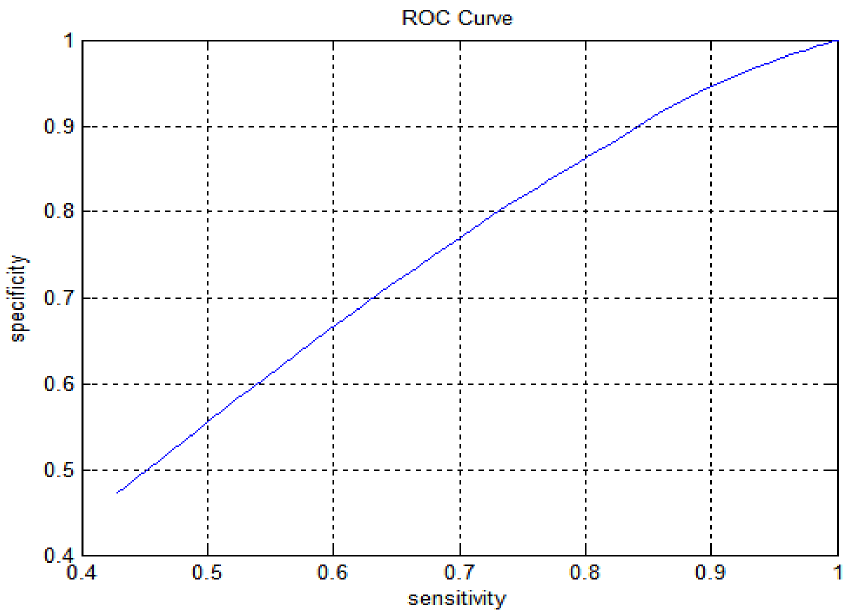


Fig 6.33-ROC for AF059734

(5)The ROC of the data set F56F11.4

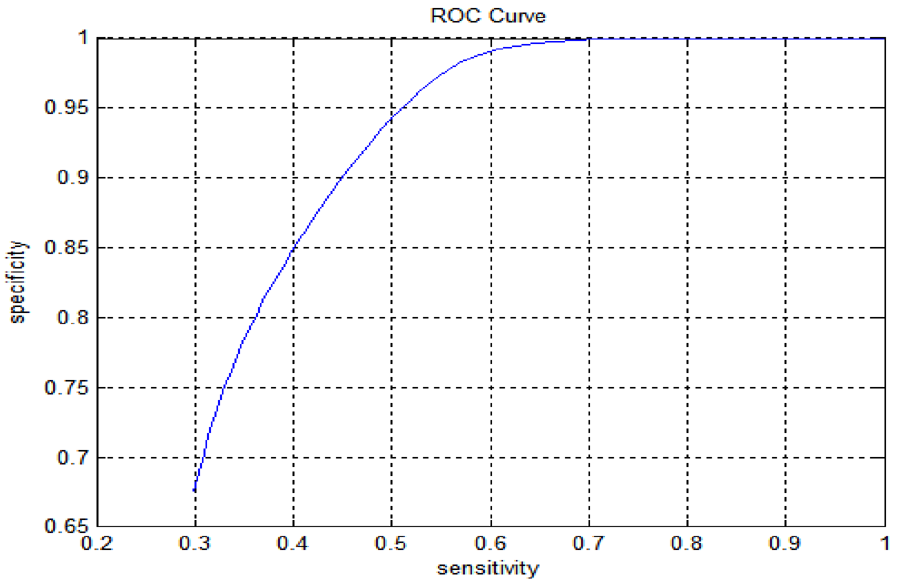


Fig 6.34-ROC for F56F11.4

The results of short time Fourier transform are compared with the results of maximum entropy method. The output of the maximum entropy method is better as compared to ST-FT. The area under the curve in case of maximum entropy method will be more accurate. The hanning window is used with the maximum entropy method:

(1)The ROC of the data set AF007189

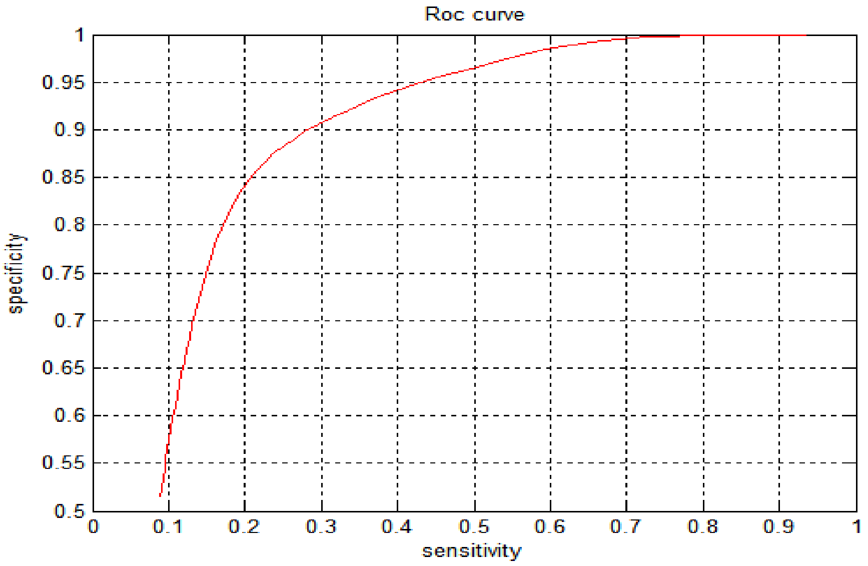


Fig 6.35-ROC for AF007189

(2)The ROC of the data set AF058762

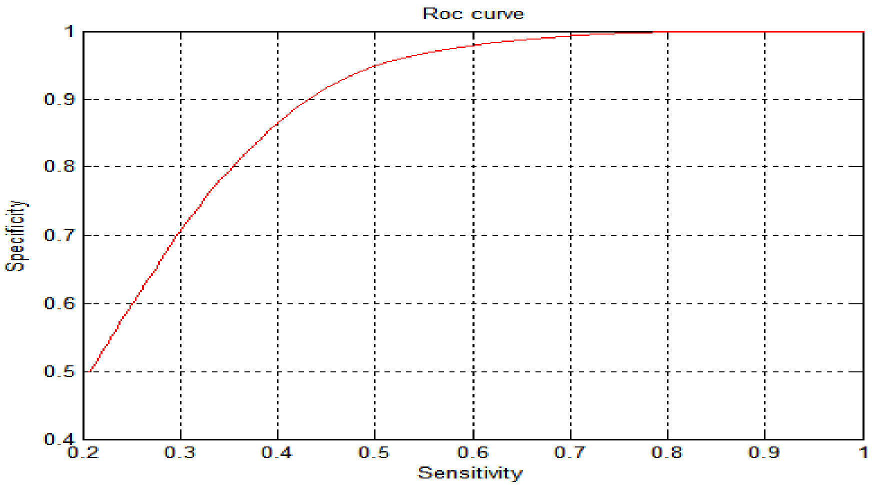


Fig 6.36-ROC for AF058762

(3)The ROC of the data set AF0282233

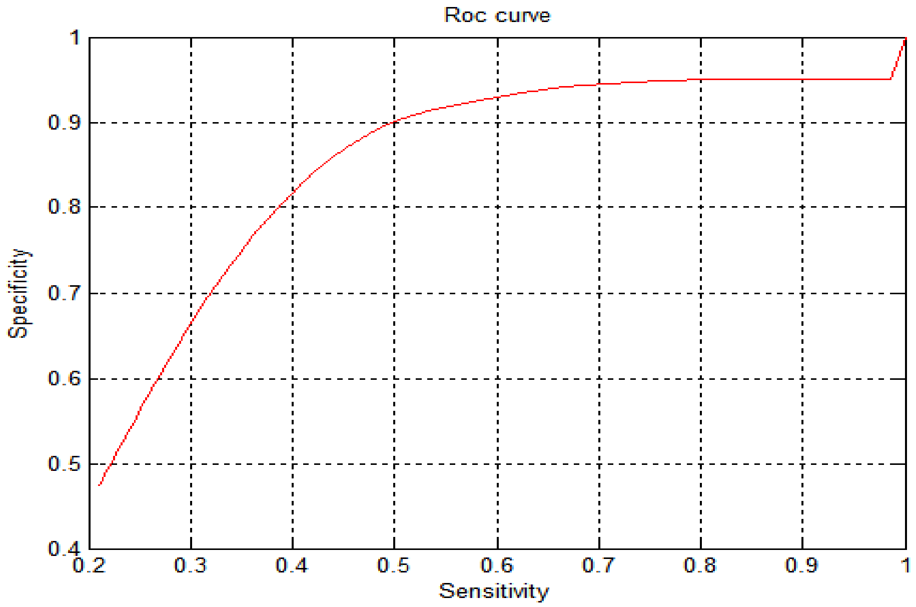


Fig 6.37-ROC for AF0282233

(4)The ROC of the data set AF059734

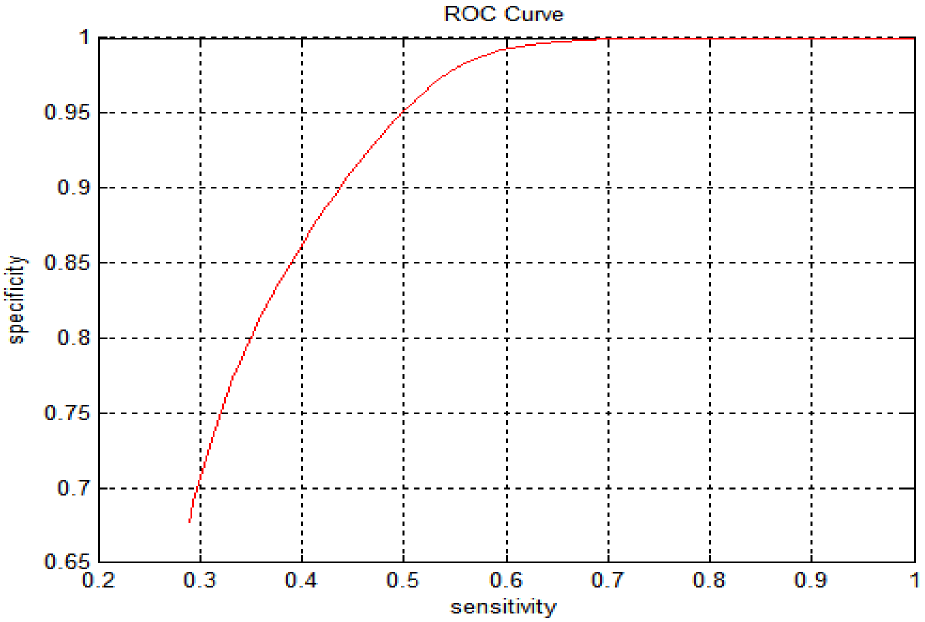


Fig 6.38-ROC for AF059734

(5)The ROC of the data set F56F11.4

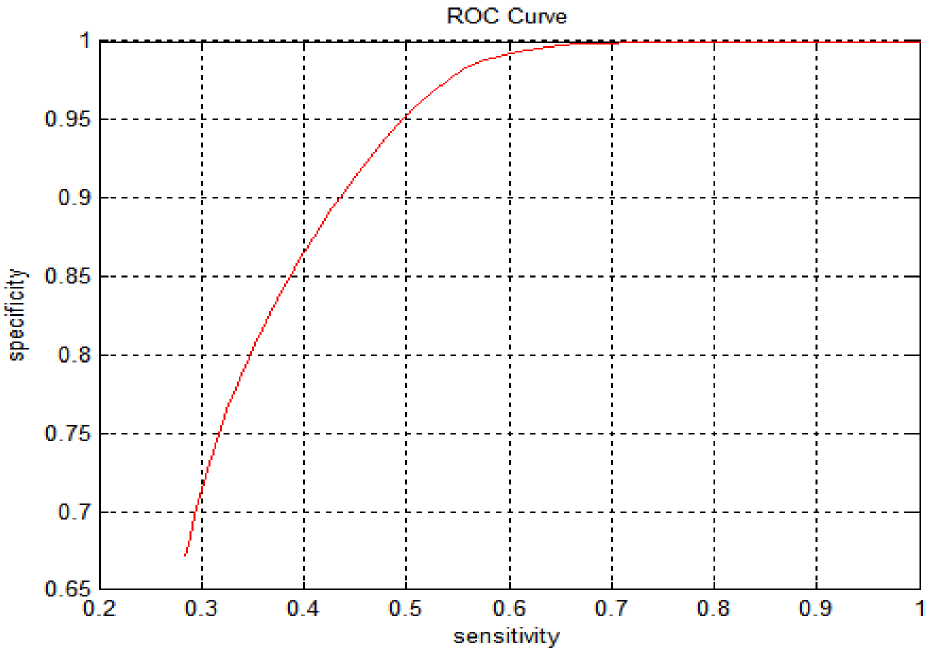


Fig 6.39-ROC for F56F11.4

The maximum entropy method provides better results as compared with the short time discrete Fourier transform. The hanning window is very efficient in case of large data sets that provide a near perfect power spectrum in maximum entropy algorithm. The area under the curve in case of short time discrete Fourier transform varies for different data sets but the maximum value it provides is 0.700 but in maximum entropy method the value is above 0.800.Hence the performance is better in case of maximum entropy method.

CONCLUSION

After the analysis it is concluded that the estimation of the protein coding region is dependent on various parameters that are the window function in which we have seen that the hanning window provide good performance in large data sets and has less impact on the small data sets. Then we make use of digital signal processing algorithm in which the short time discrete Fourier transform provide better performance with the combination of proper window with it but the maximum entropy method provides much better performance with the hanning window. This combination is suitable for the identification of protein coding region in genomes.

PUBLICATIONS

“Performance Analysis of Window Functions for Exon Prediction in DNA Sequences” Recommended for publication in international Conference on Computing Communication and Automation,ICCCA-2017,[3rd: New Delhi: 5-6 May,2017]

REFRENCES

[1] Swarna bai Arniker , Hon Keung Kwan, “*Advanced Numerical Representation of DNA Sequences*” **,** International Conference on Bioscience, Biochemistry and Bioinformatics, *pp.1-5,2012*

[2] Wei Hua, Jiasong Wang and Jian Zhao ,“ *Discrete Ramanujan Transform for Distinguishing the Protein Coding Regions From Other Regions*”, journel Molecular and Cellular Probes vol. 28,pp. 228-236, 2014

[3] Sanjeev N. Sharma, Rahul pachuri, Rajiv Saxena, “*Fixed Window in Fractional Fourier Domain*”, *I.J image graphic and signal processing*, pp. 1-13, 2014.

[4] Heba Mohamed. Wassfy, Mustafa M. Abd Elnaby, Mohamed Labib Salem, Mai S. Mabrouk, Abdel-Aziz Awad Zidan, “*Advanced DNA Mapping Schemes for Exon Prediction Using Digital Filters*”, American Journal of Biomedical Engineering,vol.8,pp. 25-31, 2016

[5] SAJID A. MARHON and STEFAN C. KREMER, “*Gene Prediction Based on DNA Spectral Analysis A Literature Review*”, journal of computational biology, vol. 18,pp. 433-445, 2011

[6] Inbamalar T M and Sivakumar R , “ *DNA Sequence Analysis Using DSP Techniques*”, Journal of Automation and Control Engineering vol. 1,pp.336-342, 2013

[7] Rajasekhar Kakumani, Vijay Devabhaktuni and M. Omair Ahmad, “*Prediction of Protein-Coding Regions in DNA Sequences Using a Model-Based Approach*”, IEEE international symposium on circuit and system ,pp.1918-1921,2008

[8] Malaya Kumar Hota and Vinay Kumar Srivastava, “*Performance Analysis of Different DNA to Numerical Mapping Techniques for Identification of Protein Coding Regions Using Tapered* ”, International Journal of Advances in Engineering & Technology,vol.3,pp.561-569,2002

[9] Alexander, Ed, and D. Poularikas. "*The Handbook of Formulas and Tables for Signal Processing*." Boca Raton, FL, USA: CRC Press vol.3, pp. 73-79 1998

[10] Mahmood Akhtar, Julien Epps, and Eliathamby Ambikairajah, “*Signal Processing in Sequence Analysis Advances in Eukaryotic Gene Prediction*”, IEEE journal of selected topics in signal processing, vol. 2, no. 3,pp.310-321**,** 2008

[11] S. D. Sharma, Rajiv Saxena, S.N. Sharma, A. K. Singh "*Short Tandem Repeats Detection in DNA Sequences Using Modified S-Transform",* International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 4, No. 2, June, 2011

[12] CHANG CHUAN YIN and STEPHEN S.-T. YAU, "*A Fourier Characteristic of Coding Sequences Origins and a Non-Fourier Approximation*", Journal of computational biology, vol.12, pp. 1153–1165, 2005

[13] Mahmood Akhtar, Julien Epps, Eliathamby Ambikairajah, "*Signal Processing in Sequence Analysis Advances in Eukaryotic Gene Prediction*", IEEE journal of Selected Topics in Signal Processing,vol.4,pp.468-472, 2008

[14] T. M. Inbamalar and R. Sivakumar ,"*Improved Algorithm for Analysis of DNA Sequences Using Multi Resolution Transformation*", The Scientific World Journal, vol.6,pp.678-680, 2015

[15] Hayes, Monson H.*," Statistical digital signal processing and modeling",*. John Wiley & Sons, 2009.

[16] D. K. Shakya, Rajiv Saxena, S. N. Sharma, "*Identification of Eukaryotic Genes with Improved Noise Suppression*" International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 4, No. 2, June, 2011

[17] Ahmad Rushdi , Jamal Tuqanand ,Thomas Strohmer, "*Map-Invariant Spectral Analysis for the Identification of DNA Periodicities*", pp.1-652,2011

[18] Mahmood Akhtar, Julien Epps, Eliathamby Ambikairajah, "*On DNA Numerical Representations for Period-3 Based Exon Prediction*", IEEE International ,vol.4,pp.456-459,2007.

[19] Michael Q. Zhang," *Computational Prediction of Eukaryotic Protein-Coding Genes*" Nature Publishing Group, pp.267-278,2002

[20] Mohammed Abo-Zahhad, Sabah M. Ahmed, Shimaa A. Abd-Elrahman, " *Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Techniques*" I.J. Information Technology and Computer Science ,pp. 22-36, 2012

[21] M Mahmood Akhtar, Julien Epps, and Eliathamby Ambikairajah, "*Signal Processing in Sequence Analyses Advances in Eukaryotic Gene Prediction*", IEEE journal of selected topics in signal processing, vol. *2,pp.* 333-344, June 2008.

[22] D. Anastassiou, “Genomic Signal Processing,” *IEEE Signal Processing Magazine*, pp. 8–20, 2001.

[23] 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.

[24] Fredric J. Harris, “*On the Use of Windows for Harmonic Analysis with a Discrete Fourier Transform”,* Proceeding of the IEEE, 1978.

[25] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, “*Prediction of Probable Genes by Fourier analysis of genomic sequences,”CABIOS*, pp. 263-270, 1997.

[26] T. M. Inbamalar and R. Sivakumar, “*Improved Algorithm for Analysis of DNA Sequences Using Multi resolution Transformation*”, The Scientific World Journal, vol.2, pp.324-333, 2015.

[27] H. K. Kwan and S. B. Arniker, “*Numerical Representation of DNA Sequences*”, IEEE Inter, Conf. on Electro/Information Technology, EIT '09, Windsor, pp.307-310, 2009.

[28] J. Tuqan, and A. Rushdi, “*A DSP Approaches For Finding the Codon Bias in DNA Sequences*,” IEEE Journal of Selected Topics in Signal Processing, vol. *2*, pp. 343-356, 2008.

[29] Gene Finding”. Online. [Available] https://www.wormbase.org

[30] “Gene Finding”. Online. [Available] ncbi.nlm.nih.gov

[31] http://www.wormbase.org/species/c_elegans/transcript/ F56F11.4b#06--1