IMPLEMENTATION AND DEVELOPMENT OF THE DSP ALGORITHM FOR SPLICE SITE PREDICTION IN DNA SEQUENCE

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

MASTERS OF TECHNOLOGY

IN

ELECTRONICS & COMMUNICATION ENGINEERING

By

Kanika Sandal

Enrollment No.: 152009

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 THE DSP ALGORITHM FOR SPLICE SITE PREDICTION IN DNA SEQUENCE"** 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.

( )

Kanika Sandal
Enrollment no. 152009
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 THE DSP ALGORITHM FOR SPLICE SITE PREDICTION IN DNA SEQUENCE"** which is being submitted by **Kanika Sandal** 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

I am greatly indebted to my guide Mr. Pardeep Garg, Asst. Prof at Jaypee University of Information and Technology for introducing me to this topic, suggesting this work and associated key paper for background reading and giving various opportunities during my study. Words cannot express my gratitude to his invaluable guidance and encouragement throughout my study. He is always accessible to me whenever I need help, in spite of his busy schedule. 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 dissertation 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 **Kanika Sandal**

ABSTRACT

Now a days genomics signal processing play an important role in bioinformatics area and one of the important problem of this area is gene identification. 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 algorithms. This project report gives the full study of numerical mapping for protein coding region using short time Fourier transform further the study is focused on exons (protein coding region) as to find its correct location in DNA sequence. Further the evaluation parameter that is (Sensitivity, Specificity and the area under the curve) are considered and are calculated. Twenty mappings are implemented and are compared with each other by comparing their area under the curve. This work provides an accessible study and a review of various Mapping techniques and prediction of exons with Digital signal processing tools and further to find exact location of protein coding region. Further this work provides the splice site of the sequence and calculation of score for Acceptor region with PSSM method.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

DNA	Deoxyribonuclic Acid
RNA	Ribonuclic Acid
DSP	Digital Signal Processing
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
FM	Fixed Mapping
PCPBM	Physico chemical property based mapping methods
SPBM	Statistical property based mapping methods
EIIP	Electron Ion Interaction Potential
WMM	Weight Matrix Method
WAM	Weight Array Method
WWAM	Windowed Weight Array Method
STFT	Short Time Fourier Transform
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

CHAPTER-1

INTRODUCTION

1.1 Biological Background

Cell is the structural and functional unit for all living organism. DNA (Deoxyribonucleic acid) encodes all the necessary information to run a cell. That can be viewed as the blue print for cell machinery, cell can be classified into two parts Prokaryotes and Eukaryotes. Further, eukaryotes are separated into small protein coding region called exons, interrupted by non coding region named as introns. In DNA numerical sequence each nucleotide is converted to numerical values through various mapping methods. Mapping of the sequence is necessary to apply a extensive range of tools, including Digital Signal Processing and machine learning methods. In recent years, many mapping methods are introduced to map nucleotides into numerical values of each nucleotide.

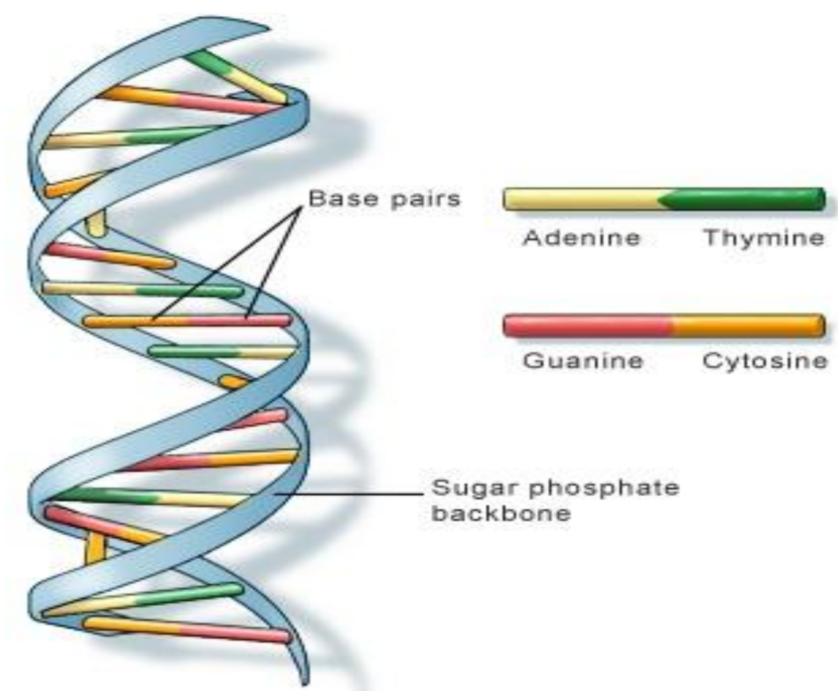


Figure 1.1 Basic Diagram of DNA

A number of exons prediction methods are projected and are discussed. Further to apply DSP tools to sequences, representative nucleotides of DNA must first be transformed into numerical values. Many representative mapping schemes are used to represent DNA nucleotides. All biological properties are taken into consideration in order to safeguard its biological meaning. In general, avoiding decadence and redundancy are considered in numerical mapping scheme. In addition, there is management between the mapping and the mathematical tools applied to the DNA sequence. The representation of DNA into numerical sequence can be divided into three modules that is The Fixed mapping, the Physico chemical property based mapping, and the Statistical property based mapping.

The eukaryotic is further divided into genes and intergenic spaces. A gene is divided further into two sub- region called exons and introns as shown in Figure.1.2

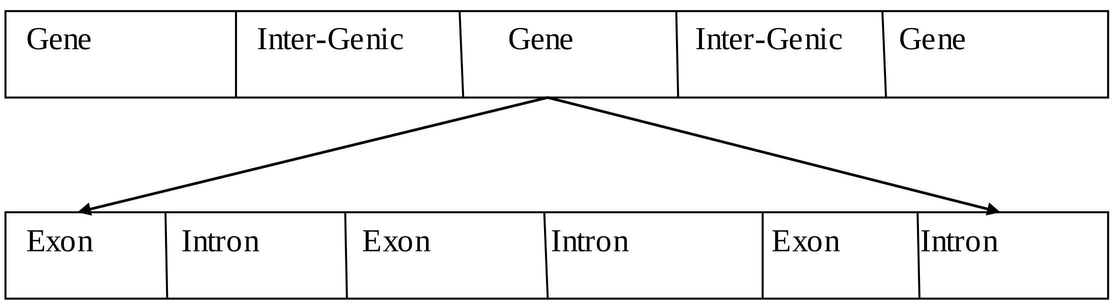


Figure 1.2 DNA structure of Eukaryotes

1.2 Numerical Representation

Some achievable characteristics of a numerical mapping include [2]:

(1) Compressed Representation

(2) Least redundancy

(3) Each nucleotide has equal amount

(4) Complementary structure of nucleotide pairs preserved

(5) Distance between all nucleotide brace are equal

(6) Natural and arithmetic information is captured or well modelled in numerical properties

(7) Skill to acquire information in various reading frames

(8) Depiction should not introduce any bias or spurious results

(9) Feasible to reconstruct the Sequence

(10) Compatibility with different mathematical analysis or DSP tools.

The numerical mapping can be divided into three major classification [1]:

(i) Fixed mapping methods (FM)

(ii) Physico chemical property based mapping methods (PCPBM).

(iii) Statistical property based mapping methods (SPBM)

1.2.1 Fixed mapping

In this type of technique, the DNA nucleotides of DNA are changed into a chain of capricious numerical sequences. The Fixed mapping include Voss mapping[2], Tetrahedron mapping [3], 2- Bit Binary mapping[4], 3-Bit Binary mapping[5], 4-Bit Binary mapping[6], Paired Nucleotide mapping[7]-[8], Integer Number mapping[9], Real Number mapping[10]-[11], Complex Number mapping[9], Pentanary Code mapping[12], Quaternion mapping[13]-[14], 12-Letter Alphabet mapping[15], and 18-Letter Alphabet mapping[16].

1.2.1.1 Methodology

The Voss representation is used to maps the DNA sequence C, G, A, and T into four display sequences as C_n, G_n, A_n, and T_n which shows the existence with 1 or nonexistence with 0 of the nucleotide [3]. In the Tetrahedron representation mapping is done by mapping its vertices which further reduce its indicator sequence [3]. The 2-Bit Binary mapping representation of the the nucleotides T, G, A, and C into two-bit binary as, 11, 01, 10, 00 respectively resultant into a 1-D sequence [4].The 3-Bit Binary mapping representation is a 1-Dimensional mapping of DNA nucleotides which can be achieved by mapping the sequence C, G, A, and T as 100, 010, 001 and 000 respectively [5]. Further, in the 4-bit binary the nucleotides C, G, A, and T are represented as, 0010, 0001, 1000 and 0100 respectively giving result into a 1-D display [6]. Paired Nucleotide mapping takes a DNA sequence and assigns binary values to it. Firstly gathering of T, A nucleotide is assigned 0 and nucleotide G, C are given 1, secondly T, C nucleotide are given 0 and nucleotide G, A are given 1, and in final gathering nucleotide T, G are given 0 and C, A nucleotide are assigned 1 resulting in 1-D display sequence [7]-[8]. The Integer mapping is represented as T=0, C=1,A=2, and G=3[9]. In Real Number mapping representation, A nucleotides is given as -1.5, T as 1.5, C as 0.5 and G as -0.5, which bear matching property and is efficient in finding the flattering strand into a DNA sequence [10]-[11]. The Pentanary Code representation can be obtain by mapping complex numbers {j, -j, 1, -1, 0} to the four nucleotides [12]. The Quaternion mapping representation estimates the sequence problem and also detects the protein coding region [13]-[14]. Coding regions may be described more absolutely with the 12-symbol alphabet due to the inbuilt codon bias in exons [15]. The 18-Symbol Alphabet mapping representation is the addition of the *A12* representation that takes into account the non-uniform distribution of stop codons along the three phases $p \in \{0, 1, 2\}$[16]. The Complex mapping representation reflect the harmonizing nature of A-T and C-G pairs as A as 1+j, C as -1+j, G as -1-j, and T as 1-j [9]. This results in a 1, 2 or 4 dimensional mapping of DNA bases

TABLE 1.1

Merits and Demerits of Fixed Mapping

S. No	Representation	Merit	Demerit
1.	Voss Representation	It offers a graphical and numerical representation, base distribution has efficient	idleness, demonstration is linearly dependent

		spectral detector [3].	[17].
2.	Tetrahedron Representation	Consist a periodicity detection and analysis of power spectrum [3].	Reduced idleness [3].
3.	2-Bit Binary Representation	Neural network Gene identification [6].	Linearly dependent
4.	3-Bit Binary Representation	Beneficial for inductive interference [5].	Various training and planning is required.
5.	4-Bit Binary Representation	Having identical Hamming distance [17].	Various training and planning is required.
6.	Paired Nucleotide Representation	Locate pattern and sequences in genomes. [7]-[8].	
7.	Integer number Representation	Simple and identification of protein coding region [9].	Mathematical problems that are not used in DNA sequence.
8.	Real Number Representation	Nucleotide AT and CG are complement [10]-[11]. Identification of protein coding region.	Mathematical problems that are not used in DNA sequence.
9.	Pentanary Code Representation	Nucleotide AT and CG are complex conjugate [12].	
10.	Quaternion Representation	Estimate the sequence problem [13]-[14].	Training is required
11.	12-Letter Alphabet Representation	Identify Protein coding Region	Working with DFT only [15].

12.	18-Letter Alphabet Representation	Detect borders between coding and non coding region [16].	Training is required
13.	Complex Representation	Complementary Features [9].	Bias in time domain analysis [11].

1.2.2 DNA physico chemical property based mapping

DNA physico type of mapping use to calculate the property of DNA sequence that is used for map the nucleotides that are used to search the various genetic ideology and structure in DNA. This mapping includes DNA walk[17] , EIIP[18], Z-curve[17], The 3-D Z-signals[19], Phase Specific Z-curve[20], Atomic Number[17], Paired Numeric[17], Molecular Mass[21], The Paired Nucleotide Atomic Number [22], The Simple Z [17], The Genetic Code Context[23].

1.2.2.1 Methodology

Methodology which allows envisaging the fluctuations directly of the purine content in a sequence is known ad DNA physico chemical property [17]. The slopes which are positive should match the highest level of pyrimidine, whereas the negative slopes correspond to the higher value of purine [17]. EIIP (Electron–ion interaction potential), a numerical sequence can be assigned to it such that the nucleotides are equal to the value of EIIP.

The EIIP values for the nucleotides are G=0.0806, A=0.1260, T=0.1335, C=0.1340 [18]. The Z-curve is a 3-D curve which provides a exclusive mapping of a DNA sequence in that the nucleotides (T-A, G-C) are given values of -1 and +1 are to be used to denote A-T and C-G nucleotide base pairs respectively [17]. The molecular mass representation mapping is a 1-dimensional display sequence shaped by mapping the molecular mass of the nucleotides A=134, C=110, G=150, and T=125 in a DNA sequence [21]. In paired nucleotide the representation of a nucleotide in a DNA sequence is represented as C, T=42 and A,G=62 respectively resulting in a 1- dimensional display sequence[22]. The simple DNA sequence and the Z-curve can each be individually reconstructed [17]. Therefore it carries all the information related to DNA sequence. The shape of the curve is zigzag, hence it is known as Z-curve. Digital Z-signals decomposes the DNA sequence into triple series of digital signals, based on Z-curves [19]. Phase-specific Z-curves detect the allocation of bases at first, second and third positions in a sequence, in order resulting in (9-D) mapping representation. The phase-specific Z curves contain three components, as compared to normal z-curve representation [20]. Atomic number representation display sequence is formed by handing over the atomic number to each base as C=58, A=70, G=78 and T=66 in a sequence [17]. The paired numeric representation Z (SZ) representation mapping is obtained by performing the maximum process on the 9-dimensional phase-specific Z-curves resulting in diminution of the size of Simple Z-curve features to one-third of the phase specific Z-curves [17]. In GCC, every uninterrupted nucleotide from the three reading frames in a DNA sequence is changed to an amino acid and each amino acid in turn is represented by a exclusive feature of DNA that is protein coding region [23].

TABLE 1.2

Merits and Demerits of DNA Physico Chemical Property Mapping

S. No.	Representation	Merit	Demerit
14.	DNA walk Representation	Offer numerical and graphical visualization Providing long range correlation information [17].	Not suitable for lengthy sequences [17].
15.	EIIP Representation	Better results for identifying protein coding region with hanning window [18].	Fail to detect in some genomes [18].

16.	Z-curve Representation	Understandable biological version, reduced computation, offers statistical and graphical Representation [17].	Not good for long and extended sequences [17].
17.	Digital Z-signals Representation	Detects small length and coding regions, clear biological meaning [19].	unsuccessful to perceive in some genomes
18.	Phase specific Z-curve Representation	Good recognition rate [20].	Higher number of Features [20].
19.	Atomic number Representation	Nucleotide fluctuations in genes	Requiring further Analysis [17].
20.	Paired Numeric Representation	Reflecting DNA structural property, improved coding region identification correctness over other methods [17].	Training is required
21.	Molecular Mass Representation	A consistent molecular weight is applied in nucleotide identification [21].	Requiring additional Analysis
22.	Paired Nucleotide Atomic Number Representation	Fractal dimension analysis of nucleotide [22].	Training is required
23	Simple Z Representation	Good identification rate using fewer features [17].	Functional for short length sequences
24.	Genetic code context Representation	Unique spectral analysis [23].	Training is required [23].

1.2.3 Mapping based on Statistical Property

In this mapping the DNA nucleotides are mapped in a binary form having different properties like the distance between the nucleotides, the definite data, and the nucleotide bias in terms of the coding and non coding information, the point count function, the codon file based on reappearance time. This mapping include Inter-Nucleotide Distance[24], Single Nucleotide Probability Indicators[25], Correlation Function[26], Binucleotide Distance[27], CCP[28], PCF[29], Codon Index based on repetition of Time[30], Ratio-R[28], Galois Field[31], Complexity[32], Frequency of Nucleotide Occurrence[14].

1.2.3.1 Methodology

In this mapping representation of each nucleotide is replaced with a number N which is the nucleotide space connecting the next analogous [24. This is known as one dimensional binary sequence. The single nucleotide bias probability indicator is the ratio of the normalize frequencies of DNA nucleotides C, G, A, and T in the coding and non coding regions of the dataset which Incorporates genome statistics and further can be represented as A=0.19, G=0.20, C=0.27, T=0.36 respectively [25]. Further this type of mapping Displays regular patterns in DNA sequences. In binucleotide mapping every base 'A' is represented as N which is the detachment to the next base 'T', every base 'T' by the detachment to the subsequent base 'A', every base 'C' by the detachment to the next base 'G' and every base 'G' by the detachment to the next base 'C'[26], If in some case a nucleotide is not present then the sequence value is the length of the last base in the sequence which further brings out the existence of spectrum in protein coding region [27]. CCP measures the subsistence of pairs of identical elements at a space of k base pairs [28]. The position count function (PCF) measures the number of times the nucleotides C, G, A, and T appear in the three positions within a codon along a DNA sequence [29]. It is computationally proficient and faster than STFT-based algorithms. Codon Index based on repetition of Time [30] represents a genomic DNA sequence hierarchically by quantify repeating patterns in a genomic for characterize period 3 feature in genome. The ratio-R representation [28] is the ratio of the count of bases (C or T) to count of bases (A or G) in an interval of window size defined by the user and repeat the process for the complete DNA sequence. The Galois field indicator is formed by giving the numerical values to the nucleotides as G=3, A=0, C=1and T=2 in a DNA sequence which is used to display regions which are exhibiting periodicity in a DNA sequence [31]. Complexity mapping representation is Suitable method for visualize various intricacy domains [32].

TABLE 1.3

Merits and Demerits of Statistical Property

S. No	Representation	Merit	Demerit
25.	Inter-Nucleotide Distance Representation	Depends upon distance Measure [24].	Not beneficial for biological information [24]
26.	Single Nucleotide Probability Indicators Representation	Incorporates genome Statistics [25].	Model dependent [25].
27.	Correlation Function Representation	Display usual patterns in DNA sequences [26].	Training is required
28.	Binucleotide Distance Representation	Depends upon distance Measure [27].	Not beneficial for biological information [27].
29.	CCP Representation	Identification of period-3 property and protein coding region [28].	Un successful to detect period-3 in some genes [28].

30.	PCF Representation	Provides an programmed DFT based approach for predicting protein coding regions[29].	Window cannot calculate the authentic boundaries of protein coding region [29].
31.	Codon Index based on repetition of Time Representation	Identify protein-coding Regions [30].	Slightly lower accuracy [30].
32.	Ratio-R Representation	Based on ratio of nucleotides [28].	Based on ratio of nucleotides [28].
33.	Galois Field Representation	DNA sequence analysis [31].	Requiring additional Examination [31].
34.	Complexity Representation	Efficient and earlier than STFT-based algorithms [32].	inability to detect very small coding regions [32].
35.	Frequency of Nucleotide Occurrence Representation	Identifies leading features first Before compressing [14].	Lower accuracy to find period-3 detection method [14].

CHAPTER-2

SPLICE SITE PREDICTION

Splice site prediction plays an important role to detect correct location of protein coding region, the boundary of Intron and exon are known as ‘AG’ Acceptor splice site whereas boundary of exon and intron are refer to as ‘GT’ that are known as Donor splice site. ‘AG’ Nucleotide is only present in exons know as protein coding segments.

The intergenic and intronic regions of humans often make up more than 95% of their genomes. Codons in exons instruct 3 terminator signals and 20 amino acids, known as stop codons. Initially protien starts with start codon refer to as ‘AG’ [35].

High calculation accuracy can often be recognized to affable instruction and test sequences [36] sequences that include of one complete gene with consensus intronic dinucleotide “*GT*” and “*AG*,”.

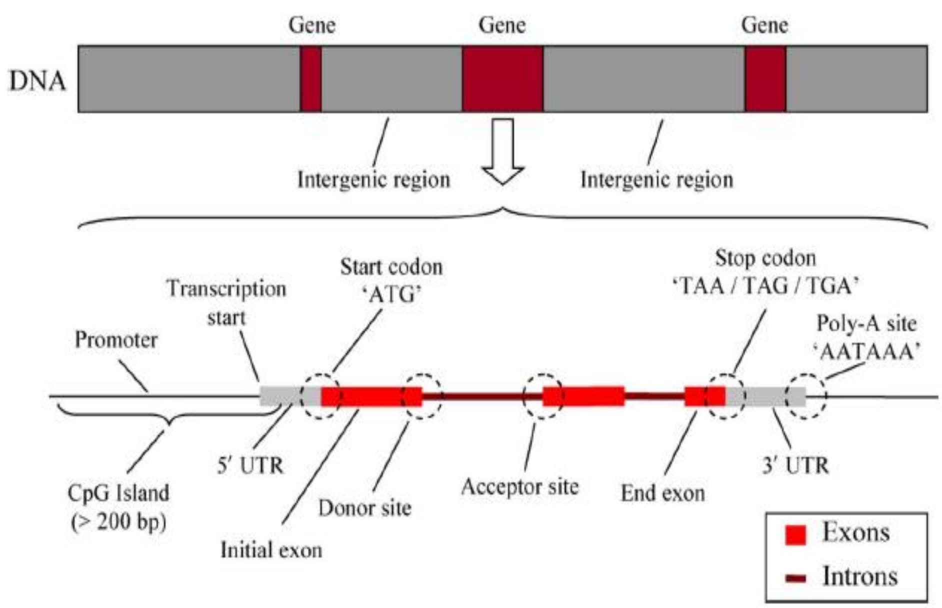


Figure 2.1 Eukaryotic diagram for acceptor and donor site.

2.1 Acceptor Splice Site Detection Method

There are number of methods used for splice site prediction as the accurate prediction for protein is very important as well as to detect the donor splice site prediction and also very important to detect the end points of DNA[35]. Due to the regular occurrence of nucleotide at locations other than acceptor sites throughout a sequence, detection is very difficult. In order to apply various methods, firstly the candidate should extract the 'AG' nucleotide as windows of 140 nucleotides around each consent dinucleotide "*AG*".

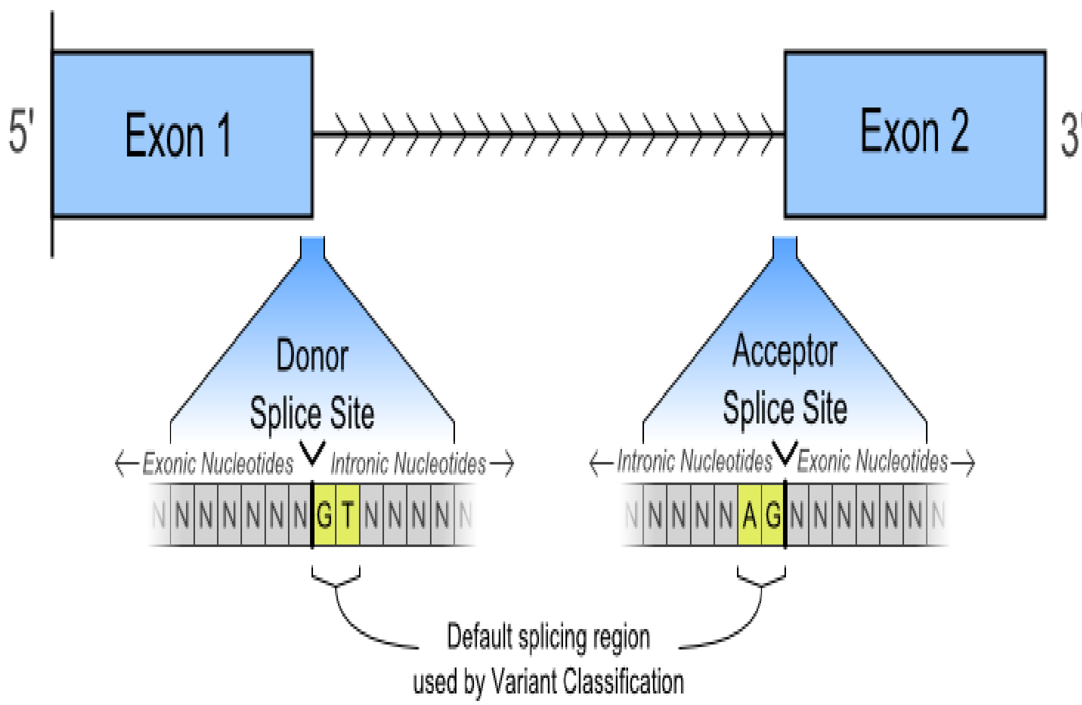


Figure 2.2 Basic Splice Site diagram

2.1.1 Weight Matrix Method (WMM)

This method the probability of each nucleotide is calculated and each nucleotide is independent to each other [35]. The probabilities of generating a signal *k* of length *N* under negative and positive wmm's of the pyrimidine-rich acceptor region are

$$\mathrm{wmm}\{k\} = \prod_{K=1}^{N} \mathrm{P}_{m}^{s}$$

Where p_m^s refer to as probability of generating nucleotide m at point of the signal, which is estimated from the positional frequencies of nucleotides from the false and true acceptor site sequence. Therefore, the negative and positive probabilistic models are calculated using false and true acceptor site sequences, respectively.

2.1.2 Weight Array Method (WAM)

The Weight Array Method and capture the dependency between closest positions, in contrast to the WMM, which considers each position independently. [38]The probability of generating a indicator of length under negative and positive WAM of the acceptor region which can be computed as

$$\mathrm{WAM}\{k\} = P_{k_1}^{1} \prod_{\mathrm{K}=2}^{\mathrm{N}} \mathrm{P}_{m,n}^{s-1,s}$$

Where $\mathrm{P}_{m,n}^{s-1,s}$ is the conditional probability of the nucleotides.

2.1.3 Windowed Weight Array Method (WWAM)

The WWAM Method is a second-order weight array Method model, in this method conditional point branch is generated on the nucleotides of the previous positions of a sequence [40].

a) Score Calculations for 'AG' Acceptor Site:

$$SCORE(S) = \log_2\left(\frac{\sum_{Lc}^{Uc} P3}{\sum_{Lnc}^{Unc} P3}\right)$$

Score of each Acceptor site is defined as the ratio of the sum of 3-base periodicity features which is denoted by 'P3', in non coding region as well as for coding region. Where L and U are the upper and lower indices of the period 3 features, coding region is denoted by 'c' whereas
Non coding region is denoted by 'nc'.

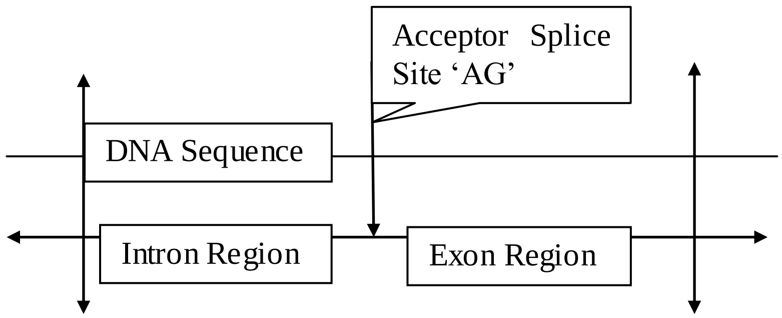


Figure 2.3 Acceptor Splice Site Region

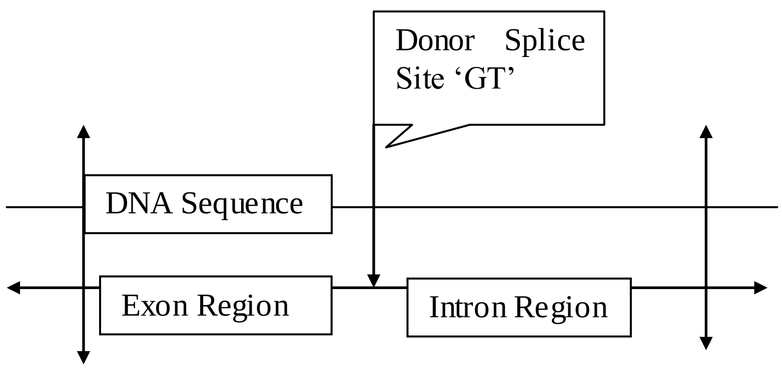


Figure 2.4 Donor Splice Site Region

2.1.4 Markov Models for Splice Site

Markov Model 1, Markov Model 2 is the most accepted methods in use for splice site detection [41]. They aim to study the preserved sequence pattern at downstream and upstream regions surrounding the splice site region (GT-AG). Firstly the markov model one processes the DNA input sequence and generate some position probability parameters which are also known as emission probabilities [42].

2.1.5 Support vector machines

Support Vector Machines is the best method used for detection of Splice Site [43]. Firstly DNA sequence should be converted into numerical values and this will the input for support vector machine. So, therefore converting DNA sequence into numerical form is a basic and important task for Splice Site Prediction [44]. The accurateness of the Support Vector Machines (svm) depends on the proper parameters. The Support Vector Machines aim to locate a maximal margin hyper plane to separate classes.

2.1.6 Estimation of Distribution Algorithms (EDA)

EDA is also a good method for calculating splice site prediction. Basically it try to surmount difficulties by providing a more numerical analysis of the selected individuals (AG nucleotide for acceptor region and GT nucleotide for donor region), thereby explicitly modeling the relationships among the variables is done.

2.1.7 Position specific scoring matrix (PSSM)

PSSM (position specific scoring matrix) is a proposed method for splice site prediction. Position specific scoring matrix is also known as Position specific weight matrix. Basically it is used to represent sequence. They are the set of the functionally aligned sequence to calculate the score in a sequence and to study each nucleotide in a sequence. Basically the sequence is changed into a matrix form or in a 2- dimensional array. Its main advantage is it read the biological sequence and it is also an essential component for the modern algorithms.

Sequence to PSSM

Firstly the string of the DNA sequence is arranged and are displayed individually in rows and column representing different nucleotide (A,T,G,C), now the next step is to position the frequency matrix by counting its occurrence individually in a pattern at each position. Further the position probability matrix is generated now to normalize the sequence divide the former base sequence at each position by the no of sequence present.

Normalizing Score =Raw Score/ Overall frequency of given nucleotide.

Windowing in PSSM (Position Specific Scoring Matrix)

Size of the window plays an important role in calculating the acceptor site in a DNA sequence. It is given that smaller the window length more accurate will be the location of acceptor site. Its main advantage is that it is easy to compute and can be used in comprehensive evaluations and also having a minimal assumption.

CHAPTER-3

OBJECTIVE

The principle objective of this dissertation is to investigate and develop efficient prediction of exon and to attain a proper splice prediction of protein coding region with the help of various mapping techniques, DSP tools, different windowing and further the implementation of splice site prediction techniques to get a exact location of AG nucleotides (acceptor splice sites). The exact location of exons is very important to detect so to calculate the desired boundary range splice site prediction is important to implement after calculation of protein coding region from STFT algorithm. Further to implement PSSM to calculate the score to find 'AG' in a DNA sequence.

CHAPTER-4

LITREATURE REVIEW

Nowadays 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 digital signal processing tools to DNA sequence, mapped nucleotides of DNA must be transformed into a numerical sequence and these are affecting the performance of the algorithm.

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. Each Nucleotide is independent and plays an important rule that each nucleotide has an equal weight as all nucleotide in a sequence is independent r even though some researchers have studied the parameters corresponding to the nucleotides (A G T C) [47]. One more property that has been considered in some mapping schemes is the categorization of pyrimidine (C,T) and purine (A,G) and also the potency of the hydrogen bond. This property is used in Z-Curve and complex mapping representation [48]. Generally, removing degeneracy and redundancy is taken into contemplation in mapping schemes. There is always synchronization between the mapping scheme and the digital signal processing tool applied to the numerical data. In 1992 Voss a technique was proposed in which numerical values is composed of K symbols that can be decomposed into ‘K’ different binary sequences. [49]-[50]In this way problem of association can be solved easily which occurs between the nucleotides of DNA sequence when the dimension is less than one. One is represented as where the given nucleotide is present rest it will be represented by zero in a DNA sequence.

In 1994 Zhang and Zhang [43] proposed a representing mapping scheme into series of three signals known as the Z- curve mapping representation, also known as 3D representation of

DNA sequence. Values of these series consist of +1 and -1 [51].The initial series characterize the allocation of the nucleotide with various hidden features. The following series characterize the classification of amino\ kito type of the nucleotide in a sequence. Finally the last series of the hydrogen bond (weak/strong) and also its strength of nucleotide in a DNA sequence. Therefore, the digital values are represented with three series of digital values.

The frequency occurrence mapping is another method which is comparable to the Voss mapping. Complex mapping is also related to Voss mapping the only difference is the output is in complex format. Various Tools are been analyzed by M. Akhtar (2008). Singular valued decomposition was also used to analyse the protein coding region. Sanjay Verma in 2015 gave a new algorithm for protein coding region with Goertzel Algorithm which is also giving good results for the prediction.

Further the Vyan Syus suggested a method which gives the new feature regarding splice site with EDA method. It gives a scoring value for acceptor and donor site, which further tells the boundary of exons and intron region.

CHAPTER-5

METHODOLGY

Genomic DNA sequence is calculated using digital signal processing (DSP) techniques such as filtering, alteration and information compression that has been attracting the interest of researchers. DSP is an significant region of engineering which comprehend the management of numerical mapping to produce a eminence signal than the original one [54]. The DSP tool in a genomic sequence is the new field use to accomplish goals such as gene prediction, locations of proteins. This refers to understand things in a proper way.

There are a array of Digital signaling processing tools which are used for exon prediction like [55] DFT, STFT, FFT and Wavelet Transform. DSP methods intend gene prediction methods that calculate 3- base periodicity distinguish non coding and coding regions. DSP tool is used to calculate the control range peak at frequency f=1/3 in sequence. [49]It processes a DNA sequence based on a window also called as a sliding window. Firstly DNA sequence is changed to numerical sequence, and then a sliding window is applied which is shifted along the sequence. As each time the shifting is done by a window, the power spectrum is calculated frequency k/3 is calculated in order to extort the 3-base periodicity property in the sequence.

After calculating the 3-basr periodicity [49] sequences are classified as non coding and coding regions. Arrangement can be performed using various thresholds. Thresholding is considered as the main challenge in the field of genomics since the assortment of its best value varies from one DNA sequence to another.

Short Time Fourier Transform

STFT (short time Fourier transform) is defined as a signal that has frequency content which is changing over time. It is basically a Fourier related transform [51]. The basic procedure to calculate the STFT is to divide the larger time signal into small signals of equal length and then calculate the Fourier transform of signal separately on short length [52]

$$X(m,k) = \sum_{n=\infty}^{\infty} x[m+n]w[n]e^{-jkm}$$

Fast Fourier Transform

FFT compute the DFT that is discrete Fourier transform of a given sequence, and is used to calculate its inverse [53]. Fourier converts a signal from its original domain into time and frequency domain.

$$X(k) = \sum_{n=0}^{N-1} x_n \; e^{-\frac{j2\pi kn}{N}} \qquad k = 0, \ldots\ldots.., N-1$$

5.1 Gene Prediction Process

a) This process includes firstly to collect a DNA sequence then further to change the symbolic representation into numerical representation.

b) The next step is to apply a window function and before that do the zero padding such that the sequence comes in between and all sequence is read that is all nucleotides are considered.

c) All the methods that are used to calculate the protein coding region and further to calculate the period 3 component in a DNA sequences.

d) 3 Base periodicity is calculated which tells the coding and non coding region. Further Threshold is set which is considered as one of the difficult challenge which varies from one sequence to another.

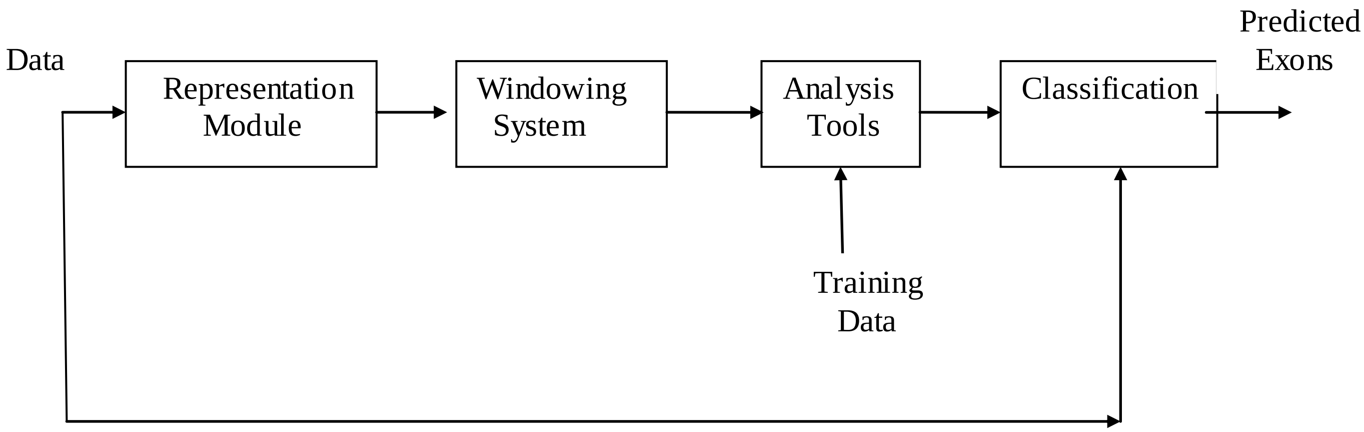


Figure 5.1 Block diagram for gene prediction process

e) The DNA sequence is converted into the numerical representation with various mapping, after that a hanning window (351) is applied to the sequence and further STFT approach is applied to the sequence to predict protein coding region [34].

$$X[t] = \sum_{n=0}^{N-1} x(n).w(n)e^{-j2\pi nk/N} \quad , 0 \leq n \leq N\text{-}1 \qquad (1)$$

as $w(n)$ is a Hanning window,

$$w(n) = cos^2\left(\frac{n}{N}\pi\right) = \frac{1}{2}\left[1 + \cos\left(\frac{2n}{N}\pi\right)\right] \quad n = -\frac{N}{2}, \dots \dots - 1,0,1 \dots \frac{N}{2}$$

The Hanning window is good as it forces end to zero, wave is being analyzed with the amplitude modulation. It usually selects a subset of a series of sample in order to apply Fourier transform. It has decreased trade off and has a low aliasing which is its advantage.

$$S[t] = |X[t]|^2 \qquad (2)$$

When S[t] is plotted against t, it gives a peak at N/3 for a sequence. Hanning windows is considered with a window length of 351 are used for calculating the STDFT of the F56F114 (8100bp) gene in C.elegans [34].

f) The sensitivity, specificity and AUC, are calculated as these are the evaluation parameters for exon prediction and to measure the effect of different numerical mapping representation and also to calculate the overall efficiency for the prediction of exons.

Sensitivity (Sn) =TP/ (TP+FN)

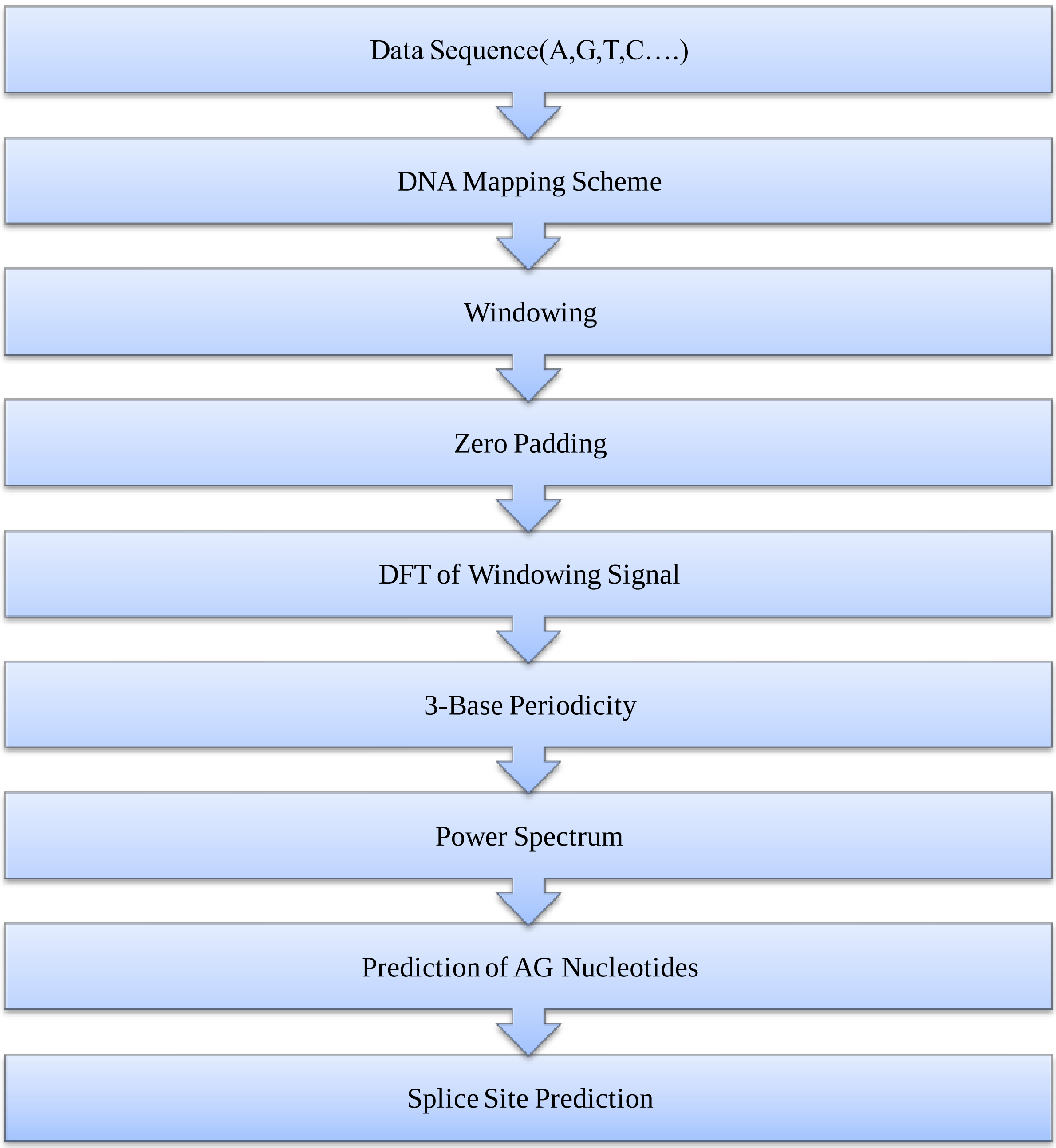
Specificity (Sp) =TN/ (TN+FN)

ROC (Receiver operating characteristics)

ROC (Receiver operating characteristics) curves have been used for the performance analysis of a method on the basis of AUC (area under the curve) of ROC [41]. AUC of ROC curve defines the accuracy of the method and it is directly proportional to AUC of ROC curve.

ROC curve depicts the performance of TP and FP at different Threshold [35].

5.2 Procedure for Exon Prediction and Splice Site Prediction



CHAPTER-6

RESULTS

Firstly the data is collected from the [53] and [54] the data is present in fasta format and further the sequence is present in spliced or unspliced form, but we only take unspliced DNA sequence for the exon prediction. There are number of sequences for different type of species for example C- elegan (f56f11.4a).

In order to achieve the desired result, we have applied the prediction technique and various mapping on C-elegan sequence F56F114 (8100bp) testing genes for exons prediction downloaded from worm base dataset.

```
TTGAATTCAA TTAAAACATG CTTTTTTGGG GGTAAAAAGA GCAACAAAAA
ATTTTTTCAA ACTGGGGAAA TCCGTCTTGG GCTCAATTTT GCTCCGAACT
TAGTGCCGTT TTTTGCTCCA CCGTGGGGCT AAATATTTCT AGTAGGATTT
CAAATATTAG AACATGAAGT CACACGGCTC TGGAGTTATT AACGAAAACG
AAAGGGGACA TTTTTTCGCA AGCCAAAAAA AACGCGAAAA AACGCGAAAA
AGGGGCGGAG TCGCCACACT CGGCATTTAT TAGAGGCTGC TTGGCGTTTT
TCCTTGGAAT GTCCAGTTTG TTTCTTAAAT TTAGTCATTT TCAAGATTTT
TCCTATTAAA AATCTGAAAA TTTTTAAAAA TTATTTTTAC TGTAGAAGTA
CACCGCGACG CAAAATTGCG TACCAGCGGG ATTTTTTGAT TATAATTATA
ATCGATTTTA ATAATTTTTT TGCTGTTTTG TTATGATTTT TGTTGATTAG
TCCGCAATTT TGCACGATTT TCATTCATTT TTTACGAAAA TCTAGTTAAT
TTTATCAAAA CTCTCATTAA AAATCATATT TTTAGCCTCC TTTGTAACCA
AATTCAAAAA ACACGAAACA AAAAATTGTG TTCACTCATT TTTTACTGAA
AAATTAGAAT TTTCACATTA TTTTATGTTA AAACATCAAA ATTCCACTTA
TTTTCTGGAA TTTCCCGCTC GAAATCTTTA AAAATTAAAT GAGGGTTACT
GTAGAAGTAC ACCGCGACGC AAAATTGCGT ACCCGCGGGA TTTTTTGATT
TTATTTATTT TTGTGCGGGT TTTTGCGGTT GTGTCCCCAT TTTTGTCGAT
TTTCATGTGA TTTTGCCCAA TTTTAAGGTT TTTCCAGCCA TTTTAATTTA
ATTTTTCATA AAAATTGCAA TTTTCAGAAT GGGTCCAGCT CCAGTTTTCC
CAAATTCGCG GAAGCCGGCG ACACTTGACG CGAACTCTGA CGAGCAGACA
CTTCGTCCGT ATTTTAAGAC GAAAGTTGAA CAAGCGGAGG TGAATTTTCG
AGAAAATAAC TGGAAAAAAC GAATTATTTC
```

Figure 6.1 Basic DNA sequence

```
Columns 1 through 13

   0    0    0    0    0    1    1    0    1    1    0    0    1
   0   -1    0   -1    0    0    0   -1    0    0    0   -1    0
  -1    0   -1    0    0    0    0    0    0    0    0    0    0
   0    0    0    0   -1    0    0    0    0    0   -1    0    0


 Columns 14 through 26

   1    1    1    1    0    0    0    1    1    0    0    0    0
   0    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0   -1   -1    0    0    0    0    0    0
   0    0    0    0   -1    0    0    0    0   -1   -1   -1   -1
```

Figure 6.2 Numerical Mapping of a DNA Sequence

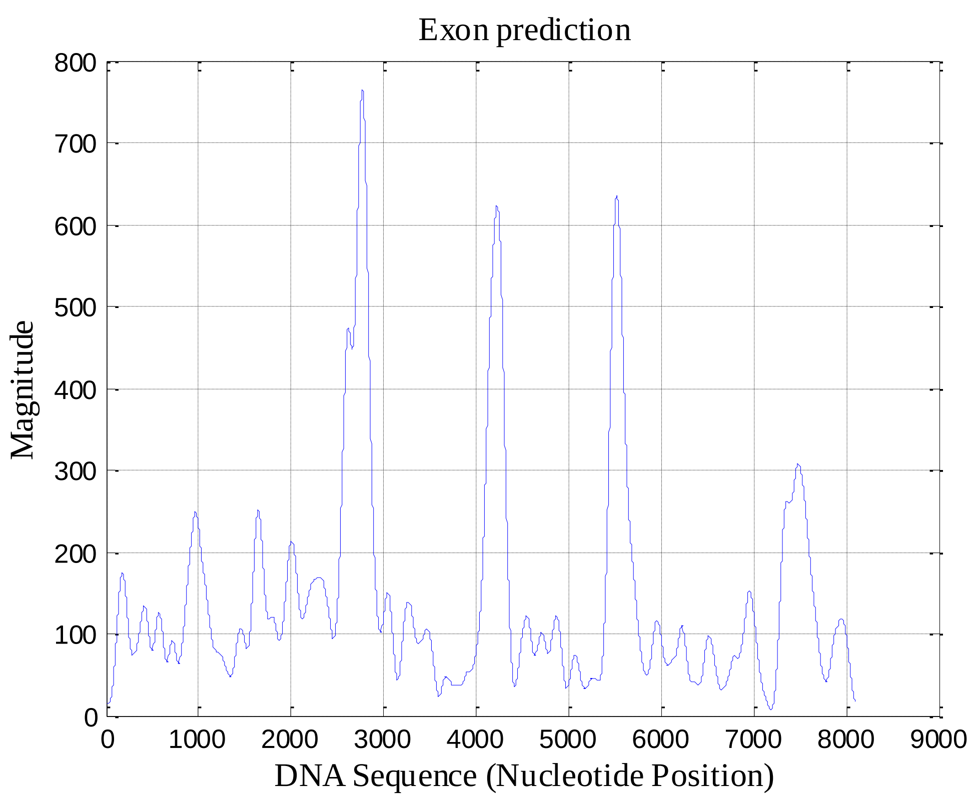


Figure 6.3 Prediction of exons from Paired Numeric Mapping for sequence F56F114 (8100bp)

TABLE 6.1

Actual location of Exons

Exons	Start	End	Length
1	929	1137	208
2	2528	2857	329
3	4114	4377	263
4	5465	5644	179
5	6342	7605	1263

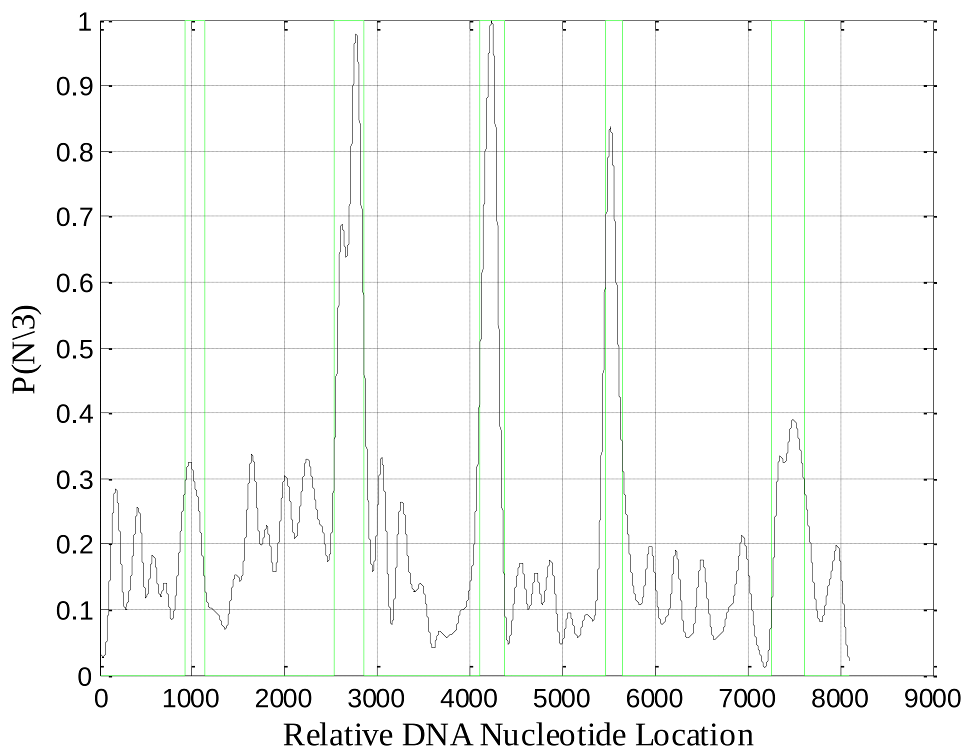


Figure 6.4 Comparison for sequence F56F114

Comparison for sequence F56F114 is shown in figure 6.4. The actual locations of exons are shown by the green dashed box, and the predicted exons shown by the black. It gives the values where the exon is present in a sequence which further helps to calculate the Evaluation Measures which include Sensitivity and Specificity

TABLE 6.2

Results and Performance Analysis for various Mapping Techniques

S. No	Mapping	Numerical Representation k(n)=[AGCT]	AUC
1.	Voss Representation	A_n =[1 0 0 0], G_n=[0 1 0 0] C_n=[0 0 1 0], T_n=[0 0 0 1]	**0.8155**
2.	2- Bit Representation	[11,10,00,01]	**0.8127**
3.	3-Bit Representation	[010,001,100,000]	0.7506
4.	4-Bit Representation	[0010,0001,1000,0100]	0.7312
5.	Integer Representation	[1, 3, 2, 0]	0.7664
6.	Real Number Representation	[0.5, -0.5, -1.5, 1.5]	0.7783
7.	Complex Representation	[-1-j, -1+j, 1+j, 1-j]	0.7739
8.	EIIP Representation	[0.1340, 0.0806, 0.1260, 0.1335]	**0.8100**
9	Atomic Number Representation	[58, 78, 70, 66]	0.7272
10	Paired Numeric Representation	[-1, -1, 1, 1]	0.7746
11	Molecular Mass Representation	[110,150,134,125]	0.7720
12	Paired Nucleotide Representation	[42, 62, 62, 42]	0.7721
13	DNA Walk Representation	[1, -1, -1, 1]	0.7435
14	Inter Nucleotide distance Representation	[3, 2, 1, 0]	0.7686

15	Binucleotide Distance Representation	[1, 2, 1, 0]	0.7382
16	Single Nucleotide Probability Representation	[0.27, 0.20,0.19, 0.36]	0.7808
17	Galois Field Representation	[1, 3, 0, 2]	0.7620
18	Complexity Representation	[0.60, 0.35,0.79,0.0]	0.7362
19	Frequency Nucleotide occurrence Representation	[0.28142, 0.28179,0.23326, 0.20354]	0.7490
20	Pentanary code Representation	[-j, -1, 1, j]	0.7768

TABLE 6.3

APPLICATION OF VARIOUS MAPPING

S. No	Mapping	Application
1.	Voss Representation	Detection of protein coding Regions.
2.	2-Bit Binary Representation	Gene Identification.
3.	3-Bit Binary Representation	Power spectrum Study.
4.	4-Bit Binary Representation	Gene Identification.
5.	Paired Nucleotide Representation	Periodicity detection.
6.	Integer number Representation	Autoregressive model and element analysis of DNA sequences.
7.	Real Number Representation	Splice junction identification with neural network.
8.	Pentanary Code Representation	Wavelet transform of the 3-dimensional DNA Walk.
9	Complex Representation	Phase analysis in 2D and 3D in complex and vector sequence respectively.
10	DNA Walk Representation	It gives graphical representation for

		DNA sequence.
11	EIIP Representation	Identifying protein Coding regions.
12	Atomic number Representation	Nucleotide fluctuations in genes
13	Paired Numeric Representation	Fractal dimension Study of nucleotide is done.
14	Molecular Mass Representation	Gene identification and protein.
15	Inter-Nucleotide Distance Representation	Reveals the existence of coding region.
16	Single Nucleotide Probability Indicators Representation	Improves the intolerance capability of genes.
17	Binucleotide Distance Representation	Reveals the existence of inequitable spectral envelope.
18	Galois Field Representation	Analysis of DNA sequence.
19	Complexity Representation	To demonstrate regions which exhibit periodicity.
20	Frequency of Nucleotide Occurrence Representation	Study of long range Association of sequence.

6.1 Graphical representation of Mapping

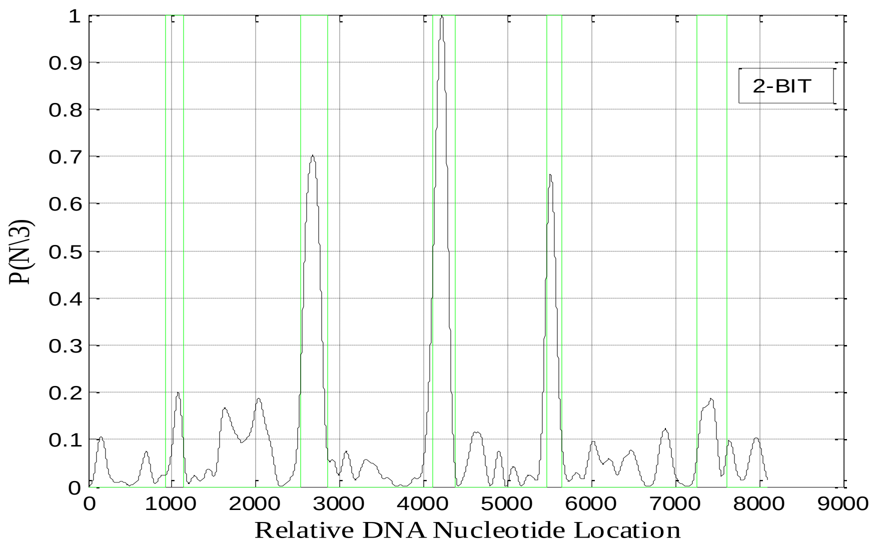


Figure 6.5 2-Bit Binary Representation

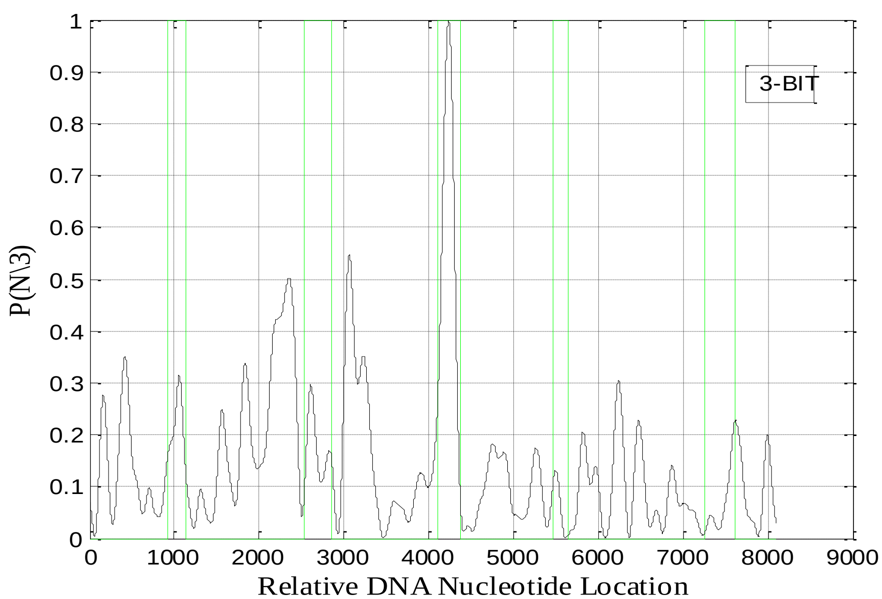


Figure 6.6 3-Bit Binary Representation

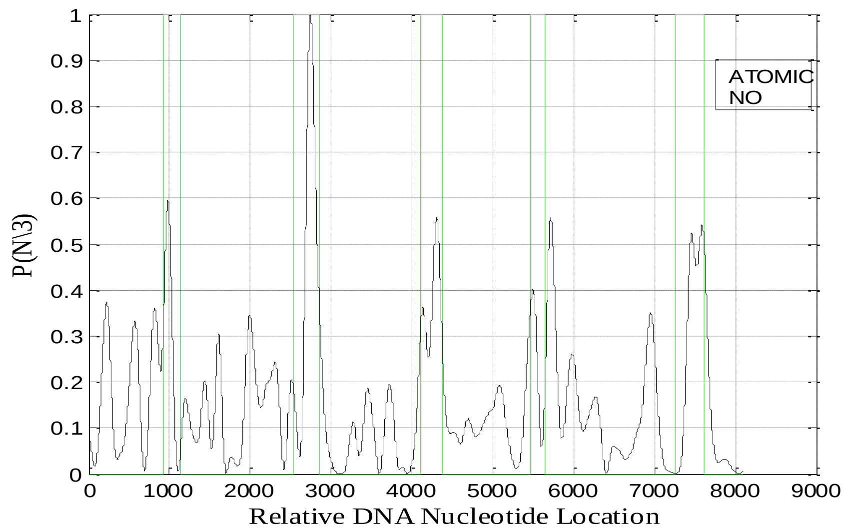


Figure 6.7 Atomic Number Representation

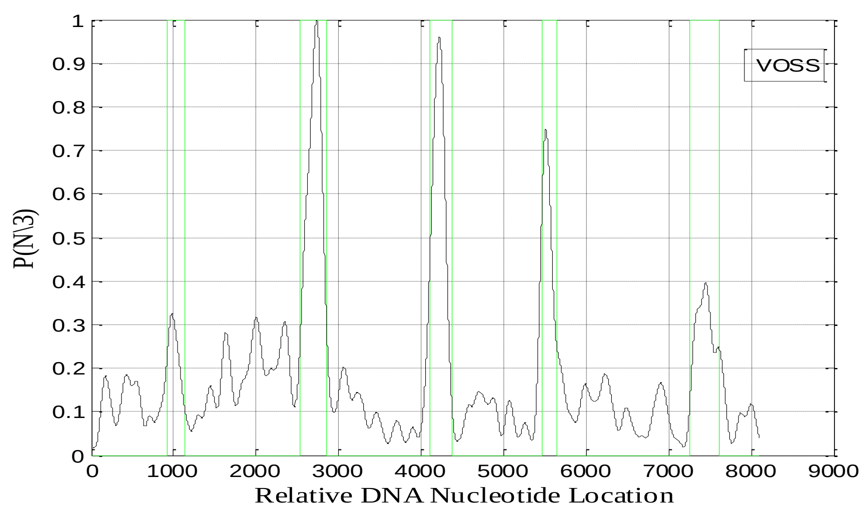


Figure 6.8 Voss Representation

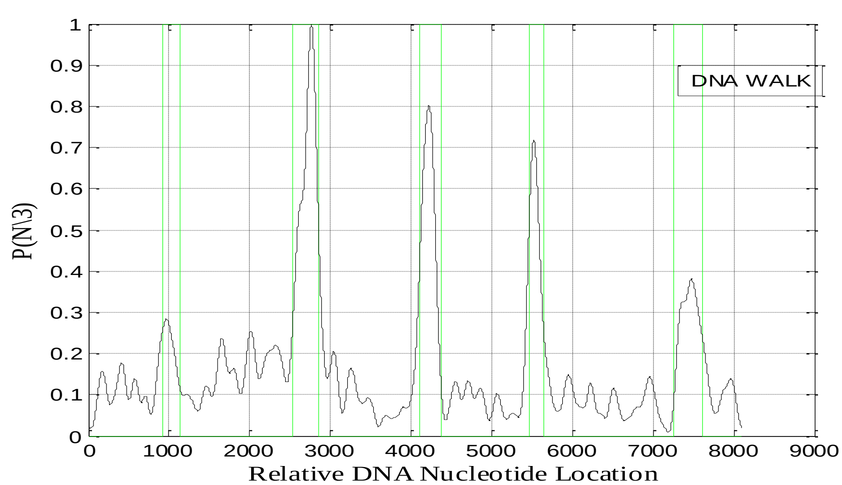


Figure 6.9 DNA Walk Representation

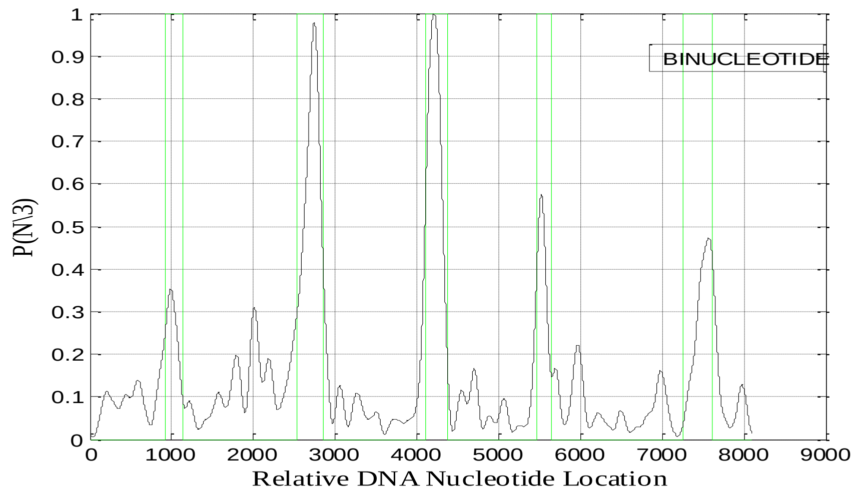


Figure 6.10 Binucleotide Representation

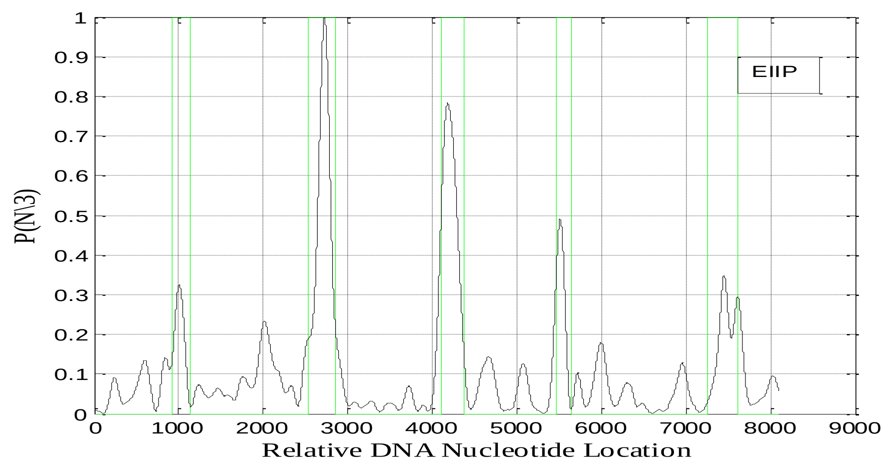


Figure 6.11 EIIP Representation

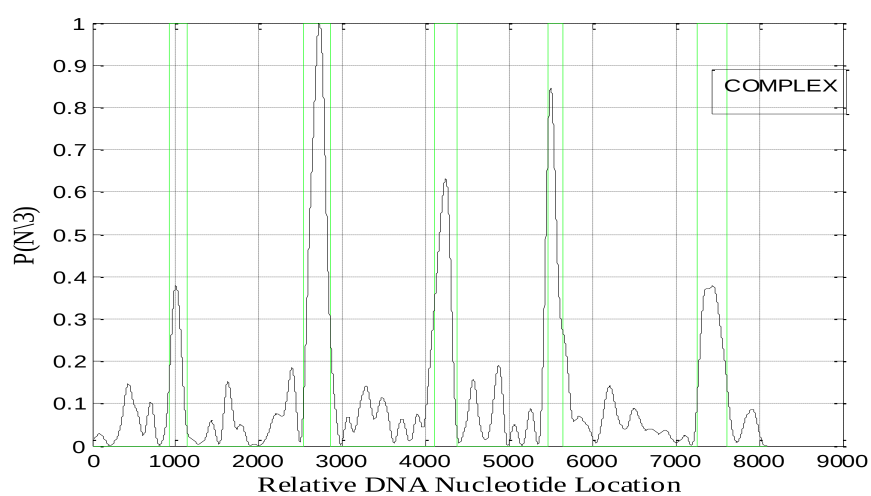


Figure 6.12 Complex Representation

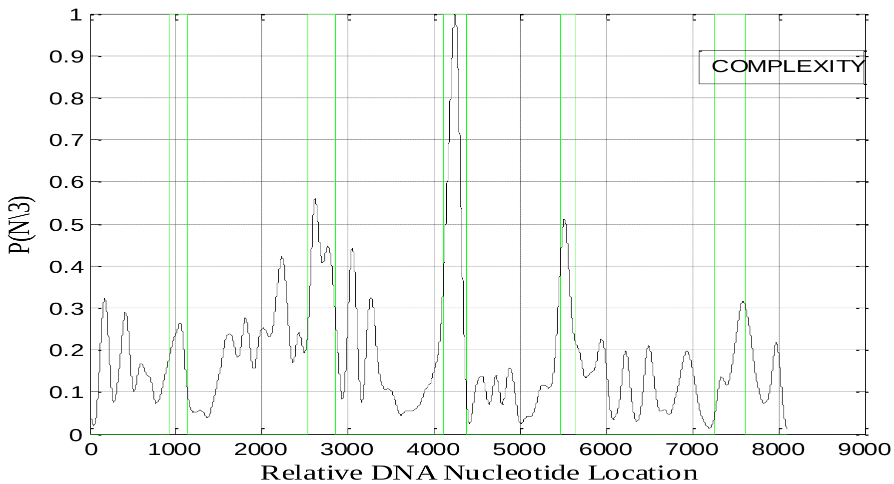


Figure 6.13 Complexity Representation

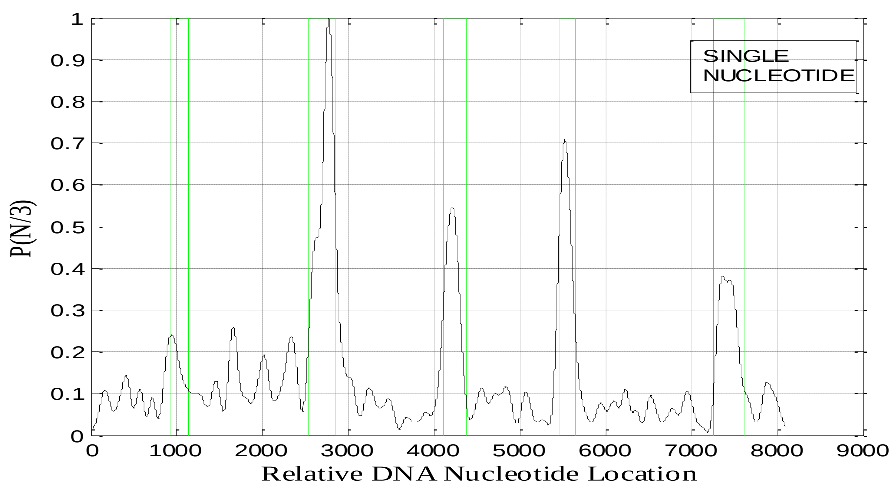


Figure 6.14 Single Nucleotide Representation

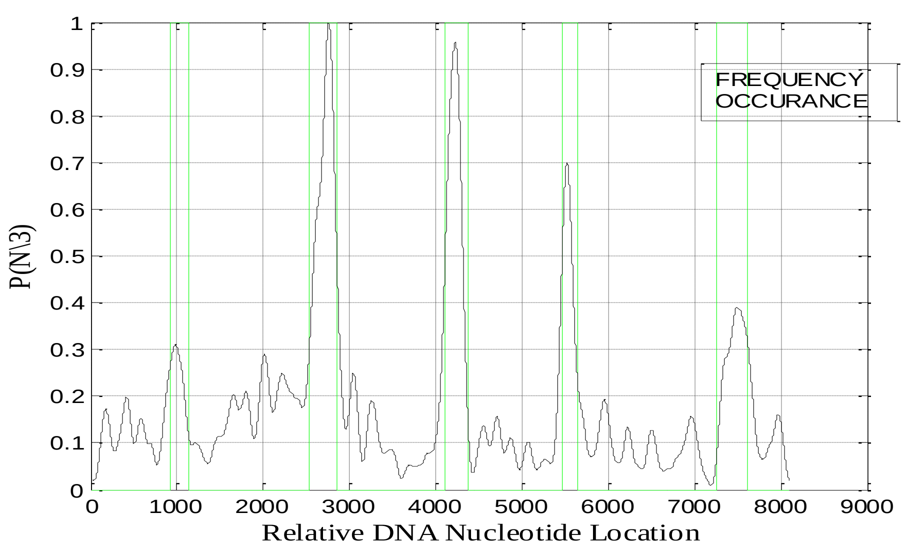


Figure 6.15 Frequency Occurrence Representation

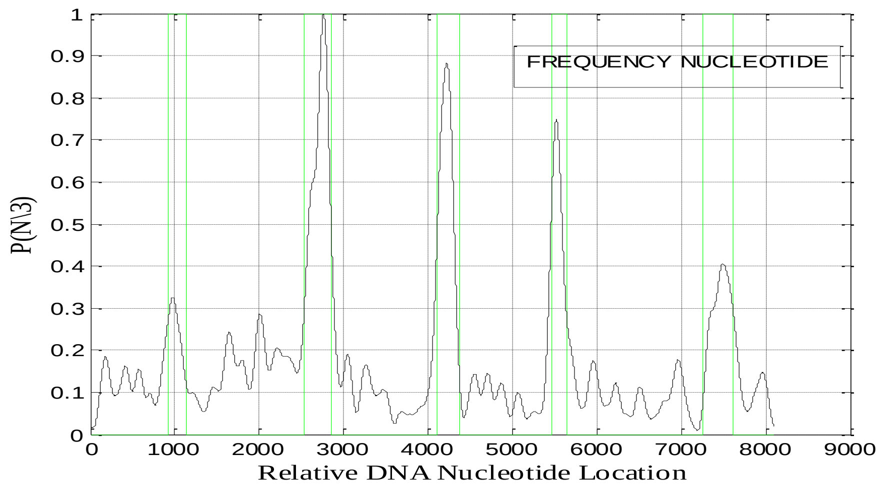


Figure 6.16 Frequency Nucleotide representation

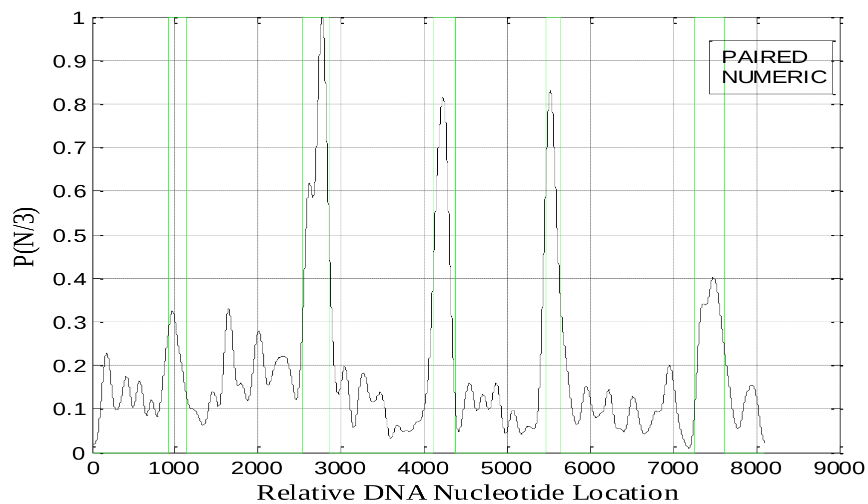


Figure 6.17 Paired Numeric representation

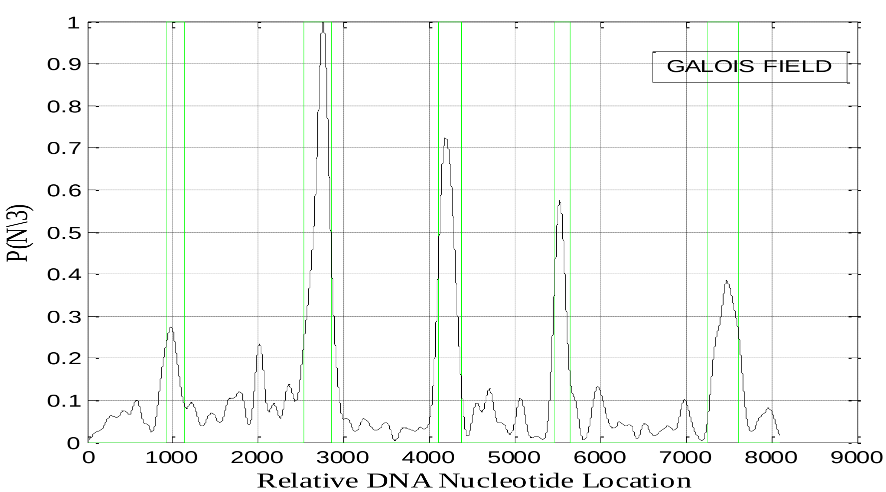


Figure 6.18 Galois Field Representation

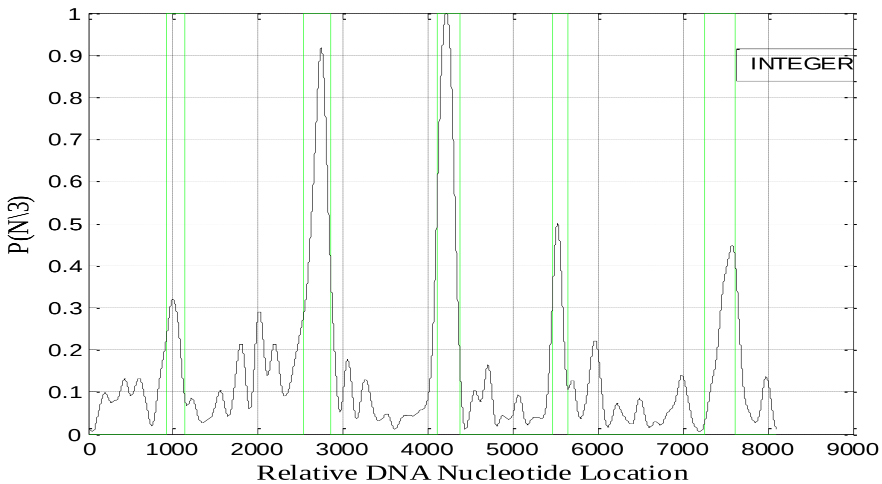


Figure 6.19 Integer Representation

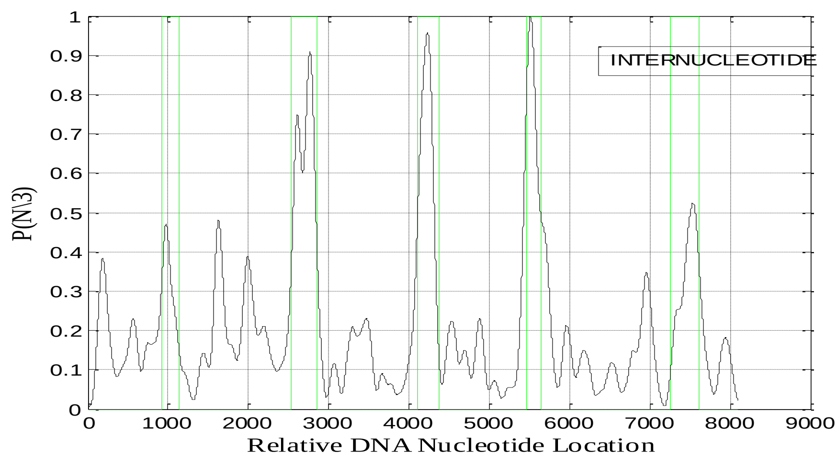


Figure 6.20 Internucleotide Representation

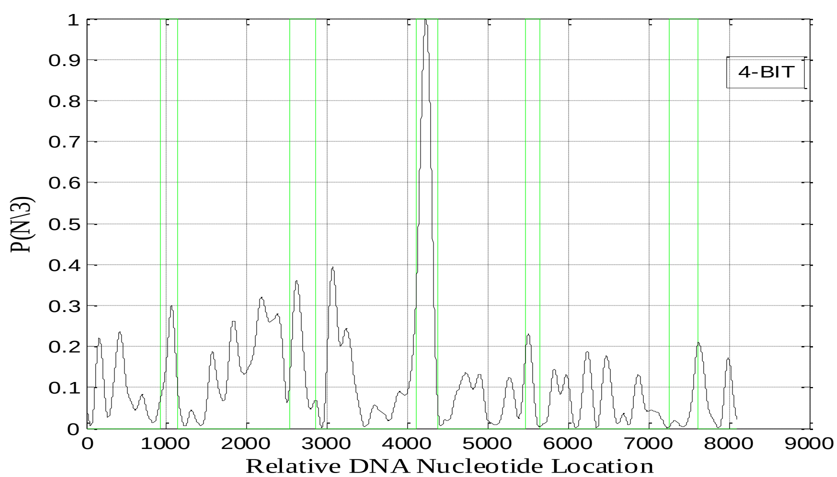


Figure 6.21 4-Bit Binary representation

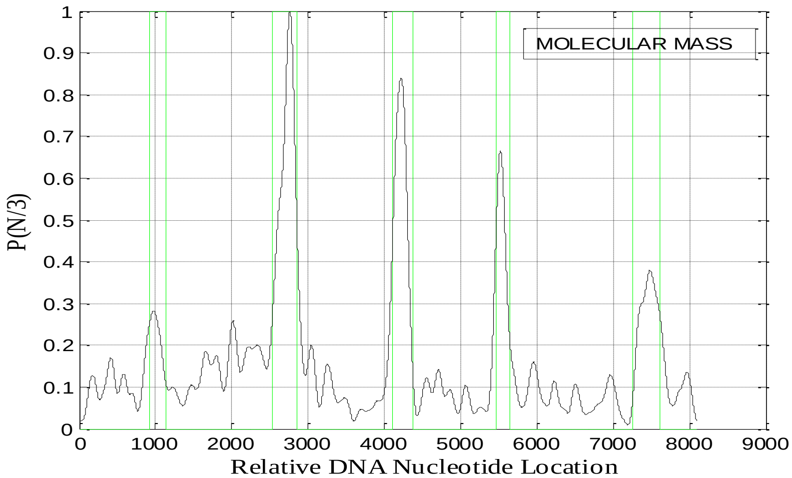


Figure 6.22 Molecular Mass representation

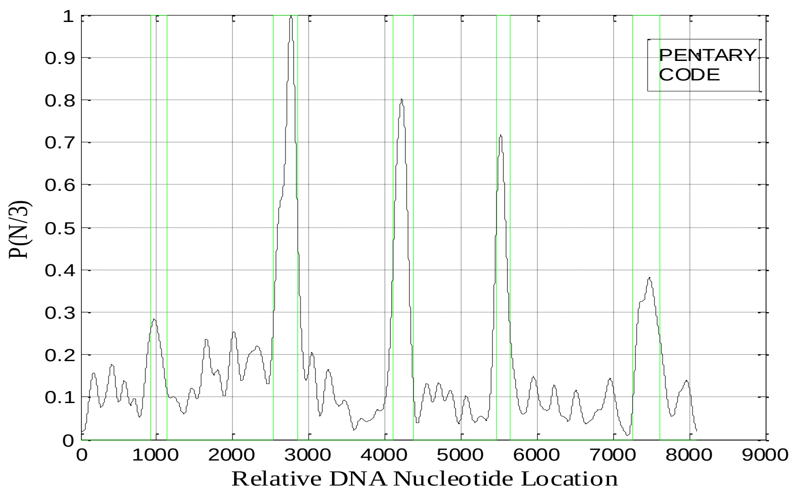


Figure 6.23 Pentary Code representation

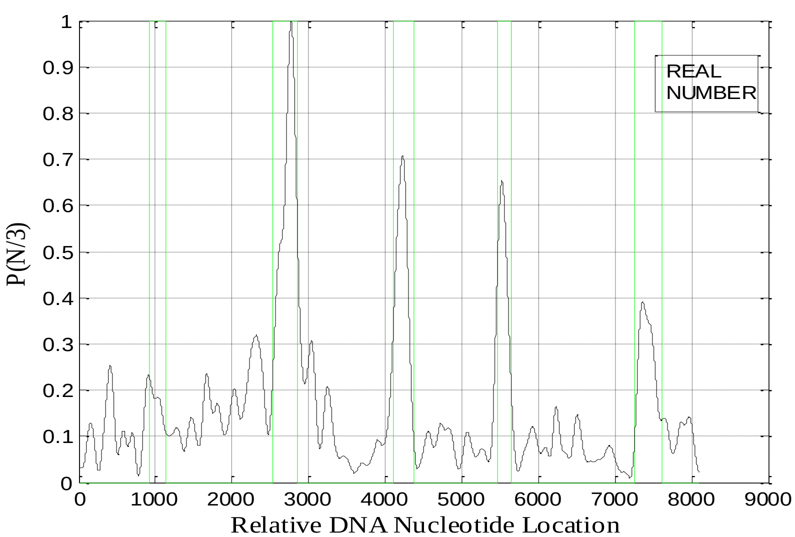


Figure 6.24 Real Number Representation

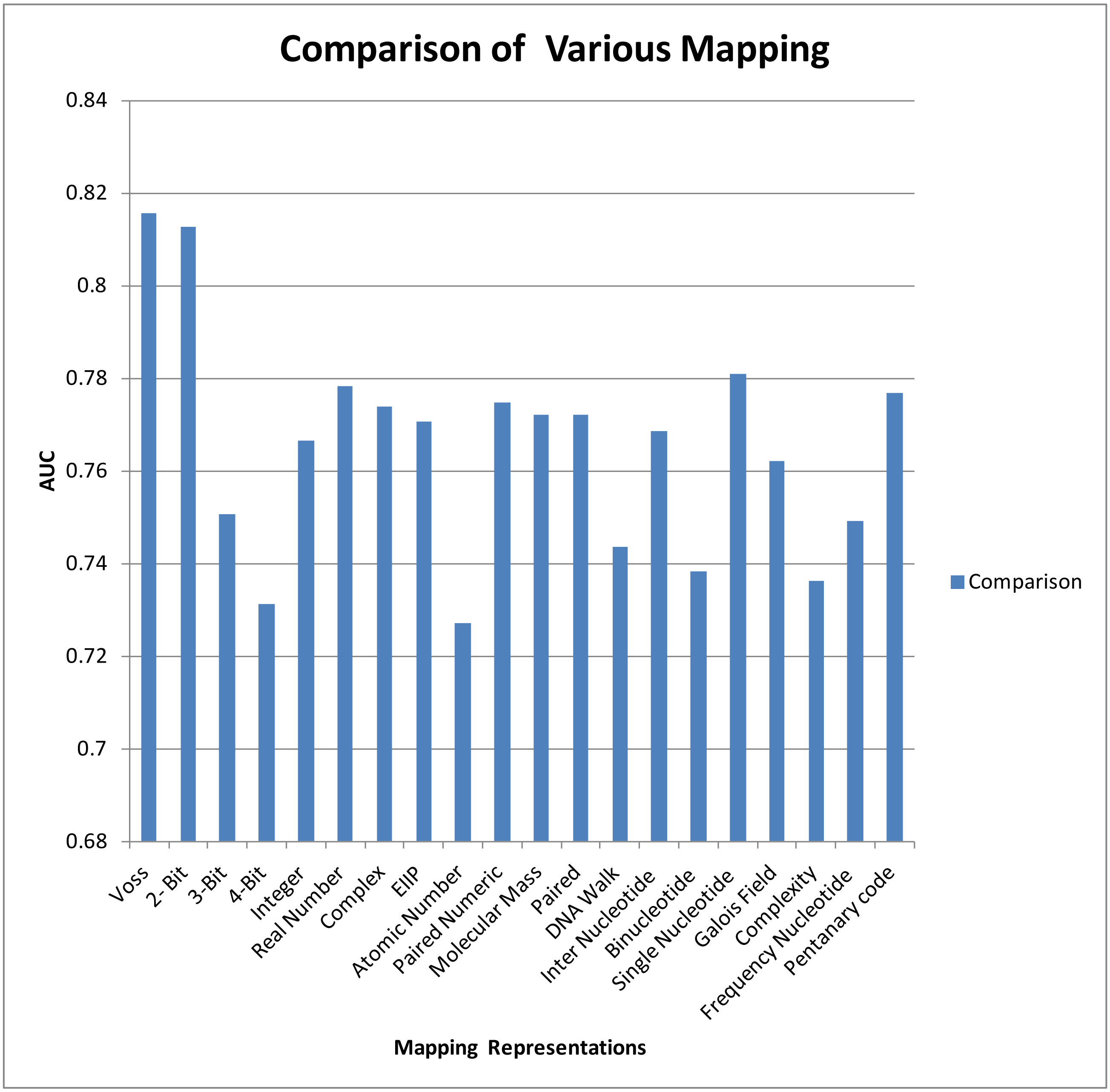


Figure 6.25 Comparison of all twenty mappings

6.2 Evaluation Measures

6.2.1 Sensitivity and Specificity

Sensitivity and Specificity are the evaluation parameter which is represented as represented as sn and sp respectively.

Sensitivity (Sn) =TP/ (TP+FN)

Specificity (Sp) =TN/ (TN+FN)

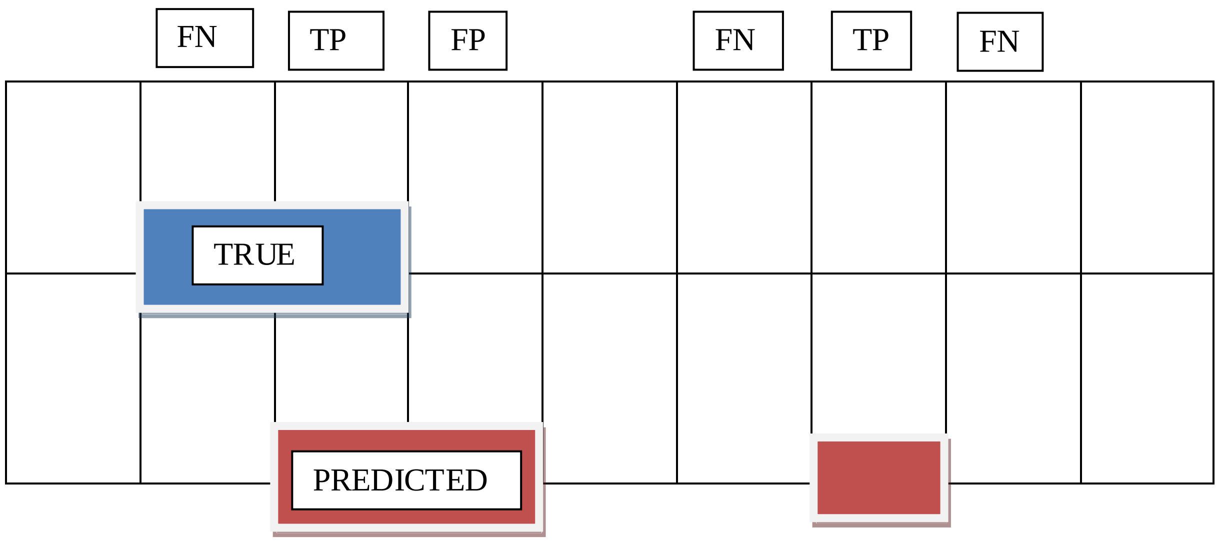


Figure 6.26 Block Diagram for True and Predicted Exons

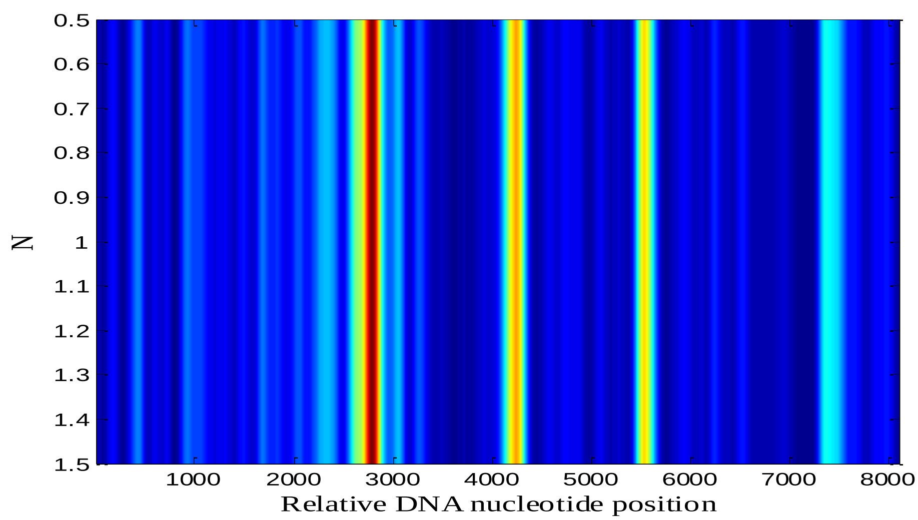


Figure 6.27 Power spectrum of exons for f56f114 sequence

Area Under the curve

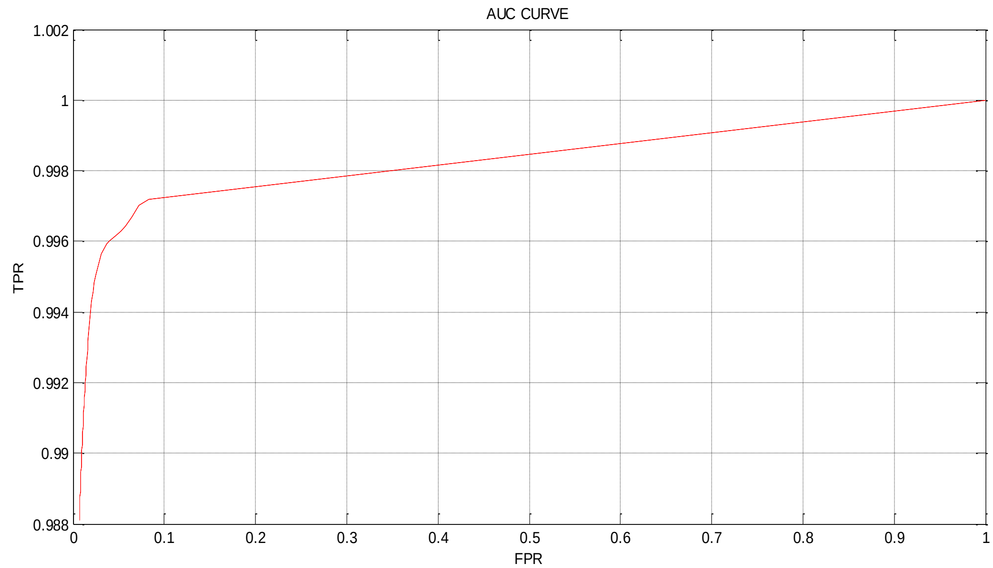
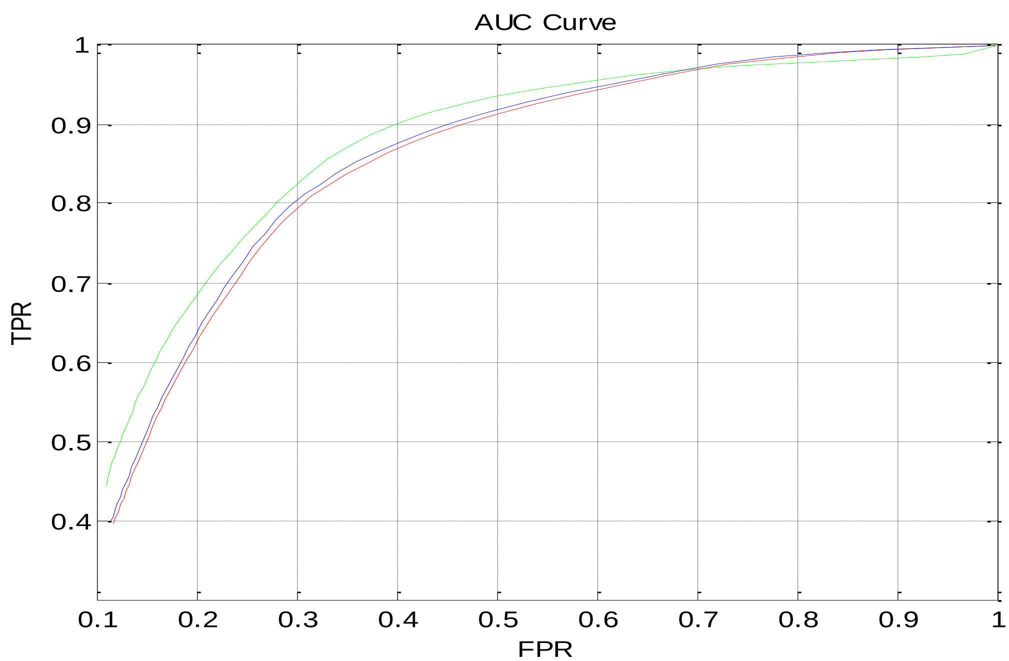


Figure 6.28 AUC for B0432 sequence



----- **Voss**

----- **2-Bit Binary**

----- **EIIP**

Figure 6.29 Comparison of AUC for F56f11.4 sequence

Figure 6.27 show the AUC of three different mapping which results that Voss mapping is gives the best result 0.8155 area under the curve then 2-Bit Binary mapping which give 0.8127 value and last with EIIP mapping which give 0.8100 value. So, Therefore Voss gives best result as larger areas indicate more accurate detection method.

TABLE 6.4

EVALUATION PARAMETERS FOR DIFFERENT SEQUENCE

Gene ID	Voss Representation		Complex Representation	
Evaluation Parameter	Sn	Sp	Sn	Sp
F56F11.4a	0.0265	0.8953	0.0347	0.9080
F56F11.1	0.3953	0.8993	0.3938	0.8730
B0432.12	0.5171	0.8323	0.3540	0.9395
F56F11.4b	0.3466	0.9935	0.3130	0.9945

6.3 Splice site prediction

The splice site prediction of a sequence can be calculated by using method position specific scoring matrix (PSSM). Score of acceptor region is calculated by applying window length to the data obtain by applying STFT to the DNA sequence.

TABLE 6.5

Spliced data for acceptor region

Sequence	Range	
'TTTGCAGGTAAT'	921	932
'TTTTCAGGTTCG'	2521	2532
'TTTCCAGGTTCC'	4107	4118
'TTTTCAGCAAAC'	5458	5469
'TTTTCAGAATGG'	7248	7259

TABLE 6.6

Score calculation

Score calculation for F56F114 Sequence.	AG
	0.0148

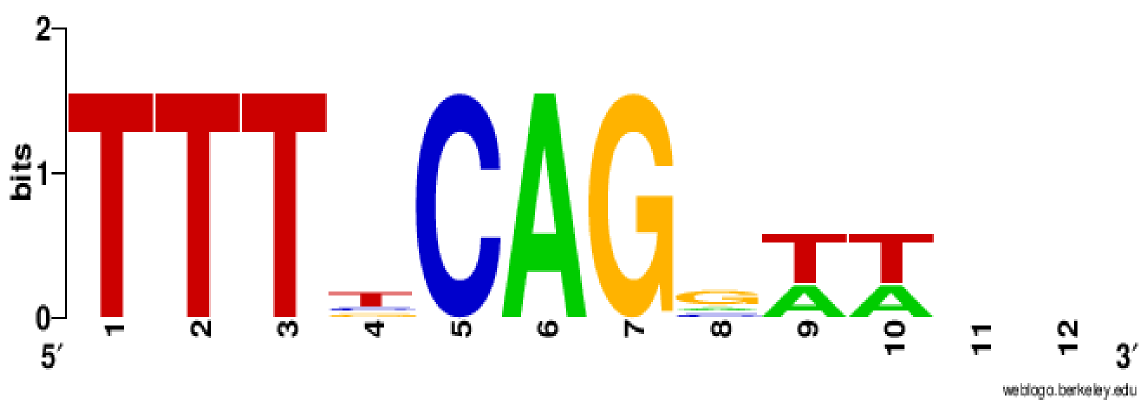


Figure 6.30 Logo diagram for acceptor region

Graph for splice site

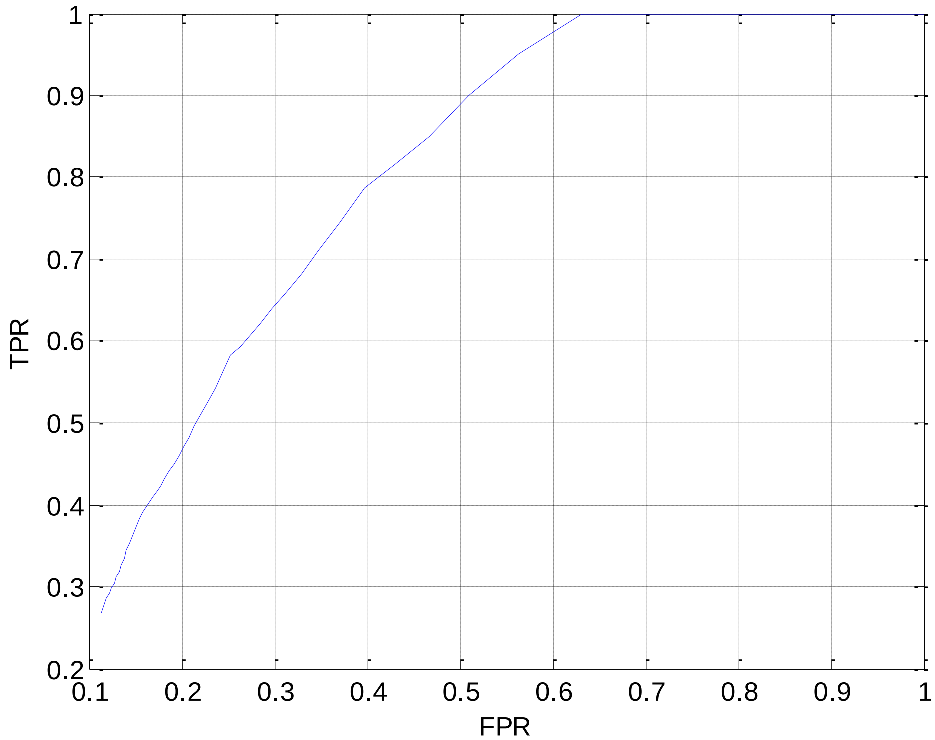


Figure 6.31 ROC curve for 'AG' acceptor site detection using PSSM

CHAPTER-7

CONCLUSION

Hence the exons are predicted and a comparison is done to get a exact location of exons. To discover the outcome of mapping on the exactness of the prediction of protein coding regions in a given sequence, different mapping techniques and methods are implemented to check the accuracy of prediction and to analysis that which mapping gives the best result. The results for various mapping are shown in table 6.2 in which area under the curve is calculated which shows that Voss, EIIP and 2 bit binary gives best performance with AUC as **0.8155**, **0.8127** and **0.8100** respectively**.** As larger the area under the curve more accurate will be the result for mapping. Further the splice site is done in which score of ‘AG’ nucleotide is calculated in a DNA sequence of 8100bp (F56F114). The score of ‘AG’ comes out to be **0.0148** which indicate that the exonic region which start when the score is 0.0148 The PSSM method gives the occurrence of each nucleotide further it calculate the score of the nucleotide.

PUBLICATIONS

1. Numerical Representation of DNA Sequences for Protein Coding Region Identification: A Study, RICSIT-2017(International conference on recent innovations in computer science and Information Technology).

   (Under review)

2. Performance analysis of window functions for exons prediction in DNA Sequence, Recommended for publication in international conference on computing communication and automation, ICCCA-2017, New Delhi 5-6 May 2017.

REFERENCES

[1] Hua W , Jiasong Wang , Jian Zhao , "*Discrete Ramanujan transform for distinguishing the protein coding regions from other regions"*,Molecular and Cellular Probes 28 (2014), pp228-236

[2] Arniker Swarna Bai, Hon Keung Kwan*, "Advanced Numerical Representation of DNA Sequences"*, International Conference on Bioscience, Biochemistry and Bioinformatics, pp.196-202, 2012.

[3] Anastassiou.D , *"Genomic signal processing,"* IEEE Signal Proc. Mag., vol. 18, pp.8-20, Jul. 2001.

[4] B. Demeler, G. W. Zhou, "Neural network optimization for E.coli promoter prediction," *Nucleic Acids Res., vol. 19*, pp. 1593-1599, Apr. 1991.

[5] S. Rampone, *"Splice-junction recognition on Gene sequences (DNA) by BRAIN learning algorithm*," in Proc. of IEEE World Congress on Computational Intelligence, Anchorage, USA, vol. 1, pp. 774-779, May 1998.

[6] S. Brunak, J. Engelbrecht, S.Knudsen*, "Prediction of human mRNA donor and acceptor sites from the dna sequence*," Journal of Molecular Biology, vol. 220, pp.49-65, Jul. 1991.

[7] P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan, J. L. Oliver*, "Study of statistical correlation in DNA sequences*," Gene,vol.300,pp.105-115,Oct.2002.

[8] P. Lio, and M. Vannucci, "*Finding pathogenicity islands and gene transfer events in genome data*," Bioinformatics, vol. 16, pp. 932-940, Oct. 2000.

[9] P. D. Cristea, "*Genetic signal representation and analysis,"* in Proc. of Society of Photo- Optical Instrumentation Engineers (SPIE) conference, San Jose, USA, vol.4623, pp. 77-84, Jan. 2002.

[10] N. Chakravarthy, A. Spanias, L. D. Lasemidis, and K. Tsakalis, *"Autoregressive modeling and feature analysis of DNA sequences*," EURASIP Journal of Genomic Signal Processing, vol. 1, pp. 13-28, Jan. 2004.

[11] J. Zhao, J. P. Li, X. W. Yang and Y. Y. Tang, "*DNA sequence classification based on wavelet packet analysis*," in Proc. of the Second International Conf. on Wavelet

Analysis and its Applications: Lecture Notes in Computer Science, vol. 2251, pp.424-429, Jan. 2001.

[12] P. D. Cristea, "*Conversion of nucleotides sequences into genomic signals*," J. Cell.Mol. Med., vol. 6, pp. 279-303, Apr.-Jun. 2002.

[13] S. K. Mitra ,J. A. Berger , M. Carli, and A. Neri, "*New Approaches to genome sequence analysis based on digital signal processing,"* in Proc. of IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS), Raleigh, USA, pp. 1-4, Oct. 2002.

[14] J. Epps, M. Akhtar and E. Ambikairajah, *"On DNA numerical representations for period-3 based exon prediction*," in Proc. of IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS), Tuusula, Finland, pp. 1-4 ,Jun. 2007.

[15] I. Grosse. P. Carpena , P. B.-Galvan, J. L. Oliver, R. R.-Roldan, H. E. Stanley, *"Finding borders between coding and non coding DNA regions by an enthropic segmentation method,"* Physical Review Letters, vol. 85, pp. 1342-1345, Aug. 2000.

[16] Daniel Nicorici and Jaakko Astola*, "Information divergence measures for detection of borders between coding and non coding DNA regions using recursive entropicse gmentation,"* IEEE Workshop on Statistical Signal Processing, St. Louis, USA,pp. 577-580, Sept.-Oct. 2003.

[17] Mohammed Abo-Zahhad, Sabah M. Ahmed, Shimaa A. Abd-Elrahman *"Genomic Analysis and Classification of Exon and Intron Mapping Techniques"* I.J Information technology and computer science, pp.22-36, 2012.

[18] Inbamalar T M and SivakumarR *"Study of DNA Sequence Analysis Using Sequences Using DNA Numerical DSP Techniques"* Journal of Automation and Control Engineering Vol. 1, *No. 4*, December 2013.

[19] M. Yan, C.-T. Zhang, Z.-S. Lin, *"A new Fourier transform approach for protein coding measure based on the format of the Z curve,"* Bioinformatics, vol. 14, pp.685-690, Sep. 1998.

[20] J. Wang and C.-T. Zhang, *"Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve*," Nuc. Acids Res., vol.28, pp. 2804-2814, Jul. 2000.

[21] G.Damiani, P. Arrigo, F. Giuliano, F.Scalia, A. Rapallo, *"Identification of a new motif on nucleic acid sequence data using Kohonen's self-organizing map,* "Computer Applications in the Biosciences (CABIOS), vol. 7, pp. 353-357, Jul. 1991.

[22] Todd Holden, E. Cheng, R. Sullivan, C. Sneider, G. Tremberger, Jr. A. Flamholz, D. H. Leiberman, R. Subramaniam and T. D. Cheung, *"ATCG nucleotide fluctuation of Deinococcus radiodurans radiation genes,"* in Proc. of Society of Photo-Optical Instrumentation Engineers (SPIE), San Deigo, USA, vol. 6694, pp. 669417-1 to 669417-10, Aug. 2007.

[23] C. Yin and S. Yau, *"Numerical representation of DNA sequences based on genetic code context and its applications in periodicity analysis of genomes,"* IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, Sun Valley, USA, pp. 223-227, Sep. 2008.

[24] T. Mahalakshmi and A. S. Nair, *"Visualization of genomic data using inter nucleotide distance signals,"* IEEE International Conference on Genomic Signal Processing, Bucharest, Romania, pp. 149-156, Jul. 2005.

[25] S. Sreenadhan and A. S. Nair, *"An improved digital filtering technique using nucleotide frequency indicators for locating exons,"* Journal of the Computer Society of India, vol. 36, pp. 54-60, Jan.–Mar. 2006.

[26] G. Dodin, L. Marcourt, P. Vandergheynst, P. Levoir, C. Cordier, *"Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences*," Journal of Theoretical Biology, vol. 206, pp. 323-326, Oct. 2000.

[27] T. Mahalakshmi and A. S. Nair, T., *"GSP using bi-nucleotide distance signals,"*13th International Conference on Advanced Computing and Communications, ADCOM, Coimbatore, India, Dec. 2005.

[28] T. Mahalakshmi and A. S. Nair, *"Are categorical periodograms and indicator sequences of genomes spectrally equivalent"* In Silico Biology, vol. 6, pp. 215-222, Aug. 2006.

[29] A. Asif and S. Datta, "*A fast DFT based gene prediction algorithm for identification of protein coding regions,"* in Proc. of IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing, Philadelphia, USA, vol. 3, pp. V-653-V-656, Mar. 2005.

[30] Y. Cao, J. Gao, Y. Qi, J. Hu, "*Building Innovative representations of DNA sequences to facilitate gene finding,"* IEEE Intelligent Systems, vol. 20, pp. 34-39,Nov.-Dec. 2005.

[31] G. L. Rosen, *"Signal processing for biologically-inspired gradient source localization and DNA sequence analysis,"* Ph.D. Dissertation, Georgia Institute of Technology, Atlanta, USA, Aug. 2006.

[32] S. K. Mitra, J. A. Berger, M. Carli, and A. Neri, *"Visualization and analysis of DNA sequences using DNA walks,"* Journal of the Franklin Institute, vol. 341, pp.37-53, Jan.-Mar. 2004.

[33] Devendra Kumar Shakya, Sanjay Verma, "Detection *of Protein Coding Regions using Goertzel Algorithm*", International journal of computer science, vol. 124, August 2015.

[34] Ambikairajah , Akhtar, M, and Epps, J. 2008b "*Digital signal processing techniques for gene finding in eukaryotes"* Lect. Notes Computer. Sci. 5099, pp 144–152, 2008.

[35] Swarna bai Arniker , Hon Keung Kwan, " *Advanced Numerical Representation of DNA Sequences"***,** 2012 International Conference on Bioscience, Biochemistry and Bioinformatics,pp1-5.

[36] Mohd A.Zahhad,S. M.Ahmed, S.A. Abd-Elrahman, " *A Novel Circular Mapping Technique for Spectral Classification of Exons and Introns in Human DNA Sequences",* I.J. Information Technology and Computer Science, 2014, 04, pp19-29.

[37] Mahmood Akhtar, J. Epps, E. Ambikairajah, "Signal Processing in Sequence Analys is Advances in Eukaryotic Gene Prediction", *IEEE Journal of selected topics in signal processing vol 2, no. 3, june 2008,*pp310-321.

[38] Mohammed Abo-Zahhad , Sabah M. Ahmed "*Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Techniques*" **I**.J. Information Technology and Computer Science, 2014, 04, pp22-29.

[39] Sajid Marhon∗ , Stefan C. Kremer1 "*Theoretical justification of computing the 3-base periodicity using nucleotide distribution variance*" BioSystems 101 (2010), pp185–186.

[40] R. Zhang and C. T. Zhang, Z curves. "*An Intuitive Tool, for Visualizing and Analyzing the DNA sequences [J]*"*J. BioMol. Struct. Dyn, 1994, 11,* pp767-782.

[41] Swarna bai Arniker , Hon Keung Kwan , **“***Advanced Numerical Representation of DNA Sequences*”, 2012 International Conference on Bioscience,pp1-5.

[42] C.-T. Zhang and J. Wang. “*Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve*” Nuc. Acids Res., 2000, 28(14):pp2804-2814.

[43] C. Yin, S. Yau.*” Numerical representation of DNA sequences based on genetic code context and its applications in periodicity analysis of genomes”* IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2008,pp 223-227.

[44] Dan Larhammar and C. A. C. Dreismann, *“Biological origins of long-rangecorrelations and compositional variations in DNA,”* Nucleic Acids Research, vol.21, pp5167-5170.

[45] Anastassiou, D. 2000 *“Frequency-domain analysis of biomolecular sequences”* Bioinformatics 16,pp1073–108.

[46] Cristea, P.D. 2002 *“Conversion of nucleotides sequences into genomic signals. J. Cell”* Mol. Med. 6,pp279–303.

[47] Zhang, Z.G., Zhang, V.W., Chan S.C., et al. 2008 *“Time-frequency analysis of click-evoked otoacoustic emissions by means of a minimum variance spectral estimation-based method”* Hearing Res. 243, pp18–27.

[48] G. L. Rosen, *“Signal processing for biologically-inspired gradient sourcelocalization and DNA sequence analysis,”* Ph.D. Dissertation, Georgia Institute of Technology, Atlanta, USA, Aug. 2006.

[49] Akhtar, M., Ambikairajah, E., and Epps, J. 2008b “*Digital signal processing techniques for gene finding in eukaryotes”*Lect. Notes Comput. Sci. 5099,pp 144–152.

[50] Ahtar, M., Ambikairajah, E., and Epps, J. 2008c “*Optimizing period-3 methods for eukaryotic gene prediction.”* Proc. IEEE Int. Conf. Acoust.Speech Signal Process (ICASSP*)* pp621–624.

[51] Yin, C., and Yau, S. 2007 *“Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence*” Bio Systems. J. Theor. Biol. 247, pp687–694.

[52] Marhon, S.,and Kremer, S.C. 2010 “*Theoretical justification of computing the 3-base periodicity using nucleotide distribution variance”Biosystem,104,*pp639-679

[53] https://www.wormbase.org/#012-34-5

[54] www.ncbi.nlm.nih.gov