DETECTION AND LOCALIZATION OF HIDDEN PATTERNS IN DNA SEQUENCES USING SIGNAL PROCESSING

Thesis submitted in fulfillment of the requirements for the Degree of

DOCTOR OF PHILOSOPHY

by

PARDEEP GARG

Department of Electronics and Communication Engineering

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
WAKHNAGHAT, SOLAN-173234, HIMACHAL PRADESH, INDIA

December, 2021

Table of Contents

DECLARATION BY THE SCHOLAR

I hereby declare that the work reported in the Ph.D. thesis entitled **“Detection and Localization of Hidden Patterns in DNA Sequences Using Signal Processing”** submitted at **Jaypee University of Information Technology, Waknaghat, India,** is an authentic record of my work carried out under the supervision of **Dr. Sunil Datt Sharma**. I have not submitted this work elsewhere for any other degree or diploma. I am fully responsible for the contents of my Ph.D. Thesis.

Pardeep Garg

Department of Electronics and Communication Engineering
Jaypee University of Information Technology, Wakhnaghat, Solan
(HP), India
Date: 31-12-2021

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 Ph.D. thesis entitled **“Detection and Localization of Hidden Patterns in DNA Sequences Using Signal Processing”**, submitted by **Pardeep Garg** at **Jaypee University of Information Technology, Waknaghat, India,** is a bonafide record of his original work carried out under my supervision. This work has not been submitted elsewhere for any other degree or diploma.

Dr. Sunil Datt Sharma

Assistant Professor (Senior Grade)
Department of Electronics and Communication Engineering
Jaypee University of Information Technology, Waknaghat, Solan
(HP), India
Date: 31-12-2021

ACKNOWLEDGEMENT

My first and foremost thanks to the **Almighty**, for his endless blessings showed upon me throughout this endeavor.

The successful completion of this research work could not have been possible without the help of these people to whom I am very thankful. First of all, I would like to express my gratitude to my supervisor **Dr. Sunil Datt Sharma**, Assistant Professor, Jaypee University of Information Technology, for his support, encouragement and guidance throughout the course of my Ph.D. His intelligence, personality and constructive criticism have always kept me in a positive path of success. I would also like to thank him for the undying support, patience and motivation during times of some rough patches. He is a person with great energy, wisdom and dynamic personality that I admire. His positive feedback, smart work and great spirit kept me encouraged to work with fresh energy and new ideas.

I gratefully acknowledge **Jaypee University of Information Technology** for offering me to perform this research successfully, for providing necessary facilities and support. I owe my gratitude to the **Vice-chancellor Prof. (Dr.) Rajendra Kumar Sharma** for his optimism and humble nature that has always been an inspiration for me. A special thanks to the **Dean (Academic & Research) Prof. (Dr.) Ashok Kumar Gupta** for imparting quality based education with ethics and values as its bedrock.

My heartfelt appreciation to **Dr. Rajiv Kumar, Head of Department** of Electronics and Communication Engineering, for his co-operation, support and constant encouragement. It is my pleasure to acknowledge the timely help and support of the faculty and lab staff of the department.

I am also very thankful to all my **DPMC members** to provide useful suggestions throughout the journey of this research work. My special thanks to **Prof. (Dr.) P. B. Barman** for his constant support, motivation, and valuable guidance throughout the tenure of this work.

Last but not the least; I want to thank my **parents** for their blessings, encouragement, moral support, personal attention and care. Most of all I owe my gratitude to my wife **Ms. Deepika** and my son **Mr. Atharva** for their extraordinary belief, love and providing me the space for completing this research work.

LIST OF ACRONYMS & ABBREVIATIONS

DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
NCBI	National Centre for Biotechnology Information
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
PCR	Protein-Coding Region
TBP	Three-Base Periodicity
TR	Tandem Repeats
STR	Short Tandem Repeats
bps	Base Pairs
mRNA	Messenger Ribonucleic Acid
UTR	Untranslated Region
TBP	3-Base Periodicity
CGI	CpG Island
MS	Microsatellite
Kbps	Kilo Base Pairs
Mbps	Mega Base Pairs
HMM	Hidden Markov Model
DSP	Digital Signal Processing
GSP	Genomics Signal Processing
AUC	Area Under Curve
ROC	Receiver Operating Characteristics
S-Golay	Savitzky-Golay
STFT	Short-Time Fourier Transform
FFT	Fast Fourier Transform
DWT	Discrete Wavelet Transform

MGWT	Modified Gabor Wavelet Transform
P-spectrum	Periodicity Spectrum
IPDFT	Integer Period Discrete Fourier Transform
CDS	Coding Sequence
DFT	Discrete Fourier Transform
STDFT	Short-Time Discrete Fourier Transform
PSWR	Paired and Weighted Spectral Rotation
AWSTFT	Adaptive Window Short-Time Fourier Transform
SONF	Statistically Optimal Null Filter
AR	Autoregressive
WSHHT	Wavelet Subspace Hilbert-Huang Transform
WRWW	Wide-Range Wavelet Window
AST-PCA	Adaptive S-Transform-Principle Component Analysis
o/e	observed/expected
IIR	Infinite Impulse Response
TLBO	Teaching Learning based Optimization
RLS	Recursive Least Square
TRF	Tandem Repeats Finder
STPT	Short-Time Periodicity Transform
MFPS	Modified Fourier Product Spectrum
EPSD	Exactly Periodic Subspace Decomposition
QPT	Quaternion Periodicity Transform
OMWSA	Optimized Moving Window Spectral Analysis
WBEMD	Wavelet-Based Empirical Mode Decomposition
S-T	S-Transform
SRF	Spectral Repeats Finder
PSE	Parametric Spectral Estimation
EMWD	Empirical Mode and Wavelet Decomposition
EMD	Empirical Mode Decomposition
CCA	Cross-Correlation Analysis
2-D	Two Dimensional

AST	Adaptive S-Transform
SVD	Singular Value Decomposition
MPSA	Modified P-Spectrum based Algorithm
ECG	Electrocardiograph
EIIP	Electron Ion Interaction Potential
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
Sn	Sensitivity
Sp	Specificity
AC	Accuracy
Prec	Precision
Rec	Recall
GGF	Gardiner-Garden Frommer
CpGPNP	CpG Island Prediction and Primer Design
ST-IPDFT	Short-Time - Integer Period Discrete Fourier Transform

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The completion of human genome sequencing project in April 2003 and subsequently next generation sequencing technology provided the direction for annotation of genome to come into existence. The huge amount of genomic raw DNA data is annotated in genome annotation by extracting useful information and the annotated data is added to the data base of genome. Genomic data available in the form of DNA sequences consists of various hidden patterns which are associated with the functioning of the organism. Protein-coding regions, introns, CpG Islands, tandem repeats, genic regions, inter genic regions, promoter regions, transcription start sites, and untranslated regions are few examples of sections of DNA which are important. Many computational approaches have been developed for the identification of these regions. However, development of accurate and efficient approaches for the detection and localization of the hidden patterns in DNA sequences has always been a challenging task. In this research work, efficient approaches have been proposed for the detection of hidden patterns such as protein-coding regions, CpG Islands, and tandem repeats in the DNA sequences. The signal processing based tools have been employed in all the proposed approaches of this research work. The platform used for the simulation of proposed algorithms in this work is MATLAB (R2013). The performance assessment has been carried out using standard evaluation metrics and the comparison has been done with recent state-of-art methods on the benchmark datasets. The proposed approaches have achieved significant improvement in detection over recent state-of-art methods.

CHAPTER 1

INTRODUCTION

One of the areas of signal processing known as genomics signal processing is gaining popularity and advancing very rapidly nowadays and it has been made possible due to the deoxyribonucleic acid (DNA) sequencing efforts made by various government and private sector organizations worldwide. A lot of research is being done in this field because of the availability of vast amount of genomics data in the form of DNA sequences made publically available by websites like National Centre for Biotechnology Information (NCBI) [1]. Although it seems to be complex to portray this field, yet it can be defined as the area of exploration which is the intersection of structural biology, molecular biology, and molecular evolution whose main aim to investigate the relationships amongst sequence, biological function, structure, and evolution using computational analysis of DNA sequences. The nucleotide bases known as Guanine 'G', Thymine 'T', Cytosine 'C', Adenine 'A' are the constituent elements of DNA sequences. It has been reported in literature that most of the part of DNA sequences comprise of repeated patterns of varying periodicity. Such hidden information regarding periodicities inside the DNA sequences needs to be explored to understand the biological relevance related to these which can be of great importance for the society. Coding and non-coding regions are the important constituents of deoxyribonucleic acid (DNA) sequences. Coding regions are commonly known as 'protein-coding regions (PCRs)' or 'exons' and the non-coding regions are called as 'introns'. It is believed that the nucleotides belonging to exonic regions contribute in protein formation. It is well known that the PCRs reveal three-base periodicity which is popularly called as TBP or period-3 property while non coding regions generally do not possess such property [2].

Another significant constituent of DNA sequences which is having great importance is CpG Islands. CpG Islands are those segments inside the DNA sequences where nucleotide 'C' is followed by nucleotide 'G' and the concentration of dinucleotides 'CG' is higher in these segments compared to region which is non CpG Island. The 'p' in CpG Islands represents the phosphodiester bond between G and C nucleotides [3]. Some of the important activities which highlight the relevance of CpG Islands are like: the detection of CpG Islands can facilitate in the identification of promoter regions and subsequently genic regions [4], inactivation of X chromosome, some human malignancies, silencing of gene, and also may be helpful for the early stage forecasting of cancers [5].

Tandem repeats are those patterns inside the DNA sequences where more than two adjacent copies of recurring pattern of a particular periodicity are present. The analysis of tandem repeats is carried out utilizing the important features like repeat pattern structure, pattern length, number of copies, location of these patterns inside the DNA sequences. On the basis of length of recurring pattern, the tandem repeats (TR) are categorized as: microsatellites, minisatellites, and satellites. Microsatellites which are also called as short tandem repeats (STR) range between 2-8 bps in their pattern size. Minisatellites' repeat pattern size ranges between 9-80 bps. The size of repeat pattern of satellites is greater than 100 base pairs (bps) [6]. The study of TRs is important because some of these TRs are accountable for a number of diseases. Moreover, TRs find applications in many additional areas such as DNA fingerprinting, population's study of a region, and DNA forensics analysis etc. [7]. Hence, the vast amount of genomic data already available needs to be analyzed properly for the benefit of society in respect of various applications corresponding to the outcome of exploration. The remaining of the chapter covers the following topics: DNA and genomics, sequencing of DNA, annotation of genome, possible approaches for genomic data processing, application of digital signal processing methods in genomics, and the organization of thesis.

1.1 DNA and Genomics

Cell is the basic fundamental, biological unit of all living organisms and all the genetic information of organisms is stored inside the nucleus of the cells. The two types of organisms, prokaryotes and eukaryotes differ on the basis that eukaryotic cells possess membrane bound nucleus whereas prokaryotic cells do not possess this property. Their composition is depicted in Figure 1.1. Eukaryotes are mostly multicellular whereas prokaryotes are unicellular. Eukaryotes accumulate their genetic information inside the nucleus in a substance called as chromatin. According to cell cycle chromatin may be found in either compressed state or uncompressed state and the compressed state of chromatin is known as chromosome. Eighty percentage composition of chromatin is made up of proteins and the remaining twenty percentage is made up of nucleic acids. Nucleic acids are essential to all living organisms and are composed of

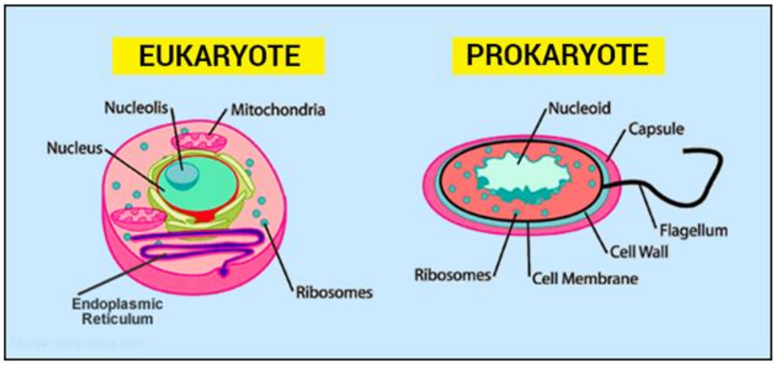


Figure 1.1: Difference between eukaryote and prokaryote cells in terms of presence/absence of nucleus [source: fitz6.wordpress.com]

nucleotides. The three basic components which comprise of nucleotides are: a phosphate group, a 5-carbon sugar, and a nitrogenous base. Depending upon the sugar type, nucleic acids is classified as DNA if the sugar type is deoxyribose and RNA (ribonucleic acid) corresponding to ribose as sugar type. All the necessary genetic information related to the functionality and development of all living organisms is contained in the DNA. The encoding of hereditary information is performed by DNA and correspondingly the one species can be distinguished from the other species. The arrangement of nucleotides inside the DNA is linear and they form a DNA strand. In general, two single strands of DNA molecule are twisted around each other and they remain in helical shape and form a double helix structure [8-11] as represented in Figure 1.2.

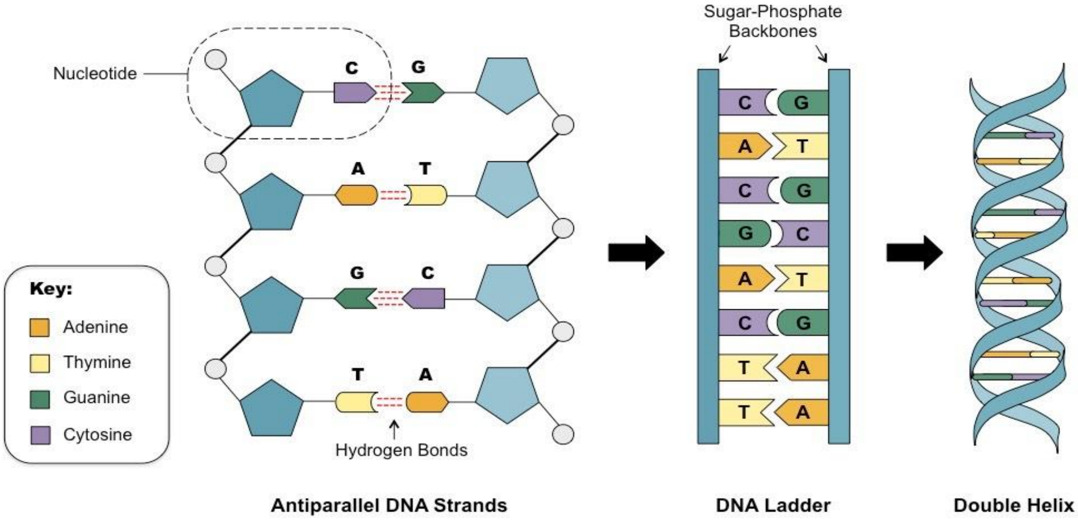


Figure 1.2: Double helix structure of DNA molecule [source: https://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/26-structure-of-dna-and-rna/dna-structure.html]

The two strands of DNA molecule are arranged in anti-parallel fashion because the two strands point in opposite directions. The backbone of each strand of DNA is composed of phosphate groups and sugars. The nucleotide nitrogenous base ‘A’ of one DNA strand is always found in pair with base ‘T’ of opposite DNA strand and ‘C’ pairs with ‘G’ always. Purine and pyrimidine are the two types of nitrogen based found in DNA molecule. ‘C’ and ‘G’ comes under pyrimidine bases category while ‘A’ and ‘T’ falls under Purine bases category. The chemical bonding between the nucleotides of one DNA strand with the nucleotides of other strand of DNA is a hydrogen bond. Because a single hydrogen bond is weak, hence these bonds altogether form a stable and double helix structure which looks like a rope. The discovery of double helix structure of DNA molecule proved to be the biggest attainment in the field of molecular biology because after its discovery it became transparent that genes are functionally definite parts of DNA molecules. And through this process cells translate the information contained in DNA to particular amino acids and which are further utilized to produce proteins.

The complete set of DNA sequence of a species is known as genome and the study of genes in a species is known as genomics. In genomics, the buried features inside the genome of a species are extracted; analyzed and useful information is obtained. In a single cell of every human body, there remain around 3 billion of DNA base pairs (bps). The cells are responsible for the formation of particular proteins using enzymes and messenger molecules. The enzymes are responsible for copying the information regarding genes from DNA to messenger RNA (mRNA) molecule. The movement of mRNA from nucleus to cytoplasm of cell is read by the ribosomes. The formation of particular protein is then governed with the help of ribosomes by providing link between mRNA and the order of amino acid. The development of various body structures like tissues, organs etc., the carriage of signal between cells, and the controlling of chemical reaction occur with the help of proteins. But if there occurs some mutation in the DNA sequence then the normal protein development changes to abnormal protein formation. Such abnormal proteins may disturb the normal functioning of the human body which may lead to development of disease such as cancer [12]. Hence the field of genomics for the analysis of such cases finds its importance.

1.2 Sequencing of DNA

The process of determination of nucleic acid sequence which is the order of four nucleotides in DNA is considered as sequencing of DNA. The great acceleration in biological & medical research and the development of various computational tools, signal processing algorithms in the field of genomics has become possible because of rapid methods of sequencing of DNA. And this all has occurred because of identification of DNA and the double helix structure of DNA; in 1869 Friedrich Miescher [13-14] provided a landmark direction in genetic research by first identifying which he named as 'nuclein' inside the nuclei of human species. The term 'nuclein' is nowadays known as nucleic acid and subsequently DNA. Another breakthrough was provided by Phoebus Levene in 1919 [15] who was the first to find out the order of three major constituents of a nucleotide as phosphate-sugar-base. Also, he was the first to find out the carbohydrate component of RNA and the carbohydrate component of DNA. Again, he was the first to correctly recognize the mode in which DNA and RNA molecules are put together. The foundation laid by Levene was strengthened by Chargaff who provided two major rules; firstly he noticed that there is a variation in nucleotide composition of DNA among different species. Secondly he concluded that irrespective of organism or type of tissue, almost all DNA maintains certain properties which is total amount of purines and pyrimidines are almost equal mostly[16]. Another big achievement in the field of generic research was provided by Watson and Crick in 1953 who derived the three-dimensional, double helical model for DNA's structure [8]. The DNA is considered as the genetic material and is responsible for the functioning, structural development of organisms. Hence, the technologies used for the sequencing of DNA have been developed to assist the biologists and medical society in broad category of applications such as medicine, forensics, and various areas of biology. The sequencing of DNA can be utilized to find out the sequence of individual genes, clusters of genes, complete chromosomes, or full genomes of all species [17-18]. The technologies used for sequencing of DNA are required to be precise, fast in processing, inexpensive, and easy-to-use. The earliest form of nucleotide sequencing was RNA sequencing and the major achievement was proposed by Walter Fiers in 1972 & 1976 by providing the sequence of first complete gene and the full genome of Bacteriophage MS2 [19-20]. The first method of DNA sequencing to determine DNA sequences was established in 1970 by Ray Wu and in this method a location-specific primer extension strategy was employed [21-

22]. Frederick Sanger in 1977 then utilized this primer-extension philosophy and developed more rapid DNA sequencing technologies and named as “DNA sequencing with chain-terminating inhibitors” [23]. This technology is popularly called as Sanger sequencing technology also and has been used extensively in various fields such as comparative and functional genomics, evolutionary genetics etc. Therefore, this method remains a popular method in various laboratories across the world. Another sequencing technology based on chemical degradation was established by Walter Gilbert and Allan Maxam [24-25]. But these sequencing technologies were very laborious. Hence, improvement in technologies were being done to make the task of sequencing automatic and its output was observed in 1987 with the development of first automatic sequencing machine known as AB370. Applied Biosystems had introduced this machine and this technique of sequencing was fast and accurate which employed capillary electrophoresis. This machine was capable of detecting 500 bases per day with read length reaching 600 bases and 96 bases at one time. The latest machine model AB3730xl capability was 2.88 M bases in a day and read length was reached around 900 bases [26]. The main tools which played very important role in the completion of human genome sequencing project in 2001 were the automatic sequencing instruments and associated software which used the capillary machines of sequencing and Sanger technology of sequencing [27]. Motivated from human genome sequencing project, the Next Generation Sequencing (NGS) technology was developed which provided high throughput by doing parallel analysis, was faster, accurate and the cost was also reduced compared to Sanger sequencing technology. The cost of sequencing has fallen so dramatically nowadays that a single laboratory can afford to sequence large genomes even. The genomics data which has been made available by various repositories has a lot of significant hidden information inside it and it has become a big data problem. This huge genomics data need to be analyzed for the benefit of medical research and society.

1.3 Annotation of Genome

Once the sequencing of genome is completed, the huge amount of genomic raw DNA data generated from that process has to be analyzed to extract important information out of it. This process is termed as genome annotation and the annotated data is added to the data base of genome. The two classifications of genome annotation are: structural and functional annotation. Structural annotation deals with identification of various elements such as introns, exons, etc.

whereas functional annotation deals with attaching biological information to the genomic elements [28]. Our emphasis in this work is on structural genome annotation and the various regions of genome annotation are represented in Figure 1.3.

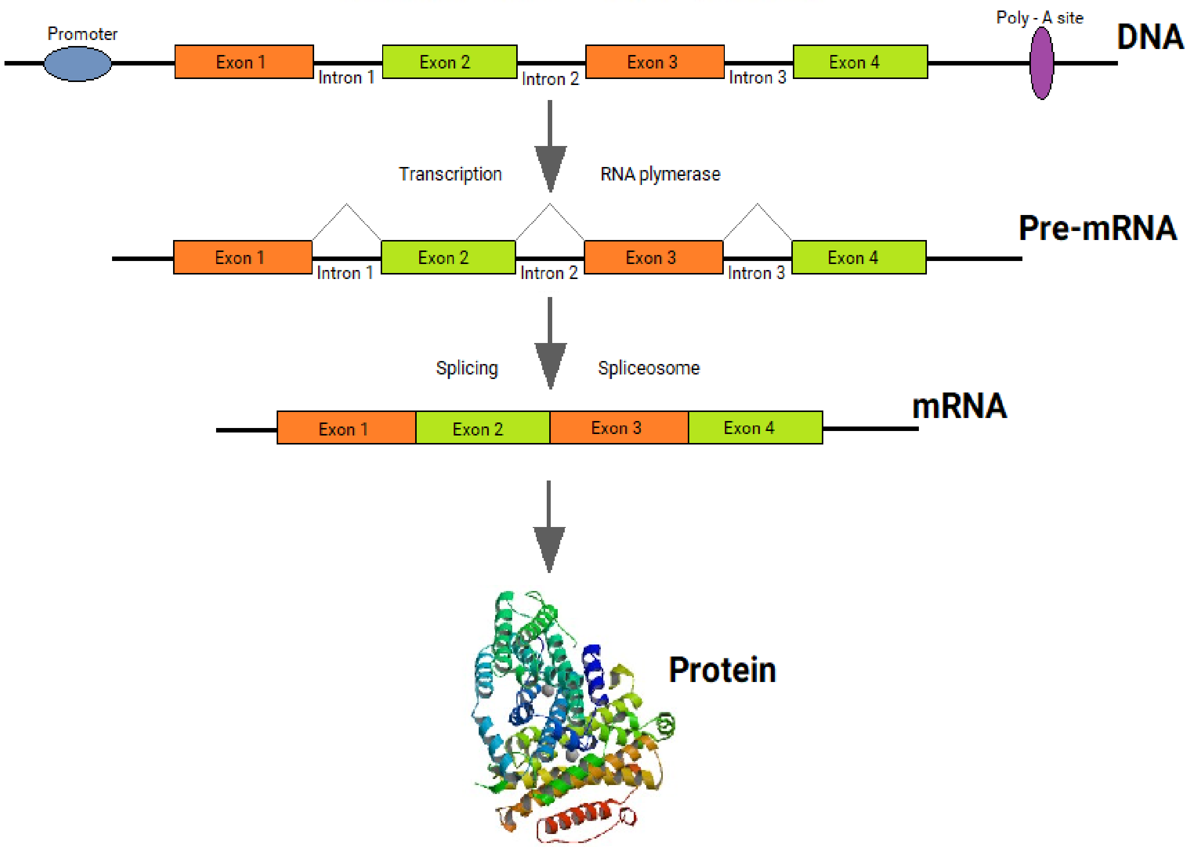


Figure 1.3: Annotation of genome

[source:https://en.wikipedia.org/wiki/Split_gene_theory#/media/File:Introductory_figure_for_transcript_and_splicingV2.png]

The various important sections of a genome are described as following:

i) Promoter regions

A promoter region usually located near the beginning of a gene is defined as a non-coding sequence of DNA in which transcription of a gene is initiated. The promoter region is responsible for controlling when and where the gene of interest is to be expressed in an organism. It is needed to turn a gene on or off. The typical length of promoters is around 100-1000 base pairs [29].

ii) Untranslated regions

Untranslated regions (UTRs) are found on the two sides of a coding sequence. If it is located on the 5’ side of on a strand of mRNA, it is termed as 5’ UTR or leader sequence and if is placed on the 3’ side, then it is known as 3’ UTR or trailer sequence. UTRs are not associated with the formation of proteins however UTRs and introns find their importance for the controlling of complex gene expressions [30].

iii) Exons

The nucleotide sequences in DNA which are associated with the formation of proteins are termed as exons or protein coding region of gene. Proteins are known to be an essential component of each cell in the body. Next to water, proteins are considered to be the most abundant type of molecules present in the body. Proteins are made up of hundreds of amino acids in the form of a long chain with the linkage of peptide bonds. 20 different amino acids are present in the body and in protein coding regions each of these amino acids are encoded as a sequence of three successive nucleotides. Approximately 3 billion base pairs are present in the human genome and out of this only 2% constitute exons whereas remaining 98% are most likely intergenic region or introns [31] as the length of exons is usually shorter than introns. The fundamental difference in prokaryotes and eukaryotes is in the organization of genes inside them. A prokaryotic gene appears as a continuous stretch of DNA which does not require any processing and gets transcribed into RNA to serve as messenger RNA (mRNA). Whereas, a eukaryotic gene has exons spread across its length in many stretches which are interrupted with introns in between. These introns are considered to have no significance in the protein synthesis and hence are also known as non coding regions. Alternative splicing process removes the introns and joins the exons to create an interrupted gene known as mRNA and finally another cellular mechanism termed as Translation converts mRNA into different proteins [32-33] as depicted in Figure 1.4. It has been revealed that there exists a well known short-range correlation in the arrangement of nucleotides in exonic regions called as period-3 property or 3-base periodicity (TBP) [34]. The researchers working in the field of genomic signal processing who focus on developing digital signal processing based methods utilize this TBP property to detect the exonic regions in DNA sequences. This TBP has a close relation with the deranged allocation of the nucleotides in the three coding positions which tell that the nucleotides in exonic regions exhibit non-uniform distribution whereas in the intronic regions nucleotides possess a balanced distribution. The non-uniform distribution of nucleotides in exonic regions exists because in these regions the usage of nucleotides is extremely biased towards special amino acids composition [35-37].

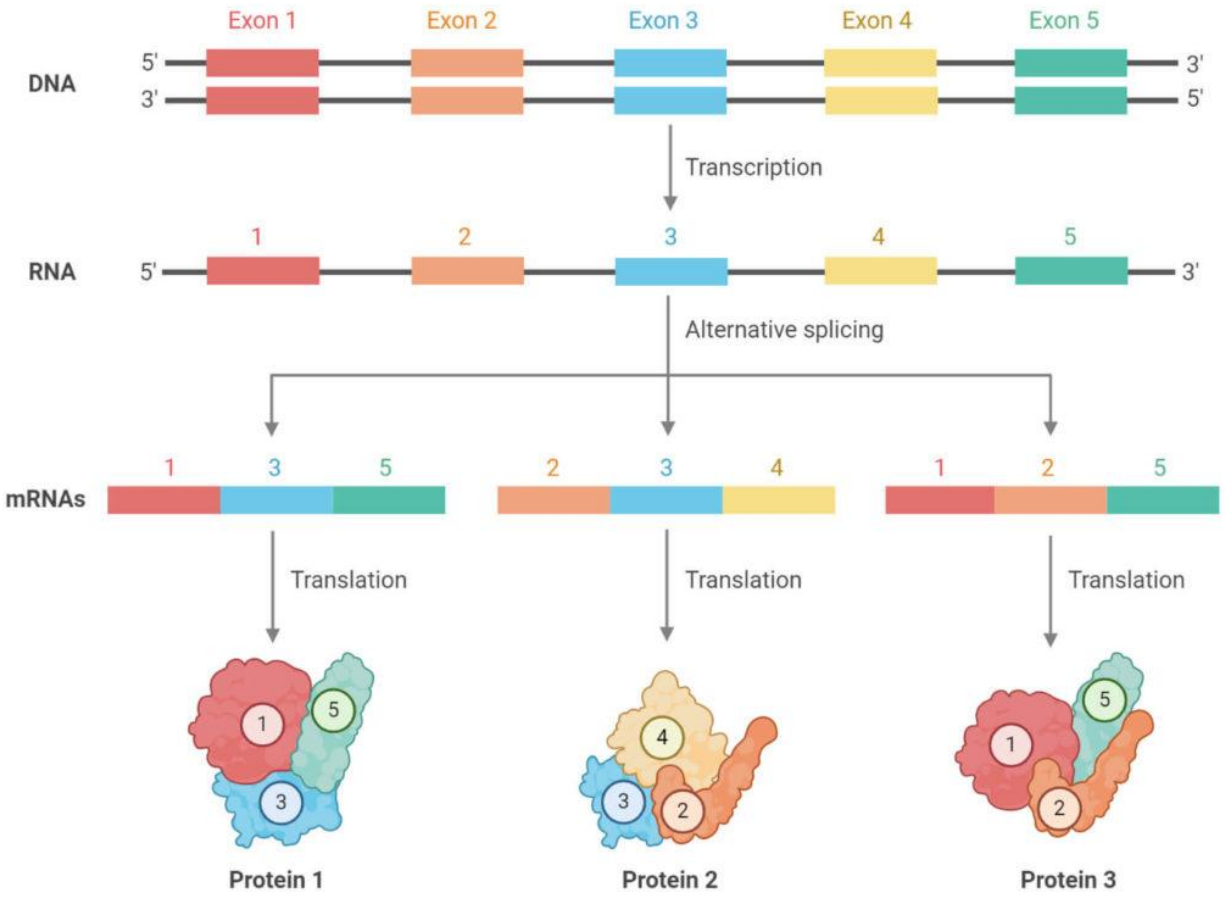


Figure 1.4: Process of different proteins synthesis through alternative splicing
[source: https://msnoller.weebly.com/transcription-and-gene-expression-72.html]

iv) Introns

Introns were discovered in 1977 [38-40]. The sections of gene which have no association in the formation of proteins are called introns or non-coding regions. Unlike exons, the feature of periodicity because of non-uniform codon bias is not possessed by introns while various other periodicities due to some recurring patterns may be possessed by introns [33]. The structure of introns inside the genes is random in which these are placed separated by exons in between. The occurrence of introns across the spectrum of a species usually varies in terms of their density and the length of intronic region. An example is the average number of introns per gene in human genome is 8.4 whereas there is no intronic gene is found in mitochondrial genome of vertebrates. And also, prokaryotic genes contain no introns but intronic genes are present in eukaryotic genomes. The introns are removed during alternative splicing process and finally functional mRNA is produced. As exons and introns remains in close proximity in the gene structure therefore, during the removal of introns a very accurate identification of boundaries connecting exons and introns is highly essential because outcomes can be misleading even if a single DNA character corresponding to exons and introns sequences is wrongly detected. The importance of introns is that these provide various significant short sequences for splicing process to be efficient like as donor sites and acceptor sites at start or end of introns respectively. The arrangement of introns along with other necessary details [41] is depicted in Figure 1.5.

v) CpG Islands

CpG Island also written as CGI is considered to be one of the important segments of DNA sequences. CGI are the regions inside the DNA sequences in which nucleotide ‘C’ is followed by nucleotide ‘G’ and which are rich in CG dinucleotides. p stands for phosphodiester bond in CGI and it is different from hydrogen bond found between C and G nucleotides within the two strands inside the double helix structure of DNA molecule. The length of the CGIs inside the DNA sequences varies from 200 bps to maximum up to 5000 bps. The motivation for the researchers working in the field of genomic signal processing to develop algorithms for the identification of CGIs in DNA sequences is the association of CGIs with many epigenetic events. CGIs are associated with promoter regions and hence these find application in the identification of the promoter regions and consequently to predict the genes in DNA sequences [4] as depicted in Figure 1.5. Also, gene silencing, cancers and many other epigenetic issues [42] are caused by the process of methylation of CGIs which happens by the addition of methyl group (CH_3) to the 5’-position of the carbon.

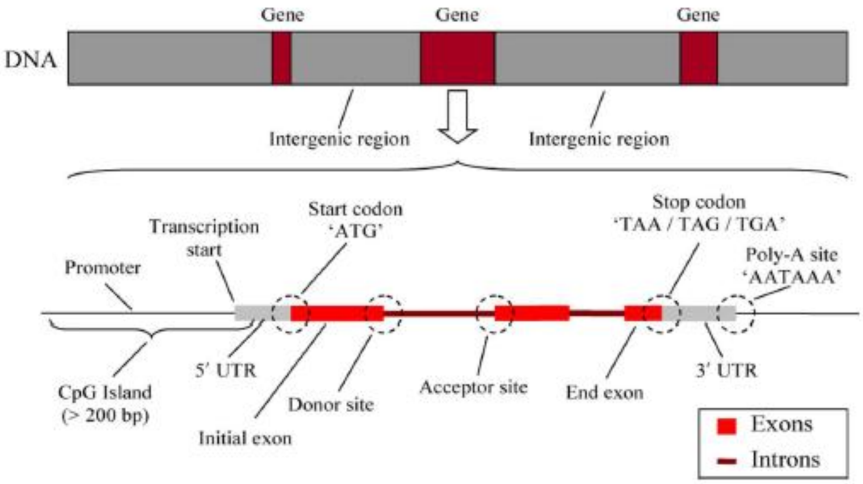


Figure 1.5: Introns and other important regions associated with them

1.4 Tandem Repeats

Most of the DNA sequences consist of recurring patterns. These sequences have the various nucleotide repeat patterns of their respective periodicity. These periodicities (repetitive patterns) are accountable for their particular functionalities in the body of the living organisms [43]. Tandem and Interspersed are the two broad categories of repeats found in DNA sequences. Contiguous repeat patterns are present in the tandem repeats whereas the interspersed repeats consist of noncontiguous repeated patterns [44-47]. An example of tandem and interspersed repeats is presented in Table 1.1.

Table 1.1: Tandem and Interspersed repeats

Tandem repeats	CGAT	CGAT	CGAT	CGAT	CGAT
Interspersed repeats	CGAT		CGAT		CGAT

Based on the size of repeat pattern, the tandem repeats can be further categorized as, satellites, minisatellites, and microsatellites (MS). The pattern size of satellites is greater than 100 bps and their length varies from 100 Kbps to 1 Mbps. The range of pattern size of minisatellites is from 9-80 bps and their length varies between 1-20 Kbps. Microsatellites which are more commonly called as short tandem repeats (STR) pattern size range between 2-8 bps and have a length less than 150 bps [47]. On the basis of mutations, tandem repeats are again categorized as perfect tandem repeats and imperfect tandem repeats. In perfect tandem repeats, the exact number of copies of repeat pattern is found. Whereas in imperfect also called as approximate tandem repeats, the repeat patterns are not present in the exact copies of patterns [48]. An example of perfect and imperfect tandem repeats is depicted in Table 1.2.

Table 1.2: Perfect and Imperfect tandem repeats

Perfect tandem Repeats	TGCA	TGCA	TGCA	TGCA	TGCA
Imperfect tandem Repeats	TGAA	TTCA	TGCC	TGCA	GGCA

The study of these repeats is helpful in various studies like DNA forensics analysis, DNA fingerprinting, and study of population of an area etc. Also, MSs amongst the three listed tandem repeats are more important because of their association with various diseases like Huntington's disease, Fragile-X syndrome, Spinocerebellar ataxia type 31, Frederick's ataxia, and 40 other neurodegenerative, neuromuscular, and neurological diseases [49-52].

1.5 Possible Approaches for Genomic Data Processing

Once the sequencing of genome of a species gets completed, first and the most important task after that to understand the molecular behavior of genome is gene finding. Gene finding was involving extremely pains taking experiment on living cells and species in the early days. The functional genomics deals with performing lab experiments and statistical analysis can be applied thereafter to find out the order of genes on a specific chromosome with the help of their rate of homologous recombination. The information gained from various experiments is combined to generate a particular map for the identification of rough position of the known genes related with each other. Nowadays, the accessibility of genome sequences of various species and availability of extensive computational tools for genomics data, the gene finding has turned up as a computational procedure [53]. Various computational gene finding tools have been introduced over a period of years like as: FGENES [54], HMM [55], HMMGene [56], GENSCAN [57], MZEF [58], Morgan [59], Genemark [60], Genie [61], Geneid [62], and AUGUSTUS [63]. Abinitio, extrinsic, and comparative are the three types of techniques employed for gene finding purpose [64]. The functioning of abinitio approaches is based upon searching of protein coding regions in DNA sequences utilizing certain properties of these regions. Some of such properties are such as statistical features, some contents, and biological signals related to the protein coding regions. The most popularly used methods based on abinitio technique are Geneid and GENSCAN. The extrinsic techniques' approach is to use the reverse translation of genetic code for the derivation of family of probable coding DNA sequences. These probable coding DNA sequences are then utilized to find out a target for matches which are partial or complete, and exact or random. The most widely utilized tool for this purpose is basic local alignment search tool. The principle of working of comparative gene finding approaches is to compare the following features in genome of associated species: length and number of coding regions, sequence similarity, position of gene, the amount of non-coding DNA in every genome, and additional vastly conserved regions.

1.6 Application of Digital Signal Processing Methods in Genomics

Although numerous gene finding algorithms exist which are data dependent but accuracy in terms of gene prediction is considered as their limitation. Their accuracy can be increased by one possible way of employing hybrid approach in which the three different types (extrinsic, abinitio, and comparative) of gene finding techniques can be combined in one single program such as AUGUSTUS+ [63]. Another possible way of enhancing accuracy is to combine different gene finding programs [65-66]. But the dependency on data would increase in both of these approaches [53]. A solution to this problem has been provided by signal processing methods in which the DNA characters are converted to numerical values by applying numerical mapping techniques and these signal processing methods have proved to be very useful in the analysis of genomic data [67-69]. The various signal processing operations like filtering of numerical data, application of time-frequency/spectral analysis tool, suitable thresholding are then applied to extract hidden features inside the genomic data and this area of signal processing is known as genomic signal processing (GSP) [70]. A general flow graph consisting of these basic blocks of GSP is depicted in Figure 1.6.

The first step in genomic signal processing is to obtain the DNA sequence from the standard database. Now to be able to apply the DSP methods on genomics data, it is essential to convert the DNA characters to numerical sequence using numerical mapping method. The numerical sequence data is passed through a filter as a pre-processing step for suppression of noise. With the help of signal processing tools such as time-frequency analysis and spectral analysis methods, the fundamental periodicities and their temporal location present in the DNA sequence are then detected [37], [71]. A suitable threshold is applied to capture the biological features related to characteristic periodicity for a particular hidden pattern present in the DNA sequence and then the performance evaluation is carried out using standard evaluation metrics. All the algorithms developed for the research work which are discussed in the subsequent chapters of this thesis are based upon the general flow graph of GSP shown in Figure 1.6.

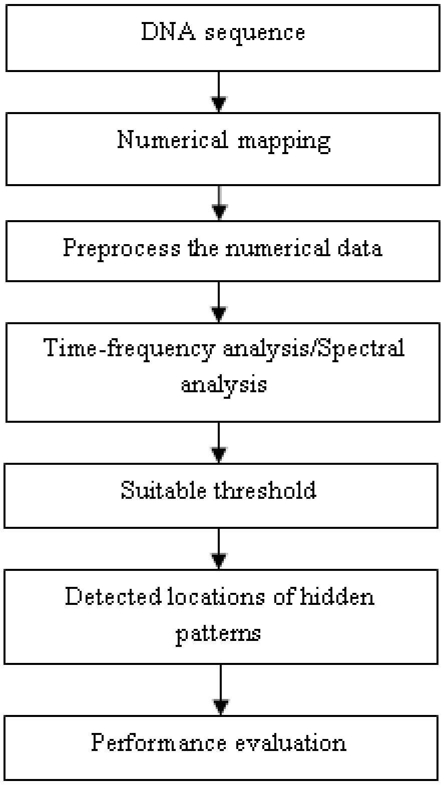

DNA sequence
Numerical mapping
Preprocess the numerical data
Time-frequency analysis/Spectral analysis
Suitable threshold
Detected locations of hidden patterns
Performance evaluation

Figure 1.6: Flow graph of GSP

1.7 Organization of Thesis

The thesis consists of total seven chapters. A brief description of each chapter is discussed as following:

Chapter 1: Introduction

The introduction of the work which has been carried out in this thesis has been discussed in this chapter. The basics of molecular biology are covered in this chapter. This area has application problems for which signal processing based algorithms have been proposed as solution.

Chapter 2: Literature Review

To formulate the problems existing in the molecular biology by developing clear understanding of this area, the existing literature was studied rigorously. The existing research gaps were identified which are described in this chapter. The computational algorithms are developed to address these research gaps and these are discussed in subsequent chapters.

Chapter 3: Identification of Protein Coding Regions in DNA Sequences using Singular Value Decomposition based Modified P-Spectrum Algorithm Employing Optimized Window and S-Golay Filter

In this chapter singular value decomposition based modified P-spectrum algorithm has been proposed for the identification of protein coding regions in the DNA sequences of eukaryotes. The area under the Receiver operating characteristics (ROC) curve has been chosen as the optimization parameter to optimize the window length. Savitzky- Golay (S-Golay) filter has been used to suppress the noise and to improve the signal-to-noise ratio by preserving the important features of signal. The performance of proposed method has been assessed on standard benchmark datasets and also has been compared with current state of art algorithms.

Chapter 4: Identification of CpG Islands in DNA Sequences using Short-Time Fourier Transform

In this chapter an algorithm based on short-time Fourier transform (STFT) has been proposed for the identification of CpG Islands (CGIs) in the DNA sequences. The periodicity features present in the CpG Islands have been detected with the help of STFT by conducting experiment on benchmark DNA sequence. Also, the performance of various existing numerical mapping methods has been assessed and then a solution based on combination of 24 mappings of integer mapping scheme has been proposed. A database consisting of 100 DNA sequences comprising of human, fish, and mouse species has been made and the performance of proposed method has been tested and compared with other state of art methods for the data set.

Chapter 5: Sensitivity Enhancement and overall improvement for the detection of CpG Islands in DNA sequences of Human Species using Modified P-Spectrum based Algorithm and Wavelet Transform based Proposed Algorithm Respectively

This chapter consists of two parts. In the first part, a modified P-spectrum based algorithm has been proposed for the sensitivity enhancement of the detection of CpG Islands in the 100 DNA sequences of human species. In the second part of the chapter, a Wavelet transform based algorithm has been proposed for the overall improvement of standard evaluation metrics on the database of 100 DNA sequences of human species.

Chapter 6: Detection of Tandem Repeats in DNA Sequences using Integer-Period Discrete Fourier Transform, and Modified Gabor Wavelet Transform based Proposed Algorithms

The emphasis in this chapter is on development of signal processing based algorithms for the tandem repeats detection in the DNA sequences. This chapter contains two parts. In the first part, an integer-period discrete Fourier transform based algorithm has been presented to detect the tandem repeats in DNA sequences. Modified Gabor Wavelet transform based algorithm has been described in the second part of chapter.

Chapter 7: Conclusion and Future Work

The conclusion of the thesis work has been presented in this chapter. The future directions in which this thesis work can be extended along with some open research problems have also been provided in this chapter.

CHAPTER 2

LITERATURE REVIEW

Genomics data comes under the category of big data because it contains a vast and huge amount of data, it becomes very confusing to apply genomic signal processing tools without having knowledge of some hidden features present in the genomics data. But the presence of various periodicities helps and enables the researchers working in this field to obtain a better understanding of the characteristic features present in the DNA sequences and develop the computational tools for their analysis subsequently. And this all is possible because it is proved in literature that most of the genomic data consists of repetitive patterns. The analysis of genomics data to extract the information about their functionality can be done by detecting and locating the periodicities in the DNA sequences. Trifonov *et al.* [72] have revealed the hidden periodicities 400, 200, 10.5, and 3 bps present in the genomic sequences. The period-3 property also known as three-base periodicity (TBP) found in the DNA sequences as a result of codon bias or non uniform codon probability has very considerable role in the identification of exons or protein-coding regions, and introns or non-coding regions [35-36], [72-75]. The percentage of exonic regions in the whole genome is approximately 2%. The percentage of repetitive sequences in human genome is about 60% whereas exonic regions are present in very small percentage. The periodic recurring patterns which the repetitive sequences possess are of varying 1/f base pairs periodicities [76-77]. Intronic regions show 10-11 bps periodicities and these are mostly associated with DNA folding, or the structure of DNA helical repeat [78-80] and these periodicities are not observed in regions where TBP is present. Along with these periodicities, various other periodicities which are reported in literature along with their feature description are presented in Table 2.1.

Detection of exons/coding sequences (CDS) in annotation of eukaryotic genome is considered as highly important. The principle of signal processing based approaches is to utilize the TBP present in these exonic regions for their identification [71]. Another very important step in annotation of genome is detection of CpG Islands (CGIs) as their detection helps in prediction of promoter regions and subsequently genes, early prediction of cancer [5] and various other important biological/medical activities. Various computational algorithms have been developed for the identification of CGIs and been reviewed by Tahir *et al.* [3]. The tandem repeats which

Table 2.1: Periodicities present in DNA sequences

S. No.	Periodicity	Feature/Repeats
1	3	Protein-coding region
2	5-6	Telomeric or Subtelomeric
3	10-11	DNA bendability/Helical repeat structure
4	48-50	Centromeric
5	68	β satellite DNA
6	102	Nucleosome structure in eukaryotes
7	105-106	Isochores/regions having low C+G concentration
8	~135	Dimeric Alu repeat structure
9	~165	A rich Homopolymeric DNA sequence in Alu repeats
10	171	α satellite DNA
11	~300	Alu
12	~680	DNA bend sites

are associated with biological functionality of organisms and have periodicities ranging from 2 to >100 bps [72], [81-84]. For the identification of existing gaps in the already reported and existing methods for identification of these periodicities, the literature has been thoroughly and rigorously reviewed and the same has been discussed in following sections.

2.1 Identification of Protein-Coding Regions in DNA Sequences

In literature, various methods have been developed, reported, and proposed for the identification of protein-coding regions in DNA sequences of eukaryotic organisms since last two decades [85-86]. There are three different categories in which these methods can be classified as suggested by Blanco *et al.* [87]. These categories are: search by signal or site, search based upon similarity, and search using content. Guigo [88] classified these methods in a different way as model-dependent and model-independent techniques. The functioning of signal and similarity based search methods is based upon the principle of trained database which is known a priori and the same is applied to train supervised classifier such as Markov models. Such methods come under the category of model dependent methods. In these approaches, the classification amongst exons or introns requires a huge amount of data trained by machine learning based models or probabilistic models. Whereas content based search approaches explore for DNA sections with particular features such as codon compositions, nucleotides' frequency, CpG islands, and the proportion of nucleotides with rich A-T or G-C contents etc. and such methods can be either model independent or model dependent as well. Discussion of various model dependent methods is presented in [89-91]. Gao *et al.* suggested a Z-curve based model dependent method in which compositional exploration has been used [89]. A method called 'GeneScout' was proposed by Yin *et al.* [90] in which specially designed hidden Markov models have been used for the prediction of exon coding potential computation, and therefore this is model-dependent approach. Borodovsky *et al.* proposed Genemark [91] in which authors have used specific Markov models for exonic and intronic regions along with Bayes' decision making function which is categorized as model dependent approach. These model dependent methods have dependency on trained datasets for their functioning and therefore their performance can suffer because of addition of new data in repositories of genomic databases. There may occur unknown genes in the training datasets which are not the part of existing databases and hence the detection performance of such methods will be affected [92-93]. Under these circumstances, a better choice can be to use model independent methods; although model dependent methods are more precise.

Nucleotides of DNA sequences get converted into aminoacids by triplets (codons) in protein-coding regions. Proteins contain only 20 aminoacids. As the number of possible codons is 64, the number of aminoacids produced by using many to one mapping of codons is only 20 [88]. The codons responsible for coding the aminoacids may not have the uniform distribution of probability in a species. Due to this codon-bias which occurs as a result of non-uniform codon usage, three-base periodicity (TBP) has been observed in the protein-coding regions of eukaryotes. Most of the DSP based algorithms developed in last many years which are considered to be in the category of model independent methods have utilized this feature of TBP for the prediction of protein-coding regions and short-time discrete Fourier transform (STDFT) has been used in these approaches like in [37], [41], [94-96]. In these approaches, the DFT is computed by sliding a fixed length window across the length of DNA sequence to identify the TBP of protein-coding regions using f=1/3. The presence of spurious spectral peaks and few artifacts observed in the spectrum of windowed DFT is considered as the drawback of these methods. The dependency of STDFT approach upon the choice of window length and shape is a major limitation. The solution of this limitation of DFT method has been proposed by Rao *et al.*

[97], Mena-Chalco *et al.* [98] & Sahu *et al.* [99], Mariapushpam *et al.* [100] employing multiresolution based transform techniques such as wavelets, modified Gabor wavelet transform (MGWT), S-transform, and discrete Wavelet transform (DWT) respectively. However, computational complexity is a limitation of these methods. Akhtar *et al.* [101] proposed the paired and weighted spectral rotation (PSWR) measure for the reduction of computational complexity and the improvement in accuracy of gene prediction. Another solution of fixed window size limitation of DFT based method has been suggested by Shakya *et al.* [102] in which the authors have proposed the adaptive window length strategy in STFT as a remedy for the choice of window size problem and the method is known as AWSTFT. Another DSP based approach in which filtering technique has been employed by Vaidyanathan *et al.* [103] is a faster approach of prediction. Ramachandran *et al.* [104] suggested a filter based method but this approach is a model dependent method. Hota *et al.* [105] proposed the use of three antinotch filters for reduction of computational complexity load and for the improvement of accuracy of prediction of protein-coding regions. The use of instantaneous matched filtering based Statistical Optimal Null Filter (SONF) for the prediction of exons by detecting TBP in DNA sequences has been proposed by Kakumani *et al.* [106] and Zhang *et al.* [107]. The use of entropy measures for detection of exons has been proposed by Ginnori *et al.* [75], Roldan *et al.* [108], and Nicorici *et al.* [109]. In these approaches, Shannon entropy measure based entropic segmentation of DNA sequences into homogeneous domain has been utilized [75][108]. Renyi divergence measure, nucleotide statistics, and stop codon statistics have been employed in another entropic measure based method to identify exons [75][109]. Autoregressive (AR) model based exon detection has been proposed by Chackravarthy *et al.* [110]. Choong *et al.* developed AR model based multiscale parametric spectral analysis for exon identification whose performance is better than DFT and earlier AR model based algorithms [111]. The Wavelet subspace Hilbert-Huang transform (WSHHT) based exons identification has been developed by Jiang *et al.* [112]. Wide-Range Wavelet Window (WRWW) method has been proposed by Marhan and Kremer [113] and their method is able to predict protein-coding regions satisfactorily across a range of length of protein-coding regions. Recently Adaptive S-transform-principle component analysis (AST-PCA) based approach has been proposed by Sharma *et al.* [114]. In this method, the authors have identified the short exonic regions associated with intronic regions during alternative splicing and have employed multiple mapping schemes. Mostly transform based approaches have been

developed and proposed so far for the identification of protein-coding regions in the DNA sequences of eukaryotes. The fundamental procedure in transform based methods is to convert the signal from time-to-frequency domain. This transformation may result in domain bias and subsequently lead to loss of some important information of signal like protein-coding regions. Also, some of the approaches proposed so far have emphasized on detection of shorter length exons only while some approaches have focused on predicting bigger length exons only.

2.2 Identification of CpG Islands in DNA Sequences

CpG Islands (CGIs) are the regions where the DNA character 'Cytosine' is followed by character 'Guanine' along the length of DNA sequences in 5' to 3' direction. As CGIs are associated with various epigenetic functions such as genes mutation, gene regulation, promoters' prediction, chromosome inactivation, DNA methylation, and cancers etc. Hence, the detection of CGIs is considered as highly important. The CGIs can be predicted experimentally by the biologists and the results of prediction are considered as accurate. However such experimental methods of detection of CGIs are extremely time consuming as the amount of genomic data is huge [115]. Therefore, computational methods which are efficient in prediction of CGIs are considered a good choice. The first computational method for prediction of CGIs in vertebrates has been proposed by Garden *et al.* [116] and this method is popularly called as GGF. In this method, a particular section of DNA sequence which satisfies the following 3 conditions is categorized as CGI and otherwise non-CGI: The length of section has to be at least 200 bps, GC content which is referred to as proportion of Cs and Gs should be minimum 50%, and the observed/expected (o/e) ratio must be at least 0.6. Takai *et al.* [117] suggested rigorous modifications in GGF criteria with a minimum section to be 500 bps, GC content as minimum 55%, and ratio of o/e as 0.65. The criterion for minimum sequence length to be 500 bps was incorporated for the prevention of Alu repeats. Depending upon the principle of working the computational methods developed for CGI identification are classified in four areas which are: window based methods, methods based upon Hidden Markov Model (HMM), methods on the basis of density, and methods developed using distance-/length criteria [3],[118].

The functioning of window dependent methods is based upon a sliding window across the genome and prediction of CGIs is performed applying already defined statistical conditions. In these methods a moving window keeps on sliding by one nucleotide and checks continuously o/e ratio and GC content in the window until the required conditions for CGI section are achieved. These methods are used very much because these methods firmly pursue the standards defined for categorization of a region of genome as CGI or non CGI. However the dependency on size of the scrolling window which is considered as a highly important factor to predict the CGI accurately is a great limitation of such methods. The smaller window size has the advantage of less computational complex but has the drawback of missing a probable CGI. The predictive accuracy is higher with larger window size but at the cost of high computational complexity [3], [118]. The CGI detection methods developed using window based approach is presented in [116-117], [119-123]. The method developed by Ponger *et al.* is known as CpGProd [119]. Rice *et al.* proposed an approach called EMBOSS [120]. Chuang *et al.* proposed a novel idea termed as CPSORL which is based on particle swarm optimization for the detection of accurate CGI followed by some parameters and fast convergence [121]. Park *et al.* proposed a technique of CGI prediction coined as CpGPNP [122]. In this method a window shifts by 1 nucleotide along the length of DNA sequence to search the probable CGI using the predefined conditions of CGI. Yang *et al.* developed an ion motion optimization based algorithm for CGI prediction known as CpGIMO [123]. In this approach the authors have used 200 to 2000 bps window for the prediction of CGI randomly.

The HMM was applied in sequence analysis earlier and thereafter the concept was successfully implemented for partition of genomes [3], [124]. The prediction of CGI applying HMM was proposed by Durbin *et al.* [125] and subsequently an extensible approach for the identification of CGI was proposed. In HMM based techniques, Markov chains based two different models for CGI and non-CGI separately are employed and according to the probability of CGI and non-CGI regions, the log-likelihood ratio is computed to show the difference between these two regions for every sequence. The prediction accuracy of these methods suffers because of variable patterns present in CGI which creates some noise and secondly lack of adequate data for training purpose. And also the computational capacity of such HMM methods is not very efficient [3], [118]. Some HMM based CGI detection methods are developed by Wu *et al.* [115], Yoon *et al.* [126]. In the approach proposed by Wu *et al.* the probability scores are generated as a result of the summary to indicate the status of CGI [115]. Yoon *et al.* have utilized the Markov chain model presented in [125] and proposed a technique based on a bank of IIR lowpass filters for CGI identification [126].

The calculation of density of CpG sites using statistical parameters just like window-based methods is the main principle of functioning of methods on the basis of density for detection of CGI. The density of CGI is computed by finding out the percentage of CpG sites in CpG Islands and the complete length of CpG Islands [3], [118]. In these methods the fundamental operation is to set the seeds initially for adjustment of the density variables on iterative basis and subsequently to expand the coverage of regions which are rich in CGI. To begin with, a low threshold value set for density is adjusted to analyze the predicted boundaries of CGIs. Subsequently a high threshold value of density is employed for the determination of range of borders of CGI where the criteria set for density is met by the DNA sequence. As the functioning of these methods is fully dependent on the thresholds of density, this is considered as a major limitation of such methods [3], [118]. The density based approach has been proposed by Ye *et al.* [127]. Ye *et al.* developed an algorithm termed as CpGIF (CGI Finder). They incorporated the distinctive features of existing methods and at the same time overcame their drawbacks. In their method, regions having high density of CGI which served as seeds are searched first and then final CGIs are computed by extension and clustering of those seeds [127].

A faster way of predicting CGIs is employed in distance-/length based methods in which clustering of data between CpG sites is performed. A newer viewpoint of understanding the phenomena of CGIs is provided in these methods by analyzing the sequence property amongst any two aligned CpG sites but this is also considered as a point of criticism of these methods. There may occur varied outputs of a same CGI under different scenarios and hence the predictive sensitivity is low which is a major drawback of these methods [3], [118]. An approach based on this distance criterion has been proposed by Hackenberg *et al.* and the method developed by them is popularly called as CpGcluster [128]. In this method the CpG clusters are determined directly on the basis of physical distance. The classification of statistically significant clusters as CGIs is done after each group has been assigned the p-value [3], [128].

In addition to these four categories of approaches for CGI detection, various other computational advanced techniques have also been reported. Kakumani *et al.* proposed a statistically optimal null filter (SONF) based CGI identification approach. In this approach the authors have proposed the combination of maximum signal to noise ratio and the criteria of least square optimization for the estimation of CGI prediction characteristic [5]. Gaussian model based algorithm termed as

GaussianCpG for the identification of CGI in human genome DNA sequences has been proposed by Yu *et al.* [118]. Gaussian model has been designed to represent the fundamentals of microscopic links in complex human genome. In this model at initial level, every CpG site's energy distribution is being investigated by scanning across the primary structure of human genome and subsequently statistical parameters are adjusted. A hybrid approach named CpGclusterTLBO in which clustering process and teaching learning based optimization (TLBO) process have been combined for the detection of CGIs in human genome has been proposed by Yang *et al.* [129]. In this approach clustering process has been employed to capture the candidate CGIs and the effect of clustering is the reduction of superfluous DNA segments out of huge volume of data. TLBO has been applied thereafter to finally capture verified CGIs out of candidate CGIs [129]. A discrete Wavelet transform (DWT) based improved algorithm for CGI detection has been developed by Mariapushpam *et al.* [130]. In this proposed approach the authors have applied DWT using Symlet 11 wavelet function for filtering and subsequently recursive least squaring (RLS) based adaptive filtering has been employed to predict the CGI in genomic sequences [130].

As the nature of DNA sequences represent the recurring patterns which indicate in the direction that CpG Islands can have some periodic patterns hidden inside them. The CpG Islands prediction approaches developed so far have not focussed on hidden periodic patterns in the CGIs.

2.3 Tandem Repeats Identification in DNA Sequences

The role of computational methods for the analysis and processing of biological signals is of great impact. The *abinitio* techniques developed for identification of repeats in DNA sequences of eukaryotes have a great significance. The repeats in DNA data are associated with a lot of diseases. Tandem and interspersed repeats are the two broad classes in which DNA repeats are categorized. If two or more than two copies of a particular period are located in a continuous manner, such repeats are termed as tandem repeats. On the other hand, the non-continuous location of two or more copies of a particular pattern in the DNA sequences correspond to interspersed repeats. Microsatellites, Minisatellites, and Satellites are the three classes in which tandem repeats can be placed according to the length of repeated pattern. If the repeated pattern's length varies between 2-8 bps, such repeats are termed as microsatellites which are also known as short tandem repeats (STR). The 9-80 bps size of repeated pattern place the repeats in category of Minisatellites; and Satellites are the repeats whose periodic pattern size is above 100 bps [47]. As mutation affects and replaces the character of DNA sequence, this effect can be observed in terms of tandem repeats as perfect and imperfect repeats. If the accurate number of copies of a particular pattern is observed in DNA data, such repeats indicate perfect tandem repeats. On the other hand, the inexact number of copies of a certain pattern gives rise to imperfect tandem repeats. In the literature, numerous computational techniques have been reported for the detection of tandem repeats. Correspondingly, researchers have designed and proposed various algorithms for the detection of periodic pattern, their location, number of copies, and structure of these periodicities in the DNA data. The two main areas in which these approaches can be broadly classified are: stochastic and deterministic [131]. A lot of probable paths occur in stochastic models for a process which starts from the known points. The popular use of sequence alignment following probabilistic models is in the prediction of microsatellites [132]. The heuristic approaches have been proposed for the reduction of run time of these algorithms [133]. However, prior information regarding the period of repeat or the fundamental pattern of the segment is considered a limitation on the applicability of these methods. The solution of many of the limitations of such methods has been provided by Benson [134]. The method developed by Benson is popularly known as Tandem Repeats Finder (TRF), and the working of TRF is based on stochastic model [134]. TRF method is able to predict repeats having larger patterns and its detection capability is governed by indel (insertion/deletion) probabilities, matching probabilities, and some model-based statistical criteria. The conversion of character sequence of DNA into numerical sequence with the help of numerical mapping schemes has opened up many directions for signal processing based algorithms to be applied and further analyze the DNA sequences [69]. Various deterministic algorithms which employ signal processing methods have been reported for the identification of tandem repeats. The behaviour of deterministic algorithms is predictable and also the advantage of such algorithms is that these can detect more number of repeats in DNA data with enhanced sensitivity towards detection of approximate repeats. Algorithms which apply correlation techniques and are based on signal processing methods have been designed to detect the TBP of protein-coding regions in DNA sequences of eukaryotes [102]. These methods can predict higher number of approximate tandem repeats because of higher sensitivity of such methods towards latent periodicities. On the other hand, it becomes very difficult task for methods based on string-matching conditions to predict approximate repeats which results from mutations because the matching conditions are very stringent of these methods. A review of string matching principle based algorithm has been provided by Lim *et al.* [135]. Some signal processing methods such as DFT (discrete Fourier transform) [44], [136], modified Fourier product spectrum [137], STPT (short-time periodicity transform) [138], EPSD (exactly periodic subspace decomposition) [46], QPT (quaternion periodicity transform) [45], OMWSA (optimized moving window spectral analysis) [140], AR (auto regressive) model [6], WBEMD (Wavelet-based empirical mode decomposition) [139], AST (adaptive S-transform) [141-142], ST (S-transform) [143] have been developed and reported in literature for the prediction of short tandem repeats. Development of Fourier transform based algorithms for the prediction of tandem repeats has remained very popular amongst the researchers working in the area of SP based methods. DFT [136], SRF (spectral repeats finder) [44], MFPS (modified Fourier product spectrum) [137] are some of the Fourier transform based tandem repeats detection approaches proposed in the literature. The detection of tandem repeat having any arbitrary periodicity present in the DNA sequences is considered an advantage of these methods. The DFT method is capable of detecting imperfect or approximate tandem repeats in the DNA sequences. SRF method which is based on DFT approach primarily detects the repetitive periodicities located inside any DNA sequence and thereafter the DNA sequence is scanned at these periods to locate the approximate segments where the repeat patterns are contained. The power spectrum containing peaks having high intensity is observed for large number of exact tandem repeats in this method. The degradation of signal's quality in the case of DNA sequence having mutations is considered as a limitation of this approach. The solution to this problem which occurs because of spectrum sum has been proposed by Tran *et al.* [137]. They proposed the use of Fourier product spectrum and detected weak approximate tandem repeats. To overcome the shortcomings of spectral methods, Buckner *et al.* have introduced a time-domain approach known as STPT (short-time periodicity transform) to localize the tandem repeats properly using periodogram [138]. However, in both the Fourier transform and STPT based algorithms, the limitation of multiple periodicities has been observed which implies that it is difficult to predict a particular detected repeat to be period p or multiples of p such as 2p, 3p, and likewise. The EPSD (exactly periodic subspace decomposition) method proposed by Gupta *et al.* [46] does not suffer from multiple periodicities problem & classifies a detected repeat to be of period p or its multiple and this method is able to detect exact and approximate tandem repeats as well. However, requirement of various window sizes with different repeat periodicities and no precise specification of criteria to decide the window size are the limitations of EPSD approach. To address the shortcomings of EPSD approach, Brodzik *et al.* proposed a solution by developing QPT (quaternion periodicity transform) method [45]. This method overcomes various shortcomings such as symbol bias, absence of criteria to detect the indels, and lack of an appropriate postprocessing stage. However, the detection capability of this method is restricted because the minimum length of repeat period has to be specified in advance. An approach named as OMWSA (optimized moving window spectral analysis) has been suggested by Liping *et al.* which is robust and perfect method in the presence of indels in DNA sequences as compared to FT based methods [140]. An AR (auto regressive) model based on parametric spectral estimation (PSE) developed by Zhou *et al.* analyzes the DNA sequences using their spectrograms [6]. This approach has been considered as an improvement/extension over OMWSA approach. The background noise is reduced to a great extent in this method and this model generates a sharp peak. However, as per the characteristics of data, the optimal order of the AR model has to be decided which is not known in advance in this method. And the working of this approach is based upon deciding the smallest frequency for the calculation of repeat periodicity when several frequencies are contained in the power spectrum of repeats. Hence, there arises the possibility of false detection of tandem repeat when the smallest frequency to be predicted is too weak. Moreover, the choice of an appropriate length of window is also a limitation of this approach. The solution of this problem of order selection and window length, Zribi *et al.* [143] has suggested a solution using S-transform based approach in which p-nuc coding has been employed. Jiang *et al.* [139] has proposed an approach named as EMWD (Empirical mode and Wavelet decomposition) in which the authors have employed wavelet algorithm in combination with EMD (empirical mode decomposition) in the pre-processing stage and a cross-correlation analysis (CCA) as post-processing stage. The power spectral density has been displayed efficiently for both short and long signals in the 2-dimensional (2-D) frequency-time plane [139]. Sharma *et al.* has proposed the AST (adaptive S-transform) based microsatellite detection [141]. The authors have optimized the standard deviation of Gaussian window kernel to be used in the S-T for integer periodicities with the help of maximization of concentration measure. This algorithm can detect microsatellites only and not able to capture minisatellites which is reported as a limitation of this method. Sharma *et al.* has also proposed another AST (adaptive S-transform) based algorithm using Kaiser window and this window function helped in the detection of both long and short repeats [142]. The authors have also detected exact and approximate tandem repeats as well.

2.4 Research Gaps Existing in Current Solutions

Having undergone the exhaustive review of the current solutions for the research problems discussed in Section 2.1 to 2.3, the potential of improvisation over existing reported methods was observed. The various number of efficient and accurate methods has been developed for the identification of protein-coding regions many of which are based on transforms. As the domain transformation may lead to biasing in context of losing of very important information, hence an approach which does not require any transformation is strongly required to detect the protein-coding regions. Also, some methods are able to capture short length exons only whereas other methods can identify bigger length exons only. Hence, an approach which can capture both shorter and bigger length exons simultaneously is highly required. The CpG Islands detection methods reported so far have not explored the hidden periodicities inside them. Therefore, an algorithm which can reveal this periodicity feature with experimental proofs is extremely desired for efficient detection of CpG Islands. The tandem repeats detection approaches developed so far suffer from some shortcomings which require to be addressed. The research problems proposed in this dissertation have been formulated based on these research gaps.

2.5 Problems Formulation

The following research problems have been formulated on the basis of detailed study of the existing work:

i) To develop a signal processing based algorithm for the detection of protein-coding regions in the DNA sequences of eukaryotes.

ii) To develop an efficient signal processing based approach for the detection of CpG Islands in the DNA sequences.

iii) To improve the sensitivity and overall performance for the CpG Islands detection in the DNA sequences of human species using signal processing based algorithms.

iv) To develop signal processing based methods for the detection of tandem repeats in DNA sequences.

CHAPTER 3

IDENTIFICATION OF PROTEIN-CODING REGIONS IN DNA SEQUENCES OF EUKARYOTES USING SINGULAR VALUE DECOMPOSITION BASED MODIFIED P-SPECTRUM BASED ALGORITHM

Mostly transform based approaches have been developed and proposed so far for the identification of protein-coding regions in the DNA sequences of eukaryotes. The fundamental procedure in transform based methods is to convert the signal from time-to-frequency domain. This transformation may result in domain bias and subsequently lead to loss of some important information of signal like protein-coding regions. The solution to this issue has been proposed in this chapter using singular value decomposition (SVD) based modified P-spectrum [144-146] based algorithm (MPSA). Also, many of the approaches developed so far emphasize on prediction of shorter length exons only while other approaches have focused on detecting bigger length exons only. Therefore, an approach which can identify both smaller and bigger length exons simultaneously is highly required. Again, this issue has also been covered by the proposed algorithm discussed in this chapter. The 24 possible combinations of integer mapping are applied to convert DNA characters to numerical values. The window length has been optimized which has been varied from value 27 to 351 in the step size of value 3 by maximizing the performance metric: area under curve (AUC). The Savitzky-Golay (S-Golay) filter has been applied as a post-processing step to filter out the noise while retaining the important features of signal.

3.1 Proposed Algorithm for Identification of Protein-Coding Regions

The flow graph of the proposed algorithm has been depicted in Figure 3.1 and the steps of the proposed algorithm are outlined as:

i) The DNA sequence in which protein-coding regions have to be identified is applied to the proposed algorithm.

ii) The value of window length is selected as 27 initially.

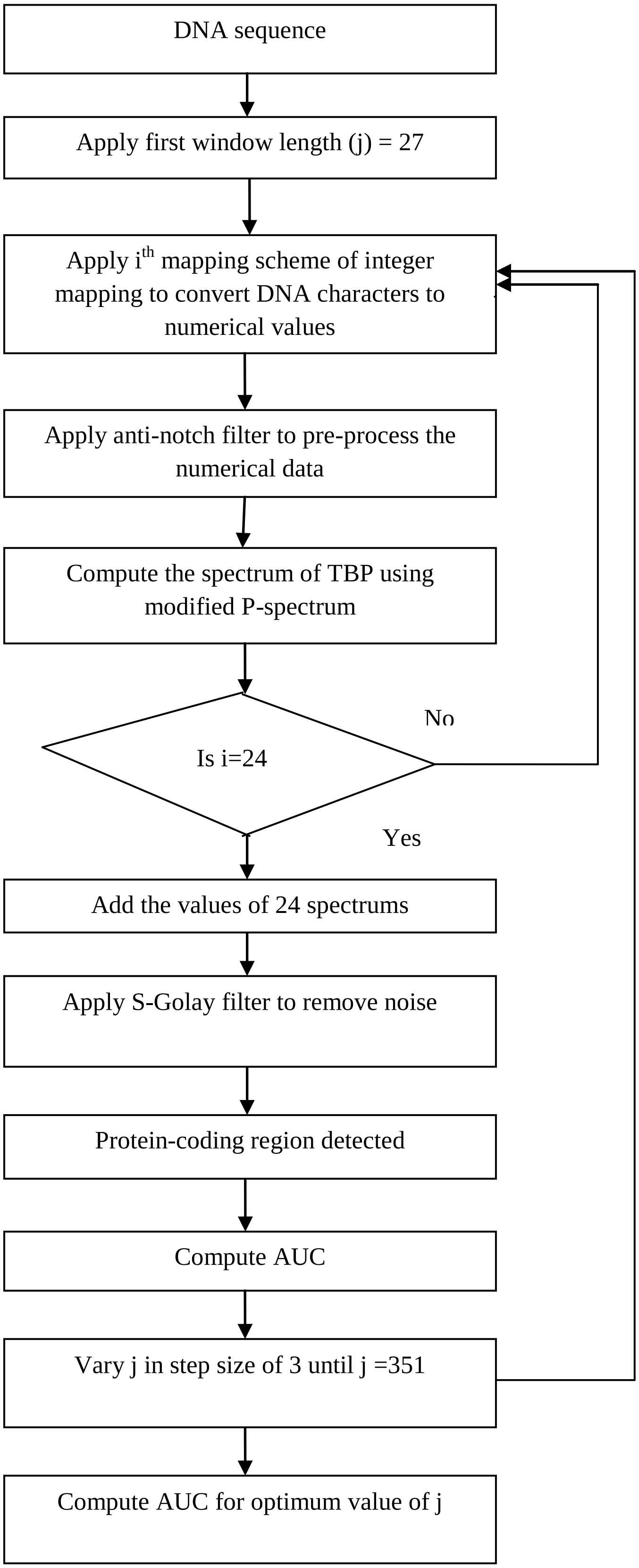

DNA sequence
Apply first window length (j) = 27
Apply ith mapping scheme of integer mapping to convert DNA characters to numerical values
Apply anti-notch filter to pre-process the numerical data
Compute the spectrum of TBP using modified P-spectrum
Is i=24
No
Yes
Add the values of 24 spectrums
Apply S-Golay filter to remove noise
Protein-coding region detected
Compute AUC
Vary j in step size of 3 until j =351
Compute AUC for optimum value of j

Figure 3.1: Flow graph of the proposed algorithm

iii) The first value of mapping out of 24 possible combinations of integer mapping scheme being chosen as numerical mapping method to convert the DNA characters to numerical values is applied to given data.

iv) The most popular anti-notch filter proposed by Vaidyanathan *et al.* [94], [103] in the area of protein-coding region identification whose centre is at corresponding to period-3 frequency which is 2π/3 is applied to numerical sequence to filter out the noisy elements from the data. The value of bandwidth control parameter to be used in the filter which is also considered as quality control parameter has been chosen as 0.992.

v) The SVD based modified P-spectrum is applied to detect the hidden TBP in the given data.

vi) After applying all 24 combinations of integer mapping, the 24 spectrums obtained are added linearly.

vii) The S-Golay filter is applied as post-processing step to the spectrum to remove the noise in the detected spectrum of TBP of protein-coding regions. The key elements of designing of S-Golay filter are the polynomial order and the frame size. It is desirable to keep the polynomial order always than the frame size to achieve better smoothing. The polynomial order value has been chosen as 3 and the frame size has been selected as 41 empirically.

viii) The AUC of detected protein-coding region of the given DNA sequence for the initial window length is computed.

ix) The window length value is now changed to 30 and similarly next time in step size of 3; the steps outlined from iii) to viii) are repeated until the last window length value as 351. The total window length iterations thus undertaken from 27 to 351 in step size of 3 are 109.

x) The AUC is computed for proposed algorithm run for all 109 window lengths and the maximum value of AUC out of 109 iterations is selected finally.

The details of the methodologies employed in the proposed algorithm are presented as follows:

3.2 Modified P-Spectrum

It is well known that the protein-coding regions reveal three–base periodicity which is popularly called as TBP or period-3 property while non coding regions generally do not possess such property [2]. Many digital signal processing based algorithms have been proposed since last two decades for the identification of the protein-coding regions whose principle of working is based upon detection of TBP. The main idea is to have the algorithm capable of detecting the TBP in the DNA sequences to correctly classify the region as protein-coding region. To detect the TBP, modified P-spectrum has been used in this paper. The use of P-spectrum for periodicity detection is not new; it has been used by Kanjilal *et al.* [144] to detect and then correspondingly separate the periodic components which are entrenched in an irregular series. Qiu *et al.* have used modified P-spectrum for the detection of QRS component in Electrocardiograph (ECG) signals [145]. Liscombe *et al.* [146] have proposed modified P-spectrum with considerable reductions in the computational complexity and processing time. Garg *et al.* have used P-spectrum to identify the tandem repeats present in the DNA sequences [147]. It has been observed from these proposed approaches that P-spectrum is an effective and robust method for the detection of periodicities present in the different types of data. Therefore modified P-spectrum has been used in the proposed algorithm and tuned to detect the TBP present in the protein-coding regions of DNA sequences. The overview of P-spectrum and subsequently its counterpart modified P-spectrum are discussed now in detail as follows:

For a probable value of periodicity ‘p’ which is TBP in protein-coding regions, a discrete-time signal C may be represented as shown in following equation:

$$C = [c_1 \;\; c_2 \;\; c_3 \; .... \;\; c_M] \tag{3.1}$$

It is necessary that the given signal is a strict multiple of the ‘p’ period for the computation of P-spectrum. To achieve the same, the number of zeros obtained by the difference of the period and remainder are added after the last element of the signal C where the remainder can be calculated by dividing the given signal C with ‘p’ period. The signal C after this rearrangement can be written now as:

$$C = [c_1 \;\; c_2 \;\; c_3 \;\; c_4 \, ... \; 0\;0\;0\;0. ...] \tag{3.2}$$

The matrix Bp is now obtained as represented in equation (3.3) whose rows are corresponding to the ‘j’ non-overlapping sections in respect of period ‘p’ generated from signal C.

$$\mathrm{B_p} = \begin{bmatrix} \mathrm{c_1} & \mathrm{c_2} & \mathrm{c_3} & \cdots \mathrm{c_n} & \mathrm{c_{n+1}} \cdots & \mathrm{c_p} \\ \mathrm{c_{p+1}} & \mathrm{c_{p+2}} & \mathrm{c_{p+3}} & \cdots \mathrm{c_{p+n}} & \mathrm{c_{p+n+1}} \cdots & \mathrm{c_{2p}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathrm{c_{jp+1}} & \mathrm{c_{jp+2}} & \mathrm{c_{jp+3}} & \mathrm{c_M} & 0 & 0 \end{bmatrix} \tag{3.3}$$

The computation of modified P-spectrum is discussed in the following steps:

The matrix $B_p^{'}$ is obtained from matrix $\mathrm{B_p}$ by considering the starting two rows of $\mathrm{B_p}$ as following [146]:

$$\mathrm{B_p^{'}} = \begin{bmatrix} \mathrm{c_1} & \mathrm{c_2} & \mathrm{c_3} \ \ldots & \mathrm{c_p} \\ \mathrm{c_{p+1}} & \mathrm{c_{p+2}} & \mathrm{c_{p+3}} \ \ldots & \mathrm{c_{2p}} \end{bmatrix} \tag{3.4}$$

Now the singular value decomposition (SVD) of matrix $B_p^{'}$ is computed to obtain signal $\mathrm{D_{svd}}$ and the first singular value is selected which is considered as the most dominating value because it indicates the presence of strong periodic component. The SVD is known as the most robust method for extracting this singular value.

$$D_{svd} = \max(SVD(B_p^{'})) \tag{3.5}$$

In the next step, all the elements of matrix $B_p^{'}$ are added and the signal obtained is named as $\mathrm{E_{sum}}$:

$$\mathrm{E_{sum}} = \mathrm{sum}\left(\mathrm{B_p^{'}}\right) \tag{3.6}$$

And now an auxiliary spectrum signal called as $\mathrm{aux_{spec}}$ is derived from signal $\mathrm{E_{sum}}$ using following equation:

$$\mathrm{aux_{spec}} = \max\left({}^{\mathrm{E_{sum}}}/_{2}\right) \tag{3.7}$$

In the final step, the spectrum corresponding to the TBP of DNA sequences is computed by multiplication of the rows of signal D_{svd} and $\mathrm{aux_{spec}}$ to obtain the signal named ($\mathrm{resultant_{spec}}$) as:

$$\mathrm{resultant_{spec}} = \mathrm{D_{svd}} \times \mathrm{aux_{spec}} \tag{3.8}$$

The modified P-spectrum is believed to detect the periodicity based on the singularity of matrix $B_p^{'}$ [145].

3.3 Savitzky-Golay Filter

S-Golay filter is a digital filter which is considered as a smoothing filter. The basic principle of working of S-Golay filter is to search the best fit of the data inside a movable window utilizing the least-squares polynomial fitting theory [148-149]. This principle helps to achieve a high value of signal-to-noise ratio and consequently important features of a signal like as height, peak, and width are retained satisfactorily. The operation of smoothing in S-Golay filter is achieved by sliding a window of length W_L (which needs to be odd value) upon the data having noise. In this process, some mathematical operation is employed to get the windowed data converged to a single value which is the window's midpoint. The following equation describes the smoothing operation:

$$B_i = \frac{\sum_{j=-n}^{j=n} a_j b_{i+1}}{W_L} \tag{3.9}$$

, where B_i indicates the smoothed data at index i and b_i corresponds to the noisy data at index i. The local indexing of data inside the window is represented by index j, the coefficient corresponding to j^{th} smoothing is shown by a_j. W_L is taken as 2j+1 and represents the data points inside the smoothing window.

3.4 Numerical Representation Method

DNA sequences are comprised of A, G, T, and C characters. In the area of development of digital signal processing based algorithm for genomic signal processing, it becomes necessary to convert the character sequence to obtain the numerical sequence employing numerical representation scheme. A lot of numerical representation schemes are reported in literature for this function [41], [69], [150-152]. One of the numerical representation schemes is integer mapping. In this mapping scheme, the numerical values A=1, G=2, T=3, C=4 are assigned to the DNA characters. But this configuration of mapping can suffer from nucleotide bias effect [153] which will affect the performance of algorithm. Hence to overcome such nucleotide bias effect, the following possible combinations of integer mapping can be obtained which are called as 24 possible combinations of integer mapping:

Table 3.1: Possible combinations of integer mapping

	Possible combinations of integer mapping for conversion of DNA characters			
	A	**G**	**T**	**C**
i=1	1	2	3	4
i=2	1	4	3	2
i=3	1	2	4	3
i=4	1	4	2	3
i=5	1	3	2	4
i=6	1	3	4	2
i=7	2	4	3	1
i=8	2	1	4	3
i=9	2	3	1	4
i=10	2	1	3	4
i=11	2	3	4	1
i=12	2	4	1	3
i=13	3	2	1	4
i=14	3	2	4	1
i=15	3	4	1	2
i=16	3	1	4	2
i=17	3	1	2	4
i=18	3	4	2	1
i=19	4	3	2	1
i=20	4	1	3	2
i=21	4	2	3	1

i=22	4	3	1	2
i=23	4	2	1	3
i=24	4	1	2	3

3.5 Applicability of Proposed Algorithm on a Benchmark DNA Sequence

A benchmark DNA sequence F56F11.4 [1], [98], [102] which is reported in literature by many researchers working in this field has been chosen as an example sequence to show the applicability of the proposed algorithm. The DNA sequence F56F11.4 consists of 8100 base pairs (bps) and there exists five protein-coding regions in this sequence at location: 928-1039, 2528-2857, 4114-4377, 5465-5644, and 7255-7605 [102]. The proposed algorithm's result obtained on the example DNA sequence is depicted in Figure 3.2:

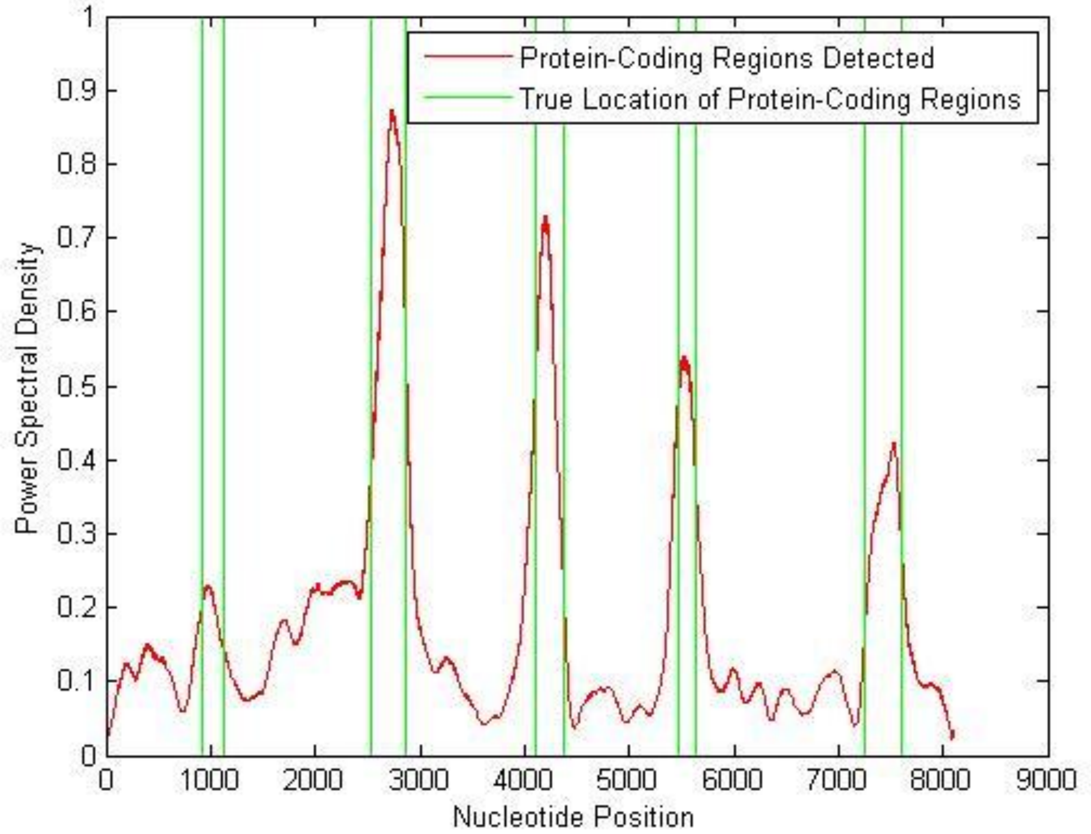


Figure 3.2: Result obtained of proposed algorithm on example DNA sequence F56F11.4

The period-3 power spectrum of protein-coding regions depicted in Figure 3.2 is computed by running the proposed MPSA for the window length varying from 27 to 351, the maximum AUC value 0.9617 is obtained at window length 78.

3.6 Experimental Analysis for Optimization of Window Length

The experiments are performed on the example DNA sequence using proposed algorithm for the window length varying from 27 to 351. The reason is that the performance parameter selected for optimization is AUC and the values of AUC obtained for varying window lengths are different. Out of these, the window length at which the maximum AUC is obtained has been then selected finally and the spectrum of protein-coding regions is plotted for that window length. The values obtained of AUC for varying window lengths for example DNA sequence F56F11.4 is tabulated in Table 3.2 and it has been observed that the optimized window length for example DNA sequence is 78.

Table 3. 2: Value of AUC for varying window lengths for example DNA sequence F56F11.4

Index	Window Length	AUC
1	27	0.9530
2	30	0.9544
3	33	0.9562
4	36	0.9575
5	39	0.9578
6	42	0.9578
7	45	0.9581
8	48	0.9582
9	51	0.9589
10	54	0.9586
11	57	0.9592
12	60	0.9599
13	63	0.9603
14	66	0.9608
15	69	0.9606
16	72	0.9614
17	75	0.9604
18	**78**	**0.9617**
19	81	0.9604
20	84	0.9613
21	87	0.9605
22	90	0.9608
23	93	0.9603
24	96	0.9601
25	99	0.9600
26	102	0.9596
27	105	0.9592
28	108	0.9589

29	111	0.9587
30	114	0.9584
31	117	0.9578
32	120	0.9571
33	123	0.9567
34	126	0.9560
35	129	0.9559
36	132	0.9556
37	135	0.9555
38	138	0.9552
39	141	0.9554
40	144	0.9551
41	147	0.9550
42	150	0.9547
43	153	0.9545
44	156	0.9542
45	159	0.9538
46	162	0.9534
47	165	0.9527
48	168	0.9521
49	171	0.9515
50	174	0.9511
51	177	0.9509
52	180	0.9503
53	183	0.9499
54	186	0.9493
55	189	0.9486
56	192	0.9480
57	195	0.9475
58	198	0.9470
59	201	0.9466
60	204	0.9462
61	207	0.9456
62	210	0.9452
63	213	0.9446
64	216	0.9440
65	219	0.9437
66	222	0.9434
67	225	0.9431
68	228	0.9429
69	231	0.9426
70	234	0.9421
71	237	0.9416
72	240	0.9412
73	243	0.9406
74	246	0.9401
75	249	0.9397
76	252	0.9391
77	255	0.9386
78	258	0.9380
79	261	0.9375
80	264	0.9370
81	267	0.9363

82	270	0.9357
83	273	0.9351
84	276	0.9345
85	279	0.9339
86	282	0.9335
87	285	0.9332
88	288	0.9328
89	291	0.9320
90	294	0.9315
91	297	0.9310
92	300	0.9304
93	303	0.9297
94	306	0.9291
95	309	0.9285
96	312	0.9276
97	315	0.9266
98	318	0.9259
99	321	0.9252
100	324	0.9248
101	327	0.9239
102	330	0.9232
103	333	0.9225
104	336	0.9219
105	339	0.9212
106	342	0.9204
107	345	0.9197
108	348	0.9192
109	351	0.9186

The reason for choosing the window length values minimum as 27 and maximum as 351 is that for many data set sequences of HMR195, BG570, and GENSCAN datasets [114], [154-155]; the maximum AUC value obtained is at minimum window length value 27 and below window length value 27 the AUC value observed is not significant. And the maximum value of window length has been chosen as 351 because again for many data set sequences of HMR195, BG570, and GENSCAN datasets; the maximum AUC value obtained is at maximum window length value 351 and above window length value 351 the AUC value observed is not significant.

3.7 Experimental Analysis for Choice of 24 Combinations of Integer Mapping

Experiments are performed on example DNA sequence using numerical representation schemes other than 24 combinations of integer mapping. The numerical representation schemes considered are integer mapping, electron-ion-interaction potential (EIIP) mapping, modified EIIP mapping, atomic number, Complex mapping, four-bit binary, pseudo EIIP, three-bit binary, two-bit binary, nucleotide frequency occurrence, real number, molecular mass, and quaternary. The value assigned to the four characters of DNA sequence using these numerical representation schemes is presented in Table 3.3, and the value of performance parameter AUC obtained corresponding to optimal window length using these mapping methods is summarized in Table 3.4. The result obtained using proposed algorithm applying these mapping schemes in place of combination of 24 mappings of integer mapping and keeping all other steps same; are represented in Figure 3.3-3.15.

Table 3. 3: Numerical values assigned to DNA characters for different mapping schemes

Numerical representation scheme	**Numerical values assigned to DNA characters**			
	A	**G**	**T**	**C**
Integer	1	3	4	2
EIIP	0.1260	0.0806	0.1335	0.1340
Modified EIIP	0.1260	1	0.1335	1
Atomic Number	70	78	66	58
Complex	1+j	-1-j	1-j	-1+j
Four-bit binary	0010	0001	0100	1000
Pseudo EIIP	0.1994	0.0123	0.1933	0.0692
Three-bit binary	010	001	000	100
Two-bit binary	11	10	01	00
Nucleotide frequency occurrence	0.28142	0.28179	0.20354	0.23326
Real number	-1.5	-0.5	1.5	0.5
Molecular mass	134	150	125	110
Quaternary	1	-1	j	-j

Table 3.4: AUC obtained on example DNA sequence using different mapping schemes

Numerical representation scheme	**AUC (Optimal window length)**
Integer mapping	0.7018 (342)
EIIP	0.8037 (351)
Modified EIIP	0.8091 (342)
Atomic number	0.9153 (297)
Complex	0.9366 (240)
Four-bit binary	0.5793 (171)
Pseudo EIIP	0.8075 (351)
Three-bit binary	0.6466 (237)
Two-bit binary	0.7967 (339)
Nucleotide frequency occurrence	0.8818 (351)
Real number	0.7878 (321)
Molecular mass	0.9155 (270)
Quaternary	0.9366 (240)
Combination of 24 mappings of integer mapping	**0.9617 (78)**

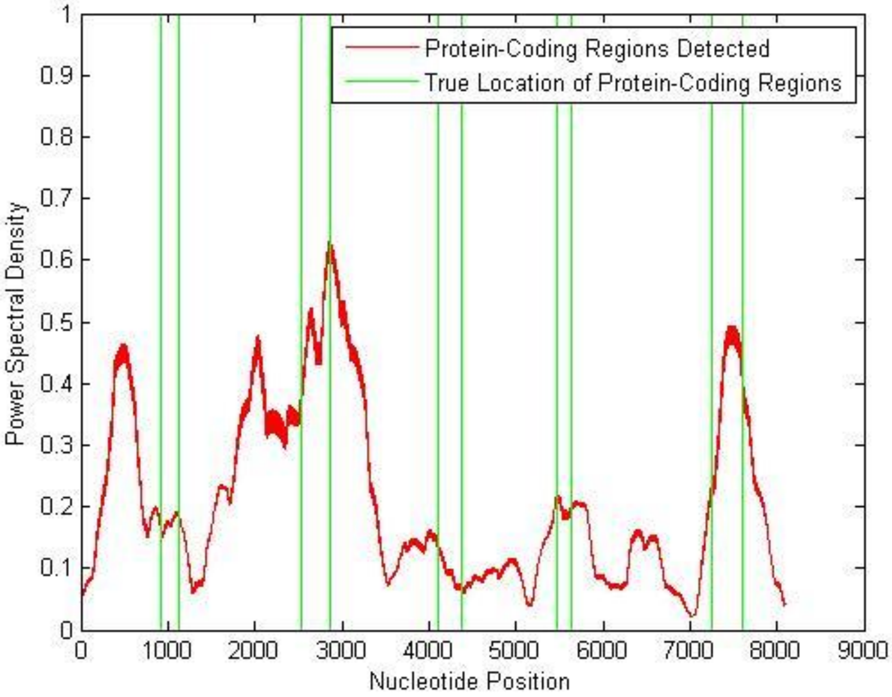


Figure 3.3: Result obtained using integer mapping on example DNA sequence F56F11.4

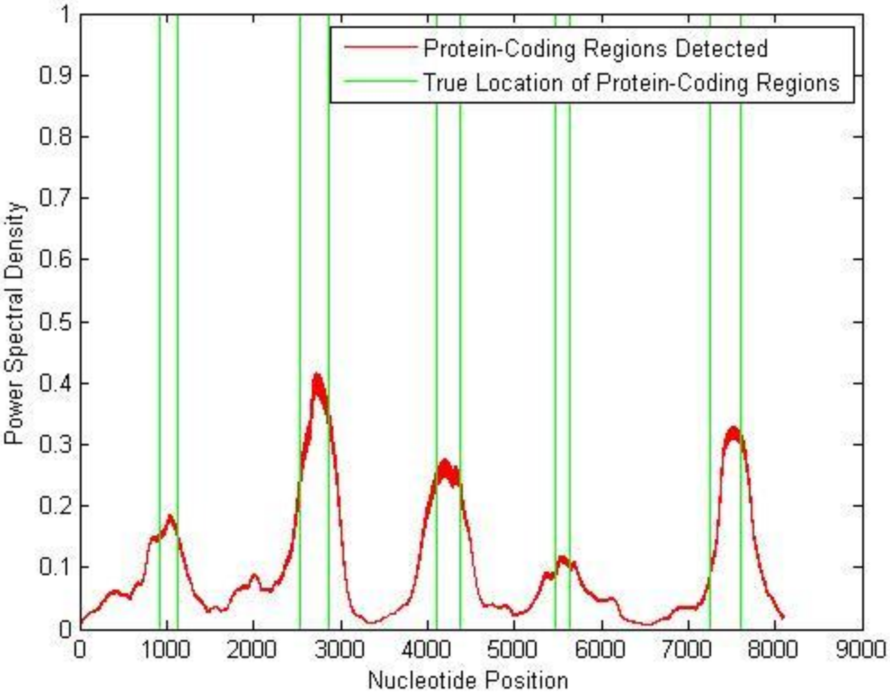


Figure 3.4: Result obtained using EIIP mapping on example DNA sequence F56F11.4

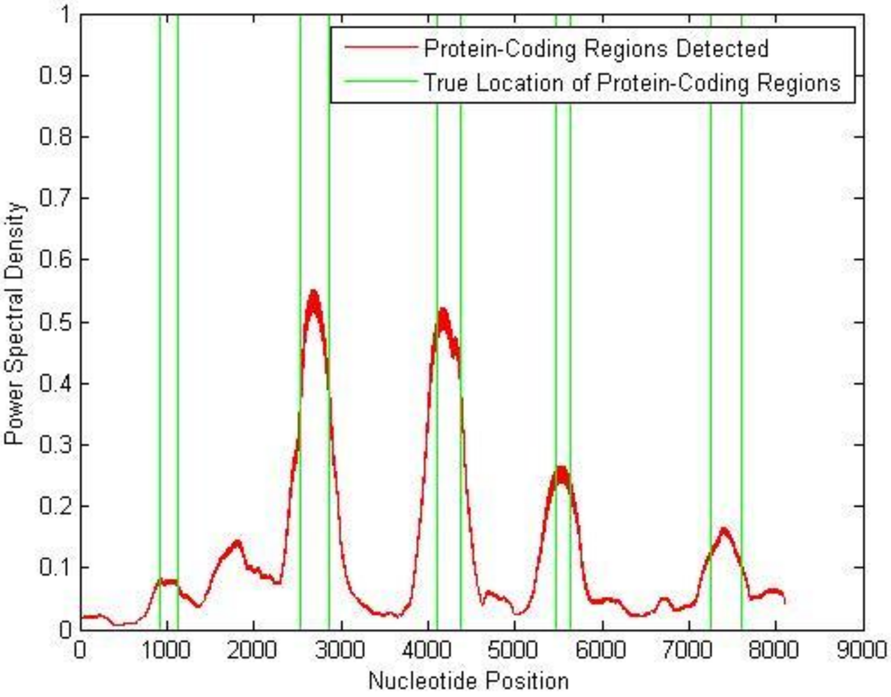


Figure 3.5: Result obtained using modified EIIP mapping on example DNA sequence F56F11.4

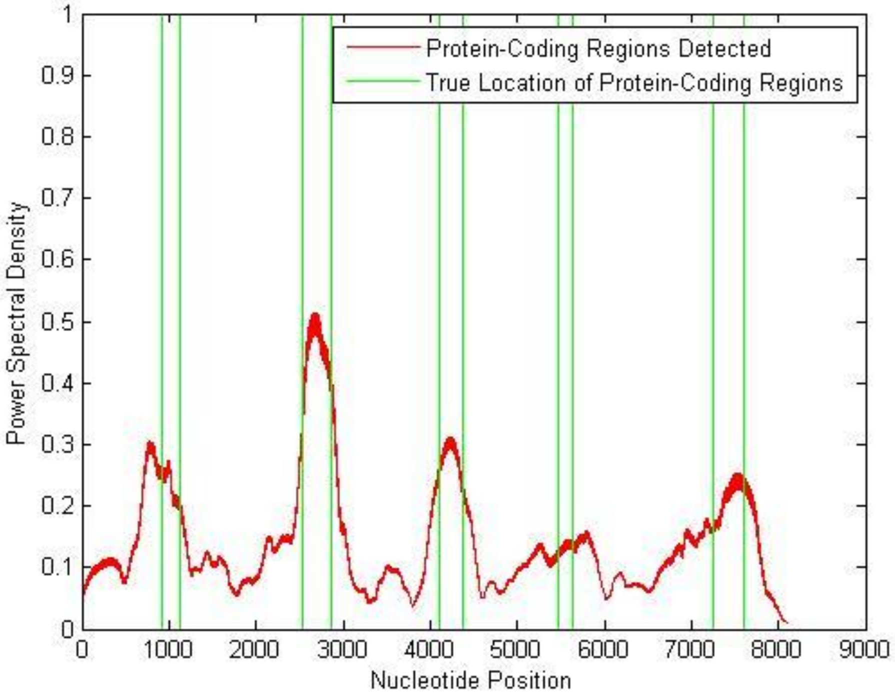


Figure 3.6: Result obtained using atomic number mapping on example DNA sequence F56F11.4

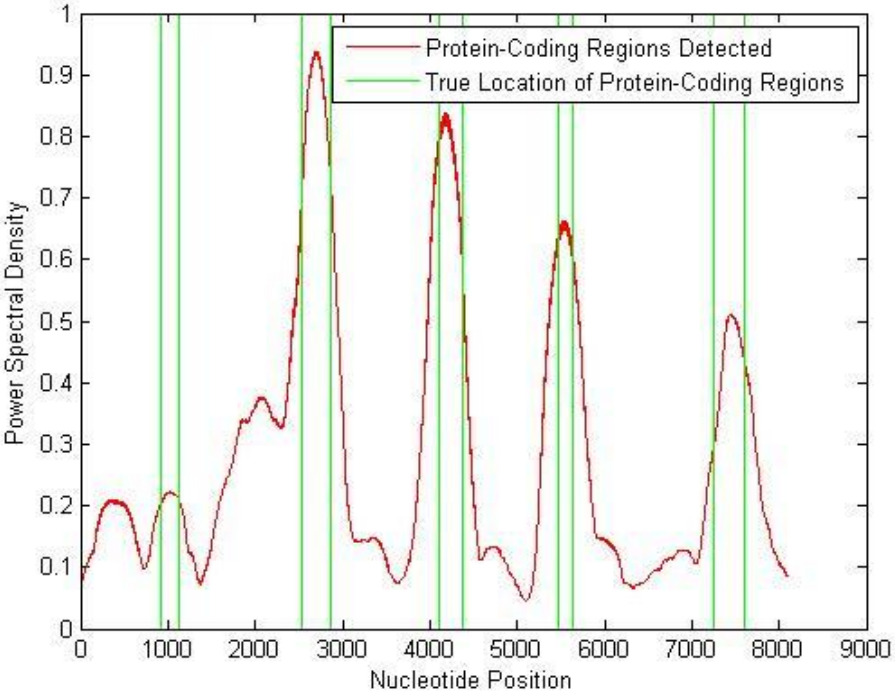


Figure 3.7: Result obtained using complex number mapping on example DNA sequence F56F11.4

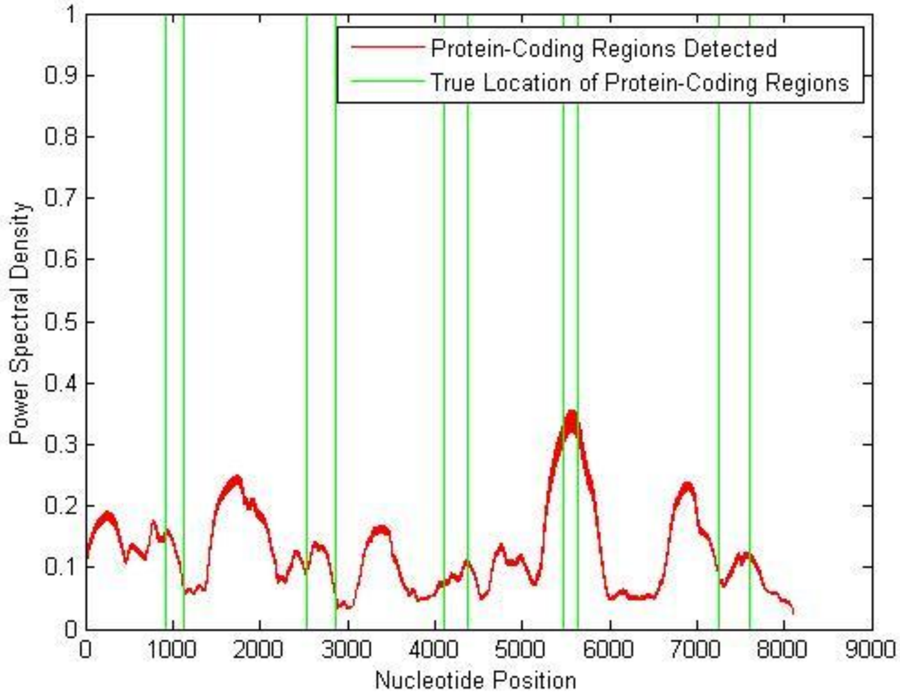


Figure 3.8: Result obtained using four-bit binary mapping on example DNA sequence F56F11.4

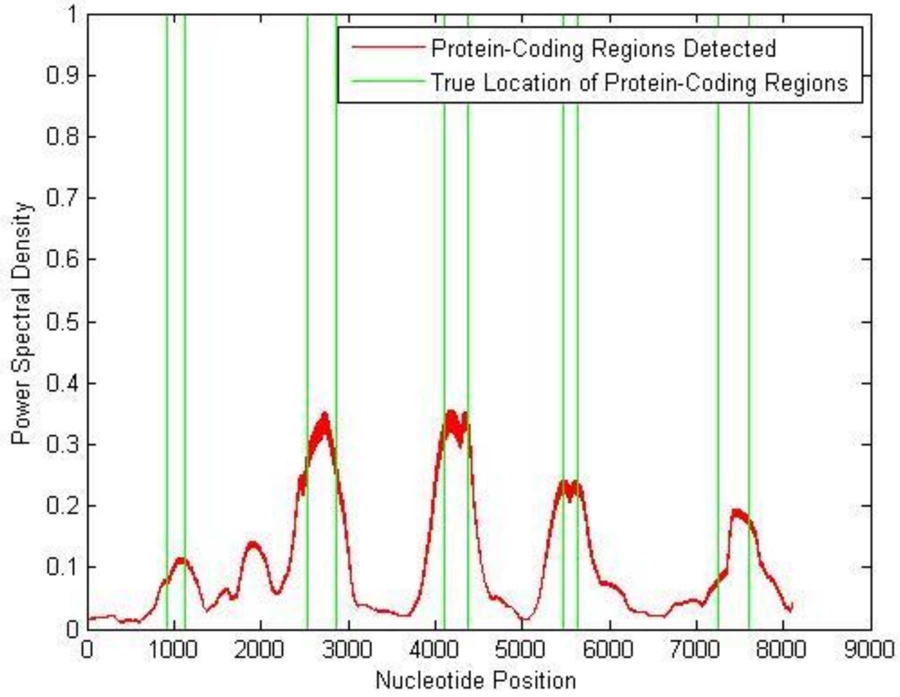


Figure 3.9: Result obtained using pseudo EIIP mapping on example DNA sequence F56F11.4

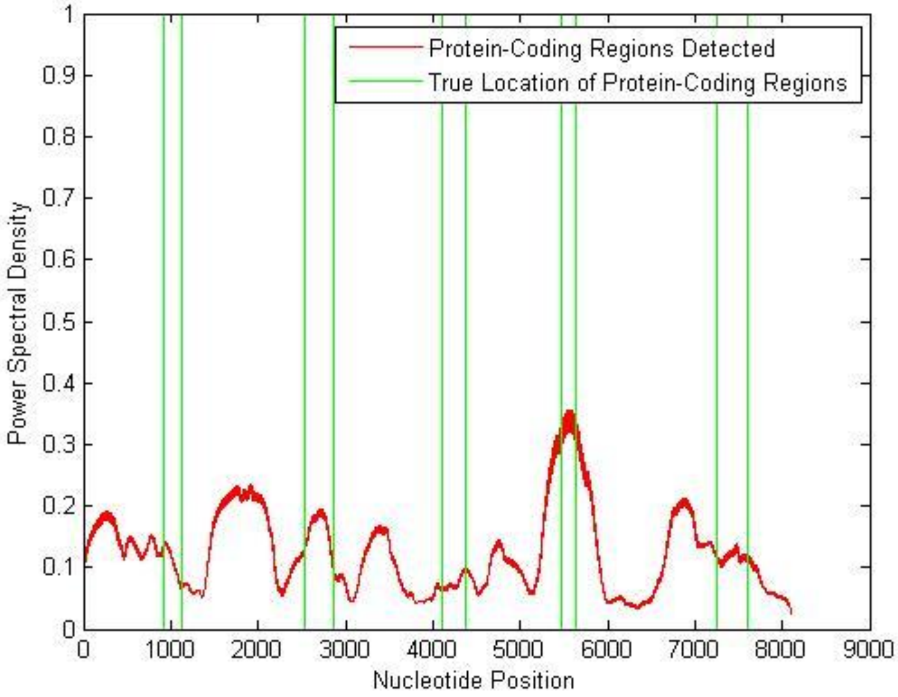


Figure 3.10: Result obtained using three-bit binary mapping on example DNA sequence F56F11.4

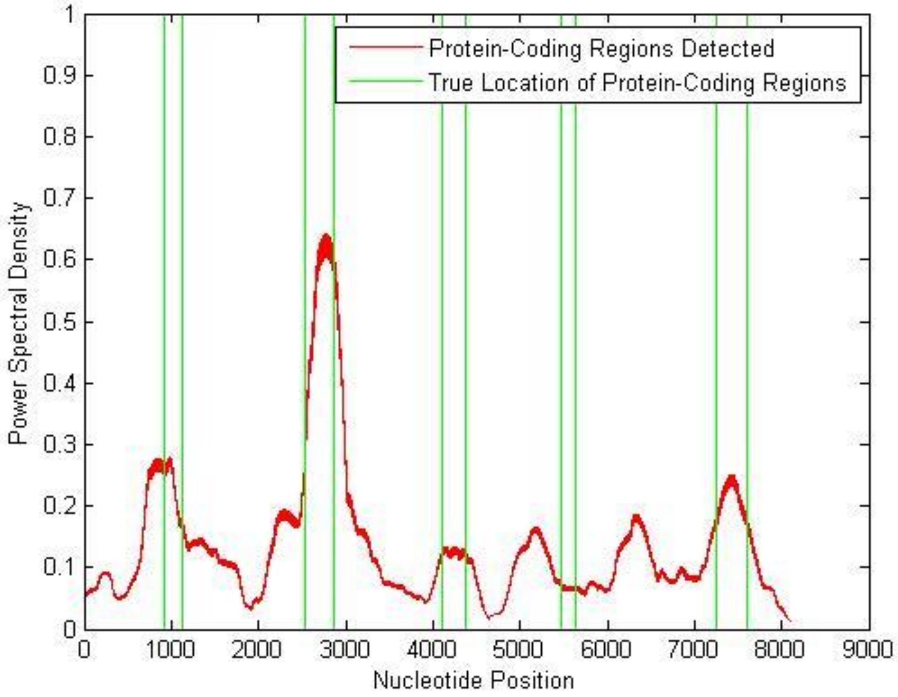


Figure 3.11: Result obtained using two-bit binary mapping on example DNA sequence F56F11.4

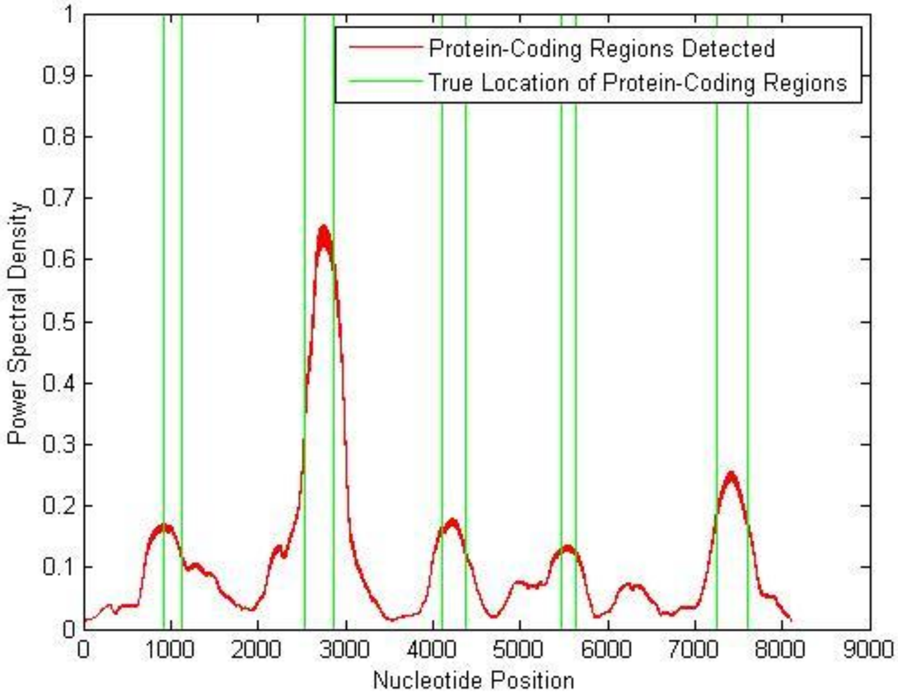


Figure 3.12: Result obtained using nucleotide frequency occurrence mapping on example DNA sequence F56F11.4

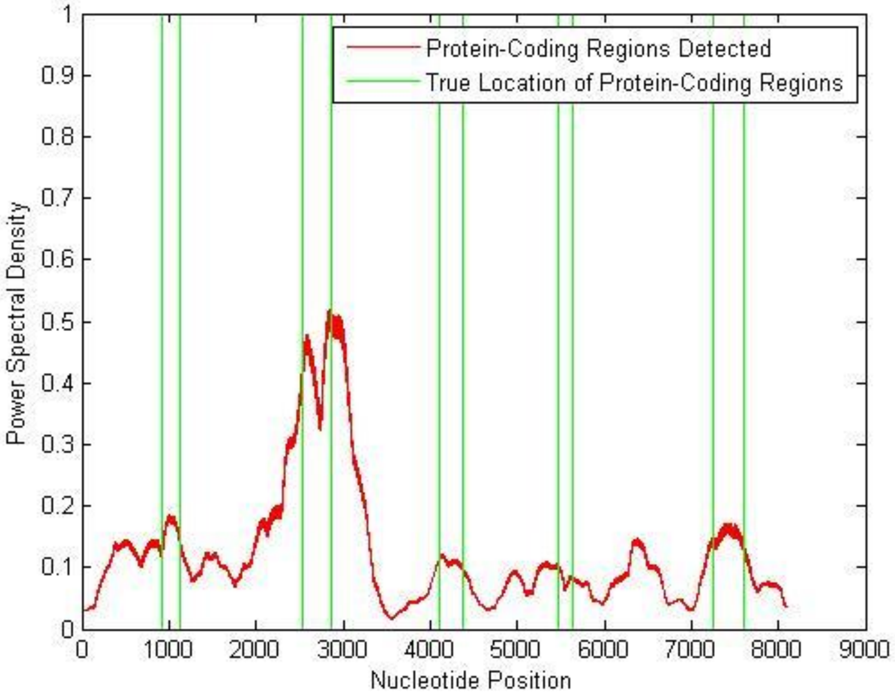


Figure 3.13: Result obtained using real number mapping on example DNA sequence F56F11.4

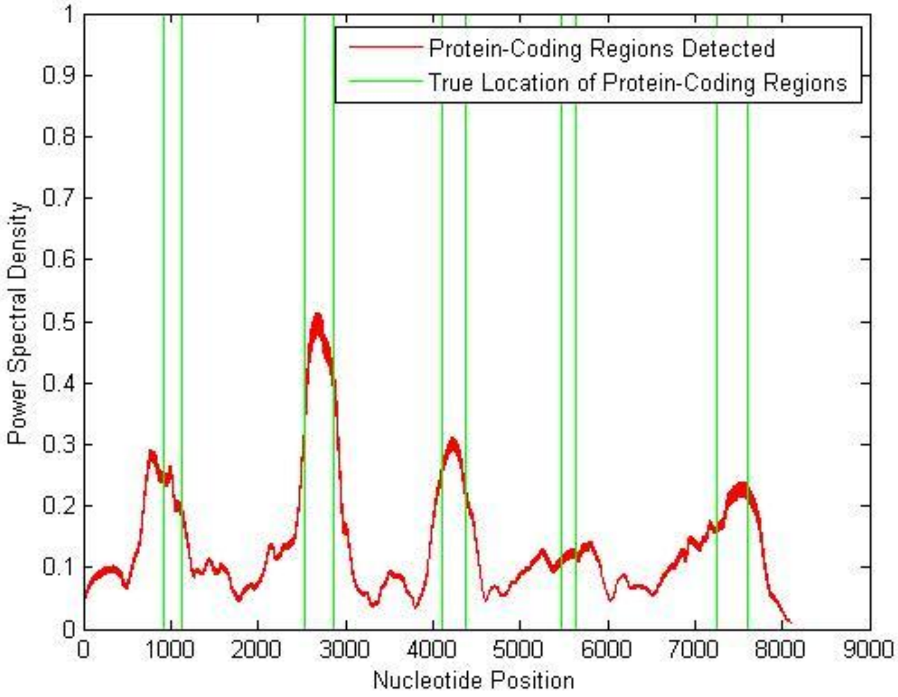


Figure 3.14: Result obtained using molecular mass mapping on example DNA sequence F56F11.4

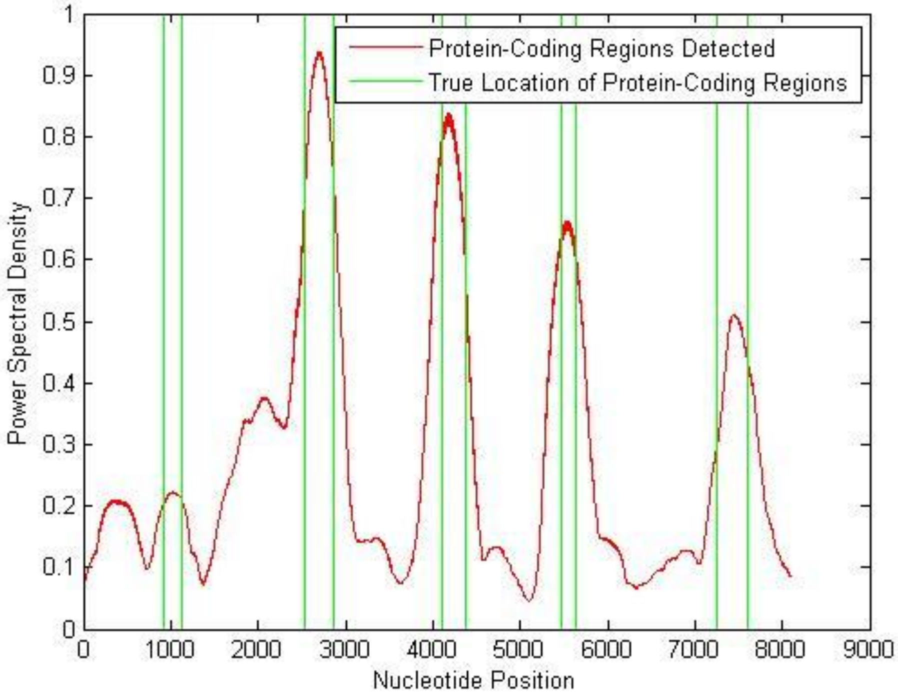


Figure 3.15: Result obtained using quaternary mapping on example DNA sequence F56F11.4

It has been observed from Table 3.4 that none of the numerical representation schemes considered in the experiment is able to achieve the value of AUC as the combination of 24 mappings of integer mapping scheme in the proposed algorithm has achieved. Therefore, the 24 combinations of integer mapping scheme has been selected as numerical representation scheme.

3.8 Results and Discussion

Many methods have been proposed in literature for the identification of protein-coding regions, some methods focus on locating short length protein-coding regions only while other methods emphasis on detecting larger length protein-coding regions only. In this research work, the protein-coding regions of any length varying from shorter to larger are identified. The benchmark datasets considered previously in literature [114], [154-155] are applied to proposed approach and other methods as well for performance comparison. These datasets are HMR195, BG570, and GENSCAN. There are 195 mammalian sequences in HMR dataset which have precisely one complete single-exon or multi-exon genes. In this dataset, human: mouse: rat sequences are in the proportion of 103:82:10. The protein-coding regions in this dataset are 948 and the average length of protein-coding regions is 208 base pairs (bps) [98]. There are 570 vertebrate multi-exon gene sequences in the BG570 dataset. This dataset contains 2649 protein-coding regions and the average length of protein-coding regions is 168 bps. GENSCAN dataset comprises of 65 selected coding sequences, and the average length of exons is 150.

The performance metric considered in the paper for evaluation and comparison purpose is area under the receiver operating characteristics (ROC) curve (AUC) [114]. The following performance parameters which are considered as the standard outcomes of any algorithm are used in the calculation of AUC. True positive (TP) depicts those locations which have been identified aptly by the algorithm where true exons are located, false positive (FP) tells those segments which have been detected erroneously by the algorithm where true exons are actually not located, true negative (TN) represents those sections which are detected precisely where true exons are not located, and those portions which are not captured by the algorithm where true exons are located are termed as false negative (FN). Using these four possible outcomes, the customary performance parameters, sensitivity (Sn), specificity (Sp), true positive rate, and false positive rate are computed. Sn (TP/(TP+FN)) highlights the details related to the proportion of TP which have been detected correctly by the algorithm. Sp (TN/(TN+FP)) gives the statistics related to the proportion of TN predicted appropriately by the algorithm. True positive rate shows the probability of correct detections which is same as Sn, and false positive rate (1-Sp) is computed from Sp. The ROC curve is calculated by plotting the values of false positive rate against true positive rate by varying values of threshold. The characteristic of ROC which is a single number obtained by calculating the area under ROC curve is known as AUC. It is always desired to have the value of AUC as maximum as achievable for a better prediction accuracy; which is governed by ROC curve. If the ROC curve is nearer to 1, the AUC will be higher and the algorithm will be better compared to that which has lesser value of AUC.

The value of AUC obtained for these datasets using proposed algorithm and the other reported methods is summarized in Table 3.5.

Table 3.5: AUC value on benchmark datasets

Dataset	Method	Value of AUC
HMR195	AST-PCA [114]	0.8285
	MGWT [98]	0.8396
	WRWW [113]	0.8317
	AWSTFT [102]	0.7917
	Proposed	**0.8407**
BG570	AST-PCA [114]	**0.8257**
	MGWT [98]	0.8203
	WRWW [113]	0.8137
	AWSTFT [102]	0.7756
	Proposed	0.8237
GENSCAN	AST-PCA [114]	0.8502
	MGWT [98]	0.8486
	WRWW [113]	0.8418
	AWSTFT [102]	0.8158
	Proposed	**0.8539**
Overall (Whole data set)	AST-PCA [114]	0.8348
	MGWT [98]	0.8353
	WRWW [113]	0.8291
	AWSTFT [102]	0.7944
	Proposed	**0.8394**

The superiority of proposed algorithm over other methods has been examined in Table 3.5. It has been observed from Table 3.5 that the proposed algorithm’s performance in terms of AUC over other methods is the highest for datasets HMR195, GENSCAN. For the dataset BG570, the AUC value of AST-PCA method is the highest whereas the proposed algorithm’s AUC value is very closer to this method. And the performance of proposed method is the best over all other methods in terms of the highest value of AUC for combined data set. The performance improvement of proposed algorithm in the value of AUC over other methods has also been computed and depicted in Table 3.6.

Table 3.6: % improvement of proposed algorithm in value of AUC over other methods

Dataset	Method	% improvement in the value of AUC
HMR195	AST-PCA [114]	1.45%
	MGWT [98]	0.13%
	WRWW [113]	1.07%
	AWSTFT [102]	5.83%
BG570	AST-PCA [114]	---
	MGWT [98]	0.41%
	WRWW [113]	1.21%
	AWSTFT [102]	5.84%
GENSCAN	AST-PCA [114]	0.43%
	MGWT [98]	0.62%
	WRWW [113]	1.42%
	AWSTFT [102]	4.46%
Overall (Whole data set)	AST-PCA [114]	0.55%
	MGWT [98]	0.49%
	WRWW [113]	1.23%
	AWSTFT [102]	5.37%

It has been observed from Table 3.6 that the proposed algorithm has achieved significant improvement in the value of AUC over other methods for HMR195, GENSCAN datasets; and the percentage improvement for BG570 dataset over MGWT, WRWW, AWSTFT methods is considerable. Also it has been observed that the percentage improvement of proposed method over all other methods on overall dataset is significantly high.

3.9 Summary

In the recent past, many transform based approaches have been proposed for the identification of protein-coding regions in DNA sequences of eukaryotes. The major limitation of the transform based approaches is that their principle of working is based on the transformation of domain of signal. This can result in loss of important information probably and hence may affect the performance of algorithm. An approach based on SVD called as modified P-spectrum algorithm (MPSA) which does not require any domain transformation is proposed in this research work. The modified P-spectrum which has been reported in literature in some other applications has been tuned in this research work to capture the TBP and identify the protein-coding regions. The window length of proposed algorithm has been varied over a range of 27 to 351 and the optimized window length corresponding to maximum AUC obtained has been selected. The benchmark datasets have been used to verify the applicability and to prove the superiority of proposed MPSA over existing methods in terms of identification of protein-coding regions of any size. The results obtained prove that the proposed algorithm is an effective and efficient approach for the identification of protein-coding regions in the DNA sequences of eukaryotes. The limitation of MPSA is its computational complexity because of optimization of window length for 109 iterations and applying 24 combinations of integer mapping scheme.

CHAPTER 4

SHORT-TIME FOURIER TRANSFORM BASED APPROACH FOR CPG ISLANDS DETECTION IN DNA SEQUENCES

CpG Islands (CGIs) are considered as significant constituent of DNA sequences. Some of the important activities which represent the significance of CGIs can be described as: the identification of CGIs helps in the identification of promoter regions and subsequently genic regions [4], inactivation of X chromosome, some human malignancies, suppression of repetitive elements, and also can be beneficial in case of prediction of cancers at an early stage [5]. Therefore, the detection of CGIs in DNA sequences is considered as very important. As the nature of DNA sequences represent the repeating patterns which points towards that CGIs can have some periodic patterns hidden inside them. The approaches developed and proposed so far for CGIs prediction have not focussed on hidden periodic patterns in the CGIs. In this research work, an approach based on short-time Fourier transform (STFT) has been proposed in which the periodicities present in the CGIs have been analysed through experimental proofs on benchmark data; and subsequently the proposed approach has been applied on a dataset of hundred DNA sequences comprising of human, fish, and mouse species.

4.1 Periodicity Feature in CGIs

It has been reported that CGIs are high frequency repeating patterns of CG dineucleotide [5] in DNA sequences. Hence, small periodicities have been considered as a feature of CGIs in this research work. For the validation of the periodicity feature, first step is to convert the DNA characters T, C, G, A into numerical sequences employing integer mapping scheme [69] and thereafter to compute the short-time Fourier transform (STFT) of all of the seventeen CGIs present in the benchmark DNA sequence having accession number L44140 [1, 130] individually. For the computation of STFT of the DNA data, DFT has been used to obtain the power spectrum of windowed sequence using moving window approach [41]. The calculation of an N-point DFT for a numerical sequence b(i) at each nucleotide position ‘i’ is performed as [41]:

$$B(k) = \sum_{i=0}^{N-1} b(i)w(i)e^{\frac{-j2\pi ik}{N}} \tag{4.1}$$

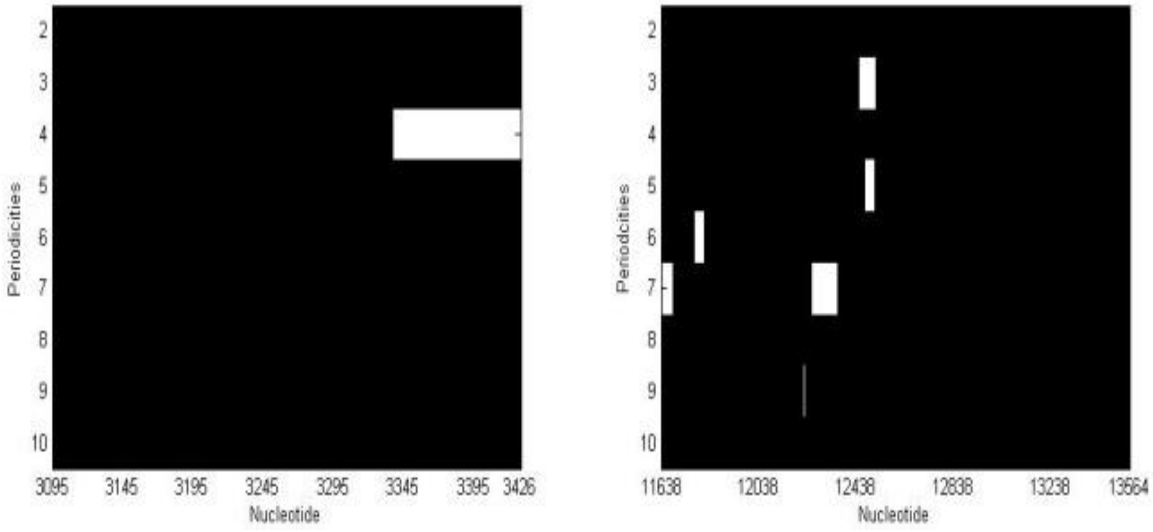
where, $w(i) = (1/\sigma\sqrt{2\pi})\exp(-i^2/2\sigma^2)$, i corresponds to Gaussian window's length, $\sigma = i/2\alpha$, α represents shaping parameter of window. In this research work, the value of parameter 'i' =210, α = 2.5, length of FFT (N) = 2520, and k = 0… N-1 have been chosen. The windowed sequence's power spectrum calculated using equation (4.1) is as follows:

$$P_1(k) = |B(k)|^2 \tag{4.2}$$

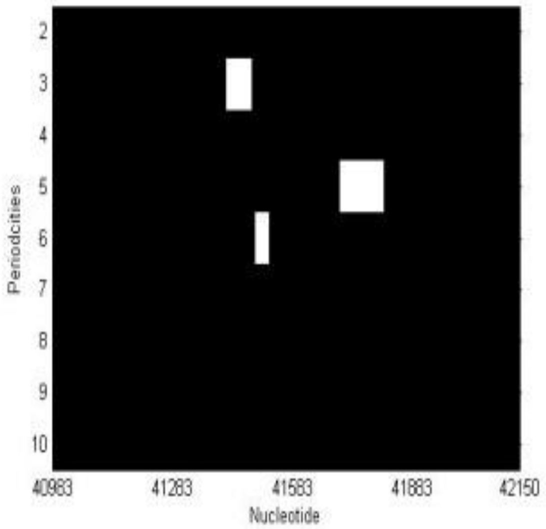
The equation (4.3) has been applied for the computation of power spectrum with respect to the periodicities ($p = 2\ to\ 10$) from windowed power spectrum $P_1(k)$ at every position of nucleotide:

$$P(i, p) = P_1(i, N/p) \tag{4.3}$$

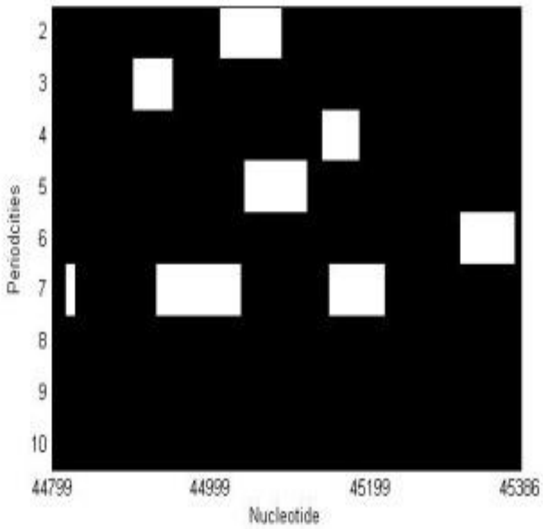
where, i corresponds to the position of nucleotide where center point of window is located; it varies from $i = 0 \ldots\ldots L$, where L shows the entire length of DNA sequence. The plots of nucleotide-position versus periodicities for seventeen existing CGI sections of DNA sequence L44140 are represented in Figure 4.1 (i-xvii):



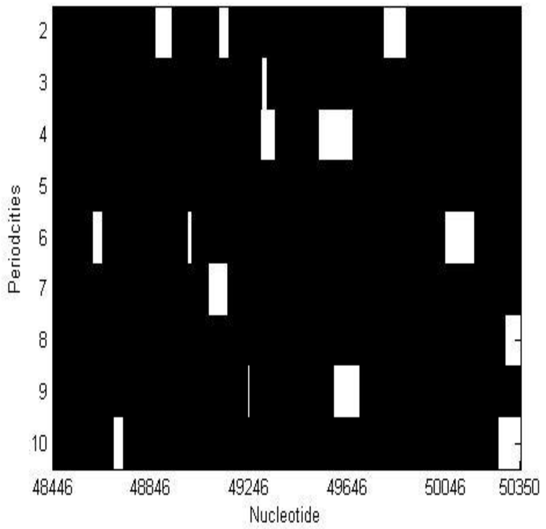
(i) **(ii)**


Periodcities
2
3
4
5
6
7
8
9
10
40983
41283
41583
41883
42150
Nucleotide

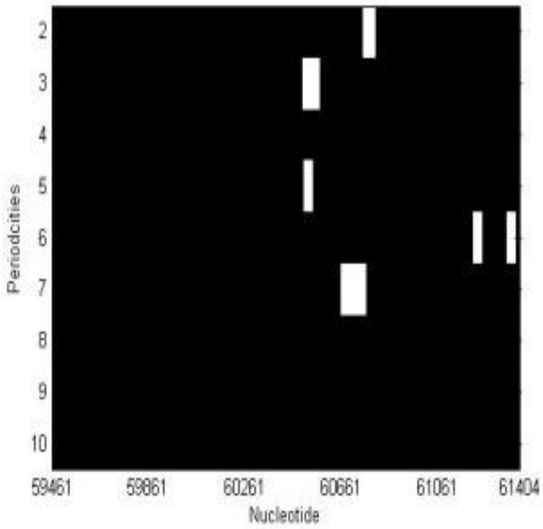
(iii)


Periodcities
2
3
4
5
6
7
8
9
10
44799
44999
45199
45386
Nucleotide

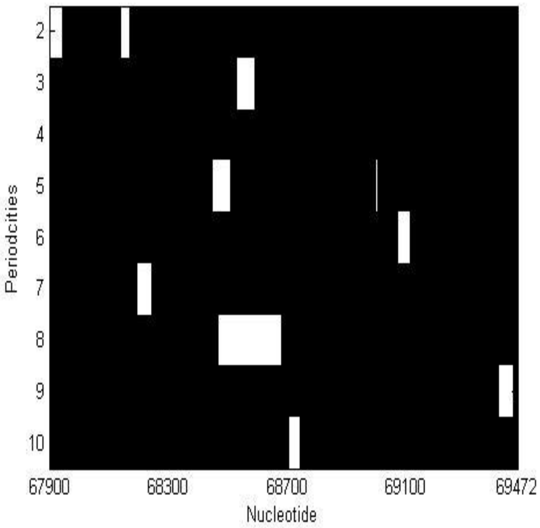
(iv)


Periodcities
2
3
4
5
6
7
8
9
10
48446
48846
49246
49646
50046
50350
Nucleotide

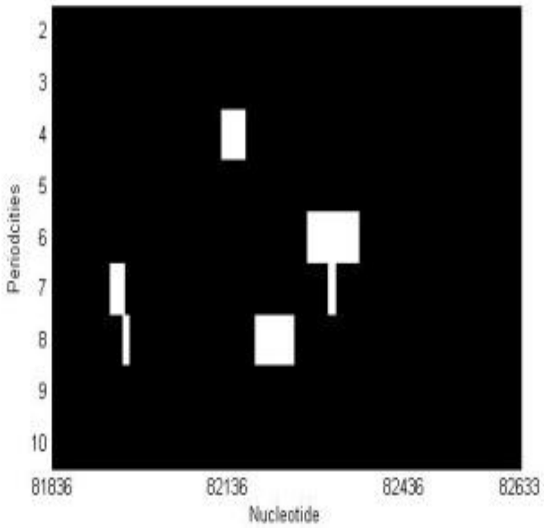
(v)


Periodcities
2
3
4
5
6
7
8
9
10
59461
59861
60261
60661
61061
61404
Nucleotide

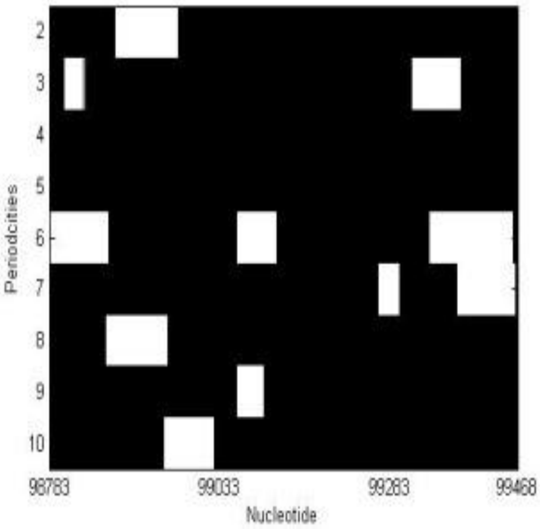
(vi)


Periodcities
2
3
4
5
6
7
8
9
10
67900
68300
68700
69100
69472
Nucleotide

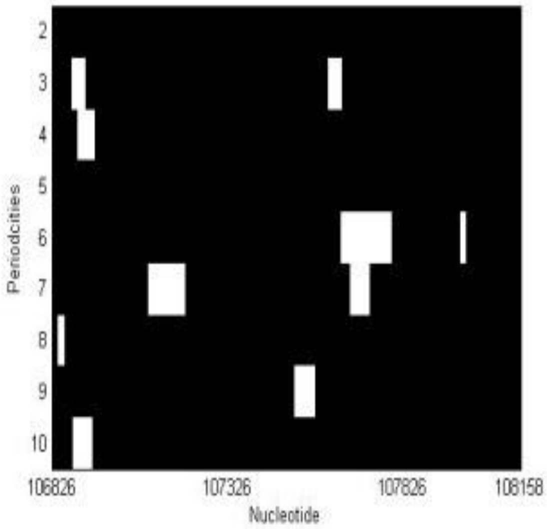
(vii)


Periodcities
2
3
4
5
6
7
8
9
10
81836
82136
82436
82633
Nucleotide

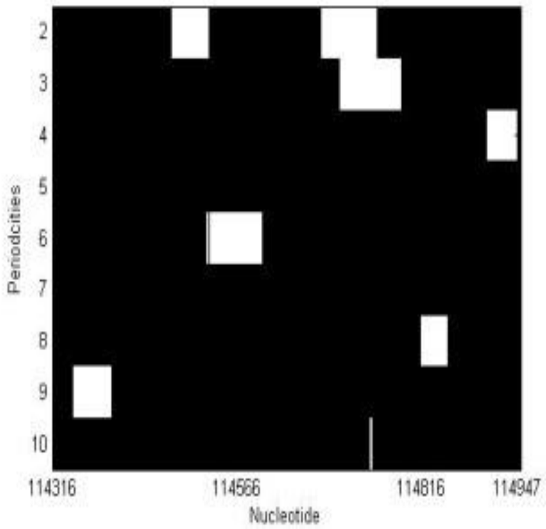
(viii)


Periodcities
2
3
4
5
6
7
8
9
10
98783
99033
99283
99468
Nucleotide

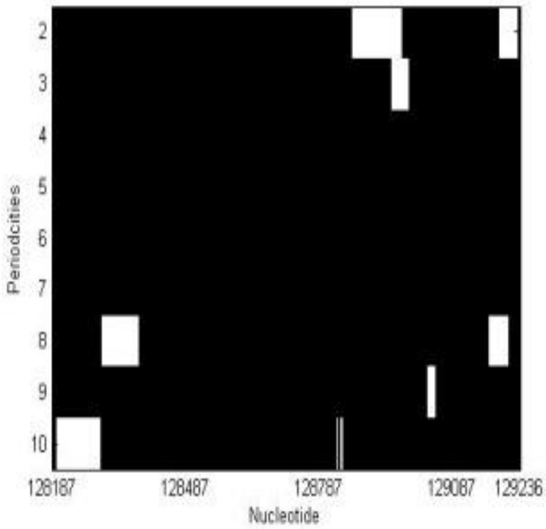
(ix)


Periodcities
2
3
4
5
6
7
8
9
10
106826
107326
107826
108158
Nucleotide

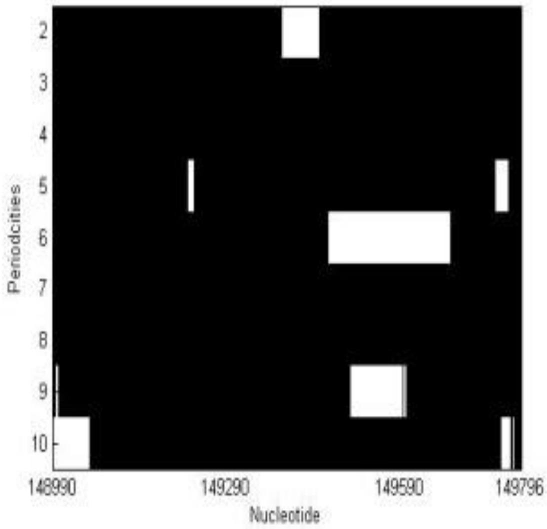
(x)



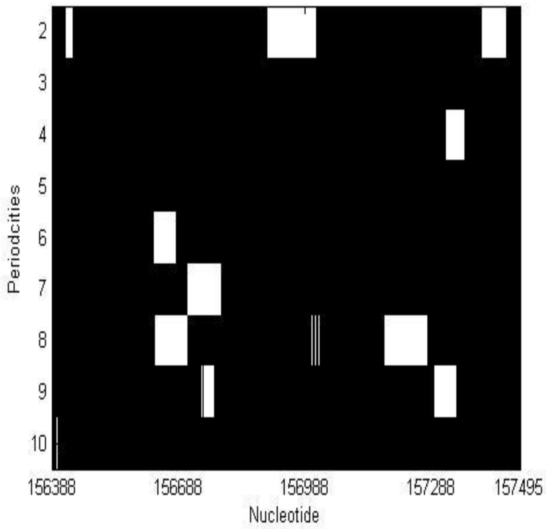
(xi)



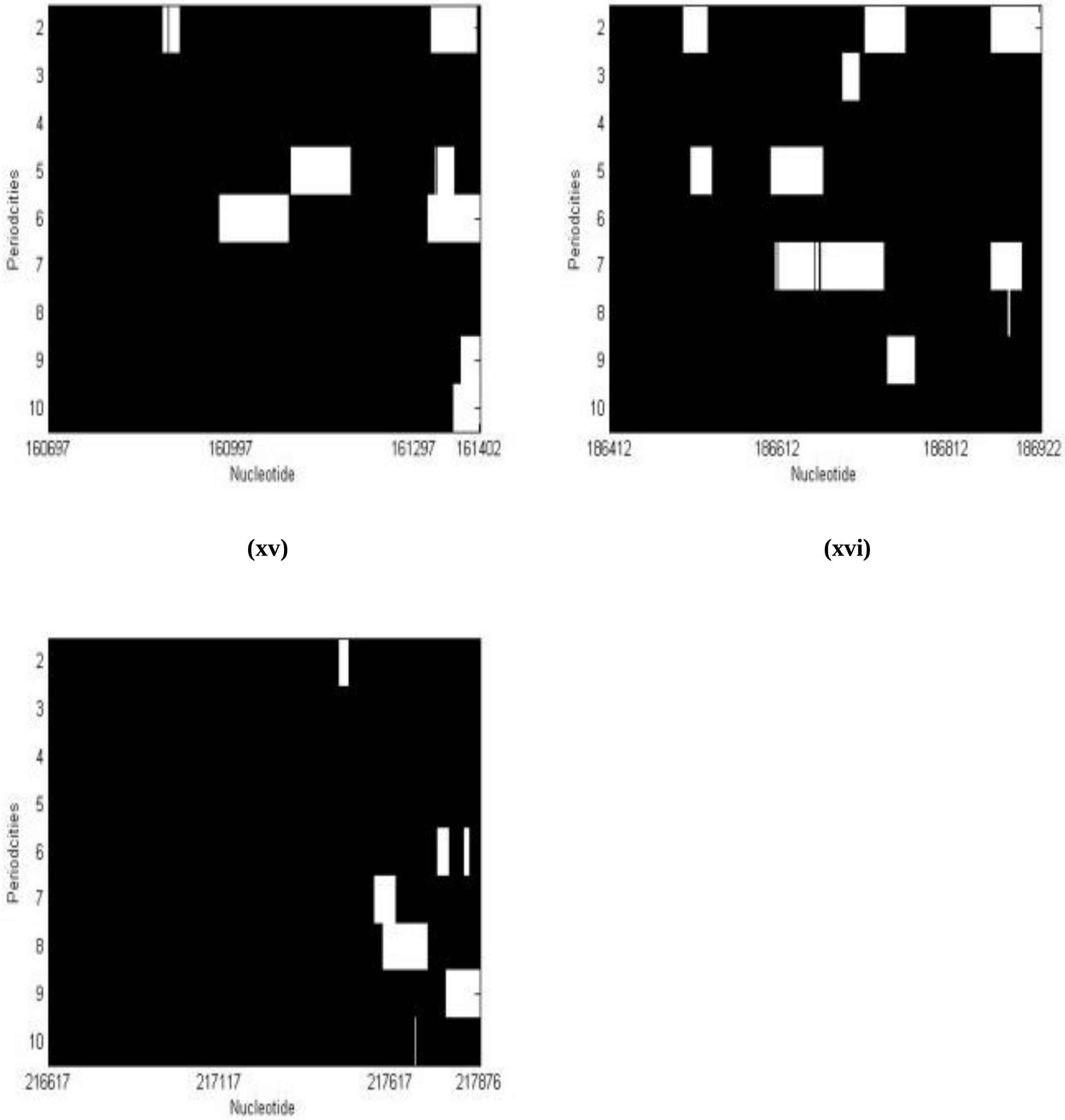
(xii)



(xiii)



(xiv)



(xv)

(xvi)

(xvii)

Figure 4.1 (i-xvii): Plot of nucleotide's position vs periodicity of seventeen CGIs of DNA sequence L44140

The criterion applied for extraction of dominant periodicities from the plots obtained in Figure 4.1(i-xvii) is as follows:

- The period whose minimum section's length is twice of the respective period has been considered as periodicity present.

- In those sections of obtained periodicities which are having overlapping with other periods, minimum period out of the overlapping periods has been selected as dominating period. For example if in some particular section, period 3, period 6, and period 9 are obtained which are overlapping then period 3 must be chosen in that particular section.

Thereafter, the verification step in which 2 necessary conditions of GGF criterion {which are i) CG % has to be at least 50%, ii) the observed/expected ratio should be above 0.6} have been applied on the predicted sections of dominant periodicities the segments of the detected dominant periodicities. Those sections of predicted periodicities which satisfy the above mentioned 2 conditions of GGF criterion required for classification of CGI have been finally selected as verified dominating periodicities; otherwise rejected.

The dominating periodicities in CGI sections of DNA sequence L44140 which are predicted and finally verified also are represented in Table 4.1:

Table 4.1: Obtained periodicities in seventeen CGI segments of DNA sequence L44140

S. No.	Start and end position of CGI in accordance with NCBI website	CGI segment's length (bps)	Periodicities acquired by proposed algorithm in CGI segments	Periodicities after verification step present in CGI segments
CGI 1	3095-3426	332	4	___
CGI 2	11638-13564	1927	3, 6, 7	3, 6
CGI 3	40983-42150	1168	3, 5, 6	3, 5, 6
CGI 4	44799-45386	588	2, 3, 4, 5, 6, 7	2, 3, 4, 5, 7
CGI 5	48446-50350	1905	2, 3, 4, 6, 8, 10	2, 3, 4, 6, 8, 10
CGI 6	59461-61404	1944	2, 3, 6, 7	3, 6, 7
CGI 7	67900-69472	1573	2, 3, 5, 6, 7, 9, 10	2
CGI 8	81836-82633	798	4, 6, 7, 8	4, 6
CGI 9	98783-99468	686	2, 3, 6, 7, 10	2, 3, 6, 7, 10
CGI 10	106826-108158	1333	3, 4, 6, 7, 8, 9	3, 6, 9

CGI 11	114316-114957	642	2, 3, 4, 6, 8, 9	2, 3, 4, 6, 8
CGI 12	128187-129236	1050	2, 3, 8, 9, 10	2, 3, 8
CGI 13	148990-149796	807	2, 5, 6, 10	2, 6, 10
CGI 14	156388-157495	1108	2, 4, 6, 7, 8	2, 6, 7, 8
CGI 15	160697-161402	706	2, 5, 6	2, 5, 6
CGI 16	186412-186922	511	2, 3, 5	2
CGI 17	216617-217876	1260	2, 6, 7	2, 6

The fact observed from experimental analysis results carried out on benchmark DNA sequence which are tabulated in Table 4.1 is that CGIs possess 2 to 10 periodicities. Therefore, with the help of these verified dominating periodicities, the proposed algorithm for the detection of CGIs is now discussed in the following sections.

4.2 Proposed Algorithm for Detection of CGIs

The flow graph of the proposed algorithm which is based on capturing the dominating periodicities present in CGIs is depicted in Figure 4.2.

The DNA sequence with accession number L44140 which belongs to Homo sapiens chromosome X region from filamin gene to glucose-6-phosphate dehydrogenase gene which is a benchmark DNA sequence has been chosen here as an example sequence for the discussion of the steps of the proposed approach. This DNA sequence consists of 219447 bps and there exists seventeen CGIs in this sequence [1]. The detailed discussion of the steps employed in the proposed approach is as follows:

4.2.1 Conversion of DNA Characters to Numerical Values

The important step in the application of DSP based methods to be applied for the analysis of DNA data is the mapping of characters of DNA to the numerical values with the help of numerical mapping scheme. For an instance, a DNA string CGATCGCGTTAA can be converted to 231423234411 using integer mapping [69].

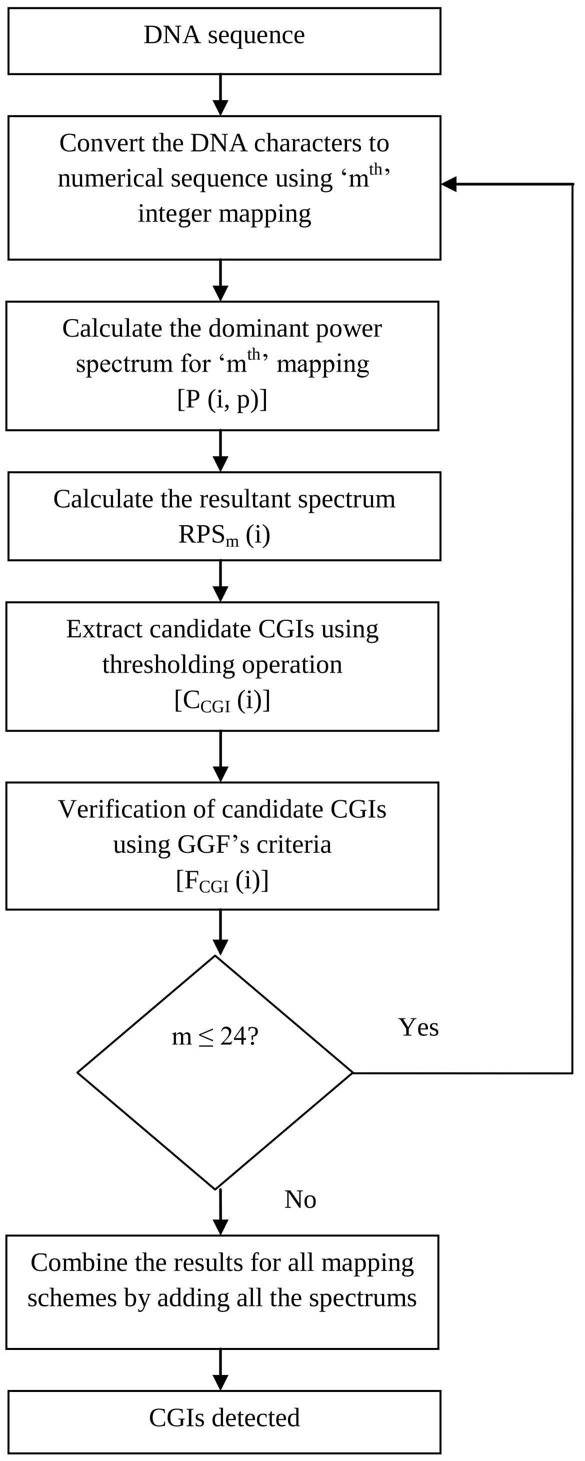

DNA sequence
Convert the DNA characters to numerical sequence using 'mth' integer mapping
Calculate the dominant power spectrum for 'mth' mapping [P (i, p)]
Calculate the resultant spectrum RPS_m (i)
Extract candidate CGIs using thresholding operation [C_{CGI} (i)]
Verification of candidate CGIs using GGF's criteria [F_{CGI} (i)]
m ≤ 24?
Yes
No
Combine the results for all mapping schemes by adding all the spectrums
CGIs detected

Figure 4.2: Flow graph of the proposed algorithm

4.2.2 Calculation of Resultant Power Spectrum

The power spectrum components with respect to every dominating value of periodicity i.e. periodicity 2 to 10 have been obtained with the application of short-time DFT in equation (4.3). The obtained value of power spectrums with respect to dominating periodicities at every nucleotide location have been added linearly then and the resultant power spectrum with respect to a mapping scheme ‘m’ has been computed as represented in equation (4.4):

$$RPS_m(i) = \sum_{p=2}^{10} P(i,p) \tag{4.4}$$

The result obtained for resultant value of power spectrum $RPS_m(i)$ on example DNA sequence L44140 has been depicted in Figure 4.3:

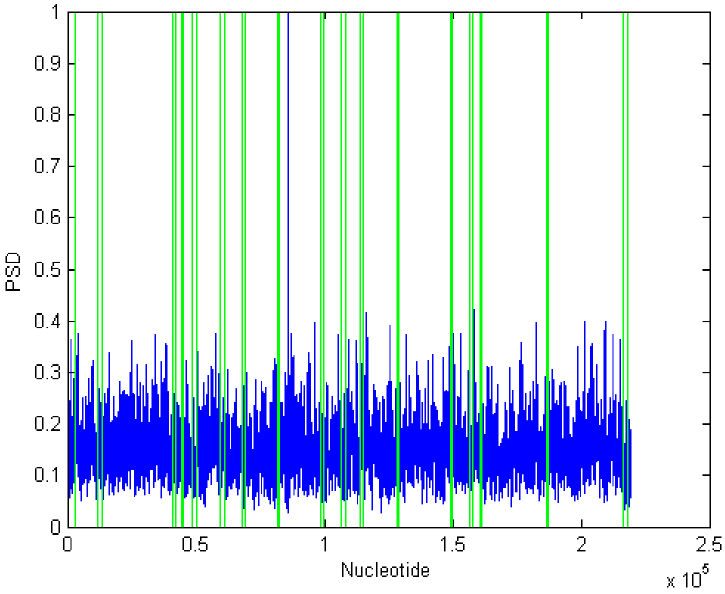


Figure 4.3: Resultant power spectrum

4.2.3 Identification of Candidate CGIs

A threshold value selected empirically as 10% of the maximum value of the resultant power spectrum $RPS_m(i)$ has been employed for the extraction of candidate CGIs from resultant power spectrum. Those segments of power spectrum whose peak value crosses the selected threshold limit have been classified as candidate CGIs.

$$C_{CGI}(i) = \begin{cases} RPS_m(i) \; if \; RPS_m(i) > Thr \\ 0, \qquad\qquad else \end{cases} \tag{4.5}$$

where, $Thr = 0.1 * \max(RPS_m(i))$

The candidate CGI spectrum $\mathrm{C_{CGI}}$(i) obtained is shown in Figure 4.4.

4.2.4 Verification of Candidate CGIs

The GGF criterion has been applied as a post processing step on the detected segments corresponding to candidate CGI to verify and classify finally such segments as detected CGIs as per equation (4.6):

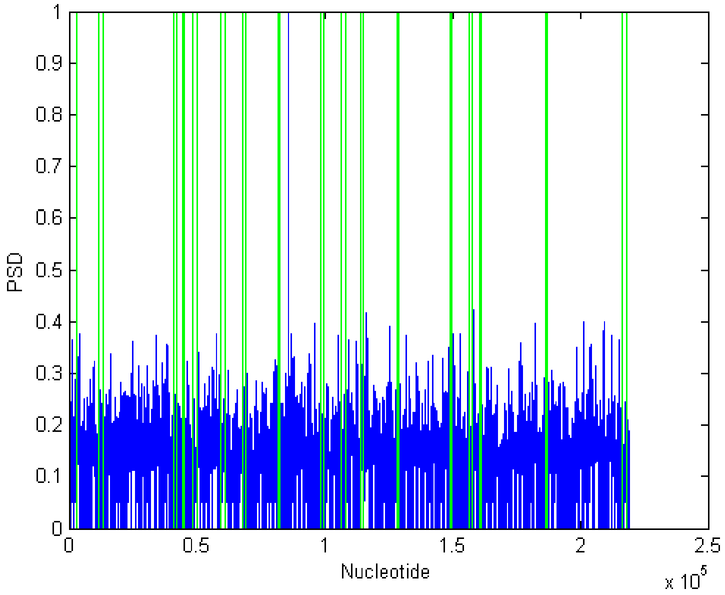


Figure 4.4: Obtained spectrum of candidate CGIs

$$F_{CGI}(i) = \begin{cases} C_{CGI}(i), & \text{segments out of } C_{CGI}(i) \text{ which meet GGF Criteria} \\ 0\,, & \text{else} \end{cases} \qquad (4.6)$$

The obtained spectrum of predicted candidate CGIs after verification step $F_{CGI}(i)$ is highlighted in Figure 4.5.

4.2.5 Combine the Mapping Results

To analyze the effect of numerical mappings on the performance of proposed algorithm, experiment has been performed with the help of 12 mapping schemes and using 24 combinations of integer mapping scheme which has been used in this research work. The results obtained in terms of standard performance metrics using all these mappings are shown in Table 4.2.

As it has been noticed from Table 4.2 that the value of performance metrics, sensitivity (Sn), and accuracy (AC) of the proposed approach with 24 combinations of integer mapping scheme is much better in comparison with other mapping schemes considered here; therefore, the final spectrum with respect to CGIs has been computed by combining the verified spectrums of 24 mapping schemes in accordance with equation (4.7).

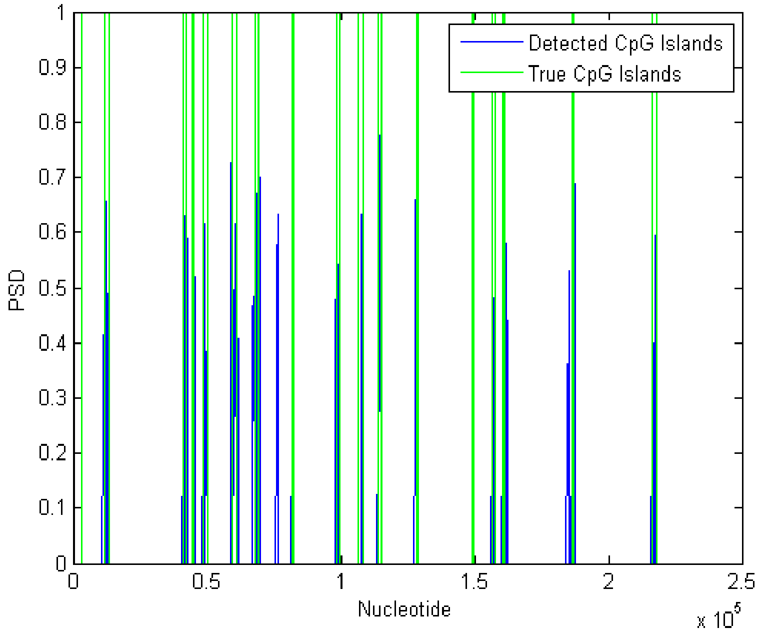


Figure 4.5: Obtained power spectrum $F_{CGI}(i)$ after verification of candidate CGIs

Table 4.2: Performance metrics obtained in DNA sequence L44140 using proposed approach employing various mappings

Mapping Scheme	**Performance Measure**		
	Sp	**Sn**	**AC**
Complex	1	0.0295	0.5148
Atomic	0.9767	0.0440	0.5104
EIIP	0.9538	0.4131	0.6834
Four-bit-binary	0.9888	0.0699	0.5293
Integer	0.9782	0.4758	0.7270
Three-bit-binary	0.9942	0.0154	0.5048
Real Number	0.9822	0.0336	0.5079
Two-bit-binary	0.9618	0.5202	0.7410
Modified EIIP	0.9492	0.5991	0.7742
Pseudo EIIP	0.9656	0.5464	0.7560
Quaternary	0.9689	0.4152	0.6920
Molecular Mass	0.9826	0.0440	0.5133
Adding 24 combinations of mappings of integer mapping	0.8285	**0.9590**	**0.8938**

$$\mathrm{FS_{CGI}}\,(\mathrm{i}) = \sum_{\mathrm{m}=1}^{24} \mathrm{F_{CGI}}\,(\mathrm{i}) \quad , \mathrm{m} \in \; [1, 24] \tag{4.7}$$

The result of final spectrum $\mathrm{FS_{CGI}}$ (i) for CGIs which has been obtained using proposed approach is depicted in Figure 4.6. The x-axis represents the nucleotides' position and the y-axis corresponds to power spectrum value with respect to nucleotides' position in Figure 4.6.

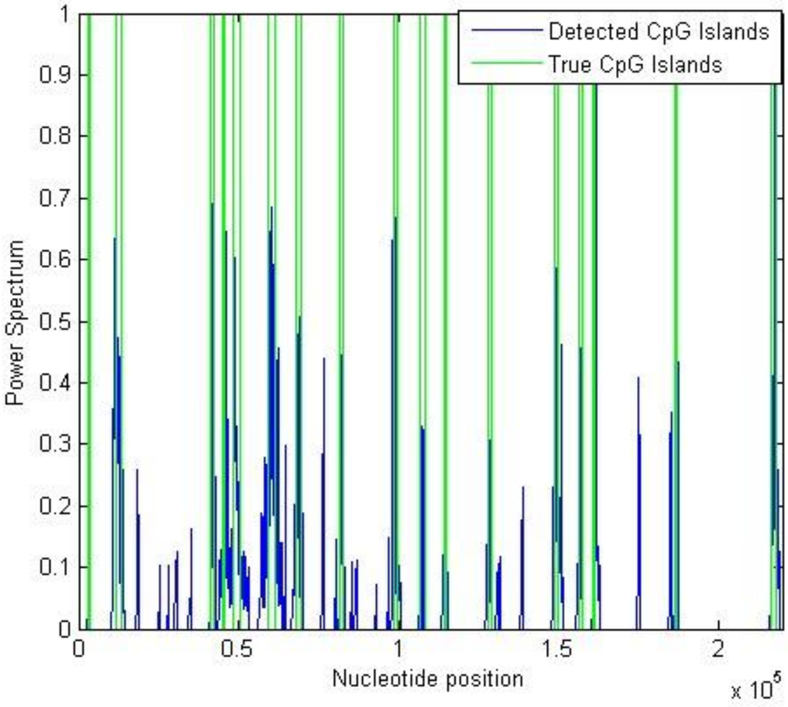


Figure 4.6: $\mathrm{FS_{CGI}}\,(\mathrm{i})$ of detected CGIs

As the length of DNA sequence L44140 is 219447 bps, it appears bit difficult to visualize the locations of the detected segments very precisely. Hence, the Figures 4.7- 4.11 are represented as a magnified view of the Figure 4.6 which is shown in smaller segments to have a better visualization of the result obtained.

Now, to get further better understanding of the locations of detected CGIs segments by proposed method, these locations are checked and tabulated in Table 4.3.

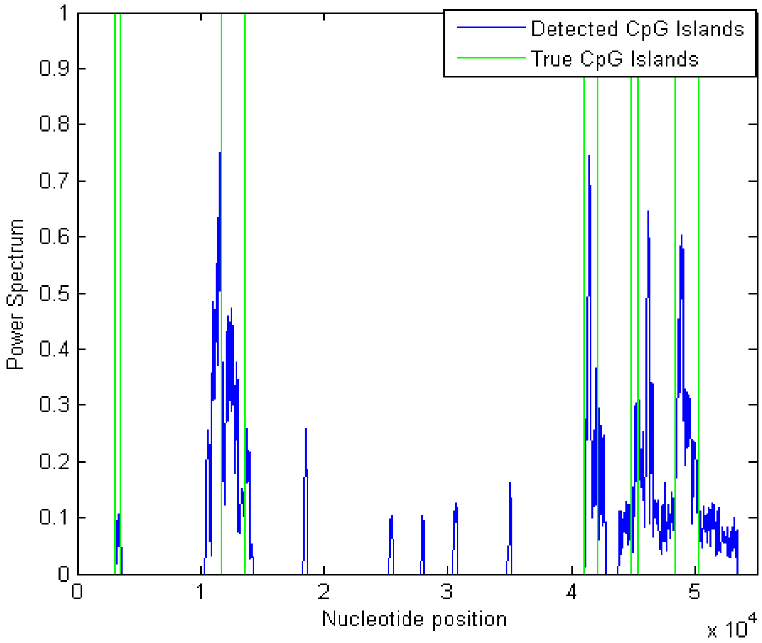


Figure 4.7: $FS_{CGI}(i)$ of detected CGIs for segment 1-55000 bps

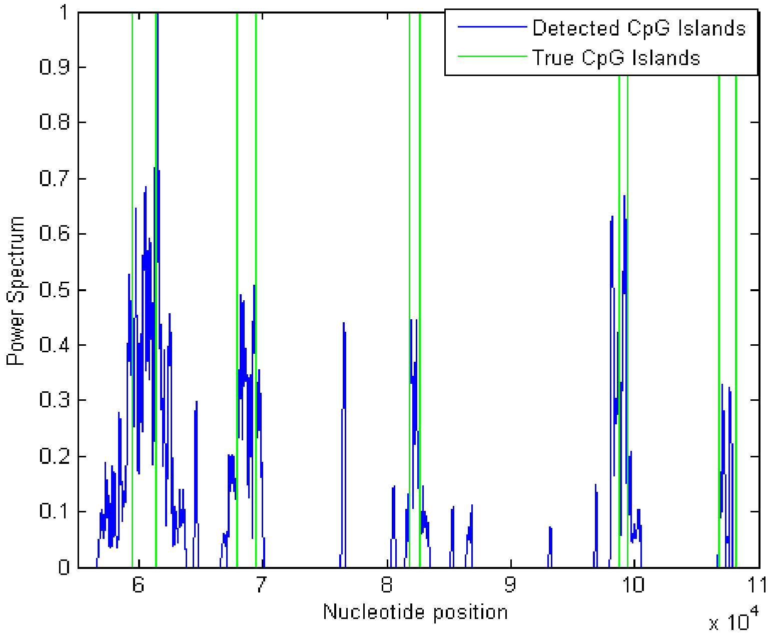


Figure 4.8: $FS_{CGI}(i)$ of detected CGIs for segment 55001-110000 bps

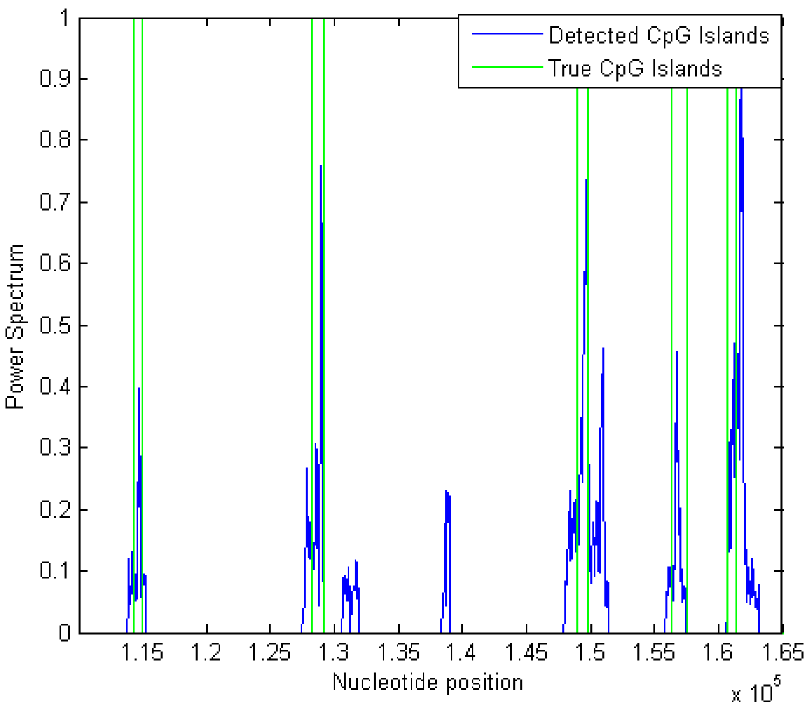


Figure 4.9: $\mathrm{FS}_{\mathrm{CGI}}(\mathrm{i})$ of detected CGIs for segment 110001-165000 bps

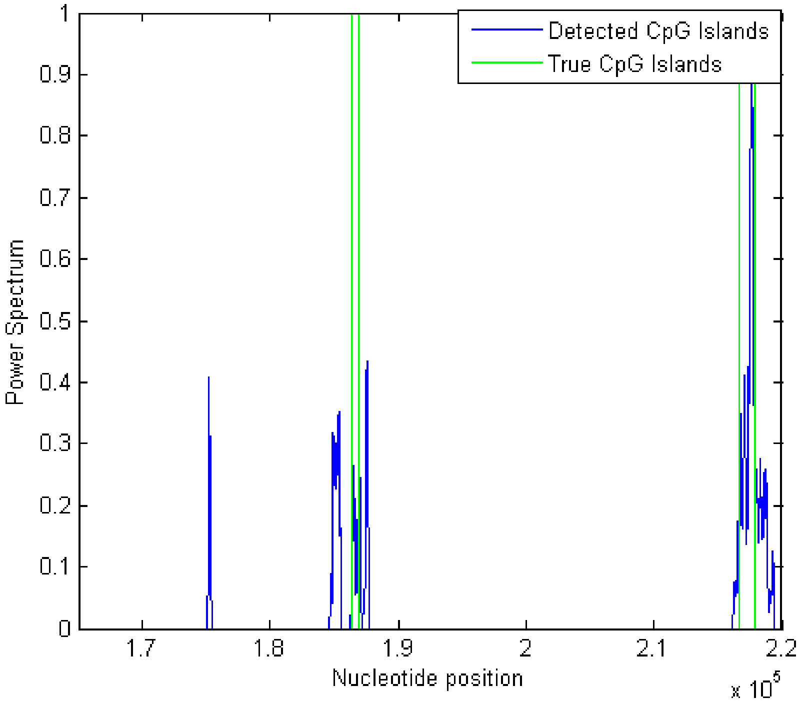


Figure 4.10: $\mathrm{FS}_{\mathrm{CGI}}(\mathrm{i})$ of detected CGIs for segment 165001-220000 bps

Table 4.3: Detected CpG Islands

DNA Sequence		Start and end location of CGI in accordance with NCBI website		Start and end location of CGI obtained using proposed approach	
L44140		Start location	End location	Start location	End location
	1	3095	3426	3192	3576
	2	11638	13564	10470	14217
				18353	18656
				25277	25597
				27863	28072
				30464	30766
				34931	35166
	3	40983	42150	41089	42737
	4	44799	45386	43840	53495
	5	48446	50350	43840	53495
	6	59461	61404	56715	63740
				64457	64720
				66726	67012
	7	67900	69472	67102	70028
				76336	76687
				80444	80658
	8	81836	82633	81493	83393
				85176	85394
				86475	86879
				93080	93286
				96768	96993
	9	98783	99468	98000	100530
	10	106826	108158	106816 107345	107300 107583

				107587	107843
	11	114316	114947	113832	115318
	12	128187	129236	127582	129155
				130652	131218
				131394	131879
				138508	139016
	13	148990	149796	147981	151460
	14	156388	157495	155887	157400
	15	160697	161402	160653	163220
				175115	175407
				184658	185511
	16	186412	186922	186327	187110
				187304	187786
	17	216617	217876	216200	219447

It has been noticed from Table 4.3 that proposed approach is able to capture all the seventeen CGIs which are contained in benchmark DNA sequence L44140. However, the proposed approach has identified some false locations of CGIs also.

If the length of a particular CGI is 200 bps then 10% of it is 20 bps, 20% comes out to be 40 bps, similarly 90% of the length of this CGI will be 180 bps, and 100% value is 200 bps. Now the performance of proposed approach has been examined on the basis of the percentage coverage of the true CGI’s length; and the performance comparison has been carried out with other recent state-of-art algorithms. Table 4.4 shows that out of seventeen CGIs present in DNA sequence L44140, which method has identified/not identified a particular CGI at 80 percent, 90 percent, and 100 percent (full length) coverage of true CGI’s length. The summary of CGIs identified by the various methods in accordance with coverage of portion of length of true CGI at 80 percent, 90 percent, and 100 percent (full length) is tabulated in Table 4.5.

Table 4.4: CGI identified/not identified by methods for seventeen CGIs at various percentage coverages

CGI and its location	**CGI identified/not-identified corresponding to percentage coverage of true CGI's length METHODS**											
	80percent				**90percent**				**100percent**			
	Pr.	**TLBO**	**PNP**	**DWT**	**Pr.**	**TLBO**	**PNP**	**DWT**	**Pr.**	**TLBO**	**PNP**	**DWT**
CGI-1 (3095-3426)	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓	✗
CGI-2 (11638-13564)	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
CGI-3 (40983-42150)	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
CGI-4 (44799-45386)	✓	✓	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
CGI-5 (48446-50350)	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
CGI-6 (59461-61404)	✓	✗	✓	✗	✓	✗	✗	✗	✓	✗	✗	✗
CGI-7 (67900-69472)	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✗	✗
CGI-8 (81836-82633)	✓	✓	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
CGI-9 (98783-	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗

99468)												
CGI-10 (106826-108158)	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
CGI-11 (114316-114947)	✓	✓	✗	✗	✓	✓	✗	✗	✓	✗	✗	✗
CGI-12 (128187-129236)	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
CGI-13 (148990-149796)	✓	✓	✗	✗	✓	✓	✗	✗	✓	✗	✗	✗
CGI-14 (156388-157495)	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
CGI-15 (160697-161402)	✓	✓	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
CGI-16 (186412-186922)	✓	✓	✗	✗	✓	✓	✗	✗	✓	✗	✗	✗
CGI-17 (216617-217876)	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗

In Table 4.4, the heading marked as Pr. represents the 'Proposed approach', TLBO corresponds to 'CpGclusterTLBO' [129], PNP corresponds to 'CpGPNP' [122], and DWT corresponds to DWT based CGI detection algorithm [130]. The symbols ✓ and ✗ in Table 4.5 signifies 'identified' and 'not identified' a particular CGI respectively.

Table 4.5: Total CGIs identified out of 17 in DNA sequence L44140

Methods	**Number of CGIs identified corresponding to percentage coverage of true CGI's length**		
	80percent	**90percent**	**100percent**
Proposed	**15**	**15**	**12**
CpGclusterTLBO	9	5	0
CpGPNP	4	3	2
DWT based method	0	0	0

It has been noticed from Table 4.5 that the performance of proposed approach in context of identification of CGIs at varying % coverage from 80% to 100% (full length of a CGI) of the length of actual CGIs is the highest compared to other recent state-of-art methods. The proposed approach has identified 15 CGIs out of total 17 CGIs present in DNA sequence L44140 at 80% & 90% coverage, and 12 CGIs at 100% (full length of CGI) coverage of actual CGI length; whereas no other recent method is able to detect these number of CGIs.

Having proved with experimental analysis the performance of proposed approach on a benchmark DNA sequence, the experiments have been performed on a big data set comprising of hundred DNA sequences. The explanation of data set, performance parameters used, and the obtained results are discussed now in following sections.

4.3 Data Set of CGIs and Performance Parameters

4.3.1 Data Set of CGIs

A CGI data set of hundred DNA sequences has been prepared by us by acquiring the details from National Centre for Biotechnology Information (NCBI) website [1]. For the testing of universal applicability of proposed approach, the data set has been prepared comprising of Human (Homo sapiens), fish, and mouse species. The complete details of the data set such as Gene bank accession number, number of base pairs (bps) in a sequence, number of CGIs, and start and end locations of CGI within a sequence are described in Table 4.6:

Table 4.6: Details of hundred DNA sequence's CGI data set

S. No.	Gene bank accession number	Number of base pairs (bps)	Number of CGIs	Start and end location of CGI as acquired from NCBI website
Data set of 85 DNA sequences of human species				
1	AL442638	188247	4	17472-17700, 22868-23148, 93250-93495, 163847-164132
2	AC073335	68275	3	31813-32080, 33619-34458, 50802-51655
3	AC073517	67706	1	35431-35977
4	AC127379	67291	2	30060-30318, 38447-39437
5	AC064843	66898	1	5531-5785
6	AC129782	66860	1	38868-40898
7	AC013270	66660	4	6075-6881, 25374-26035, 34710-36183, 48185-48621
8	AC074386	66610	2	15847-16381, 16593-16830
9	AC092103	66565	1	24844-25119
10	AC124014	66552	1	56936-57769
11	AL137791	66254	4	30724-31272, 46196-46906, 52979-53956, 61007-62096
12	AC096553	66229	1	11867-12256
13	AC105413	65958	1	50478-50751
14	AC005003	65750	1	38374-41067
15	AC145546	65625	1	52797-53645
16	AC105402	65449	2	15774-16973, 28628-28925
17	AC112698	65335	1	42309-43546
18	AC104129	65189	8	2966-3334, 8763-9020, 14023-14383, 20695-20991, 26472-26735, 28330-29188, 31762-32009, 55671-55878

19	BN000001	64961	1	895-1123
20	AC138782	64744	1	23500-24633
21	AC005021	64607	2	24663-25225, 63177-63512
22	AC093086	64601	1	58914-59518
23	AC005233	64359	1	16579-18003
24	AC013436	63823	5	12411-12652, 21066-21331, 24980-26051, 26467-26807, 60097-60448
25	AC131957	63780	1	45526-45799
26	AC004694	63749	2	9107-9494, 54481-54756
27	AC108463	63525	3	26008-26366, 26575-26982, 27079-27538
28	AC080165	63279	1	8258-8531
29	AC010890	62764	4	11407-11926, 13574-13801, 53142-53415, 53755-54041
30	AC108142	62624	1	8864-11837
31	AC080068	62623	1	535-774
32	AC093785	62466	1	31397-31665
33	AC003079	62331	1	50250-50471
34	AC078937	62035	1	38149-39359
35	AC114803	61579	7	3256-4009, 18815-19353, 32398-32647, 33247-33659, 36773-37302, 39696-39964, 55808-56144
36	AC093652	61340	1	48156-49072
37	AC093377	61056	1	729-1003
38	AC073201	60776	1	9738-11862
39	AC113611	60597	1	8638-9514
40	AC099394	60024	7	2826-4863, 10806-11866, 19723-19934, 25482-25769, 31861-32884, 36728-36931, 54994-55361
41	AC098831	59776	2	39343-39572, 51406-51689
42	AC074013	59657	3	22602-22873, 51602-52508, 53105-53331
43	AC062028	59634	1	44629-44851

44	AC106875	59580	1	4526-5382
45	AC023670	59565	1	25568-27400
46	AC079882	59427	1	39153-39736
47	AC006008	57554	1	28800-30423
48	AC108222	21776	1	21237-21776
49	AH006464	21230	1	1187-2051
50	AC093609	20710	1	7857-8257
51	AL590794	18042	1	11568-12215
52	AC136375	17863	1	16369-17534
53	BD432859	14646	2	2762-2973, 4065-5181
54	AC111201	13470	3	4327-4727, 5323-5554, 12500-13455
55	NM005876	10782	1	6154-7734
56	NM053043	10168	1	9597-9820
57	AC093460	10103	1	6951-7418
58	AC108032	9716	1	30-269
59	X86012	9541	1	335-3853
60	AC106048	8594	1	7941-8180
61	AH008870	6797	1	341-1340
62	AC079401	6568	1	3086-3935
63	AH007568	6513	3	543-803, 1212-1430, 1662-2474
64	AC105385	5952	1	2844-3080
65	AJ308559	5596	1	1228-1657
66	M92844	3889	1	3198-3889
67	AF196313	3700	1	2092-3580
68	AF281043	3662	1	1611-2734
69	U48937	3278	1	2588-3230
70	AF307776	3113	2	2334-2745, 2791-3064

71	AJ000757	3046	1	650-2840
72	AJ289875	2916	1	2325-2916
73	L07287	2704	1	1-1350
74	Z92546	73511	1	20746-21240
75	AL591222	147211	2	54605-55080, 68825-69091
76	AL513502	174636	1	116364-117432
77	AL513498	155780	1	18305-18582
78	AL357615	171446	2	56753-57030, 59607-59874
79	AL353786	139565	1	19000-19400
80	AL121926	139544	2	102641-104201, 126562-127299
81	AL049547	129811	5	27801-29311, 37094-37773, 109041-110125, 113196-114024, 126815-127265
82	AL031706	13012	1	7-552
83	AL031703	35098	4	15319-17699, 25107-26048, 30327-30736, 31615-32204
84	AJ006998	123521	1	11140-11417
85	AL031707	28707	4	6050-6520, 6693-7445, 24481-25248, 28059-28669
Data set of 9 DNA sequences of Mouse species				
86	AJ970309	7050	1	3025-4010
87	AC149868	190971	4	38226-39751, 109499-110391, 114105-114977, 167115-168150
88	AC125063	194931	4	97498-98367, 99058-100402, 106255-107246, 144134-145047
89	AC124505	222439	4	36111-37119, 132685-133458, 139610-140565, 202532-203418
90	AC145199	220892	6	29996-30867, 59938-60771, 114341-115758, 133121-133903, 204198-205934, 217247-218028
91	AC122821	220013	6	43295-44322, 59514-60693, 122943-123697, 163194-164078, 185979-186978, 218075-218923
92	AF073797	46872	4	9395-9666, 18386-18651, 32350-32477, 33946-

				34206
93	AC126029	212472	5	5851-6810, 75564-76663, 82722-84043, 152561-153650, 195134-196503
94	AF059580	36326	3	2076-3209, 2382-3017, 14983-15869
Data set of 6 DNA sequences of Fish species				
95	AL603785	89874	1	4151-4634
96	AL672065	82767	1	44999-45681
97	AL672083	111516	1	88040-88588
98	AL691521	109831	1	34191-36572
99	AL672171	114103	1	50521-51167
100	AL713869	104577	1	6954-7435

4.3.2 Performance Parameters

For the assessment and comparison of performance of proposed approach over other recent state-of-art methods, the performance parameters such as Sn (sensitivity), Sp (specificity), F-Measure [156], and AC (accuracy) [157] have been employed in this work. The following equations define these performance parameters:

$$\text{Sn (sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.8}$$

$$\text{Sp (specificity)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4.9}$$

$$\text{F} - \text{measure} = \frac{2 * (\text{prec} * \text{rec})}{(\text{prec} + \text{rec})} \tag{4.10}$$

where, $prec \text{ (precision)} = \frac{\text{TP}}{\text{TP} + \text{FP}}$, $\text{rec (recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$

$$\text{AC (accuracy)} = \frac{\text{Sn} + \text{Sp}}{2} \tag{4.11}$$

The outcome of an approach applied for detection of CGIs consists of four possible parameters and these are: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP depicts those locations of DNA sequence which are captured by the algorithm correctly and true CGIs are located at those locations. TN tells those segments where no CGIs are captured and actual

CGIs are also not located there. FP represents those erroneously identified locations of CGIs where actual CGIs are not located, and those sections of true CGI which are not captured by method are termed as FN. Using these four parameters, the evaluation metrics Sn, Sp, F-measure, and AC can be assessed. The range of value of all four evaluation metrics Sn, Sp, F-measure, and AC lies between 0-1. An approach is considered to be perfect if the value of evaluation metrics Sn, Sp, F-measure, and AC obtained using that approach is closer to 1. The parameter Sn corresponds to the percentage of TPs which have been perfectly identified by the approach; and the parameter Sp signifies the proportion of TNs which have been precisely detected by the approach. The parameter which highlights the simultaneous effect of both Sn and Sp is termed as AC. The accuracy of approach is computed using parameter F-measure which calculates the harmonic mean of prec (precision) and rec (recall). If the performance evaluation has been carried out at a single threshold only, then F-measure is s suitable parameter for assessment in place of ROC (receiver operating characteristics).

4.4 Results and Discussion

The performance comparison of the proposed approach has been carried out with other recent state-of-art methods on the data set of hundred DNA sequences. The methods used for comparison are as follows: CpGclusterTLBO [129], CpGPNP [122], and DWT based method for CGIs detection [130]. The results obtained in terms of performance parameters TP, TN, FP, FN, Sn, Sp, F-measure, and AC using all the methods for data set of 85 DNA sequences of human species, 9 DNA sequences of mouse species, 6 DNA sequences of fish species, and overall 100 DNA sequences comprising of all 3 species are highlighted in Table 4.7, 4.8, 4.9, and 4.10 respectively.

Table 4.7: Performance metrics obtained in human species' 85 DNA sequences using all methods

Evaluation metric	**CGI detection methods**			
	Proposed	**CpGclusterTLBO**	**CpGPNP**	**DWT based method**
TP	**78338**	71218	66048	65822
TN	**4456041**	4444891	4358640	1772242
FP	**130623**	136172	228024	2814422

FN	**25419**	27735	37709	37938
Sn	**0.7550**	0.7197	0.6366	0.6344
Sp	**0.9715**	0.9702	0.9503	0.3864
F-measure	**0.5010**	0.4650	0.3320	0.0441
AC	**0.8632**	0.8449	0.7934	0.5104

Table 4.8: Performance metrics obtained in mouse species' 9 DNA sequences using all methods

Evaluation metric	**CGI detection methods**			
	Proposed	**CpGclusterTLBO**	**CpGPNP**	**DWT based method**
TP	**30434**	25985	11155	17192
TN	**1262233**	1260968	1210651	703139
FP	**55750**	57015	107332	614844
FN	**3540**	7989	22819	16782
Sn	**0.896**	0.765	0.328	0.506
Sp	**0.958**	0.957	0.919	0.533
F-measure	**0.5066**	0.4443	0.1463	0.0516
AC	**0.927**	0.861	0.624	0.52

Table 4.9: Performance metrics obtained in fish species' 6 DNA sequences using all methods

Evaluation metric	**CGI detection methods**			
	Proposed	**CpGclusterTLBO**	**CpGPNP**	**DWT based method**
TP	3496	2763	3181	**3555**
TN	**595415**	579762	576127	236594
FP	**12020**	27673	31308	370842

FN	1731	2464	2046	**1672**
Sn	0.67	0.53	0.61	**0.68**
Sp	**0.98**	0.954	0.948	0.389
F-measure	**0.3371**	0.1550	0.1602	0.0187
AC	**0.825**	0.742	0.779	0.535

Table 4.10: Performance metrics obtained in overall hundred DNA sequences using all methods

Evaluation metric	**CGI detection methods**			
	Proposed	**CpGclusterTLBO**	**CpGPNP**	**DWT based method**
TP	**112268**	99966	80384	86569
TN	**6313689**	6285621	6145418	2711975
FP	**198393**	220860	366664	3800108
FN	**30690**	38188	62574	56392
Sn	**0.7853**	0.7236	0.5623	0.6055
Sp	**0.9695**	0.9661	0.9437	0.4165
F-measure	**0.4950**	0.4356	0.2725	0.0430
AC	**0.8774**	0.8448	0.7530	0.5110

The superiority of proposed approach over other state-of-art methods has been noticed from Tables 4.7, 4.8, 4.9, and 4.10. All the performance parameters TP, TN, FP, FN obtained using proposed approach for 85 DNA sequences of human species and 9 DNA sequences of mouse species are much better than other methods; and subsequently evaluation metrics Sn, Sp, F-measure, and AC are having much higher value than other methods. The performance parameters TN, FP acquired using proposed approach for 6 DNA sequences of fish species are better than all other methods, however TP and FN parameters are little lesser than DWT based method but higher than other methods. Consequently, evaluation metrics Sp, F-measure and AC obtained using proposed approach are much higher than all other methods, whereas Sn of proposed approach is little lesser than DWT based method but much higher than other methods. The overall performance of proposed method on the whole data set of hundred DNA sequences is the best in all parameters and metrics compared to all other recent state-of-art methods. The percentage improvement in terms of evaluation metrics Sn, Sp, F-measure, and AC of proposed approach over other methods has been calculated and shown in Table 4.11.

Table 4.11: Percentage improvement of proposed algorithm in value of performance metrics over other methods

Evaluation metric	**CGI detection methods**		
	CpGclusterTLBO	**CpGPNP**	**DWT based method**
Sn	7.86%	28.40%	22.90%
Sp	0.35%	2.66%	57.04%
F-measure	12%	44.95%	91.31%
AC	3.72%	14.18%	41.76%

The performance of proposed approach has also been assessed on the basis of the percentage coverage of the true CGI's length and the performance comparison has been carried out with other recent state-of-art algorithms. Table 4.12 shows the number of CGIs identified by all methods out of total 194 CGIs present in hundred DNA sequences comprised of human, mouse, and fish species at 80 percent, 90 percent, and 100 percent (full length) coverage of true CGI's length.

Table 4.12: Number of CGIs identified out of total 194 in hundred DNA sequences

Methods	**Number of CGIs identified corresponding to percentage coverage of true CGI's length**		
	80percent	**90percent**	**100percent**
Proposed	**112**	**101**	**93**
CpGclusterTLBO	108	76	50
CpGPNP	60	46	39
DWT based method	1	0	0

The superiority of proposed approach over other state-of-art methods has been noticed from Table 4.12. The number of detection of CGIs at 80 percent, 90 percent, and 100 percent (full length) coverage of true CGI's length by proposed approach is much higher than all other methods.

4.5 Summary

In this research work detection of CGI in DNA sequences with the help of STFT based approach has been developed and proposed. It has been examined and proved that application of 24 combinations of integer mapping scheme functions much better than other mapping schemes considered in this work for CGI identification. The important feature hidden in CGIs in terms of periodicities has been examined and experimentally proved on a benchmark DNA sequence. And then the self created data set of hundred DNA sequences comprising of human, fish and mouse species has been applied to test and prove the universal applicability and superiority of proposed approach over other recent state-of-art methods. The proposed algorithm’s performance has been noticed much better over other methods in terms of evaluation metrics Sn, Sp, F-measure, and AC. Also, the performance of proposed approach has been found the best amongst all other methods in the context of identification of more number of CGIs at percentage of 80 percent, 90 percent, and 100 percent (full length) of true CGI’s length.

CHAPTER 5

MODIFIED P-SPECTRUM BASED ALGORITHM AND MODIFIED GABOR WAVELET TRANSFORM BASED APPROACHES FOR CPG ISLANDS DETECTION IN DNA SEQUENCES

The main limitation of STFT is that the window length employed in it is fixed to a suitable value. This limitation may affect the performance of algorithm applied for CGIs identification in terms of missing of significant information such as CpG Islands hidden in DNA sequences. Hence the sensitivity and overall performance of algorithm may be affected which requires improvisation. To address this limitation, in this research work two approaches namely modified P-spectrum and modified Gabor Wavelet transform based algorithms for CGI identification have been proposed. A dataset of hundred DNA sequences comprising of human species has been used in both the approaches. The enhancement in sensitivity of CGI identification in context of prediction of greater number of CGIs has been achieved using modified P-spectrum based approach and an overall improvement in all performance metrics for CGI identification has been obtained using modified Gabor Wavelet transform (MGWT) based algorithm.

Part 1: Modified P-Spectrum based Algorithm for Sensitivity Enhancement of CpG Islands Detection in DNA Sequences

5.1 Proposed Approach for CGIs Identification

The important feature in terms of periodicities present in CGIs has been explored with experimental validation by the authors in [158]. It has been proved in [158] that periodic pattern corresponding to periodicities 2-10 remain hidden in CGIs of DNA sequences. This feature of periodicity has been employed in this research work. Using this feature, the flow graph of the approach proposed in this research work has been depicted in Figure 5.1:

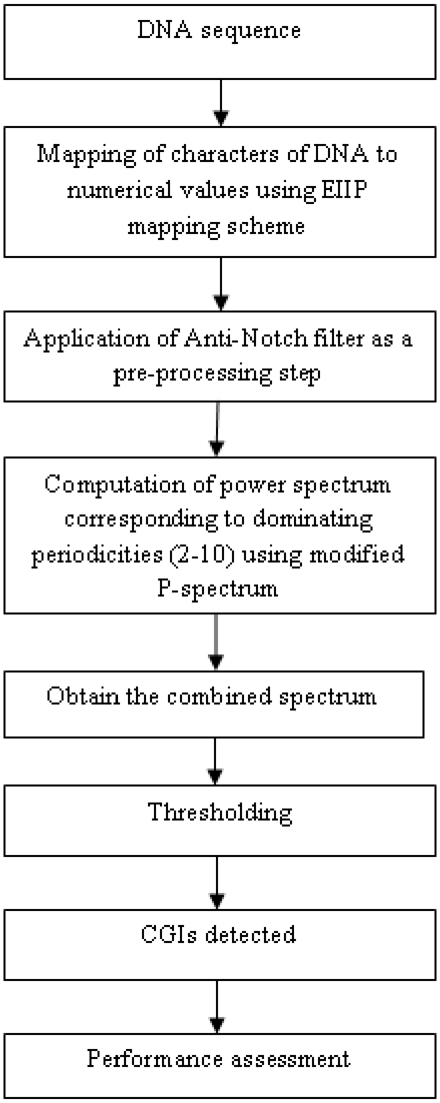

DNA sequence
Mapping of characters of DNA to numerical values using EIIP mapping scheme
Application of Anti-Notch filter as a pre-processing step
Computation of power spectrum corresponding to dominating periodicities (2-10) using modified P-spectrum
Obtain the combined spectrum
Thresholding
CGIs detected
Performance assessment

Figure 5.1: Flow graph of the proposed approach

The periodicity-spectrum (p-spectrum) as reported in literature has been considered to be conceptually, computationally, and theoretically a highly robust technique to detect the periodic components. This is possible because the p-spectrum uses the LS estimation of the most significant periodic pattern in a sequence of given signal [144]. This property of p-spectrum makes it applicable for the identification of buried periodic features in signals [144-145]. It is well known that the DNA sequences of genomic data contains a lot of buried periodic patterns, CGI is an example of such periodic pattern. Hence, to capture the hidden dominating periodicities in CGIs a modified P-spectrum based approach has been developed and proposed in this research work for the detection of CGIs in DNA sequences. The detailed discussion of the steps of proposed approach is presented as follows:

a) The DNA sequence in which the CGIs have to be detected is taken from standard

database and fed to the algorithm.

b) The alphabets of DNA sequence are then mapped to numerical values with the help of EIIP (electron ion interaction potential) numerical conversion scheme. The numerical values A= 0.1260, C=0.1340, T= 0.1335, G=0.0806 are given to the alphabets of DNA data and numerical sequence is thus obtained.

c) For the purpose of filtering the noisy components, an anti notch filter has been utilized as a pre-processing step. The transfer function of a second order anti notch filter used in this work is represented as:

$$\mathrm{H}\,(\mathrm{z}) = (1-\mathrm{z}^{-2})/(1-2\mathrm{R}\cos\left(\frac{2\pi}{3}\right)\mathrm{z}^{-1}+\mathrm{R}^2\mathrm{Z}^{-2}) \tag{5.1}$$

As noticed in equation (5.1) that the center of anti notch filter is located at an angular

frequency 2π/3. The value of R has been selected as 0.992 empirically.

d) The dominating 2 to 10 periodicities are then extracted with the help of modified P-spectrum. The functioning of p-spectrum and subsequently modified P-spectrum for any arbitrary period ‘p’ which has been employed in this work is now described in the following points:

1) The discrete version of a data (D) can be represented as:

$$\mathrm{D} = [\mathrm{d}_1\ \ \mathrm{d}_2\ \ \mathrm{d}_3 \ldots\ \mathrm{d}_\mathrm{M}] \tag{5.2}$$

2) The data signal D is desired to be an exact multiple of period ‘p’. Therefore, the data can be rearranged by inserting adequate zeros in the last row. A matrix R_p has been created which consists of ‘a’ segments having length ‘p’ and are non-overlapping as shown in equation (5.3).

$$R_p = \begin{bmatrix} d_1 & d_2 & d_3 & \cdots d_i & d_{i+1} \cdots & d_p \\ d_{p+1} & d_{p+2} & d_{p+3} & \cdots d_{p+i} & d_{p+i+1} \cdots & d_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ d_{ap+1} & d_{ap+2} & d_{ap+3} & d_M & 0 & 0 \end{bmatrix} \quad (5.3)$$

3) The presence of dominating periodic component in matrix R_p can be obtained by first singular value for which the most robust technique i.e. SVD (singular value decomposition) has been employed in next step to calculate the signal shown in following equation:

   $$\text{highest\_singular} = \max(\text{SVD}(R_p)) \quad (5.4)$$

4) In the next step, all the elements of matrix R_p are added to obtain signal 'summate'; and another signal 'modified_s' is acquired as shown in following equation:

   $$\text{modified\_}s = \max\left({}^{\text{summate}}/_{2}\right) \quad (5.5)$$

5) To obtain the modified P-spectrum, the last step is to compute the 'mod_p_spec' signal which is obtained as per following equation:

   $$\text{mod\_p\_spec} = \text{highest\_singular} \times \text{modified\_}s \quad (5.6)$$

   The× operation in equation (5.6) has been performed by taking the row-wise multiplication of the elements of two signals.

e) The combined spectrum corresponding to dominating periodicities is then computed by adding the spectrums of these periodicities by applying the steps of modified P-spectrum marked in d) point.

f) An empirical value chosen as 10% corresponding to maximum value of the combined spectrum is then applied as threshold.

g) Those regions of combined spectrum which crosses the chosen threshold limit are finally categorized as CGIs and the performance assessment has been carried out for such regions.

Now to verify the applicability of proposed approach, a DNA sequence having Genebank accession number AC005003 [1] (consisting of 65750 bps and having one CGI located at position 38374-

41067) has been selected as an example sequence and the plot of combined spectrum acquired using proposed approach is depicted in Figure 5.2:

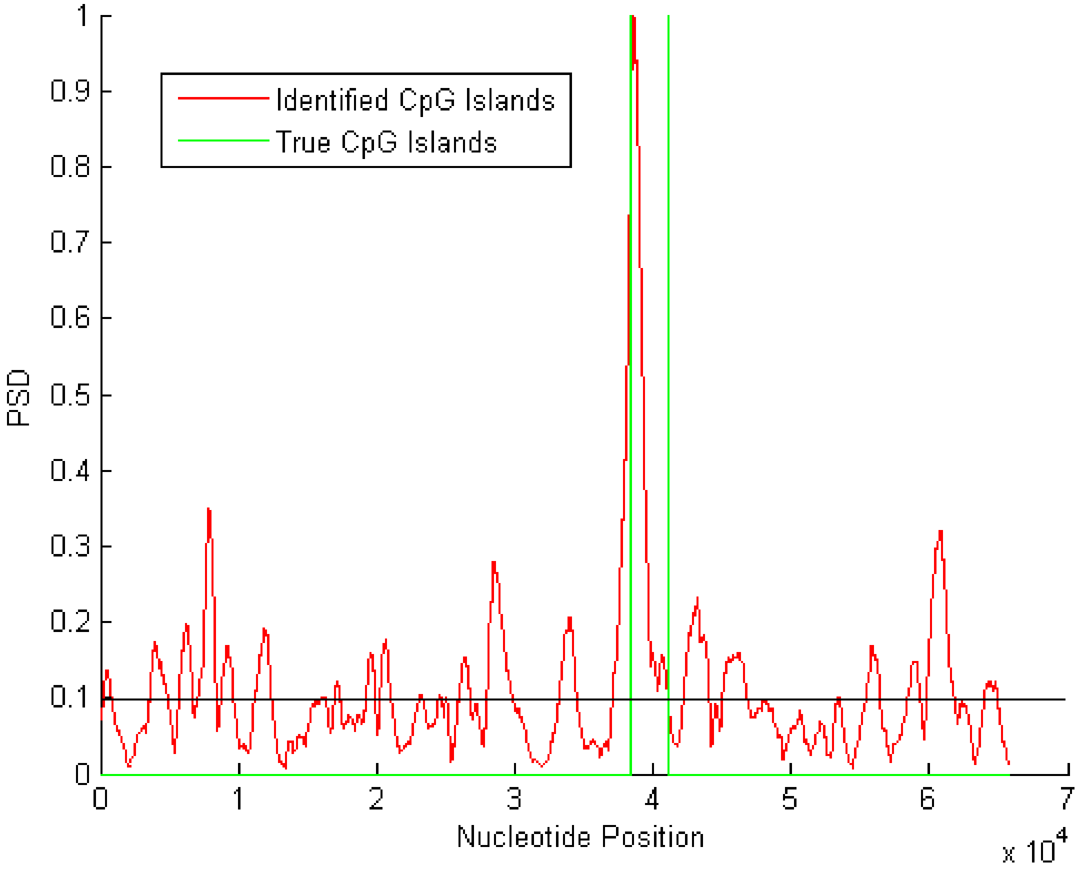


Figure 5.2: Proposed approach's result obtained of combined spectrum for example DNA sequence AC005003

The result in terms of probable prediction outputs and correspondingly the evaluation parameters Sn (sensitivity), Sp (specificity) which are obtained using proposed approach and other state-of-art methods for example DNA sequence is tabulated in Table 5.1:

Table 5.1: Evaluation parameters obtained for example DNA sequence AC005003

Performance parameter	**Methods**				
	Proposed approach	**STFT based method [158]**	**CpGPNP [122]**	**CpGclusterTLBO [129]**	**DWT based method [130]**
TP	**2694**	2603	2363	2223	1700
TN	40165	61350	57943	60135	24395
FP	22890	1705	5112	2920	38660
FN	**0**	91	331	471	994
Sn	**1**	0.9662	0.8771	0.8252	0.6310
Sp	0.6370	0.9730	0.9189	0.9537	0.3869

It has been noticed from Table 5.1 that the proposed approach is capable to enhance the sensitivity of CGI detection for considered example DNA sequence AC005003. As the length of CGI present in this sequence is 2694 which means that the number of TPs (true positive) in this sequence is 2694; the proposed approach is able to capture all the TPs in this sequence. Correspondingly, the value of Sn (sensitivity) obtained using proposed approach is 1. Whereas, the other state-of-art methods have not detected all the TPs in this sequence and hence the value of Sn obtained using these methods is not achieved as theoretically desired value of 1. Hence, it can be presumed from the experiment carried out on an example DNA sequence that the capability of proposed approach in terms of detection of number of CGIs is enhanced compared to other recent state-of-art methods. At the same time, it has been noticed from Table 5.1 that the number of FPs (false positive) obtained using the proposed approach are also on the higher side compared to STFT, CpGPNP, and CpGclusterTLBO based approaches for CGIs identification but lesser than DWT based algorithm for CGIs identification. Subsequently, the value of Sp (specificity) obtained using proposed approach is smaller compared to STFT, CpGPNP, and CpGclusterTLBO based approaches; however the value of Sp is higher compared to DWT based algorithm of CGIs identification.

Having verified the applicability and sensitivity enhancement of proposed approach using an example DNA sequence, the performance of proposed approach and other recent state-of-art methods have now been tested using a large data set of hundred DNA sequences of human species. The example DNA sequence used in this section has been considered in the whole data set of hundred DNA sequences for the computation of performance metrics in Results section.

5.2 Data Set of CGIs and Performance Metrics

5.2.1 Data Set of CGIs

A CGI data set of hundred DNA sequences of human species has been prepared by us by acquiring the details from publically available database website: National Centre for Biotechnology Information (NCBI) [1]. The complete details of the data set like Gene bank accession number, number of base pairs (bps) in a sequence, number of CGIs, and start & end locations of CGIs within a sequence are presented in Table 5.2:

Table 5.2: Details of hundred DNA sequence's CGI data set

S. No.	Gene bank accession number	Number of base pairs (bps)	Number of CGIs	Start and end location of CGI as acquired from NCBI website
1	AL024496	27210	5	1284-1927, 9755-10674, 13099-13615, 15578-16126, 21132-21595
2	AL109743	96006	2	31713-33048, 56464-57695
3	AC027644	188207	3	27115-27651, 51380-51705, 130590-131909
4	AC110076	105211	1	93622-94410
5	AC073271	117930	1	102756-103541
6	AC005282	98219	2	8323-9168, 79507-80293
7	AC110787	7335	1	11-1165
8	L47124	6996	1	3226-4068
9	AC010990	6708	2	2347-2685, 4079-4357
10	AF129290	6324	6	2026-2238, 2436-2679, 2730-3021, 3033-3353, 3355-3637, 4479-4891
11	D13370	3730	1	226-1645
12	AH004914	5426	1	1018-1636
13	AC079588	4249	1	1137-2422
14	AH009772	4240	2	1-555, 656-1588
15	AL132818	38860	1	33379-33940
16	AL442638	188247	4	17472-17700, 22868-23148, 93250-93495, 163847-164132
17	AC073335	68275	3	31813-32080, 33619-34458, 50802-51655
18	AC073517	67706	1	35431-35977
19	AC127379	67291	2	30060-30318, 38447-39437
20	AC064843	66898	1	5531-5785
21	AC129782	66860	1	38868-40898

22	AC013270	66660	4	6075-6881, 25374-26035, 34710-36183, 48185-48621
23	AC074386	66610	2	15847-16381, 16593-16830
24	AC092103	66565	1	24844-25119
25	AC124014	66552	1	56936-57769
26	AL137791	66254	4	30724-31272, 46196-46906, 52979-53956, 61007-62096
27	AC096553	66229	1	11867-12256
28	AC105413	65958	1	50478-50751
29	AC005003	65750	1	38374-41067
30	AC145546	65625	1	52797-53645
31	AC105402	65449	2	15774-16973, 28628-28925
32	AC112698	65335	1	42309-43546
33	AC104129	65189	8	2966-3334, 8763-9020, 14023-14383, 20695-20991, 26472-26735, 28330-29188, 31762-32009, 55671-55878
34	BN000001	64961	1	895-1123
35	AC138782	64744	1	23500-24633
36	AC005021	64607	2	24663-25225, 63177-63512
37	AC093086	64601	1	58914-59518
38	AC005233	64359	1	16579-18003
39	AC013436	63823	5	12411-12652, 21066-21331, 24980-26051, 26467-26807, 60097-60448
40	AC131957	63780	1	45526-45799
41	AC004694	63749	2	9107-9494, 54481-54756
42	AC108463	63525	3	26008-26366, 26575-26982, 27079-27538
43	AC080165	63279	1	8258-8531
44	AC010890	62764	4	11407-11926, 13574-13801, 53142-53415, 53755-54041

45	AC108142	62624	1	8864-11837
46	AC080068	62623	1	535-774
47	AC093785	62466	1	31397-31665
48	AC003079	62331	1	50250-50471
49	AC078937	62035	1	38149-39359
50	AC114803	61579	7	3256-4009, 18815-19353, 32398-32647, 33247-33659, 36773-37302, 39696-39964, 55808-56144
51	AC093652	61340	1	48156-49072
52	AC093377	61056	1	729-1003
53	AC073201	60776	1	9738-11862
54	AC113611	60597	1	8638-9514
55	AC099394	60024	7	2826-4863, 10806-11866, 19723-19934, 25482-25769, 31861-32884, 36728-36931, 54994-55361
56	AC098831	59776	2	39343-39572, 51406-51689
57	AC074013	59657	3	22602-22873, 51602-52508, 53105-53331
58	AC062028	59634	1	44629-44851
59	AC106875	59580	1	4526-5382
60	AC023670	59565	1	25568-27400
61	AC079882	59427	1	39153-39736
62	AC006008	57554	1	28800-30423
63	AC108222	21776	1	21237-21776
64	AH006464	21230	1	1187-2051
65	AC093609	20710	1	7857-8257
66	AL590794	18042	1	11568-12215
67	AC136375	17863	1	16369-17534
68	BD432859	14646	2	2762-2973, 4065-5181
69	AC111201	13470	3	4327-4727, 5323-5554, 12500-13455
70	NM005876	10782	1	6154-7734

71	NM053043	10168	1	9597-9820
72	AC093460	10103	1	6951-7418
73	AC108032	9716	1	30-269
74	X86012	9541	1	335-3853
75	AC106048	8594	1	7941-8180
76	AH008870	6797	1	341-1340
77	AC079401	6568	1	3086-3935
78	AH007568	6513	3	543-803, 1212-1430, 1662-2474
79	AC105385	5952	1	2844-3080
80	AJ308559	5596	1	1228-1657
81	M92844	3889	1	3198-3889
82	AF196313	3700	1	2092-3580
83	AF281043	3662	1	1611-2734
84	U48937	3278	1	2588-3230
85	AF307776	3113	2	2334-2745, 2791-3064
86	AJ000757	3046	1	650-2840
87	AJ289875	2916	1	2325-2916
88	L07287	2704	1	1-1350
89	Z92546	73511	1	20746-21240
90	AL591222	147211	2	54605-55080, 68825-69091
91	AL513502	174636	1	116364-117432
92	AL513498	155780	1	18305-18582
93	AL357615	171446	2	56753-57030, 59607-59874
94	AL353786	139565	1	19000-19400
95	AL121926	139544	2	102641-104201, 126562-127299
96	AL049547	129811	5	27801-29311, 37094-37773, 109041-110125, 113196-114024, 126815-127265

97	AL031706	13012	1	7-552
98	AL031703	35098	4	15319-17699, 25107-26048, 30327-30736, 31615-32204
99	AJ006998	123521	1	11140-11417
100	AL031707	28707	4	6050-6520, 6693-7445, 24481-25248, 28059-28669

5.2.2 Performance Metrics

To examine and compare the performance of proposed approach over other recent state-of-art methods, the standard performance metrics such as Sn (sensitivity), and Sp (specificity) are used in this work. The following equations define these performance parameters:

$$\text{Sn (sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.7}$$

$$\text{Sp (specificity)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5.8}$$

The possible outcome of any algorithm which is applied for detection of CGIs consists of four parameters and these are: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP depicts those locations of DNA sequence which are captured by the algorithm correctly and true CGIs are located at those locations. TN tells those segments where no CGIs are captured and actual CGIs are also not located there. FP represents those erroneously identified locations of CGIs where actual CGIs are not located, and those sections of true CGI which are not captured by method are termed as FN. Using these four parameters, the performance metrics Sn, and Sp can be obtained. The range of value of these performance metrics Sn, and Sp lies between 0-1. If the value of performance metrics Sn, and Sp obtained using that algorithm is closer to 1, that algorithm is considered to be perfect. The parameter Sn corresponds to the percentage of TPs which have been perfectly predicted by the algorithm; and the parameter Sp signifies the proportion of TNs which have been accurately captured by the algorithm.

5.3 Results and Discussion

The performance comparison of the proposed approach has been carried out with four recent state-of-art methods of CGI detection on the data set of hundred DNA sequences of human species. The methods which have been used for comparison are as follows: STFT [158], CpGPNP [122], CpGclusterTLBO [129], and DWT based method for CGIs detection [130]. The value of sensitivity and specificity obtained using all methods for hundred sequences have been tabulated in Table 5.3 and Table 5.4 respectively.

Table 5.3: Value of sensitivity on data set of hundred DNA sequences

		Sensitivity (Sn) Methods				
S. No.	**Accession number**	**Proposed algorithm**	**STFT based method**	**CpGPNP**	**CpGclusterTLBO**	**DWT based method**
1	AC110076	1	1	0	0.8479	0.8517
2	AC073271	1	1	0	0.6539	0.584
3	AC005282	1	0	0.5003	0.9314	0.4507
4	AC110787	1	1	1	0.9671	0.6641
5	L47124	1	1	1	0.7248	0.8422
6	AC010990	1	0.9385	1	0.8641	0.5955
7	AF129290	1	1	1	0.4161	0.5314
8	D13370	1	1	1	0.6458	0.6958
9	AH004914	1	1	1	0.5347	0.6817
10	AC079588	1	1	1	0	0.8336
11	AH009772	1	1	0.9926	0.3239	0.4153
12	AL132818	1	1	0.1975	0.4484	0.7954
13	AL024496	0.1774	0.2608	0.818	0.4399	0.4337
14	AL109743	0.4283	0.8026	0.6347	0.7741	0.5565
15	AC027644	0	0.9647	0.2455	0.6899	0.4219
16	AL442638	0.5058	0.0518	0.0691	0.8196	0.5307
17	AC073335	0.8634	0.4246	0.3089	0.8711	0.6646
18	AC073517	1	1	0.6618	0.7587	0.7916
19	AC127379	0.448	0.5312	0.7344	1	0.6776
20	AC064843	1	1	0.0941	1	0.6353
21	AC129782	1	1	0.7962	0.6760	0.5362
22	AC013270	0.9077	0.5485	0.6698	0.7553	0.5518
23	AC074386	0.6572	1	0.8202	0.6792	0.6223
24	AC092103	1	1	0.0362	0.8551	0.5616
25	AC124014	1	1	0.1691	0.9029	0.9053
26	AL137791	0.8287	0.7879	0.3618	0.8236	0.512
27	AC096553	1	0.5744	0.5	0.9615	0.4333
28	AC105413	0	0	0	0.6350	0.5985
29	AC005003	1	0.9622	0.8771	0.8252	0.631
30	AC145546	0.3498	0.6078	0.3062	1	0.4511
31	AC105402	1	0.9506	0.8825	1	0.713
32	AC112698	1	0.9814	0.7294	0.5412	0.622
33	AC104129	0	0.4763	0.5667	0.7730	0.515
34	BN000001	0.1572	1	1	0.6157	0.7991
35	AC138782	1	0.6376	0.8148	1	0.7681
36	AC005021	1	0.9789	0.4483	0.7175	0.7408
37	AC093086	1	1	0	0.5736	0.5041
38	AC005233	1	0.5074	0.7782	0.8618	0.4751
39	AC013436	0.883	0.7387	0.6876	0.9987	0.6045
40	AC131957	1	1	0	0.8431	0.6314

41	AC004694	0.5994	0.6898	0.4925	0.7199	0.5105
42	AC108463	1	0.2926	0.3676	1	0.7196
43	AC080165	1	1	0	0.9453	0.7263
44	AC010890	0.2872	0.2315	0.3972	0.8411	0.6516
45	AC108142	0.9976	1	0.8783	0.6338	0.649
46	AC080068	1	1	1	0.6542	0.5958
47	AC093785	1	0	0.2937	0.8773	0.6022
48	AC003079	1	1	0	1	0.7703
49	AC078937	1	1	0.7201	1	0.6565
50	AC114803	0.9191	0.7332	0.5679	0.968	0.6536
51	AC093652	1	1	0.6336	1	0.6347
52	AC093377	0	0.9709	1	0.8945	0.2473
53	AC073201	0.8306	1	0.984	0.4739	0.7275
54	AC113611	1	0.7514	0.8198	0.6602	0.52
55	AC099394	0.8397	0.6341	0.6986	0.5935	0.6258
56	AC098831	0.4475	0.4475	0	0.6595	0.8171
57	AC074013	0.84	0.9787	0.3741	0.9239	0.5597
58	AC062028	0.8117	0	0	0.9283	0.5336
59	AC106875	1	1	1	1	0.7398
60	AC023670	0.934	1	0.7763	0.9902	0.6454
61	AC079882	1	1	0.2911	0.8545	0.512
62	AC006008	1	1	0.7482	0.2654	0.7211
63	AC108222	0.6889	1	0.5325	1	0.613
64	AH006464	1	1	1	1	0.7087
65	AC093609	1	0.4763	1	0.995	0.399
66	AL590794	1	0.5216	0.9784	0.9398	0.696
67	AC136375	1	1	0.8902	0.8259	0.711
68	BD432859	0.7208	0.5342	1	0.7013	0.7035
69	AC111201	0.8295	0.5935	0.8798	0.2102	0.8049
70	NM005876	1	0	0.4902	0.3276	0.7989
71	NM053043	0	1	0.8884	1	0.4063
72	AC093460	1	0.1517	1	0.5662	0.6709
73	AC108032	1	0.9	1	0.7375	0.7667
74	X86012	0.8008	1	0.6641	0.2577	0.6095
75	AC106048	1	0	0	1	0.75
76	AH008870	1	1	1	0.628	0.765
77	AC079401	1	0.6918	1	0.8106	0.6812
78	AH007568	1	0.7981	0.7981	0.6234	0.4524
79	AC105385	1	0.9705	0	1	0.8523
80	AJ308559	1	1	1	1	0.7116
81	M92844	1	1	1	0.5818	0.5303
82	AF196313	1	1	1	0	0.7918
83	AF281043	1	1	1	0	0.6922
84	U48937	1	1	1	0.7341	0.6283
85	AF307776	1	1	1	0.3411	0.7289
86	AJ000757	1	1	1	0	0.8203
87	AJ289875	1	0.4484	0.9949	0.8443	0.4949
88	L07287	1	1	1	0.36	0.6519
89	Z92546	1	1	0.6303	1	0.7313
90	AL591222	1	0.6406	0	0.7914	0.6608
91	AL513502	1	0.9149	0	0.9645	0.6146
92	AL513498	1	1	0.259	0.9892	0.4137
93	AL357615	0.4908	1	0	0.7766	0.7106

94	AL353786	0	0	0.6185	0.8155	0.5062
95	AL121926	0.9004	1	0.0857	1	0.5646
96	AL049547	0.3317	0.8090	0.2794	0.7608	0.6354
97	AL031706	1	1	0.4176	0	0.6667
98	AL031703	0.9357	0.5508	0.7252	0.3944	0.6155
99	AJ006998	0.2878	0.9532	0.7158	0.8669	0.5971
100	AL031707	0.6646	0.2347	0.4802	0.3077	0.6669

Table 5.4: Value of specificity on data set of hundred DNA sequences

		Specificity (Sp)				
		Methods				
S. No.	**Accession number**	**Proposed algorithm**	**STFT based Method**	**CpGPNP**	**CpGclusterTLBO**	**DWT based method**
1	AC110076	0.4001	0.9768	0.9628	0.9765	0.3451
2	AC073271	0.4521	0.9822	0.9795	0.9814	0.3517
3	AC005282	0.2111	0.9864	0.8425	0.9305	0.4452
4	AC110787	0.0333	0.6399	0.9205	0.9675	0.2517
5	L47124	0.2588	0.7689	0.8123	1	0.2736
6	AC010990	0.0191	0.8487	0.7982	0.9905	0.2422
7	AF129290	0.0553	0	0.6793	0.9159	0.3805
8	D13370	0.0922	0.7003	0.8523	0.9701	0.2326
9	AH004914	0	0.7588	0.9255	0.9536	0.335
10	AC079588	0	0	0.7171	0	0.2407
11	AH009772	0.1578	0.1807	0.5594	0.992	0.4638
12	AL132818	0.186	0.9643	0.9744	0.9825	0.2871
13	AL024496	0.9405	0.9091	0.7704	0.969	0.4864
14	AL109743	0.3101	0.9209	0.9145	0.9364	0.3607
15	AC027644	0.9869	0.9672	0.9381	0.9501	0.5536
16	AL442638	0.3287	0.9972	0.982	0.9874	0.3624
17	AC073335	0.1097	0.9661	0.9656	0.9735	0.3071
18	AC073517	0.2149	0.9788	0.9523	0.9656	0.3568
19	AC127379	0.3741	0.9965	0.957	0.9498	0.3445
20	AC064843	0.4289	0.9901	0.9153	0.9268	0.3849
21	AC129782	0.4033	0.9752	0.9924	0.9995	0.3803
22	AC013270	0.3683	0.982	0.9405	0.9741	0.3822
23	AC074386	0.5019	0.9824	0.9812	0.9677	0.4454
24	AC092103	0.138	0.9853	0.9654	0.9524	0.3025
25	AC124014	0.5209	0.9791	0.9623	0.961	0.5132
26	AL137791	0.1881	0.9717	0.9177	0.9582	0.4978
27	AC096553	0.4214	0.991	0.9772	0.9872	0.4198
28	AC105413	0.1725	0.9968	0.9751	0.983	0.317
29	AC005003	0.637	0.973	0.9189	0.9537	0.3869
30	AC145546	0.2747	0.9716	0.9526	0.9716	0.4136
31	AC105402	0.4867	0.9508	0.9705	0.9757	0.4104
32	AC112698	0.3394	0.9552	0.902	0.948	0.4796
33	AC104129	0.9474	0.9359	0.7631	0.97	0.4313
34	BN000001	0.3544	0.9546	0.9002	0.9272	0.4317
35	AC138782	0.1509	0.9999	0.9547	0.9536	0.3326
36	AC005021	0.2251	0.9647	0.9682	0.9762	0.3092
37	AC093086	0.105	0.9775	0.9603	0.9776	0.373
38	AC005233	0.2301	0.9808	0.9443	0.9364	0.3681
39	AC013436	0.1887	0.9234	0.8985	0.9564	0.3996

40	AC131957	0.1651	0.9945	0.986	0.9924	0.3201
41	AC004694	0.5042	0.991	0.9589	0.9541	0.329
42	AC108463	0.665	0.9982	0.9797	0.9776	0.4553
43	AC080165	0.3284	0.9923	0.9769	0.9777	0.2883
44	AC010890	0.4875	1	0.9802	0.9834	0.3227
45	AC108142	0.1235	0.9353	0.9404	0.9699	0.3203
46	AC080068	0.1638	0.9839	0.977	0.9819	0.2981
47	AC093785	0.2562	0.9874	0.9519	0.9561	0.2993
48	AC003079	0.4059	0.9914	0.9809	0.9874	0.4515
49	AC078937	0.3972	0.9678	0.9753	0.9695	0.3559
50	AC114803	0.2099	0.9299	0.9215	0.9427	0.3891
51	AC093652	0.1795	0.9314	0.959	0.9726	0.3285
52	AC093377	0.833	0.9859	0.9727	0.9755	0.4276
53	AC073201	0.1744	0.9822	0.9861	0.9752	0.3693
54	AC113611	0.2596	0.9879	0.9399	0.9682	0.5701
55	AC099394	0.2681	0.9149	0.8447	0.9792	0.4369
56	AC098831	0.1673	0.9962	0.9591	0.9753	0.3071
57	AC074013	0.1084	0.9224	0.8953	0.9394	0.3115
58	AC062028	0.1531	0.9966	0.9355	0.9465	0.338
59	AC106875	0.4289	0.9436	0.954	0.9686	0.4711
60	AC023670	0.214	0.9562	0.9217	0.9465	0.3269
61	AC079882	0.3346	0.9876	0.8858	0.9416	0.4239
62	AC006008	0.1289	0.9573	0.9537	0.9673	0.336
63	AC108222	0.1939	0.9245	0.9595	0.9402	0.3377
64	AH006464	0.5067	0.9036	0.9679	0.9685	0.3131
65	AC093609	0.1067	0.9821	0.9806	0.9607	0.3283
66	AL590794	0.0981	1	0.983	0.9872	0.3154
67	AC136375	0.0767	0.9517	0.9408	0.9873	0.3377
68	BD432859	0.1612	0.9794	0.9313	1	0.2474
69	AC111201	0.1853	0.9944	0.9143	0.9997	0.4914
70	NM005876	0.0668	0.3609	0.394	0.8936	0.2498
71	NM053043	0.3628	0.9293	0.7535	0.8128	0.3787
72	AC093460	0.1133	0.9592	0.9077	0.926	0.3028
73	AC108032	0.0356	0.966	0.9484	0.9614	0.2602
74	X86012	0.4405	0.3385	0.9666	1	0.5026
75	AC106048	0.0038	0.9595	1	0.9719	0.2645
76	AH008870	0	0.9332	0.9795	1	0.26
77	AC079401	0.1786	1	0.9542	0.9851	0.2405
78	AH007568	0	0.8178	0.9276	0.9615	0.4794
79	AC105385	0.0201	0.9942	1	0.9804	0.2331
80	AJ308559	0.1001	0.6256	0.8712	0.9601	0.2492
81	M92844	0	0	0.3935	1	0.3025
82	AF196313	0	0	0.8742	0	0.3805
83	AF281043	0.1143	0	0.6496	0	0.2081
84	U48937	0	0.7065	0.8246	0.9882	0.2532
85	AF307776	0	0	0.7951	0.9802	0.2675
86	AJ000757	0	0	0.1753	0	0.4056
87	AJ289875	0	1	1	1	0.225
88	L07287	0	0	0.6179	1	0.2365
89	Z92546	0.2989	0.9713	0.9291	0.958	0.4015
90	AL591222	0.2801	0.997	0.9862	0.9926	0.3405
91	AL513502	0.4576	0.9911	0.9869	0.9893	0.3578
92	AL513498	0.7217	0.9906	0.9854	0.985	0.6358

93	AL357615	0.4363	0.9974	0.9936	0.9944	0.4234
94	AL353786	0.5656	0.9965	0.9742	0.9723	0.4504
95	AL121926	0.2369	0.9683	0.9604	0.9714	0.4051
96	AL049547	0.9653	0.9654	0.9091	0.9654	0.3799
97	AL031706	0	0.9245	0.8073	0.9247	0.3449
98	AL031703	0.1959	0.9653	0.7463	0.9555	0.4786
99	AJ006998	0.2448	0.9973	0.9946	0.9948	0.3255
100	AL031707	0.1184	0.9748	0.8734	0.9449	0.3298

The overall value obtained of performance metrics (TP, TN, FP, FN, Sn, and Sp) for the whole data set of hundred DNA sequences using all the methods are depicted in Figure 5.3-5.8 and has been tabulated in Table 5.5.

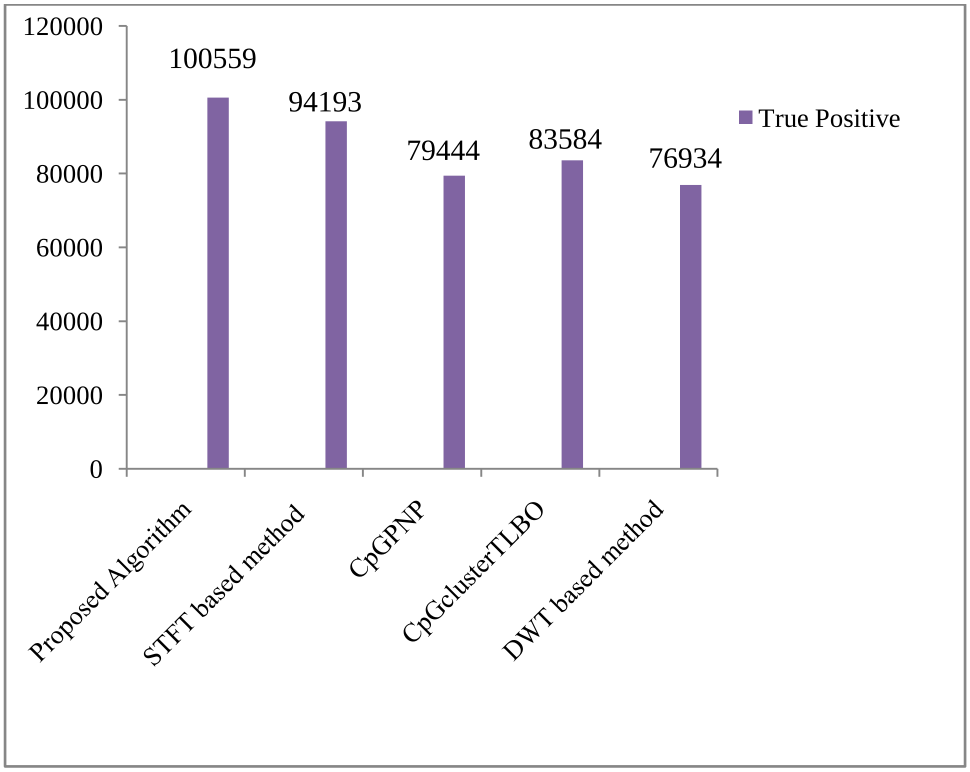


Figure 5.3: Value of True Positive obtained using all methods

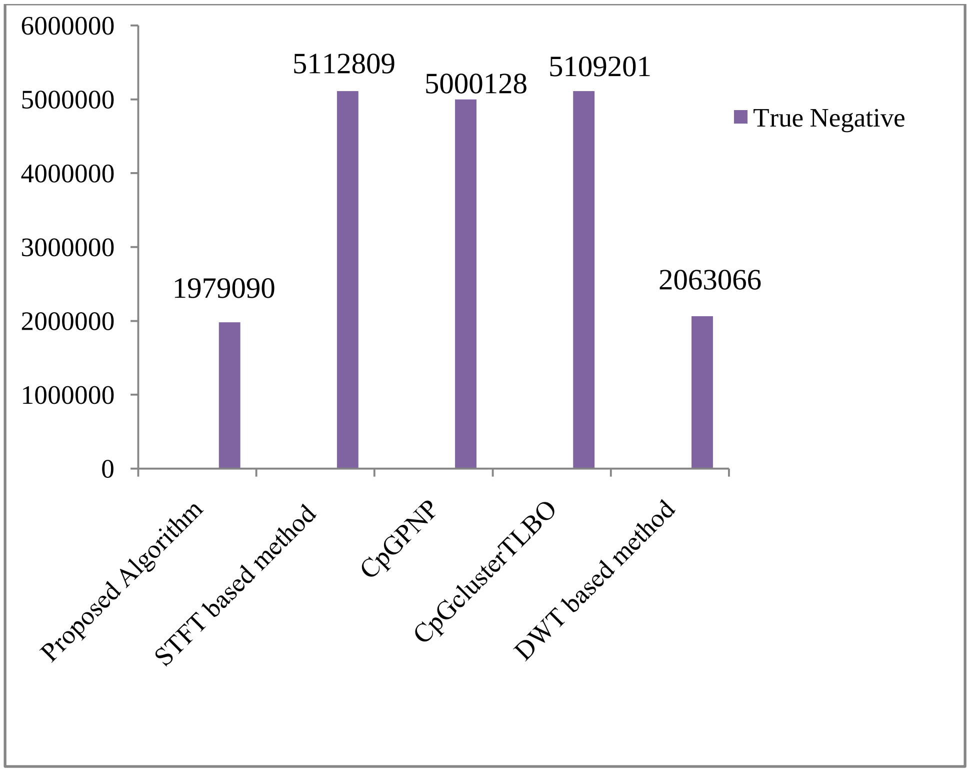


Figure 5.4: Value of True Negative obtained using all methods

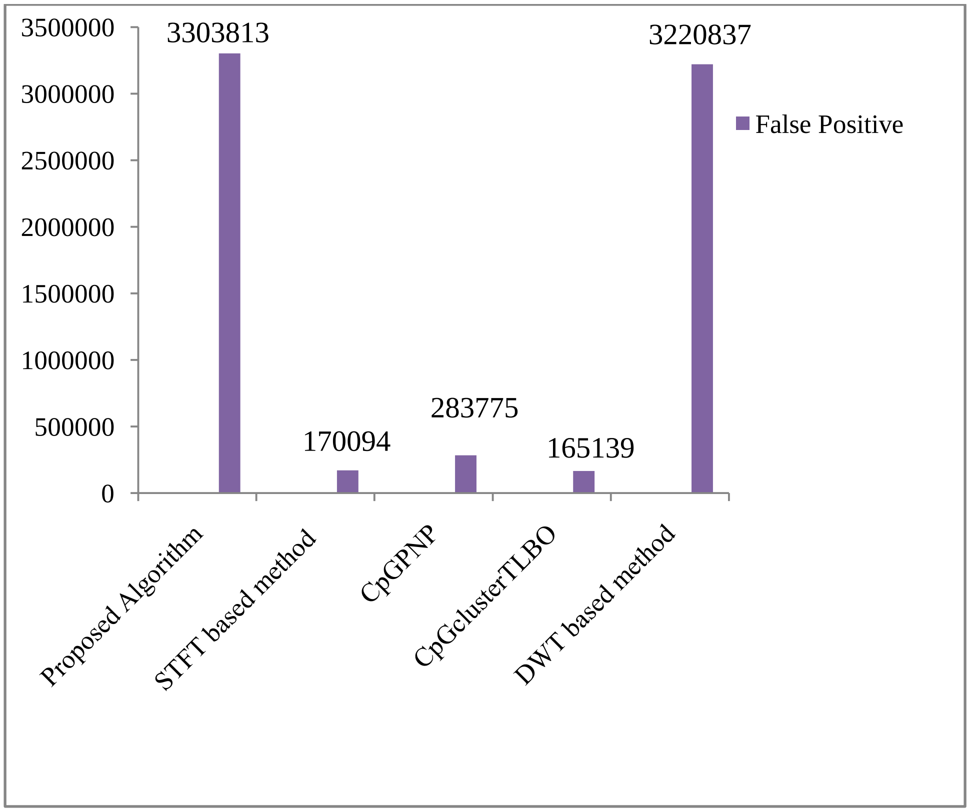


Figure 5.5: Value of False Positive obtained using all methods

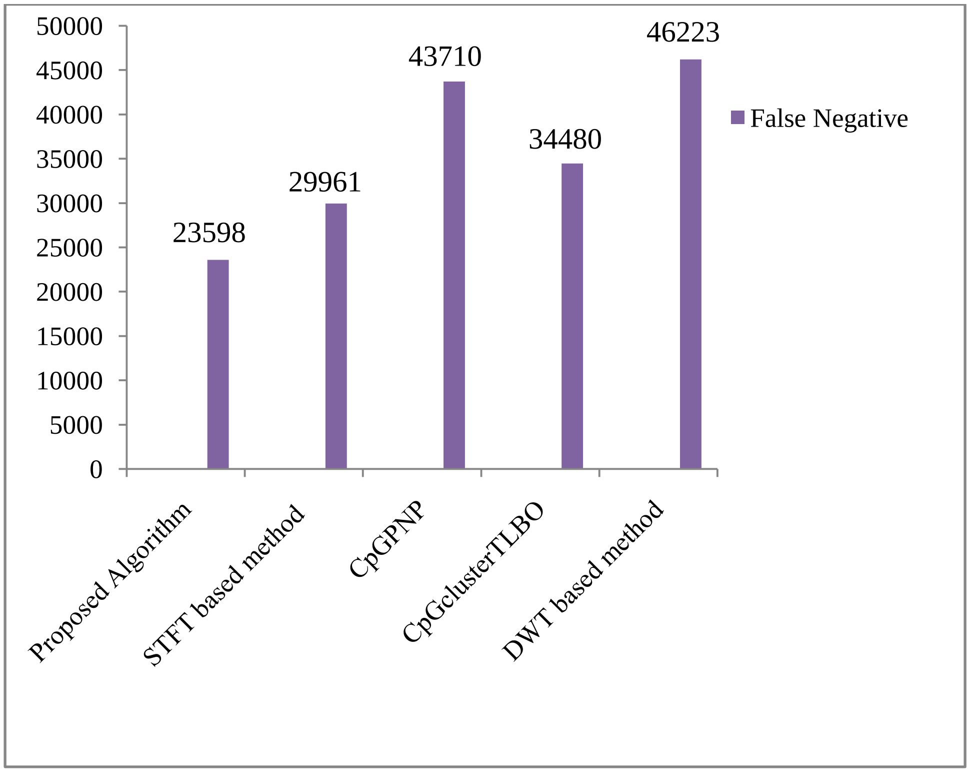


Figure 5.6: Value of False Negative obtained using all methods

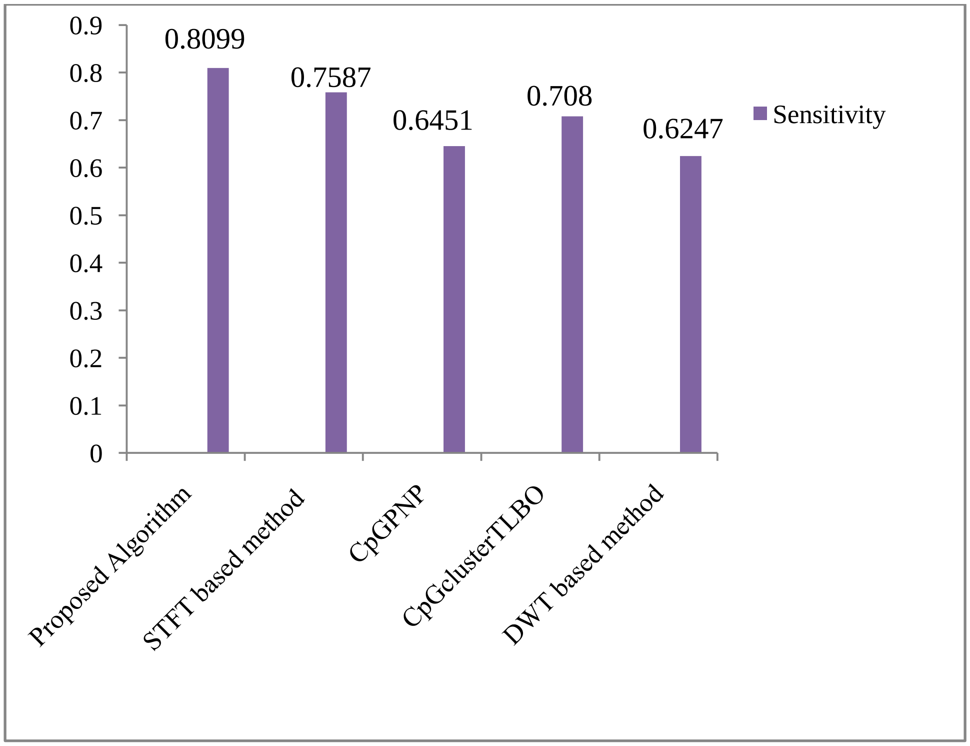


Figure 5.7: Value of Sensitivity obtained using all methods

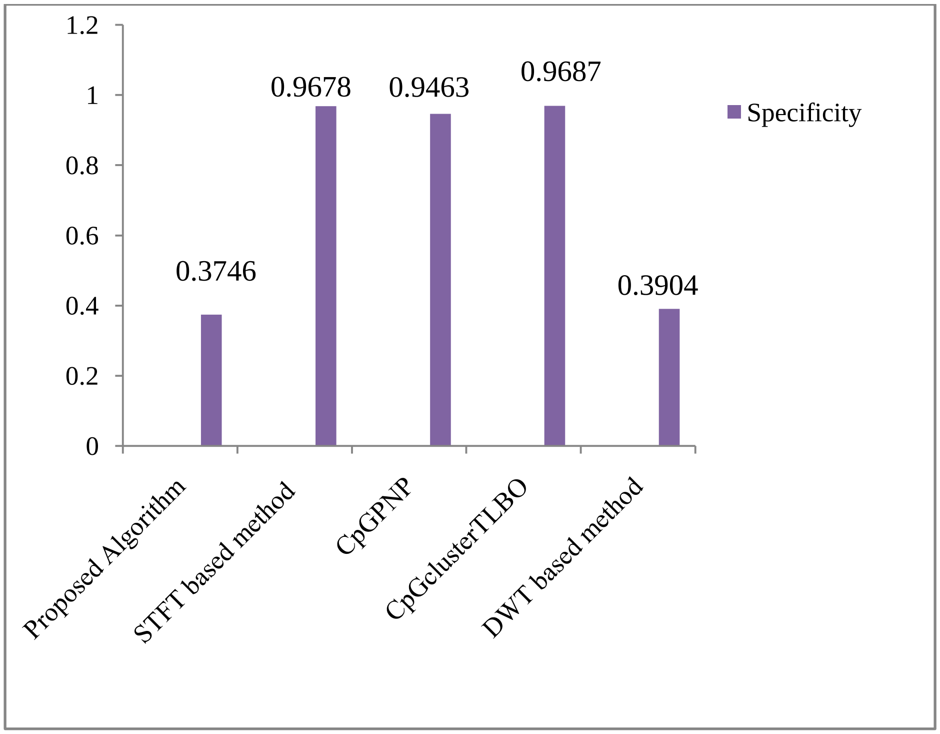


Figure 5.8: Value of Specificity obtained using all methods

Table 5.5: Performance parameters obtained using all methods for hundred DNA sequences

Performance parameter	**Methods**				
	Proposed approach	**STFT based method [158]**	**CpGPNP [122]**	**CpGclusterTLBO [129]**	**DWT based method [130]**
TP	**100559**	94193	79444	83584	76934
TN	1979090	5112809	5000128	5109201	2063066
FP	3303813	170094	283775	165139	3220837
FN	**23598**	29961	43710	34480	46223
Sn	**0.8099**	0.7587	0.6451	0.7080	0.6247
Sp	0.3746	0.9678	0.9463	**0.9687**	0.3904

The proposed approach's superiority in the context of prediction of greater number of CGIs compared to other recent state-of-art methods has been noticed from Figure 5.3, 5.6, 5.7 and Table 5.5. The proposed approach has detected the greatest number of TPs in hundred DNA sequences of human species amongst all other methods, and hence the sensitivity of proposed approach is the greatest amongst all methods. As the detection capability of any approach applied for CGIs identification is reflected by the value of Sn, hence the greatest value of Sn obtained using proposed approach clearly shows that the proposed approach is able to identify higher number of CGIs in hundred DNA sequences of human species compared to all other recent methods. However, the proposed approach has identified more number of FPs in the hundred DNA sequences as compared to other methods and subsequently the specificity of proposed approach is lesser compared to other methods.

The enhancement in sensitivity of CGIs detection obtained with the help of the proposed approach has also been examined on the basis of the % coverage of the true CGIs length. At 60%, the proposed approach has detected **134/181** CGIs, STFT based method has detected 116/181 CGIs, CpGPNP method has detected 95/181 CGIs, CpGclusterTLBO method has detected 127/181 CGIs, and DWT based method has detected 1/181 CGIs. At 70%, the proposed approach has detected **130/181** CGIs, STFT based method has detected 108/181 CGIs, CpGPNP method has detected 85/181 CGIs, CpGclusterTLBO method has detected 111/181 CGIs, and DWT based method has detected 1/181 CGIs. At 80%, the proposed approach has detected **129/181** CGIs, STFT based method has detected 105/181 CGIs, CpGPNP method has detected 69/181 CGIs, CpGclusterTLBO method has detected 98/181 CGIs, and DWT based method has detected 1/181 CGIs. At 90%, the proposed approach has detected **125/181** CGIs, STFT based method has detected 100/181 CGIs, CpGPNP method has detected 61/181 CGIs, CpGclusterTLBO method has detected 68/181 CGIs, and DWT based method has detected 1/181 CGIs. At full percentage of coverage of true CGI length i.e. 100%, the proposed approach has detected **123/181** CGIs, STFT based method has detected 91/181 CGIs, CpGPNP method has detected 50/181 CGIs, CpGclusterTLBO method has detected 40/181 CGIs, and DWT based method has detected 1/181 CGIs. The performance of proposed approach is the best in terms of detection of more number of CpG Islands at high percentage coverage of true CGIs length from 60 to 100%. However, the proposed algorithm has detected lesser number of CGIs at lower % coverage varying from 10% to 50%.

5.4 Summary

In this research work, an approach employing SVD based modified P-spectrum has been developed and proposed for the identification of CGIs in DNA sequences. The approach has been applied and compared with recent state-of-art methods on a data set of hundred DNA sequences comprising of human species downloaded from NCBI website. The sensitivity obtained using proposed approach on the whole data set is the highest amongst all methods with value 0.8099 and the proposed approach is able to capture larger number of CGIs at value of percentage coverage ranging high from 60 percent to 100 percent (full length) of true length of CGI. Therefore, the conclusion drawn is that the proposed approach has enhanced the sensitivity of CGIs detection for the data set of hundred sequences of human species in comparison with other recent state-of-art methods. However as the detection of number of false positives is little higher, the value of specificity of proposed approach is lower than other methods.

Part 2: Modified Gabor Wavelet-Transform based Algorithm for Overall Performance Improvement of CpG Islands Detection in DNA Sequences

An approach based on MGWT has been developed and proposed for the identification of CGIs in the DNA sequences in this section now. The proposed approach has been applied to overcome the limitation ‘fixed size of window’ of recent STFT based algorithm for CGIs identification. The threshold selection process has been done optimally with the help of experimental analysis. And an overall enhancement has been achieved in all performance metrics using the proposed approach.

5.5 Proposed Approach for CGIs Identification

The important feature in terms of periodicities present in CGIs has been explored with experimental validation by the authors in [158]. It has been proved in [158] that periodic pattern corresponding to periodicities 2-10 remain hidden in CGIs of DNA sequences. This feature of periodicity has been employed in this research work and the CGIs have been identified. Using this feature, the steps employed in the approach proposed in this research work have been shown in Table 5.6:

TABLE 5.6: Steps of proposed approach for CGIs detection

```
        Input: DNA sequence
1)      For imc = 1:24   % imc: integer mapping combinations
2)      For periodicities = 2:10
                Compute the power spectrums of dominating periods
                with the help of MGWT
          End (loop ended for periodicities)
                  ➢ Compute the addition of power spectrums obtained
                    corresponding to periodicities.
                ➢ Application of appropriate threshold for the
                  selection of probable CGIs.
                ➢ Post processing step using GGF criteria for the
                  verification of CGIs.
3)      Save the final spectrum for every imc^th iteration.
4)      End (loop ended for imc) compute the sum of all 24 final spectrums obtained.
```

Output: CGIs detected and evaluate the performance.

A DNA sequence possessing accession number AC105413 [1] has been chosen as an example sequence to discuss the steps and applicability of proposed approach for CGIs detection. This sequence consists of 65958 bps and possesses a CGI of length 274 bps located at 50478-50751.

5.5.1 Conversion of DNA Characters to Numerical Values

The first and necessary step after obtaining the DNA sequence from standard database is to map the characters of DNA data to numerical values. Then the DSP operations can be applied on the numerical sequence conveniently. In this research work, the 24 possible representations of integer mapping have been applied to convert the four alphabets of DNA to numerical values. The representation of these 24 combinations of integer mapping is tabulated in Table 5.7:

Table 5.7: Possible combinations of integer mapping

	Possible combinations of integer mapping for conversion of DNA characters			
	A	**G**	**T**	**C**
i=1	1	2	3	4
i=2	1	4	3	2
i=3	1	2	4	3
i=4	1	4	2	3
i=5	1	3	2	4
i=6	1	3	4	2
i=7	2	4	3	1
i=8	2	1	4	3
i=9	2	3	1	4
i=10	2	1	3	4
i=11	2	3	4	1
i=12	2	4	1	3

i=13	3	2	1	4
i=14	3	2	4	1
i=15	3	4	1	2
i=16	3	1	4	2
i=17	3	1	2	4
i=18	3	4	2	1
i=19	4	3	2	1
i=20	4	1	3	2
i=21	4	2	3	1
i=22	4	3	1	2
i=23	4	2	1	3
i=24	4	1	2	3

5.5.2 Modified Gabor Wavelet Transform (MGWT)

To capture the spectrums of dominating 2-10 periodicities, the tuning of MGWT has been done in this research work. The MGWT can be represented with the help of a numerical sequence f(u) as following:

$$F(n, a)_P = \int f(u) e^{\frac{-(u-n)^2}{2a^2}} e^{j\omega_0(u-n)} \, du \qquad (5.9)$$

The spectrums of different periodicities ‘p’ (which are 2 to 10 in this work) have been computed applying equation (5.9) and a fixed value of $\omega_0 = S/p$ has been kept to predict the periodic ‘p’ segments, where S represents the length/size of the DNA section which is under analysis. The equation (5.10) has been applied for the computation of squared complex modulus corresponding to coefficients of MGWT and the power spectrum of sequence has been obtained.

$$C(n, p)_P = |F(n, a)_P|^2 \qquad (5.10)$$

The 40 analyzing functions equivalent to scale values of 40 which are exponentially separated from 0.1 to 0.7 for every periodicity value ‘p’ have been employed in this research work. A linear addition of obtained spectrums in response to 2-10 periodicities as shown in equation (5.11) has been performed to calculate the resultant spectrum $\mathrm{RS_i}(\mathrm{n})$, where 'i' corresponds to a particular numerical mapping as shown in Table 5.7.

$$RS_i\ (n) = \sum_{\mathrm{p}=2}^{10} \mathrm{C}(\mathrm{n}, \mathrm{p}) \tag{5.11}$$

5.5.3 Application of Threshold

To select the probable CGIs spectrum from resultant spectrum $\mathrm{RS_i}(\mathrm{n})$, the experiments have been performed to obtain the optimal threshold by varying its value form 10% to 50% in a step size of 5%. The obtained value of performance metrics with respect to (w.r.t.) varying values of thresholds on the example DNA sequence AC105413 is highlighted in Table 5.8:

Table 5.8: Obtained values of performance metrics using proposed approach w.r.t. varying thresholds on example DNA sequence AC105413

Evaluation metric	**Thresholds**								
	10%	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**
TP	0	**274**	235	274	260	226	200	0	0
TN	65474	65323	65678	65022	65362	65179	65683	65683	65683
FP	209	360	5	661	321	504	0	0	0
FN	274	**0**	39	0	14	48	74	274	274
Sn	0	**1**	0.858	1	0.949	0.825	0.73	0	0
Sp	0.997	0.995	0.999	0.99	0.995	0.992	1	1	1
AC	0.498	**0.997**	0.929	0.995	0.972	0.909	0.865	0.5	0.5

The observation carried out from Table 5.8 is that the proposed approach's performance for example sequence at 15% threshold value is better in the reference of performance metrics Sn (sensitivity) and AC (accuracy) than other values of threshold considered. Although the value of Sn is 1 at threshold value of 25% which is same as value of Sn at 15% threshold as observed from Table 5.8 but value of other performance metrics Sp, and AC are lesser at 25% threshold than value at 15% threshold. Therefore, in this research work the 15% threshold value has been finalized to carry out analysis work of the proposed approach. Based upon this value of threshold, those segments of the resultant spectrum $\mathrm{RS_i}(\mathrm{n})$ whose peak value is able to cross the threshold limit have been classified as probable CGIs as represented in equation (5.12):

$$Pr_{CGI}(n) = \begin{cases} RS_i(n)\ if\ RS_i(n) > Thr \\ 0, \qquad\qquad else \end{cases} \tag{5.12}$$

where, $Thr = 0.15 * \max(RS_i(n))$

5.5.4 Verification of Probable CGIs

The GGF criterion has been applied as a post processing step for the reduction of false spectrum of probable CGIs and to finally categorize the predicted segments of probable CGIs as detected CGIs after verification step. The equation (5.13) shows the calculation of power spectrum w.r.t. verified CGIs obtained from probable CGIs.

$$\mathrm{Ve_{CGI}}\,(\mathrm{n}) = \begin{cases} \mathrm{Pr_{CGI}}\,(\mathrm{n}), \text{those segments of } \mathrm{Pr_{CGI}}\,(\mathrm{n}) \text{ which satify GGF Criteria} \\ \quad 0\,, \qquad\qquad\qquad\qquad\qquad \text{else} \end{cases} \tag{5.13}$$

5.5.5 Combine the Mapping Results corresponding to 24 combinations

The steps of the proposed approach outlined in 5.5.1 to 5.5.4 are applied to obtain the verified power spectrums of CGIs w.r.t. 24 possible combinations of integer mapping scheme. The 24 power spectrums thus obtained are then linearly added and the final power spectrum $\mathrm{Final_{CGI}}\,(\mathrm{n})$ is computed according to following equation:

$$\mathrm{Final_{CGI}}\,(\mathrm{n}) = \sum_{\mathrm{m}=1}^{24} \mathrm{Ve_{CGI}}\,(\mathrm{n})$$

The result of final power spectrum $\mathrm{Final_{CGI}}\,(\mathrm{n})$ obtained using proposed approach on example DNA sequence AC105413 is depicted in Figure 5.9.

To better understand the applicability of the proposed approach, the locations of detected CGIs segments by proposed method are checked w.r.t. true CGIs location and the obtained location results are tabulated in Table 5.9.

From Table 5.9, it has been interpreted that MGWT based proposed approach has detected the complete CGI present in the DNA sequence AC105413; however the approach has detected some false positives. The experiment performed on example DNA sequence AC105413 show the applicability of the proposed approach. Now, to prove the superiority of the proposed approach over other recent state-of-art methods of CGIs detection, the proposed approach's performance has been compared using standard evaluation metrics and the results obtained have been depicted in Table 5.10.

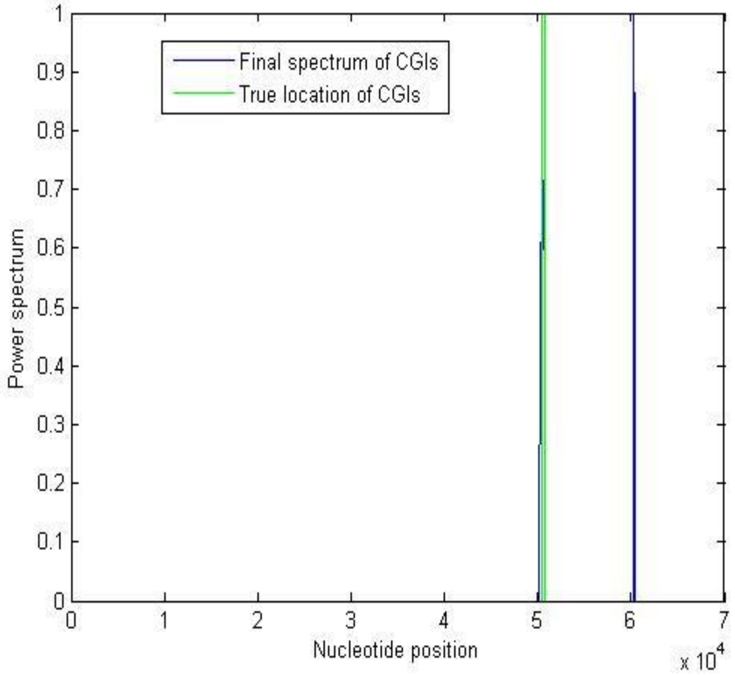


Figure 5.9: $\mathrm{Final}_{\mathrm{CGI}}\ (\mathrm{n})$ of detected CGIs

Table 5.9: Detected CGIs by proposed approach for example DNA sequence

DNA Sequence		**Start and end location of CGI in accordance with NCBI website**		**Start and end location of CGI obtained using proposed approach**	
AC105413		Start location	End location	Start location	End location
	1	50478	50751	50341	50762
				60218	60433

Table 5.10: Evaluation metrics comparison on example DNA sequence AC105413

Performance parameter	**Methods**					
	Proposed approach	**STFT based method [158]**	**CpGPNP [122]**	**CpGclusterTLBO [129]**	**DWT based method [130]**	**Modified P-spectrum based method [159]**
TP	**274**	0	0	174	164	0
TN	65323	65473	64045	64569	20822	11333
FP	360	210	1638	1114	44861	54350
FN	**0**	274	274	100	110	274
Sn	**1**	0	0	0.6350	0.5985	0
Sp	0.9945	**0.9968**	0.9751	0.9830	0.3170	0.1725
F-measure	**0.6035**	0	0	0.2228	0.0072	0
AC	**0.9973**	0.4984	0.4875	0.8090	0.4578	0.0863

The evaluation metrics on example DNA sequence AC105413 shown in Table 5.10 clearly prove that the proposed approach is much better than other recent state-of-art methods of CGIs detection in terms of sensitivity (Sn), F-measure, and accuracy (AC). Now having verified the applicability and observed the improvement in evaluation metrics of proposed approach on an example DNA sequence, the performance of proposed approach and recent state-of-art methods have been tested using a large data set of hundred DNA sequences of human species. The example DNA sequence used in this section has been considered in the whole data set of hundred DNA sequences for the computation of performance metrics in Results section.

5.6 Data Set of CGIs and Performance Metrics

5.6.1 Data Set of CGIs

A CGI data set of hundred DNA sequences of human species which is presented in Table 5.2 has been utilized in this work also.

5.6.2 Performance Metrics

The comprehensive assessment of the proposed approach and the other recent state-of-art algorithms has been carried out with the help of the evaluation parameters such as Sn (sensitivity), Sp (specificity), F-Measure [156], and AC (accuracy) [157]. The following equations describe these performance parameters:

$$\mathrm{Sn\ (sensitivity)} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{5.14}$$

$$\mathrm{Sp\ (specificity)} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \tag{5.15}$$

$$\mathrm{F-measure} = \frac{2*(\mathrm{prec}*\mathrm{rec})}{(\mathrm{prec}+\mathrm{rec})} \tag{5.16}$$

where, $prec\ \mathrm{(precision)} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$, $\mathrm{rec\ (recall)} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$

$$\mathrm{AC\ (accuracy)} = \frac{\mathrm{Sn+Sp}}{2} \tag{5.17}$$

True positive (TP), true negative (TN), false positive (FP), and false negative (FN) are the four possible performance parameters corresponding to the outcome of an approach applied for detection of CGIs. TP represents those locations of DNA sequence which are captured by the algorithm correctly and true CGIs are located at those locations. TN depicts those sections where no CGIs are captured and actual CGIs are also not located there. FP represents those erroneously identified locations of CGIs where actual CGIs are not located, and those sections of true CGI which are not captured by method are termed as FN. Using these four parameters, the evaluation metrics Sn, Sp, F-measure, and AC can be assessed. The range of value of all four evaluation metrics Sn, Sp, F-measure, and AC lies between 0-1. An approach is considered to be perfect if the value of evaluation metrics Sn, Sp, F-measure, and AC obtained using that approach is closer to 1. The parameter Sn corresponds to the percentage of TPs which have been perfectly identified by the approach; and the parameter Sp signifies the proportion of TNs which have been precisely detected by the approach. The accuracy of approach is computed using parameter F-measure which calculates the harmonic mean of prec (precision) and rec (recall). If the performance evaluation has been carried out at a single threshold only, then F-measure is s suitable parameter for assessment in place of ROC (receiver operating characteristics). The parameter which highlights the simultaneous effect of both Sn and Sp is termed as AC.

5.7 Results and Discussion

The performance comparison of the proposed approach has been carried out with five recent state-of-art methods of CGI detection on the data set of hundred DNA sequences of human species. The methods which have been used for comparison are as follows: STFT [158], CpGPNP [122], CpGclusterTLBO [129], DWT [130], and modified P-spectrum based approach for CGIs detection [159]. The value obtained of performance metrics (TP, TN, FP, FN) for the whole data set of hundred DNA sequences using all the methods has been shown in Table 5.11.

Table 5.11: Performance metrics obtained using all methods for 100 sequences of human species

Evaluation parameter	**Methods**					
	Proposed approach	**STFT based method [158]**	**CpGPNP [122]**	**CpGclusterTLBO [129]**	**DWT based method [130]**	**Modified P-spectrum based method [159]**
TP	**102443**	94193	79444	83584	76934	100559
TN	5101651	5112809	5000128	5109201	2063066	1979090
FP	181252	170094	283775	165139	3220837	3303813
FN	**21714**	29961	43710	34480	46223	23598

The observation which has been made from Table 5.11 is that the proposed approach’s value of TP is the highest compared to all other methods and consequently the value of performance metric FN is the lowest amongst all methods. However, the proposed approach's value of FP is slightly greater than STFT based method and CpGclusterTLBO method whereas this value is lesser compared to CPGPNP, DWT based method, and modified P-spectrum based method of CGIs detection. Subsequent to it, the value of TN obtained for proposed approach is slightly lesser compared to STFT based method and CpGclusterTLBO method whereas this value is greater than CPGPNP, DWT based method, and modified P-spectrum based algorithm of CGIs detection. With the help of these four performance parameters, the evaluation metrics Sn, Sp, F-Measure, and AC are computed for all methods and these are depicted in Figure 5.10-5.13:

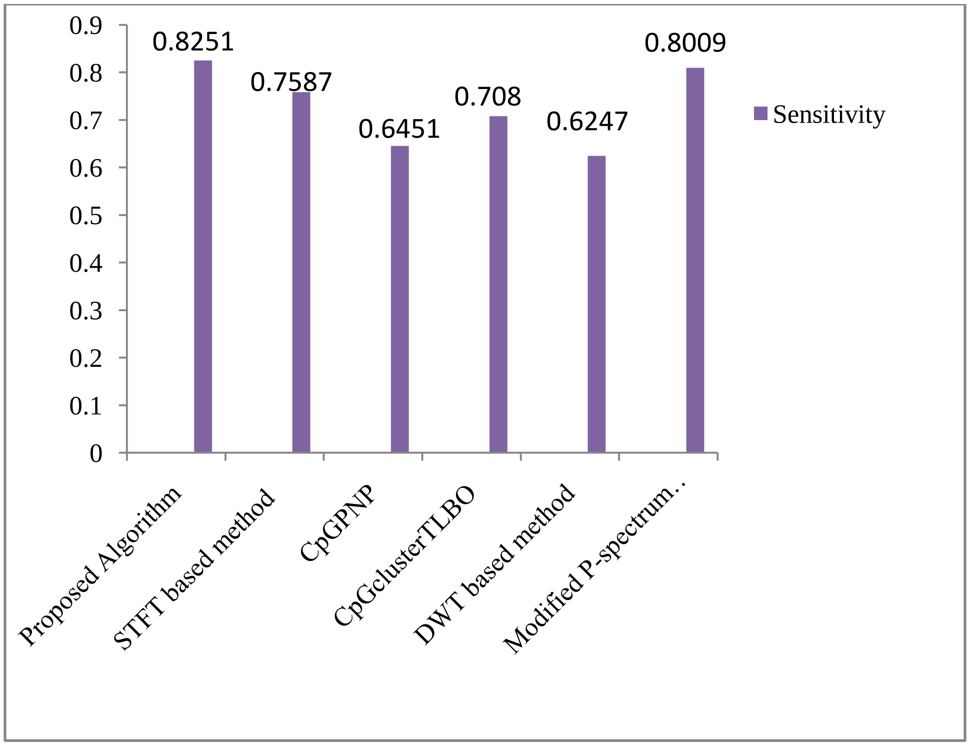


Figure 5.10: Value of Sensitivity obtained using all methods

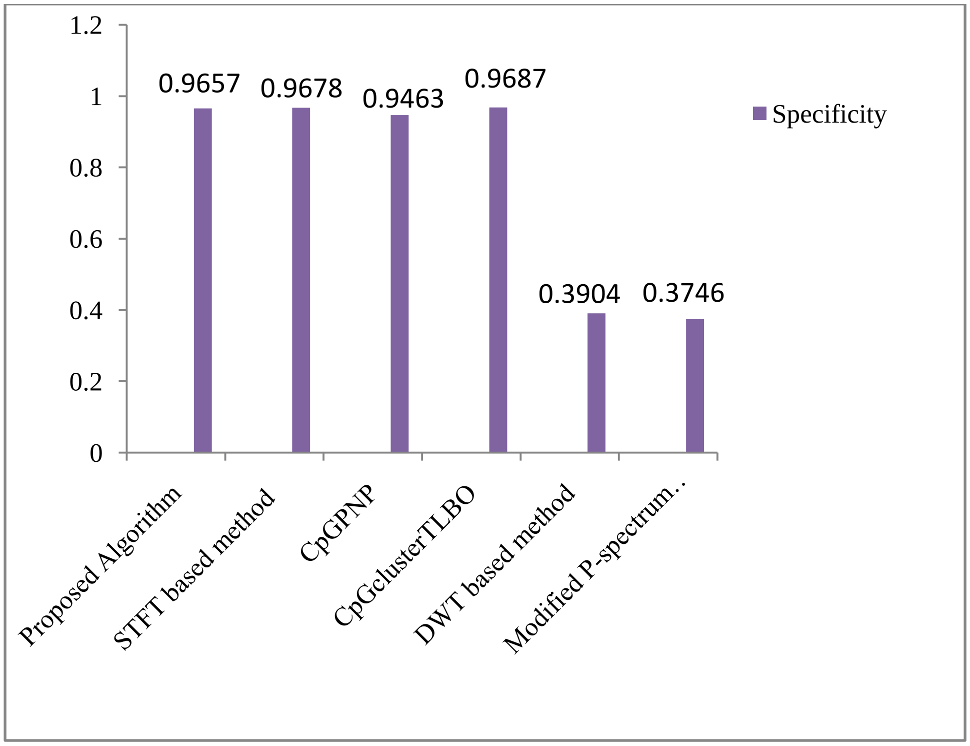


Figure 5.11: Value of Specificity obtained using all methods

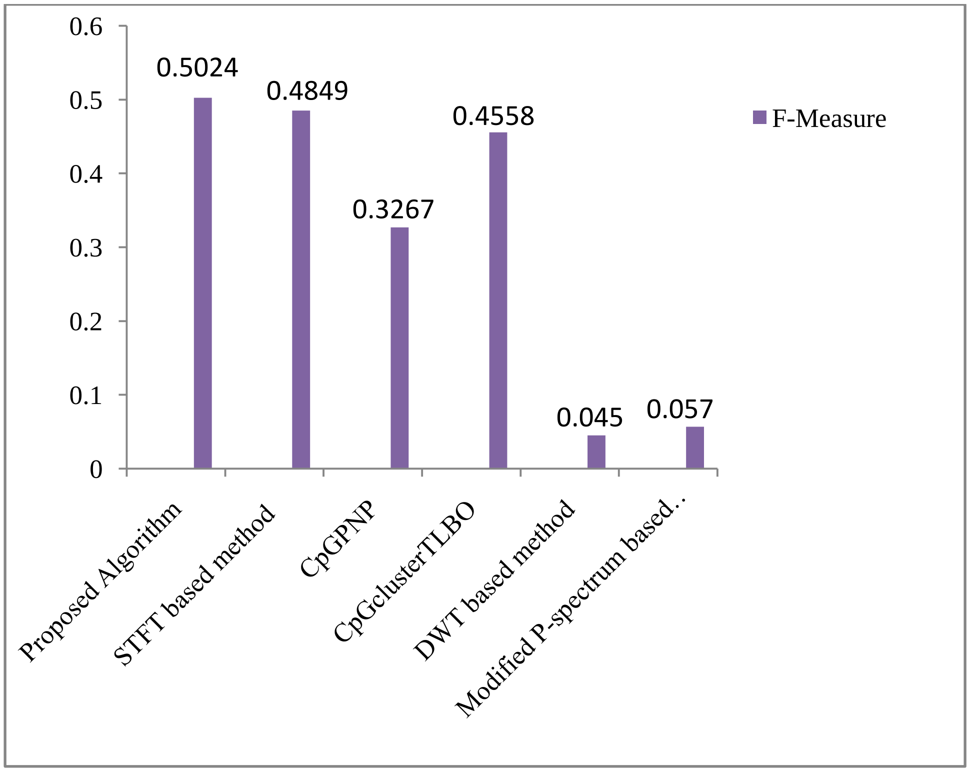


Figure 5.12: Value of F-Measure obtained using all methods

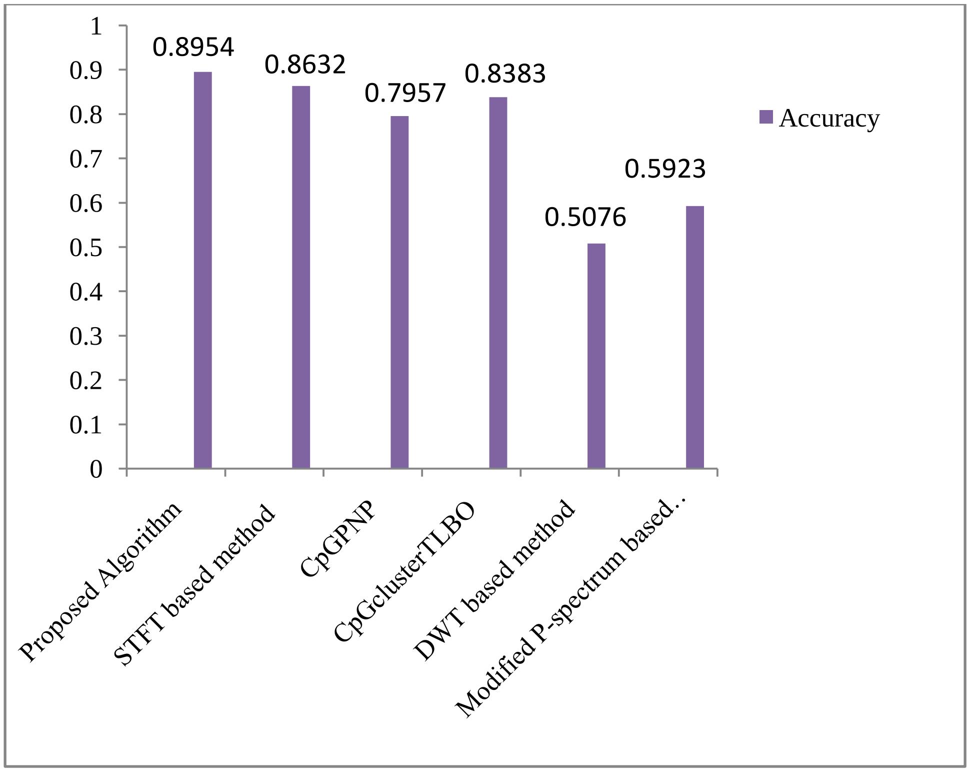


Figure 5.13: Value of Accuracy obtained using all methods

The superiority of proposed approach in terms of overall improvement in performance metrics over other state-of-art methods has been noticed in Figure 5.10, 5.12, and 5.13. The indication of detection is examined by parameter 'Sn' and the value of Sn obtained using proposed approach is the largest with value 0.8251 amongst all approaches as observed from Figure 5.10. The other performance metrics F-measure and AC obtained using the proposed approach are also observed to be the highest having value 0.5024 and 0.8954 respectively which has been noticed from Figure 5.12 and 5.13 respectively. However, the value of metric performance Sp obtained using the proposed approach (0.9657) is slightly lower than STFT based method (0.9678) and CpGclusterTLBO (0.9687) method and much higher than CpGPNP (0.9463), DWT based approach (0.3904), and modified P-spectrum (0.3746) based approach of CGIs identification as noticed from Figure 5.11.

The percentage improvement w.r.t. evaluation metrics Sn, F-measure, and AC of proposed approach over other methods has been computed as shown in Table 5.12.

Table 5.12: % improvement of proposed algorithm in value of performance metrics (Sn, F-measure, and AC) over other methods

Evaluation metric	CGI detection methods				
	STFT based method [158]	**CpGPNP [122]**	**CpGclusterTLBO [129]**	**DWT based method [130]**	**Modified P-spectrum based method [159]**
Sn	8.05%	21.82%	14.2%	24.29%	2.93%
F-measure	3.48%	34.97%	9.28%	91.04%	88.65%
AC	3.6%	11.13%	6.38%	43.31%	33.85%

5.8 Summary

In this research work, an approach employing MGWT has been developed and proposed for the identification of CGIs in DNA sequences. The approach has been applied and compared with recent state-of-art methods on a data set of hundred DNA sequences comprising of human species downloaded from NCBI website. It has been noticed that the overall improvement in the performance metrics Sn, F-measure, and AC has been obtained using proposed approach over other recent state-of-art methods. However, the value of specificity of proposed approach is almost same as that of STFT based method and CpGclusterTLBO method and much higher than CpGPNP, DWT and modified P-spectrum based methods of CGIs identification. Therefore, the conclusion drawn is that the proposed approach has improved the overall performance of CGIs detection for the data set of hundred sequences of human species over other recent state-of-art methods.

CHAPTER 6

DETECTION OF TANDEM REPEATS IN DNA SEQUENCES USING SIGNAL PROCESSING BASED APPROACHES

In this chapter the tandem repeats in DNA sequences have been detected using signal processing based algorithms. Two algorithms have been developed and proposed to identify tandem repeats. ST-IPDFT (Short-time integer period discrete Fourier transform) based proposed approach has been discussed in part 1 of the chapter. Tandem repeats detection using MGWT based proposed approach has been presented in part 2 of the chapter.

Part 1: Algorithm based on IPDFT for Identification of Tandem Repeats in DNA Sequences

6.1 Proposed Approach for Identification of Tandem Repeats

The flow graph of the approach proposed in this part of research work has been depicted in Figure 6.1:

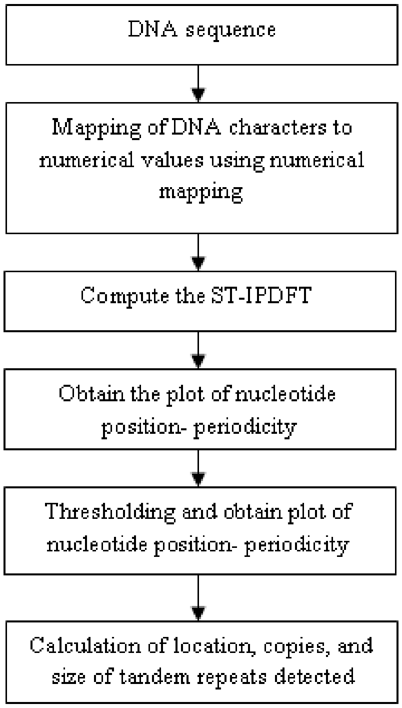

DNA sequence
Mapping of DNA characters to numerical values using numerical mapping
Compute the ST-IPDFT
Obtain the plot of nucleotide position- periodicity
Thresholding and obtain plot of nucleotide position- periodicity
Calculation of location, copies, and size of tandem repeats detected

Figure 6.1: Flow graph of the proposed approach

The description of the steps of proposed approach is as following:

a) The DNA sequence in which the tandem repeats have to be detected is taken from standard database and fed to the proposed algorithm.

b) The 4 characters of DNA sequence are then mapped to numerical values with the help of EIIP (electron ion interaction potential) numerical conversion scheme. The numerical values A= 0.1260, T= 0.1335, C=0.1340, G=0.0806 are given to the characters of DNA data and numerical sequence is thus obtained.

c) The ST-IPDFT has been then computed. For a signal s(n), the equation to represent the IPDFT [160] is as follows:

$$s_{IP}(p) = \sum_{n=0}^{N-1} s(n)e^{\frac{-j2\pi n}{p}} \ , \ p = 1, 2, 3, 4, \ldots P < N \tag{6.1}$$

where P represents the maximum period. There exists a linear relation of IPDFT with periodicity ‘p’, on the contrary there exists a linear relation of DFT with frequency.

The following equation has been then applied to calculate the ST-IPDFT for the purpose of localization of the TRs situated in the DNA Sequences.

$$s_{IP}(p, m) = \sum_{n=0}^{N-1} s(n) * w(n-m)e^{\frac{-j2\pi n}{p}} \tag{6.2}$$

where, w(n) corresponds to Hanning window whose centre at initial level is nucleotide position m=0 and thereafter it is moved by one (1) nucleotide till the last nucleotide of the DNA sequence. The length of window has been chosen as 20*p in this research work. The DNA sequence having Genbank Id X64775 [1] has been preferred as an example sequence to show the applicability of proposed approach. The nucleotide position-periodicities plot obtained has been depicted in Figure 6.2:

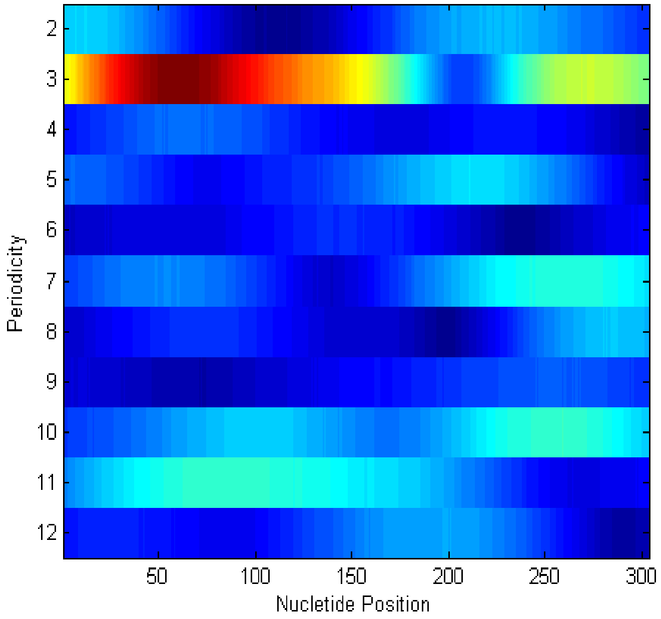


Figure 6.2: Nucleotide-position versus periodicities plot for DNA sequence X64775

d) A suitable threshold (Thr) has been applied using the thresholding equation (6.3) for the identification of location of tandem repeats of a specific periodicity.

$$\mathrm{Thr} = \mathrm{mean}(\frac{s_5(p)}{\max\left(s_5(p)\right)}) \tag{6.3}$$

where, s_5 corresponds to the sum of power spectrum as represented in equation (6.4):

$$s_5(p) = \sum_{m=0}^{M} s_{IP}(p, m) \tag{6.4}$$

The nucleotide position-periodicities plot (post thresholding) obtained after applying equation (6.5) has been shown in Figure 6.3.

$$S_{IP}(p, m) = \begin{cases} 1, & \text{if } s_{IP}(p, m) \geq \mathrm{Thr} \\ 0, & \text{if } s_{IP}(p, m) \leq \mathrm{Thr} \end{cases}$$

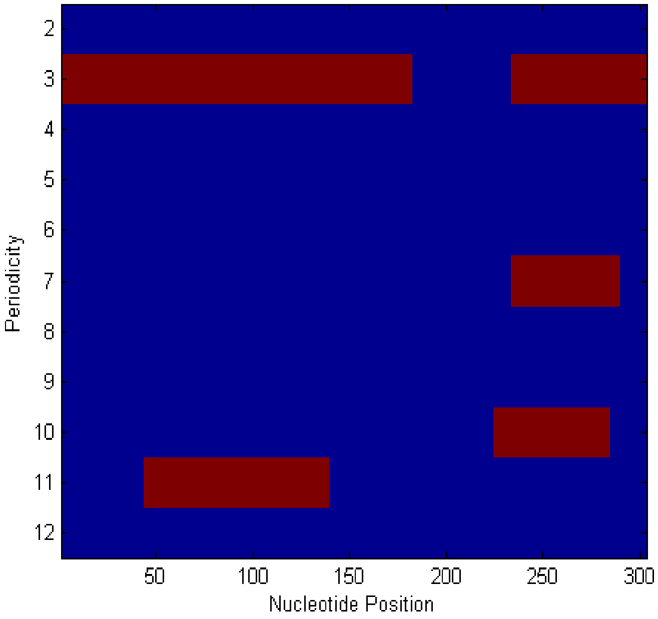


Figure 6.3: Nucleotide-position versus periodicities plot after thresholding for DNA sequence X64775

The periodicities 3, 7, 10, and 11 have been noticed from Figure 6.3 as probable tandem repeats. Periodicity 3 is located at nucleotide position 1-182 & 234-303, periodicity 7 is noticed at location 234-289, periodicity 10 is noticed at location 225-284, and nucleotide position 44-139 corresponds to periodicity 11.

e) The probable tandem repeats captured after thresholding step are then verified using verification step employing the approach proposed by Boeva *et al.* [161]. The details of tandem repeats after verification step is represented in Table 6.1 as follows:

Table 6.1: Result of verification step for probable tandem repeats

Sr. No.	**Periodi -city**	**Probable tandem repeats captured after thresholding step**		**Verification of captured probable tandem repeats**		
		Position of base pairs	**Pattern**	**Location**	**Patterns**	**No. of copies**
1	3	1-182	ATGGAGAGCGACTGC CAGTTCTTGGTGGCGC CGCCGCAGCCGCACA TGTACTACGACACGGC GGCGGCGGCGGTGGA CGAGGCGCAGTTCTTG CGGCAGATGGTGGCC	19-24	GTT CTT	02
				25-30	GGT GGC	02
				31-45	GCC GCC GCA	05

			GCGGCGGATCACCACGCGGCCGCCGCTGGGAGAGGAGGCGGCGACGGCGACGGCGGCGGCGGCGGCGGCGGCG		GCC GCA	
				50-58	TAC TAC GAC	03
				60-77	CGG CGG CGG CGG CGG TGG	06
				78-83	ACG AGG	02
				89-94	TTC TTG	02
				102-107	TGG TGG	02
				108-116	CCG CGG CGG	03
				117-123	ATC ACC	02
				125-136	GCG GCC GCC GCT	04
				142-183	AGG AGG CGG CGA CGG CGA CGG CGG CGG CGG CGG CGG CGG CGG	14
		234-303	AGACGCGTTCCACGCGCGGCGGGCCAAGCTGGAGCCGCGGGAGAAGGCGGACGTGGCGCGGGAGCTCGGG	250-255	CGG CGG	02
				268-273	CCG CGG	02
				274-279	GAG AAG	02
2	7	234-289	AGACGCGTTCCACGC	Discarded	Discarded	Discarded

			GCGGCGGGCCAAGCTGGAGCCGCGGGAGAAGGCGGACGTGG			
3	10	225-284	GGTCGCTGGAGACGCGTTCCACGCGCGGCGGGCCAAGCTGGAGCCGCGGGAGAAGGCGGA	Discarded	Discarded	Discarded
4	11	11-139	GACTGCCAGTTCTTGGTGGCGCCGCCGCAGCCGCACATGTACTACGACACGGCGGCGGCGGCGGTGGACGAGGCGCAGTTCTTGCGGCAGATGGTGGCCGCGGCGGATCACCACGCGGCCGCCGCTGGG	Discarded	Discarded	Discarded

It has been noticed from Table 6.1 that the proposed algorithm has captured periodicity 3 correctly whereas other probable periodicities 7, 10, and 11 are false and hence have been discarded after verification step.

6.2 Performance Comparison of Proposed Approach with Other Methods

The performance assessment of the proposed approach has been done on DNA sequence X64775 [1]. The comparison of performance of proposed approach with other methods has also been computed and the comparison result has been represented in Table 6.2.

Table 6.2: Comparison of results of proposed approach with other state-of-art methods

Periodicity	Method	Position of base pairs post thresholding step	Positions of base pairs post verification step	Consensus Pattern	Copies	Total Copies
3	Proposed approach	1-182	19-24	GTT CTT	02	53
			25-30	GGT GGC	02	
			31-45	GCC GCC GCA GCC GCA	05	
			50-58	TAC TAC GAC	03	

			60-77	CGG CGG CGG CGG CGG TGG	06	
			78-83	ACG AGG	02	
			89-94	TTC TTG	02	
			102-107	TGG TGG	02	
			108-116	CCG CGG CGG	03	
			117-123	ATC ACC	02	
			125-136	GCG GCC GCC GCT	04	
			142-183	AGG AGG CGG CGA CGG CGA CGG CGG CGG CGG CGG CGG CGG CGG	14	
			250-255	CGG CGG	02	
			268-273	CCG CGG	02	
			274-279	GAG AAG	02	
	Adaptive S-transform [141]	19-44	20-25	TTC(TTG)	02	48
			25-42	GCC	06	
		61-86	61-79	GGC	07	
		89-104	89-94	TTC(TTG)	02	
			94-99	GCG(GCA)	02	
		108-122	108-116	CGG	03	
			117-122	ATC(ACC)	02	
		125-135	125-135	CCG	03	
		141-149	141-149	GAG	03	
		160-186	160-186	CGG	09	
		194-207	194-199	AGG(AAG)	02	
			199-204	GCG	02	
		211-223	211-219	GGA	03	

		274-283	274-279	GAG(AAG)	02	
	EMWD [162]	57–72	57–72	CGG	5.5	21
		140–187	140–187	GGC	15.5	
	Parametric Spectral Estimation [6]	45-90	49-57	TAC	03	24.7
			59-76	CGG	06	
		140-200	141-188	GGC	15.7	

The comparison of proposed approach with other state-of-art methods in terms of detection of number of copies of periodicity 3 has been represented in Figure 6.4.

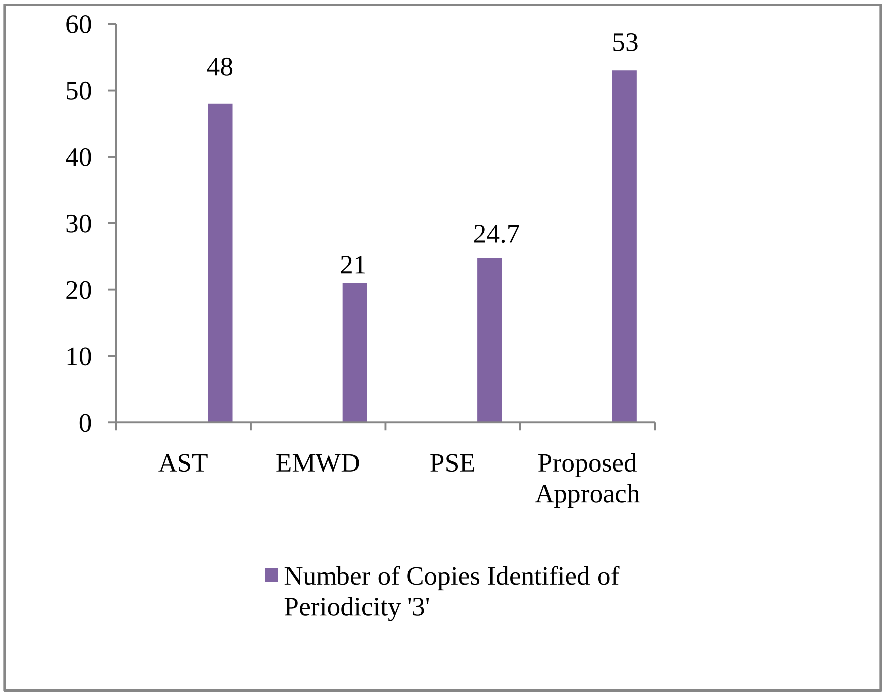


Figure 6.4: Proposed approach's comparison in terms of detection of number of copies with state-of-art methods

It has been noticed from Figure 6.4 and Table 6.2 that the proposed approach's performance is better in terms of detection of more number of copies of periodicity 3 in comparison with all other state-of-art methods in DNA sequence X64775.

6.3 Summary

The tandem repeats situated in the DNA sequences have been detected successfully using the approach presented in this part of the chapter. The conclusion drawn is that the proposed approach's performance in terms of identification of number of copies is better as compared to other state-of-art methods considered in this part of research work for comparison purpose. The fixed value of length of window is considered as the limitation of the proposed approach.

Part 2: Algorithm based on MGWT (Modified Gabor Wavelet Transform) for Identification of Tandem Repeats in DNA Sequences

6.4 Proposed Approach for Identification of Tandem Repeats

The flow graph of the approach proposed in this part of research work has been depicted in Figure 6.5:

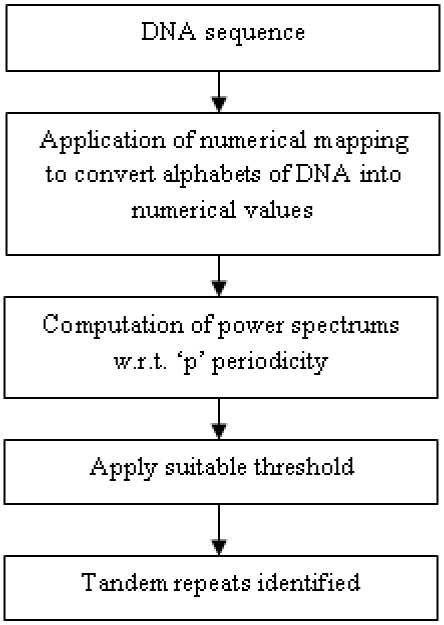

DNA sequence
Application of numerical mapping to convert alphabets of DNA into numerical values
Computation of power spectrums w.r.t. 'p' periodicity
Apply suitable threshold
Tandem repeats identified

Figure 6.5: Flow graph of the proposed approach

The description of the steps of proposed approach is as following:

a) The DNA sequence in which the tandem repeats have to be detected is taken from standard database and fed to the proposed algorithm.

b) The 4 characters of DNA sequence are then mapped to numerical values with the help of binary numerical conversion scheme [69].

c) The MGWT has been then applied for the computation of component of periodicity ‘p’ spectrum at each nucleotide’s position. For a numerical sequence ‘b(x)’, the MGWT can be represented as [98] :

$$B(n,a)_P = \int b(x) e^{\frac{-(x-n)^2}{2a^2}} e^{j\omega_0(x-n)}\, dx \qquad (6.5)$$

The spectrums of different periodicities ‘p’ (which are 2 to 12 in this work) have been computed applying equation (6.5) and a fixed value of $\omega_0 = \mathrm{Len}/p$ has been kept to predict the periodic ‘p’ segments, where ‘Len’ represents the length/size of the DNA section which is under analysis. The equation (6.6) has been applied for the computation of squared complex modulus corresponding to coefficients of MGWT and the power spectrum of sequence has been obtained.

$$P(n,p)_P = |B\,(n,a)_P|^2 \qquad (6.6)$$

The spectrum computed in equation (6.6) has to be projected on the position axis to detect the periodicity ‘p’ component at each nucleotide position. The equation (6.7) has been then utilized to compute this projection spectrum for every periodicity ‘p’ component for a DNA sequence having length ‘Len’.

$$C(n)_P = \sum_a |B(n,a)_P|^2 \text{ , n=1……Len} \qquad (6.7)$$

The DNA sequence having Genbank Id X64775 [1] has been preferred as an example sequence to show the applicability of proposed approach. The nucleotide position-periodicities plot obtained for the visualization of tandem repeats of varying periodicities in DNA sequence X64775 has been depicted in Figure 6.6:

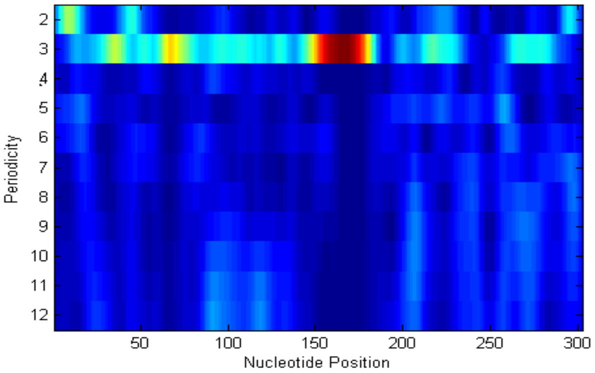


Figure 6.6: Nucleotide-position versus periodicities plot for DNA sequence X64775

d) The information regarding the starting and ending location of the tandem repeats is not obtained from the nucleotide position –periodicities plot depicted in Figure 6.6. Hence, a suitable fixed value (0.35) of threshold selected empirically has been then applied to binarize the plot obtained in Figure 6.6 and the plot obtained after thresholding step is represented in Figure 6.7.

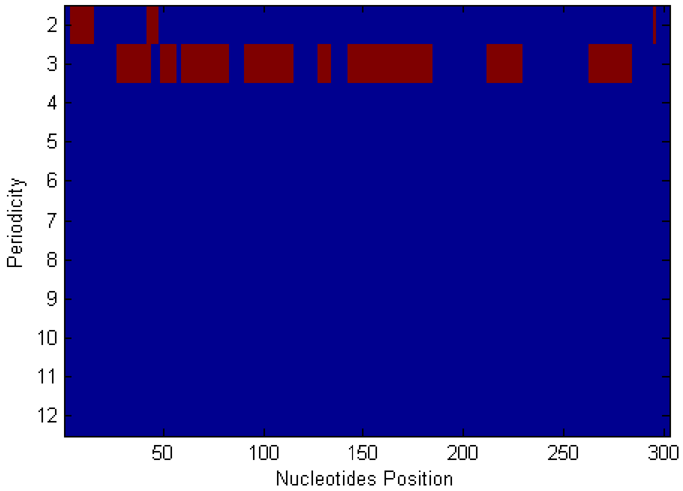


Figure 6.7: Nucleotide-position versus periodicities plot for DNA sequence X64775 after thresholding step (threshold value fixed as 0.35)

6.5 Discussion of Results and Performance Comparison of Proposed Approach with Other Methods

Forty (40) analyzing functions equivalent to scale values of 40 which are exponentially alienated from 0.2 to 0.7 for every periodicity value 'p' have been employed in this research work. The limit of these functions is set to 120 sequence points in length. The result obtained on DNA sequence X64775 shown in Figure 6.6 reveals that periodicities 2 and 3 have been captured using proposed MGWT based algorithm. Various patterns of tandem repeats having perfect and imperfect patterns corresponding to periodicity 2 and 3 have been detected using proposed approach as noticed from Figure 6.7 which has been obtained using fixed threshold value of 0.35. The exact location of these detected tandem repeats is presented in Table 6.3. Also, the performance assessment and comparison of proposed approach with other state-of-art methods has been computed and the comparison result has been represented in Table 6.3.

Table 6.3: Comparison of results of proposed approach with other state-of-art methods

Method	**Periodicity**	**Location of detected periodicity in DNA sequence ‘X64775’**	**Number of copies corresponding to periodicity**	**Nucleotides’ pattern**
Proposed approach	2	4-15 42-47 294-296	6 3 2	GA CA GA
	3	27-43 49-56 59-82 91-114 127-133 142-184 212-229 263-283	6 3 8 8 2 15 6 7	GCC TAC CGG GGC GCC GGC GGA/GGT GGA (consensus pattern)
IPDFT based method [163]	3	19-24 25-30 31-45 50-58 60-77 78-83 89-94 102-107 108-116 117-123 125-136 142-183 250-255 268-273 274-279	2 2 5 3 6 2 2 2 3 2 4 14 2 2 2	GTT/CTT GGT/GGC GCC/GCC/GCA/GCC/GCA TAC/TAC/GAC CGG/CGG/CGG/CGG/CGG/TGG ACG/AGG TTC/TTG TGG/TGG CCG/CGG/CGG ATC/ACC GCG/GCC/GCC/GCT AGG/AGG/CGG/CGA/CGG/CGA/CGG/CGG/CGG/ CGG/CGG/CGG/CGG/CGG CGG/CGG CCG/CGG GAG/AAG
Parametric Spectral Estimation [6]	3	49-57 59-76 141-188	3 6 15.7	TAC CGG GGC
Tandem Repeats Finder [134]	3	145-188	14 .33	GGC
S-transform based method [143]	3 3 3	27-37 59-71 146-183	4 4 13	CGC CGG GGC

It has been noticed from Table 6.3 that the proposed approach has detected periodicities 2 and 3 in DNA sequence X64775, whereas other state-of-art methods such as IPDFT [163] based approach, Parametric Spectral Estimation [6], Tandem Repeats Finder [134], and S- transform based approach [143] have identified periodicity 3 only and these methods have not captured period 2 in DNA sequence X64775. Moreover, the proposed approach has identified total 55 copies of period 3; whereas IPDFT based approach[163], Parametric Spectral Estimation [6], Tandem Repeats Finder [134], and S- transform based approach 143] have detected 53, 24.7, 14.33, and 21 number of copies of period 3 respectively. The proposed approach has captured total 11 number of copies corresponding to periodicity 2 whereas none of the state-of-art methods has detected period 2 in DNA sequence X64775.

6.6 Summary

In this research work, an MGWT based approach has been developed and proposed for the detection of tandem repeats and the patterns of repeats with reference to their periodicity and exact position have been visualized. The proposed approach has identified perfect and imperfect tandem repeats both. The proposed approach has captured one extra periodicity corresponding to period 2 which remained undetected by other state-of-art methods. Also, the number of total copies of periods identified by proposed approach is more in comparison with other state-of-art methods.

CHAPTER 7

CONCLUSION AND FUTURE WORK

The main aim of this research work is to develop accurate and efficient signal processing based approaches for the detection and localization of hidden patterns in the DNA sequences. Sequencing of genome and annotation thereafter generates a large amount of annotated genomic data. Development of computational approaches to extract the useful information inside the hidden patterns of annotated genomic data is a great help for the medical society. An important region of gene which is responsible for the synthesis of various proteins in organisms is termed as protein-coding regions or exons. But the process of mutation in the DNA sequence may change the normal protein formation to aberrant protein synthesis and that may lead to development of dangerous diseases. Therefore, the accurate identification of exons is considered highly important. Most of the signal processing based approaches developed so far are transform based. The transformation of domain may result in the loss of very important feature hidden in the signal such as exons. The solution to this issue has been provided in this work by developing a modified P-spectrum based algorithm for the identification of exons. Also, the selection of an appropriate length of window has always remained a challenging task in the detection of exons. This issue has been resolved in the proposed algorithm by developing an optimal window based algorithm in which optimal window length according to the characteristics of DNA sequence has been chosen for every sequence. Moreover, some approaches developed till now have identified short exons only and some other are able to detect large size exons only. The proposed algorithm is able to detect exons of short and large size as well.

Detection of CpG Islands accurately in the DNA sequences is highly essential as the contribution of CpG Islands in finding the epigenetic reasons of cancer is of great significance. The important contribution in terms of revealing the periodicities present in the CpG Islands with experimental proofs is being provided in this research work employing short-time Fourier transform based approach. Also, the selection of a particular numerical mapping technique affects the performance of detection of CpG Islands. Experiments have been performed using existing mapping schemes and thereafter a mapping scheme employing 24 possible combinations of integer mapping to reduce the nucleotide bias effect has been used in the proposed algorithm for detection of CpG Islands. A self created data set of hundred DNA sequences for CpG islands identification has been further contributed in the proposed algorithm. Further improvisation in the detection of CpG Islands by enhancing the sensitivity has been proposed by modified P-spectrum based algorithm and an overall improvement has been achieved with the help of modified Gabor Wavelet transform based approach proposed for the detection of CpG Islands.

Another important hidden pattern in DNA sequences which is associated with various neurodegenerative diseases, useful in the prediction of social behavior and DNA forensic analysis is short tandem repeats known as microsatellites. Microsatellites are characterized by regions having 2 to 8 bps periodicities. Approaches based upon integer period discrete Fourier transform and modified Gabor Wavelet transform have been proposed in this research work for the detection of microsatellites. The proposed approaches have identified the microsatellites successfully.

There exists a potential for expansion and improvisation of the algorithms proposed in this research work. The future directions in which the research work can be pursued are as following:

1) Classification of detected CpG Islands as methylated or non-methylated.
2) Identification of single nucleotide polymorphism in DNA sequences.
3) Identification of splice sites in DNA sequences.
4) Identification of hot spots in proteins.

LIST OF PUBLICATIONS

Journals:

1. P. Garg and S. D. Sharma, "Identification of CpG Islands in DNA Sequences Using Short-time Fourier Transform*," Interdisciplinary Sciences: Computational Life Sciences*, Springer, vol. 12, issue 3, September 2020, pp. 355-367.

   [SCIE & SCOPUS], IF:2.233

2. P. Garg and S. D. Sharma, "Modified P-Spectrum based Approach to Enhance Sensitivity for the detection of CpG Islands detection in Human DNA Sequences," *Biomedical Engineering: Applications, Basis and Communications*, World Scientific Publisher, vol. 34, no. 1, p. 2150052, February 2022. **[ESCI & SCOPUS]**

3. P. Garg and S. D. Sharma, "CpG Islands Detection in DNA Sequences Using Wavelet Transform", *International Journal of Computing and Digital Systems*, University of Bahrain Publisher, vol. 11, no. 1, pp. 1093-1105, March 2022. **[SCOPUS]**

4. P. Garg and S. D. Sharma, "Sensitivity Enhancement of DWT based Algorithm for detection of CpG Islands in DNA Sequences," *Procedia Computer Science*, Elsevier, 2020; 167 (2020), pp. 1829-1838, 2020. **[SCOPUS]**

5. P. Garg and S. D. Sharma, "Optimum Window- based Modified P-Spectrum Method for the Identification of Protein-Coding Regions in DNA Sequences," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. **[SCI], IF: 3.71**

Current status: With Associate Editor

Conferences:

1. P. Garg and S. D. Sharma, "CpG Island Identification in DNA Sequences using Modified P-Spectrum based Algorithm," *Journal of Physics: Conference Series (ICASSCT 2021)*, 1921 (2021) 012042, IOP Science, pp. 1-13. **(Got the Best Paper Award)**

   [SCOPUS]

2. P. Garg and S. D. Sharma, "MGWT based Algorithm for Tandem Repeats Detection in DNA Sequences" *in ISPCC 2019*, 10-12 October 2019, pp. 196-199. **[SCOPUS]**

<u>Book Chapter:</u>

1. S. D. Sharma and P. Garg, “Integer Period Discrete Fourier Transform based Algorithm for the Identification of Tandem Repeats in DNA Sequences”, in *Machine Learning, Big Data, and IoT for Medical Informatics*, Elsevier, June 2021 pp. 311-325.
[SCOPUS]

REFERENCES

[1] National Centre for Biotechnology Information (NCBI), Available online: www.ncbi.nlm.nih.gov. 2021.

[2] A. A. Tsonis, J. B. Elsner, and P. A. Tsonis, “Periodicity in DNA Coding Sequences: Implications in Gene Evolution”, *Journal of Theoretical Biology*, vol. 151, issue 3, pp. 323-331, 1991.

[3] R. A. Tahir, D. Zheng, A. Nazir, and H. Qing, “A Review of Computational Algorithms for CpG Islands Detection,” *Indian Academy of Sciences*, vol. 44, issue 6, pp. 1-11, 2019.

[4] Y. Wang, and F. Leung, “An Evaluation of New Criteria for CpG Islands in the Human Genome as Gene Markers,” *Bioinformatics*, vol. 20, issue 7, pp. 1170-1177, 2004.

[5] R. Kakumani, O. Ahmad, and V. Devabhaktuni, “Identification of CpG Islands in DNA Sequences using Statistically Optimal Null Filter,” *Eurasip Journal on Bioinformatics and System Biology*, vol. 2012, issue 1, pp. 1-14, 2012.

[6] H. Zhou, L. Du, and H. Yan, “Detection of Tandem Repeats in DNA Sequences based on Parametric Spectral Estimation,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 5, pp. 747-755, 2009.

[7] H. Ellegren, “Microsatellites: Simple Sequences with Complex Evolution,” *Nature Reviews Genetics*, vol. 5, no. 6, pp. 435-445, 2004.

[8] J. D. Watson, and F. H. Crick, “Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid,” *Nature*, vol. 171, pp. 737-738, 1953.

[9] P. P. Vaidyanathan, and B. J. Yoon, “Genomics and Proteomics: A Signal Processing Tour,” *IEEE Circuits and Systems Magazine*, vol. 4, issue 4, pp. 6-29, 2004.

[10] D. Anastassiou, “Genomic Signal Processing,” *IEEE Signal Processing Magazine*, vol. 18, issue 4, pp. 8-20, 2001.

[11] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, “Molecular Cell Biology,” 4th Edition, New York: W. H. Freeman, 2000.

[12] A Brief Guide to Genomics-National Human Genome Research Institute available at https://www.genome.gov/18016863, 2011.

[13] E. Lamm, O. Harman, and S. J. Veigl, “Before Watson and Crick in 1953 Came Friedrich Miescher in 1869,” *Genetics*, vol. 215, issue 2, pp. 291-296, 2020.

[14] R. Dahm, “Friedrich Miescher and the discovery of DNA,” *Developmental Biology*, vol. 278, issue 2, pp. 274-288, 2005.

[15] P. A. Levene, “The Structure of Yeast Nucleic Acid: IV. Ammonia Hydrolysis,” *Journal of Biological Chemistry*, vol. 40, issue 2, pp. 415-424, 1919.

[16] E. Chargaff, “Chemical Specificity of Nucleic Acids and Mechanism of Their Enzymatic Degradation,” *Experientia*, vol. 15, issue 6, pp. 201-209, 1950.

[17] International Human Genome Sequencing Consortium. Initial Sequencing and Analysis of the Human Genome. *Nature*, vol. 409, pp. 860-921, 2001.

[18] J. Shendure, R. D. Mitra, C. Varma, and G. M. Church, “Advanced Sequencing Technologies: Methods and Goals,” *Nature Reviews Genetics*, vol. 5, issue 5, pp. 335-344, 2004.

[19] W. J. Min, G. Haegeman, M. Ysebaert, and W. Fiers, “Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein,” *Nature*, vol. 237, pp. 82-88, 1972.

[20] W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaet, W. J. Min, F. Molemans, A. Raeymaekers, A. B. Van Den, G. Volckaet, and M. Ysebaert, “Complete Nucleotide Sequence of Bacteriophage MS2 RNA: Primary and Secondary Structure of the Replicase Gene,” *Nature*, vol. 260, pp. 500-507, 1976.

[21] R. Wu, “Nucleotide Sequence Analysis of DNA:I. Partial Sequence of the Cohesive Ends of Bacteriophage λ and 186 DNA,” *Journal of Molecular Biology*, vol. 51, issue 3, pp. 501-521, 1970.

[22] R. Wu, “Nucleotide Sequence Analysis of DNA,” *Nature New Biology*, vol. 236, pp. 198-200, 1972.

[23] F. Sanger, S. Nicklen, and A. R. Coulson, “DNA Sequencing with Chain-Terminating Inhibitors,” *Proceedings of the Natural Academy of Sciences USA*, vol. 74, issue 12, pp. 5463-5477, 1977.

[24] A. M. Maxam, and W. Gilbert, “A New Method for Sequencing DNA,” *Proceedings of the Natural Academy of Sciences USA*, vol. 74, issue 2, pp. 560-564, 1977.

[25] W. Gilbert, “DNA Sequencing and Gene Structure,” Nobel Lecture, 8 December 1980.

[26] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law, “Comparison of Next-Generation Sequencing Systems,” *Journal of Biomedicine and Biotechnology*, vol. 2012, pp. 1-11, 2012.

[27] F. S. Collins, M. Morgan, and A. Patrinos, “The Human Genome Project: Lessons from Large-Scale Biology,” *Science*, vol. 300, pp. 286-290, 2003.

[28] M. Yandell, and E. Daniel, “A Beginner’s Guide to Eukaryotic Genome Annotation,” *Nature Reviews Genetics*, vol. 13, no. 5, pp. 329-342, 2012.

[29] S. B. Arniker, “Human Promoter Prediction using Numerical Representation,” *Biology*, 2010.

[30] L. W. Barrett, S. Fletcher, and S. D. Wilton, “Regulation of Eukaryotic Gene Expression by the Untranslated Gene Regions and other Non-Coding Elements,” Cellular and Molecular Life Sciences, vol. 69, no. 21, pp. 3613-3634, 2012.

[31] E. S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, “Initial Sequencing and Analysis of the Human Genome,” *Nature*, vol. 409, no. 6822, pp. 860-921, 2001.

[32] T. W. Nilsen, and B. R. Graveley, “Expansion of the Eukaryotic Proteome by Alternative Splicing,” *Nature Review Insight*, vol. 463, no. 7280, pp. 457-463, 2010.

[33] D. K. Shakya, R. Saxena, and S. N. Sharma, “Improved Exon Prediction with Transforms by De-noising Period-3 Measure,” *Digital Signal Processing*, vol. 23, no. 2, pp. 499-505, 2013.

[34] E. N. Trifonov, and J. L. Sussman, “The Pitch of Chromatin DNA is reflected in its Nucleotide Sequence,” *Proceedings of the National Academy of Sciences*, vol. 77, no. 7, pp. 3816-3820, 1980.

[35] C. Yin, and S. S. T. Yau, “Prediction of Protein-Coding Regions by the 3-Base Periodicity Analysis of a DNA Sequence,” *Journal of Theoretical Biology*, vol. 247, no. 4, pp. 687-694, 2007.

[36] J. W. Ficket, “Recognition of Protein Coding Regions in DNA Sequences,” *Nucleic Acids Research*, vol. 10, no. 17, pp. 5303-5318, 1982.

[37] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, “Prediction of Probable Gene by Fourier Analysis of Genomic Sequences,” *Computer Applications in the Biosciences: CABIOS*, vol. 13, no. 3, pp. 263-270, 1997.

[38] S. M. Berget, C. Moore, and P. A. Sharp, “Spliced Segments at the 5’ Terminus of Adenovirus 2 Late mRNA,” *Proceedings of the Natural Academy of Sciences USA*, vol. 74, pp. 3171-3175, 1977.

[39] L. T. Chaw, R. E. Gelinas, J. R. Broker, and R. J. Roberts, “An Amazing Sequence Arrangement at the 5’ ends of Adenovirus 2 messenger RNA,” *Cell*, vol. 12, pp. 1-8, 1977.

[40] A. J. Jeffreys, and R. A. Flavell, “The Rabbit Beta-Globin Gene contains a Large Insert in the Coding Sequence,” *Cell*, vol. 12, pp. 1097-1108, 1977.

[41] M. Akhtar, J. Epps, and E. 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.

[42] P. Feng, W. Chen, and H. Lin, “Prediction of CpG Island Methylation Status by Integrating DNA Physiochemical Properties,” *Genomics*, vol. 104, pp. 229-233, 2014.

[43] M. Thangam, and B. Vanniappan, “Mining Association Rules in Dengue Gene Sequence with Latent Periodicity,” *Computational Biology Journal*, vol. 2015, pp. 1-10, 2015.

[44] D. Sharma, B. Issac, G. P. S. Raghava, and R. Ramaswamy, “Spectral Repeat Finder (SRF): Identification of Repetitive Sequence using Fourier Transformation,” *Bioinformatics*, vol. 20, no. 9, pp. 1405-1412, 2004.

[45] A. K. Brodzik, “Quaternionic Periodicity Transform: An Algebraic Solution to the Tandem Repeat Detection Problem,” *Bioinformatics*, vol. 23, no. 6, pp. 694-700, 2007.

[46] R. Gupta, D. Sarthi, A. Mittal, and K. Singh, “A Novel Processing Measure to Identify Exact and Inexact Tandem Repeat Patterns in DNA Sequences,” *Eurasip Journal on Bioinformatics and System Biology*, vol. 2007, pp. 1-7, 2007.

[47] H. Zhou, L. Du, and H. Yan, “Detection of Tandem Repeats in DNA Sequences based on Parametric Spectral Estimation,” *IEEE Transactions on Information Technology in Medicine*, vol. 13, no. 5, pp. 747-755, 2009.

[48] S. D. Sharma, R. Saxena, S. N. Sharma, and A. K. Singh, “Short Tandem Repeats Detection in DNA Sequences using Modified S-Transform,” *International Journal of Advances in Engineering & Technology*, vol. 8, issue 2, pp. 233-245, 2015.

[49] S. M. Mirkin, “Expandable DNA Repeats and Human Disease,” *Nature*, vol. 447, no. 7147, pp. 932-940, 2007.

[50] K. Usdin, “The Biological Effects of Simple Tandem Repeats: Lessons from the Repeat Expansion Diseases,” *Genome Research*, vol. 18, no. 7, pp. 1011-1019, 2008.

[51] Z. Yu, and N. M. Bonini, “Modeling Human Trinucleotide Repeat Diseases in Drosophila,” *International Review of Neurobiology*, vol. 99, pp. 191-212, 2011.

[52] U. Polak, E. McIvor, S. Y. Dent, R. D. Wells, and M. Napierala, “Expanded Complexity of Unstable Repeat Diseases,” *Biofactors*, vol. 39, no. 2, pp. 164-175, 2013.

[53] M. Akhtar, “Genomic Sequence Processing: Gene Finding in Eukaryotes,” *PhD Dissertation, The University of New South Wales*, 2008.

[54] V. V. Solovyev, A. A. Salamov, and C. B. Lawrence, “Identification of Human Gene Structure using Linear Discriminant Functions and Dynamic Programming,” *in the Proceedings of International Conference on Intelligent Systems for Molecular Biology*, vol. 3, pp. 367-375, 1995.

[55] D. Kulp, D. Haussler, M. G. Reese, and A. FH. Eeckman, “A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA,” *in the Proceedings of International Conference on Intelligent Systems for Molecular Biology,”* vol. 4, pp. 134-142, 1996.

[56] A. Krogh, “Two Methods for Improving Performance of an HMM and their Application for Gene Finding,” *in the Proceedings of International Conference on Intelligent Systems for Molecular Biology*,” vol. 5, pp. 179-186, 1997.

[57] C. Burge, “Identification of Genes in Human Genomic DNA,” *PhD Thesis, Stanford University*, 1997.

[58] M. Q. Zhang, “Identification of Protein Coding Regions in the Human Genome by Quadratic Discriminant Analysis,” *Proceedings of the National Academy of Sciences*, vol. 94, no. 2, pp. 565-568, 1997.

[59] S. L. Salzberg, “A Method for Identifying Splice Sites and Translational Start Sites in Eukaryotic mRNA,” *Computer Applications in Biosciences*, vol. 13, no. 4, pp. 365-376, 1997.

[60] A. V. Lukashin, and M. Borodovsky, “Genemark. HMM: New Solutions for Gene Finding,” *Nucleic Acids Research*, vol. 26, no. 4, pp. 1107-1115, 1998.

[61] M. G. Reese, D. Kulp, H. Tammana, and D. Haussler, “Genie- Gene Finding in Drosophila Melanogaster,” *Genome Research*, vol. 10, no. 4, pp. 529-538, 2000.

[62] G. Parra, E. Blanco, and R. Guigo, “Geneid in Drosophila,” *Genome Research*, vol. 10, no. 4, pp. 511-515, 2000.

[63] S. Mario, R. Steinkamp, S. Waack, B. Morgenstern, “AUGUSTUS: A Web Server Gene Finding in Eukaryotes,” *Nucleic Acids Research*, vol. 32, no. 2, pp. 309-312, 2004.

[64] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden, “NCBI BLAST: A Better Web Interface,” *Nucleic Acids Research*, vol. 36, no. 2, pp. 5-9, 2008.

[65] K. Mukarami, and T. Takagi, “Gene Recognition by Combination of Several gene Finding Programs,” *Bioinformatics*, vol. 14, no. 8, pp. 665-675, 1998.

[66] V. Pavlovic, A. Garg, and S. Kasif, “A Bayesian Framework for Combining Gene Predictions,” *Bioinformatics*, vol. 18, no. 1, pp. 19-27, 2002.

[67] P. D. Cristea, “Conversion of Nucleotide Sequences into Genomic Signals,” *Journal of Cellular and Molecular Medicine*, vol. 6, issue 2, pp. 279-303, 2002.

[68] H. K. Kwan, and S. B. Arniker, “Numerical Representation of DNA Sequences,” *IEEE International Conference on Electro-Information Technology*, pp. 307-310, 2009.

[69] S. D. Sharma, D. K. Shakya, and S. N. Sharma, “Evaluation of DNA Mapping Schemes for Exon Detection,” in *IEEE International Conference on Computer, Communication, and Electrical Technology*, pp. 71-74, 2011.

[70] E. R. Dougherty, A. Datta, and C. Sima, “Research Issues in Genomic Signal Processing,” *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 46-68, 2005.

[71] S. A. Marhon, and S. C. Kremer, “Gene Prediction based on DNA Spectral Analysis: A Literature Review,” *Journal of Computational Biology*, vol. 18, no. 4, pp. 639-676, 2011.

[72] E. N. Trifonov, H. Hanspeter, and O. Weiss, “Sequence Periodicity in Complete Genomes of Archaea Suggests Positive Super Coiling,” *Journal of Biomolecular Structure and Dynamics*, vol. 16, no. 2, pp. 341-345, 1998.

[73] V. R. Chechetkin, and A. Y. Turygin, “Size-Dependence of Three-Periodicity and Long-Range Corrections in DNA Sequences,” *Physics Letters A*, vol. 199, issues 1-2, pp. 75-80, 1995.

[74] V. V. Lobzin, and V. R. Chechetkin, “Order and Correlations in Genomic DNA Sequences: The Spectral Approach,” *Physics Uspekhi*, vol. 43, no. 1, pp. 55-78, 2000.

[75] J. V. L. Ginori, A. R. Fuentes, R. G. Abalo, and R. S. Rodriguez, “Digital Signal Processing in the Analysis of Genomic Sequences,” *Current Bioinformatics*, vol. 4, no. 1, pp. 28-40, 2009.

[76] W. Wang, and D. H. Johnson, “Computing Linear Transforms of Symbolic Signals,” *IEEE Transactions on Signal Processing*, vol. 50 , issue 3, pp. 628-634, 2002.

[77] D. Sussillo, A. Kundaje, and D. Anastassiou, “Spectrogram Analysis of Genomes,” *EURASIP Journal on Advances in Signal Processing*, vol. 2004, pp. 29-42, 2004.

[78] H. Herzel, and P. Schiega, “Periodicities of 10-11 bp as Indicators of the Supercoiled Sate of Genomic DNA,” *Journal of Molecular Biology*, vol. 343, issue 4, pp. 891-901, 2004.

[79] L. Kumar, M. Futschik, and H. Herzel, “DNA Motifs and Sequence Periodicities,” *In Silico Biology*, vol. 6, no. 1-2, pp. 71-78, 2006.

[80] L. Galleani, and R. Garello, “The Minimum Entropy Mapping Spectrum of a DNA Sequence,” *IEEE Transactions on Information Theory*, vol. 56, no. 2, pp. 771-783, 2010.

[81] Y. W. Kiyama, and R. Kiyama, “Conservation and Periodicity of DNA Bend Sites in Eukaryotic Genomes,” *DNA Research*, vol. 3, no. 1, pp. 25-30, 1996.

[82] R. Kiyama, “Periodicity of DNA Bend Sites in Eukaryotic Genomes,” *Gene Therapy & Molecular Biology*, vol. 1, pp. 641-647, 1998.

[83] A. Fukushima, M. Kinouchi, Y. Kudo, S. Kanaya, H. Mori, and T. Ikemura, “Statistical Analysis of Genomic Information: Various Periodicities in DNA Sequence,” *Genome Informatics*, vol. 12, pp. 435-436, 2001.

[84] F. Salih, B. Salih, and E. N. Trifonov, “Sequence Structure of Hidden 10.4-Base Repeat in the Nucleosomes of C. Elegans,” *Journal of Bimolecular Structure and Dynamics*, vol. 26, no. 3, pp. 273-281, 2008.

[85] M. Q. Zhang, “Computational Prediction of Eukaryotic Protein-Coding Coding Genes,” *Nature Reviews Genetics*, vol. 3, no. 9, pp. 698-709, 2002.

[86] J. H. Do, and D. K. Choi, “Computational Approaches to Gene Prediction,” *Journal of Microbiology*, vol. 44, no. 2, pp. 137-144, 2006.

[87] E. Blanco, and R. Guigo, “Predictive Methods using DNA Sequences,” *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, third edition, John Wiley & Sons, 2004.

[88] R. Guigo, “DNA Composition, Codon Usage and Exon Prediction,” *Genetic Databases*, Academic Press, pp. 53-80, 1999.

[89] F. Gao, and C. T. Zhang, “Comparison of Various Algorithms for Recognizing Short Coding Sequences of Human Genes,” *Bioinformatics*, vol. 20, no. 5, pp. 673-681, 2004.

[90] M. M. Yin, and J. T. Wang, “Genescout: A Data Mining System for Predicting Vertebrate Genes in Genomic DNA Sequences,” *Information Sciences*, vol. 163, issue 1-3, pp. 201-218, 2004.

[91] M. Borodovsky, and J. McIninch, “Genemark: Parallel Gene Recognition for both DNA Strands,” *Computer and Chemistry*, vol. 17, no. 2, pp. 123-133, 1993.

[92] A. Piovesan, F. Antonaros, L. Vitale, P. Strippoli, M. C. Pelleri, and M. Caracausi, “Human Protein-Coding Genes and Gene Feature Statistics,” *BMC Research Notes*, vol. 12, 2019.

[93] S. A. Marhon, and S. C. Kremer, “A Dynamic Representation-based, De-Novo Method for Protein-Coding Region Biological Information Detection,” *Digital Signal Processing*, vol. 46, pp. 10-18, 2015.

[94] P. P. Vaidyanathan, and B. J. Yoon, “The Role of Signal-Processing Concepts in Genomics and Proteomics,” *Journal of the Franklin Institute*, vol. 341, no. 1, pp. 111-135, 2004.

[95] D. Kotlar, and L. Yizhar, “Gene Prediction by Spectral Rotation Measure: A New Method for Identifying Protein-Coding Regions,” *Genome Research*, vol. 13, no. 8, pp. 1930-1937, 2003.

[96] T. S. Gunawan, E. Ambikairajah, and J. Epps, “A Signal Boosting Technique for Gene Prediction,” *IEEE 6th International Conference on Information, Communications & Signal Processing*, pp. 1-4, 2007.

[97] K. D. Rao, and M. N. S. Swamy, “Analysis of Genomics and Proteomics using DSP Techniques,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 55, no. 1, pp. 370-378, 2008.

[98] J. P. Mena-Chalco, H. Career, Y. Zana, and R. M. Cesar Jr., “Identification of Protein-Coding Regions using the Modified Gabor-Wavelet Transform,” *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 5, no. 2, pp. 198-207, 2008.

[99] S. S. Sahu, and G. Panda, “Identification of Protein-Coding Regions in DNA Sequences using a Time-Frequency Filtering Approach,” *Genomics Proteomics & Bioinformatics*, vol. 9, no. 1-2, pp. 45-55, 2010.

[100] I. T. Mariapushpam, and R. Sivakumar, “Improved Algorithm for Analysis of DNA Sequences Using Multiresolution Transformation,” *Scientific World Journal*, vol. 2015, pp. 1-9, 2015.

[101] M. Akhtar, E. Ambikairajah, and J. Epps, “Optimizing Period-3 Methods for Eukaryotic Gene Prediction,” *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’08)*, pp. 621-624, 2008.

[102] D. K. Shakya, R. Saxena, and S. N. Sharma, “An Adaptive Window Length Strategy for Eukaryotic CDS Prediction,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 10, no. 5, pp. 1241-1252, 2013.

[103] P. P. Vaidyanathan, and B. J. Yoon, “Digital Filters for Gene Prediction Applications,” in *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computer*, USA, pp. 306-310, 2002.

[104] P. Ramachandran, W. S. Lu, and A. Antoniou, “Filter-Based Methodology for the location of Hotspots in Proteins and Exons in DNA,” *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 6, pp. 1598-1609, 2012.

[105] M. K. Hota, and V. K. Srivastava, “Identification of Protein-Coding Regions using Antinotch Filters,” *Digital Signal Processing*, vol. 22, no. 6, pp. 869-877, 2012.

[106] R. Kakumani, V. Devabhaktuni, and O. Ahmad, “Prediction of Protein-Coding Regions in DNA Sequences using a Model-Based Approach,” *in Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 1918-1921, 2008.

[107] L. Zhang, F. Tian, and S. Wang, “A Modified Statistically Optimal Null Filter Method for Recognizing Protein-Coding Regions,” *Genomics, Proteomics & Bioinformatics*, vol. 10, no. 3, pp. 166-173, 2012.

[108] R. R. Roldan, P. B. Galvan, and J. L. Oliver, “Sequence Compositional Complexity of DNA through an Entropic Segmentation Method”, *Physical Review Letters*, vol. 80, no. 6, pp. 1344-1347, 1998.

[109] D. Nicorici, and J. Astola, “Segmentation of DNA into Coding and Non-Coding Regions based on Recursive Entropic Segmentation and Stop-Codon Statistics,” *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 1, pp. 81-91, 2004.

[110] N. Chackravarthy, A. Spanias, L. D. Iasemidis, and K. Tsakalis, “Autoregressive Modeling and Feature Analysis of DNA Sequences,” *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 1, pp. 13-28, 2004.

[111] M. K. Choong, and H. Yan, “Multi-Scale Parametric Spectral Analysis for Exon Detection in DNA Sequences based on Forward-Backward Linear Prediction and Singular Value Decomposition of the Double-Base Curves,” *Bioinformation*, vol. 2, no. 7, pp. 273-278, 2008.

[112] R. Jiang, and H. Yan, “Studies of Spectral Properties of Short Genes using the Wavelet Subspace Hilbert-Huang Transform (WSHHT),” *Physica A*, vol. 387, no. 16-17, pp. 4223-4247, 2008.

[113] S. A. Marhon, and S. C. Kremer, “Prediction of Protein-Coding Regions using a Wide-range Wavelet Window Method,” *IEEE Transactions on Computational Biology and Bioinformatics*,” vol. 13, no. 4, pp. 742-753, 2016.

[114] S. D. Sharma, R. Saxena, S. N. Sharma, and A. K. Singh, “Identification of Short Exons Disunited by a Short Intron in Eukaryotic DNA Regions,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*,” vol. 17, no. 5, pp. 1660-1670, 2020.

[115] H. Wu, B. Caffo, H. A. Jaffee, R. A. Irizarry, and A. P. Feinberg, “Redefining CpG islands using Hidden Markov Models,” *Biostatistics*, vol. 11, issue 3, pp. 499-514, 2010.

[116] M. G. Garden, and M. Frommer, “CpG Islands in Vertebrate Genomes,” *Journal of Molecular Biology*, vol. 196, issue 2, pp. 261-282, 1987.

[117] D. Takai, and P. A. Jones, “Comprehensive Analysis of CpG Islands in Human Chromosomes 21 and 22,” *Proceedings of the Natural Academy of Sciences of USA*, vol. 99, issue 6, pp. 3740-3745, 2002.

[118] N. Yu, X. Guo, A. Zelikovsky, and Y. Pan, “GaussianCpG: A Gaussian Model for Detection of CpG Island in Human Genome Sequences,” *BMC Genomics*, vol. 18, suppl. 4, 2017.

[119] L. Ponger, and D. Mouchiroud, “CpGProd: Identifying CpG Islands Associated with Transcription Start Sites in Large Genomic Mammalian Sequences,” *Bioinformatics*, vol. 18, no. 4, pp. 631-633, 2002.

[120] P. Rice, I. Longden, and A. Bleasby, “EMBOSS: The European Molecular Biology open Software Suite,” *Trends in Genetics*, vol. 16, issue 6, pp. 276-277, 2000.

[121] L. Y. Chuang, C. H. Huang, M. C. Lin, and C. H. Yang, “Particle Swarm Optimization with Reinforcement Learning for the Prediction of CpG Islands in the Human Genome,” *PLoS One*, vol. 6, issue 6, 2011.

[122] H. C. Park, E. R. Ahn, J. Y. Jung, J. H. Park, J. W. Lee, S. K. Lim, and W. Kim, “Enhanced Sensitivity of CpG Island Search and Primer Design Based on Predicted CpG Island Position,” *Forensic Science International: Genetics*, vol. 34, pp. 134-140, 2018.

[123] C. H. Yang, and Y. J. Hao, “Ion Motion Optimization Algorithm Applied to CpG Island Prediction,” *Journal of Life Sciences and Technologies*, vol. 4, no. 1, June 2016.

[124] G. A. Churchill, “Stochastic Models for Heterogeneous DNA Sequences,” *Bulletin of Mathematical Biology*, vol. 51, issue 1, pp. 79-94, 1989.

[125] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, “Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids,” *Cambridge University Press*, Cambridge, 1998.

[126] B. J. Yoon, and P. P. Vaidyanathan “Identification of CpG Islands using a Bank of IIR Lowpass Filters,” in *Digital Signal Processing Workshop and the 3rd IEEE Signal Processing Education Workshop*, pp. 315-319, 2004.

[127] S. Ye, A. Asaithambi, and Y. Liu, “CpGIF: An Algorithm for the Identification of CpG Islands,” *Bioinformation*, vol. 2, issue 8, pp. 335-338, 2008.

[128] M. Hackenberg, C. Previti, P. L. Luque-Escamilla, P. Carpena, J. Martinez-Aroza, and J. L. Oliver, “CpGcluster: A Distance-Based Algorithm for CpG-Island Detection,” *BMC Bioinformatics*, vol. 7, 2006.

[129] C. H. Yang, Y. C. Chiang, L. Y. Chuang, and Y. D. Lin, “A CpGCluster-Teaching-Learning-Based Optimization for the Prediction of CpG Islands in the Human Genome,” *Journal of Computational Biology*, vol. 24, no. 0, pp. 1-12, 2017.

[130] I. T. Mariapushpam, and R. Sivakumar, “Improved Algorithm for the Location of CpG Islands in Genomic Sequences using Discrete Wavelet Transforms,” *Current Bioinformatics*, vol. 12, pp. 57-65, 2017.

[131] A. Grover, V. Aishwarya, and P. C. Sharma, “Searching Microsatellites in DNA Sequences: Approaches used and Tools Developed,” *Physiology and Molecular Biology of Plants: An International Journal of Functional Plant Biology*, vol. 18, no. 1, pp. 11-19, 2012.

[132] O. Delgrange, and E. Rivals, “STAR: An Algorithm to Search for Approximate Tandem Repeats,” *Bioinformatics*, vol. 20, no. 16, pp. 2812-2820, 2004.

[133] M. E. Sagot, and E. W. Myers, “Identifying Satellites in Nucleic Acid Sequences,” *in Proceedings of the Second Annual ACM International Conference on Computational Molecular Biology*, pp. 234-242, 1998.

[134] G. Benson, “Tandem Repeats Finder: A Program to Analyze DNA Sequences,” *Nucleic Acids Research*, vol. 27, no. 2, pp. 573-580, 1999.

[135] K. G. Lim, C. K. Kwoh, L. Y. Hsu, and A. Wirawan, “Review of Tandem Repeat Search Tools: A Systematic Approach to Evaluating Algorithmic Performance,” *Briefings in Bioinformatics*, vol. 14, issue 1, pp. 67-81, January 2013.

[136] P. G. Pop, “Tandem Repeats Localization using Spectral Techniques,” *in 2007 IEEE International Conference on Intelligent Computer Communication and Processing*, pp. 243-246, 2007.

[137] T. T. Tran, V. A. Emanuele II, and G. T. Zhou, “Techniques for Detecting Approximate Tandem Repeats in DNA,” *in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 449-452, May 2004.

[138] M. Buchner, and S. Janjarasjitt, “Detection and Visualization of Tandem Repeats in DNA Sequences,” *IEEE Transactions on Signal Processing*, vol. 51, no. 9, pp. 2280-2287, 2003.

[139] R. Jiang, and H. Yan, “Detection and 2-Dimensional Display of Short Tandem Repeats based on Signal Decomposition,” *International Journal of Data Mining and Bioinformatics*, vol. 5, no. 6, pp. 661-690, 2011.

[140] D. Liping, H. Zhou, and H. Yan, “OMWSA: Detection of DNA Repeats using Moving Window Spectral Analysis,” *Bioinformatics*, vol. 23, no. 5, pp. 631-633, 2007.

[141] S. D. Sharma, R. Saxena, and S. N. Sharma, “Identification of Microsatellites in DNA Using Adaptive S-Transform,” *IEEE Journal of Biomedical and Health Informatics*,” vol. 19, no. 3, pp. 1097-1105, 2015.

[142] S. D. Sharma, R. Saxena, and S. N. Sharma, “Tandem Repeats Detection in DNA Sequences Using Kaiser Window based Adaptive S-Transform,” Bio-Algorithms and Med-Systems, vol. 13, issue 3, pp. 167-173, 2017.

[143] S. Zribi, I. Messaoudi, A. E. Oueslati, Z. Lachiri, "Microsatellite's Detection using the S-Transform Analysis based on the Synthetic and Experimental Coding," International Journal of Advanced Computer Science and Applications, vol. 10, no. 3, pp. 254-263, 2019.

[144] P. P. Kanjilal, J. Bhattacharya, and G. Saha, "Robust Method for Periodicity Detection and Characterization of Irregular Cyclical Series in terms of Embedded Periodic Components," *Physical Rev. E.*, vol. 59, no. 4, pp. 4013-4025, April 1999.

[145] P. Qiu, and K. J. R. Liu, "A Robust Method for QRS Detection based on Modified P-Spectrum," *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08),* pp. 501-504, 2008.

[146] M. Liscombe, and A. Asif, "A New Method for Instantaneous Signal Period Identification by Repetitive Pattern Matching," *in the proceedings of 2009 IEEE 13th International Multitopic Conference,* Islamabad, Pakistan, pp. 1-5, 2009.

[147] P. Garg, S. D. Sharma, and S. N. Sharma, "Tandem Repeats Detection in DNA Sequences using P-Spectrum Based Algorithm," in *the proceedings of Conference on Information and Communication Technology (CICT 2017),* pp. 1-5, 2017.

[148] A. K. Singh, and V. K. Srivastava, "Improved Filtering Approach for Identification of Protein-Coding Regions in Eukaryotes by Background Noise Reduction using S-G Filter," *Network Modeling Analysis in Health Informatics and Bioinformatics,"* vol. 10, issue 19, pp. 1-16, 2021.

[149] A. Savitzky, and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Analytical Chemistry*, vol. 36, issue 8, pp. 1627-139, 1964.

[150] N. Yu, Z. Li, and Z. Yu, "Survey on Encoding Schemes for Genomic Data Representation and Feature Learning-From Signal Processing to Machine Learning," *Big Data Mining and Analytics*, vol. 1, no. 3, pp. 191-210, 2018.

[151] I. M. E. Badawy, S. Gasser, A. M. Aziz, and M. E. Khedr, "On the use of Pseudo-EIIP Mapping Scheme for Identifying Exons Locations in DNA Sequences," *in 2015 IEEE Conference on Signal and Image Processing Applications*, pp. 244-247, 2015.

[152] M. Akhtar, E. Ambikairajah, and J. Epps, "On DNA Numerical Representations for Period-3 based Exon Prediction," *in the proceedings of 2007 IEEE International Workshop on Genomic Signal Processing and Statistics*, Finland, June 2007.

[153] S. V. Tenneti, and P. P. Vaidyanathan, "Detecting Tandem Repeats in DNA using Ramanujan Filter Bank," *in the proceedings of 2016 IEEE International Symposium on Circuits and Systems*, pp. 21-24, May 2016.

[154] S. Rogic, A.K. Mackworth, and F. B.F. Ouellette, "Evaluation of Gene finding Program on Mammalian Sequence," *Genome Research*, vol. 11, issue 15, pp. 817-832, 2001.

[155] M. Burset, and R. Guigo, "Evaluation of Gene Structure Prediction Applications," *Genomic*, vol. 34, issue 3, pp. 353-367, 1996.

[156] R. Touati, A. E. Oueslati, I. Messaoudi, and Z. Lachiri, "The Helitron Family Classification using SVM based on Fourier Transform Features Applied on an Unbalanced Dataset," *Medical & Biological Engineering & Computing*, vol. 57, pp. 2289-2304, 2019.

[157] L. Das, S. Nanda, and J. K. Das, "An Integrated Approach for Identification of Exon Locations using Recursive Gauss Newton Tuned Adaptive Kaiser Window," *Genomics*, vol. 111, no. 3, pp. 284-296, 2019.

[158] P. Garg, and S. D. Sharma, "Identification of CpG Islands in DNA Sequences Using Short-Time Fourier Transform," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 12, issue 3, pp. 355-367, 2020.

[159] P. Garg, and S. D. Sharma, "Modified P-Spectrum based Approach to Enhance Sensitivity for the Detection of CpG Islands in DNA Sequences in Human Species," *Biomedical Engineering: Applications, Basis and Communications*, vol. 34, no. 1, p. 2150052, February 2022.

[160] J. Epps, "A hybrid Technique for the Periodicity Characterization of Genomic Sequence Data," *Eurasip Journal on Bioinformatics and System Biology*, vol. 2009, pp. 1-8, 2009.

[161] V. Boeva, M. Regnier, D. Papatsenko, and V. Makeev, "Short Fuzzy Tandem Repeats in Genomic Sequences Identification and Possible Role in Regulation of Gene Expression," *Bioinformatics*, vol. 22, issue 6, pp. 676-684, 2006.

[162] P. G. Pop, and E. Lupu, "DNA Repeats Detection using BW Spectrograms," *in IEEE TTTC International Conference on Automation, Quality, and Testing, Robotics*, AQTR 2008, Tome III, Romania, pp. 408-412, 2008.

[163] S. D. Sharma, and P. Garg, "Integer Period Discrete Period Transform based Algorithm for the Identification of Tandem Repeats in DNA Sequences", *Machine Learning, Big Data, and IoT for Medical Informatics*, Academic Press, Elsevier, pp. 311-325, 2021.