2010

Human Promoter Prediction Using DNA Numerical Representation

Swarna Bai Arniker
University of Windsor

Human Promoter Prediction Using DNA Numerical Representation

by

Swarna Bai Arniker

A Dissertation
Submitted to the Faculty of Graduate Studies and Research
through the Department of Electrical and Computer Engineering
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy at the
University of Windsor

Windsor, Ontario, Canada
2010

Human Promoter Prediction Using DNA Numerical Representation

by

Swarna Bai Arniker

APPROVED BY:

Dr. M. Omair Ahmad, External Examiner
Department of Electrical and Computer Engineering
University of Concordia

Dr. Dan Wu, Outside Department Reader
School of Computer Science

Dr. Jonathan Wu, First Department Reader
Department of Electrical and Computer Engineering

Dr. Huapeng Wu, Second Department Reader
Department of Electrical and Computer Engineering

Dr. Hon Keung Kwan, Advisor
Department of Electrical and Computer Engineering

Dr. Sylvia Voelker, Chair of Defense
Department of Psychology

DECLARATION OF CO-AUTHORSHIP/PREVIOUS PUBLICATIONS

I hereby declare that this dissertation incorporates material that is the result of a joint research, as follows:

This dissertation incorporates the outcome of a joint research undertaken in collaboration with Dr. Bonnie N. F. Law, Dr. Daniel P. K. Lun under the supervision of Professor H. K. Kwan. The collaboration is covered in chapters 3, 6, and 7 of the dissertation. In all cases, the key ideas, primary contributions, experimental designs, data analysis and interpretation, were performed by the author; the contributions of the supervisor include the provision of valuable suggestions and helping in comprehensive analysis and presentation of the experimental results; the contribution of Dr. Law was primarily through the provision of useful discussions and comments; and the contribution of Dr. Lun was primarily through the provision of software verification and comments.

I am aware of the University of the Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my dissertation, and have permission from each co-author(s) to include the above material(s) in my dissertation. I certify that, with the above qualification, this dissertation, and the research to which it refers, is the product of my own work.

This dissertation includes two original papers that have been previously published in peer reviewed conferences, as follows:

Dissertation Chapter	Publication title/full citation	Publication status
Chapter 3	Hon Keung Kwan and Swarna Bai Arniker, "Numerical representation of DNA sequences," in Proc. Of IEEE Inter. Conf. on Electro/Information Technology, Windsor, Ontario, Canada, June 7-9, 2009, pp. 307-310.	Published
Chapter 4	Swarna Bai Arniker and Hon Keung Kwan, "Graphical representation of DNA sequences," in Proc. of IEEE Inter. Conf. on Electro/Information Technology, Windsor, Ontario, Canada June 7-9, 2009, pp. 311-314.	Published

I certify that the above material describes work completed during my registration as graduate student at the University of Windsor.

I declare that, to the best of my knowledge, my dissertation does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my dissertation, published or otherwise, are fully acknowledged in accordance with the standard referencing practices.

I declare that this is a true copy of my dissertation, including my final revisions, as approved by my dissertation committee and the Graduate Studies Office, and that this dissertation has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

With the emergence of genomic signal processing, numerical representation techniques for DNA alphabet set {A, G, C, T} play a key role in applying digital signal processing and machine learning techniques for processing and analysis of DNA sequences. The choice of the numerical representation of a DNA sequence affects how well the biological properties can be reflected in the numerical domain for the detection and identification of the characteristics of special regions of interest within the DNA sequence.

This dissertation presents a comprehensive study of various DNA numerical and graphical representation methods and their applications in processing and analyzing long DNA sequences. Discussions on the relative merits and demerits of the various methods, experimental results and possible future developments have also been included.

Another area of the research focus is on promoter prediction in human (*Homo Sapiens*) DNA sequences with neural network based multi classifier system using DNA numerical representation methods. In spite of the recent development of several computational methods for human promoter prediction, there is a need for performance improvement. In particular, the high false positive rate of the feature-based approaches decreases the prediction reliability and leads to erroneous results in gene annotation.

To improve the prediction accuracy and reliability, “DigiPromPred” a numerical representation based promoter prediction system is proposed to characterize DNA alphabets in different regions of a DNA sequence.

The DigiPromPred system is found to be able to predict promoters with a sensitivity of 90.8% while reducing false prediction rate for non-promoter sequences with a specificity of 90.4%. The comparative study with state-of-the-art promoter prediction systems for *human* chromosome 22 shows that our proposed system maintains a good balance between prediction accuracy and reliability.

To reduce the system architecture and computational complexity compared to the existing system, a simple feed forward neural network classifier known as "SDigiPromPred" is proposed. The SDigiPromPred system is found to be able to predict promoters with a sensitivity of 87%, 87%, 99% while reducing false prediction rate for non-promoter sequences with a specificity of 92%, 94%, 99% for *Human*, *Drosophila*, and *Arabidopsis* sequences respectively with reconfigurable capability compared to existing system.

Dedicated to my parents A. Narahari Rao and A. Savithri Bai,

for their constant love and encouragement

and

To my late brother A. Khande Rao whom we all miss

ACKNOWLEDGEMENTS

I am greatly indebted to my supervisor Prof. Hon Keung Kwan for introducing me to bioinformatics, suggesting this work and associated key papers for background reading, and an opportunity and financial support to be his research assistant during my study. Words cannot express my gratitude for his invaluable guidance and encouragement throughout my study. He is always very patient and cooperative to me while discussing various aspects of the research work. He is always accessible to me whenever I need inspite of his busy schedule. Without him it would have been impossible to complete my dissertation.

I am grateful to Prof. Jonathan Wu, Prof. Huapeng Wu, and Prof. Dan Wu for their valuable comments on my research work. I would also like to thank Prof. M. Omair Ahmad, University of Concordia, Canada for serving as my external examiner and also for providing me valuable comments on my research work as reported in this dissertation, and thanks to Prof. Sylvia Voelker for chairing the defense

I would like to thank Prof. Robin Gras, school of computer science who enlightened me on computational molecular biology. I wish to thank Dr. Daniel P. K. Lun, Dr. Bonnie N. F. Law for discussions and encouragements. Special thanks to Dr. Tai-Chiu Hsung, Dr. K. O. Cheng and Mr. S. C. Cheung for discussions.

I would like to thank all my officemates in the ISPLab for their help during the last few years. A special thank to Dr. Aimin Jiang for discussion and encouragement. I am grateful to the University of Windsor for the graduate assistantship and scholarship supports for my Ph.D. study. I wish to thank my fellow graduate students for their

encouragement and continued support during my stay at the University of Windsor. Special thanks to Andria Turner who goes that extra mile (GEM) to help graduate students like me, and thanks to Shelby Marchand, Don Tersigni, Frank Cicchello for their help and support. Thanks to my landlords Dorothy Gabriel and Recto Domingo for their encouragement.

I am grateful to R.C.I., Government of India for having granted study leave to pursue my Ph.D program at University of Windsor, Canada. Special thanks to my seniors and colleagues Dr. V. K. Saraswat, Mr. S. K. Ray, Dr. K. G. Narayanan, Mr. K. Rama Sharma, Dr. D. V. K. Sastry, Mr. K. Jayatirtha Rao, Mr. P. Nageshwar Rao, Mr. G. Vijay Sankar, Mr. V. V. Subrahmaniam, Mr M. K. Haware, and Mrs. P. Ratna kumari for their encouragement and support. I am thankful to State Bank of India, Hyderabad, India for providing me with education loan, special thanks to Mr. P. L. Narasimha Rao.

I would like to extend my gratitude to my old teachers in India, Mrs. S. Kalyani, Dr. C. V. R. N. Sarma, Dr. M. Prahlad Rao, Dr. E. V. V. Chari, Dr. E. G. R. Charyulu, Dr. J. John, Dr. B. Mazhari, and my mentors late S. V. Bapat and late G. Suryanarayana.

Last but not least, I would like to express my appreciation to my sister Dr. A. Hamsa and my brother-in-law Dr. G. Vedaparayana, my elder brother Mr. A. Raghuveer and my sister-in-law Mrs. A. Sharmila for their constant and generous love and support. A special thanks to my nieces, Ms. G. Amrita, and Ms. G. Ahalaya. Whenever I get into difficult situations, I can always gain strength from their mails and pictures.

Finally, I wish to thank God for having given me the power to believe in my passion and pursue my dreams. I could have never done my Ph.D without the faith I have in you, the Almighty.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A	Adenine
ANN	Artificial Neural Network
C	Cytosine
CCP	Cumulative Categorical Periodogram
CDS	Coding sequence
1-D	1-Dimensional
2-D	2-dimensional
3-D	3-dimensional
DBTSS	Database of Human Transcriptional Start Sites
DFT	Discrete Fourier Transform
DigiPromPred	Digital representation based promoter predictor
DragonGSF	Dragon Gene Start Finder
DNA	Deoxyribose Nucleic Acid
DPF	Dragon Promoter Finder
DQFT	Discrete Quaternion Fourier Transform
DSP	Digital Signal Processing
EID	Exon-Intron database

EIIP	Electron-Ion Interaction Potentials
EPD	Eukaryotic Promoter Database
Er	Classification Error
FirstEF	First-Exon and Promoter Prediction Program
FPROM	Human Promoter Prediction
FP	False Positive
FN	False Negative
G	Guanine
GCC	Genetic Code Context
GF	Galois Field
HGP	Human Genome Project
HIV	Human Immunodeficiency Virus
KNN	K-nearest neighbor
mRNA	Messenger Ribonucleic Acid
MultiNNProm	Multiple Neural Network based system for Promoter Recognition
NCBI	National Centre for Biotechnology Information
NNPP	Neural Network Promoter Prediction
ORF	Open Reading Frame
PCA-HPR	Principal Component Analysis-Human Promoter Predictor

PCF	Position Count Function
Poly-A	Poly-Adenylation
Pre	Precision
Pre-mRNA	Primary Messenger Ribonucleic Acid
RNA	Ribonucleic Acid
SC	Spectral Content
Sen	Sensitivity
Spec	Specificity
SZ	Simple-Z
T	Thymine
TFs	Transcription Factors
TFBSs	Transcription Factor Binding Sites
TN	True Negative
TP	True Positive
TRANSFAC	Transcription Factor Database
TSS	Transcription Start Sites
TSSG	Recognition of Human Pol II Promoter Region and Start of Transcription
TSSW	Predicts Potential Transcription Start Positions

U	Uracil
3’UTR	Three prime Untranslated Region
UTRdb	Untranslated Region Database

CHAPTER I

INTRODUCTION

Genome Informatics has been an active area of research for the past one decade and has drawn increasing interest in recent years from the neural network research community. More specifically, the field involves identifying protein/RNA (ribonucleic acid) -encoding genes, recognizing functional elements (promoters, enhancers, origins of replication) on nucleotide sequences, understanding biochemical processes and gene regulation, determining protein structures from amino acid sequences and modeling RNA structures, and performing comparative analysis of genomes and gene families. The major topics of genome informatics research include *gene recognition*, *functional analysis*, *structural determination*, and *family classification* as depicted in Fig. 1.1 [1].

This dissertation addresses a significant open problem in the field of gene recognition, namely promoter prediction in higher eukaryotes (especially *human*) genomic DNA sequences. It is possible to use the prediction of promoter sequences and transcriptional start point as a signal; i.e., by knowing the position of a promoter, one can deduce at least the approximate start of the transcript, thus delineating one end of the gene. The problem has a certain intrinsic interest that it poses a challenge to understand and define precisely the biochemical processes, functional regions, and signals involved in the pathway from DNA to protein sequences. On the other hand, with the recent completion of the human genome draft [2] and many new genomes to be investigated in the near future, the problem has taken on significant practical importance. The difficulty of the problem is that the properties of the promoter regions are different from the

properties of other functional regions (exons, introns, 3'UTR). Furthermore, unlike prokaryotic promoters, eukaryotic (e.g., human) promoter sequences are so diverse that it is not yet possible to draw any clear generalizations about them. The low sensitivity and high false positives of various available feature based promoter prediction software programs such as Eponine [3], FirstEF [4], PromoterInspector [5], DPF [6], DragonGSF [7, 8] and PCA-HPR [9], and evidence of improving the promoter prediction by DNA numerical sequence approach [10], suggest that this area is worth investigating further using non-traditional approaches and is dealt with herein by examining the role of artificial neural networks (ANN) techniques, developing the architecture of the proposed DigiPromPred system, and DNA numerical representation selection strategy to advance state of the art in human promoter prediction algorithms. ANN-based methods are attractive because of their many features and capabilities for recognition, generalization and classification, and are well suited for genome informatics studies. Genomic information is inherently discrete in nature because there is a finite number of nucleotide in the DNA alphabet, invites investigation by ANN techniques. The conversion of DNA nucleotide symbols (i.e., A, C, G, and T) into discrete numerical values enables novel and useful ANN-based applications for the solution of different sequence analysis related problems such as promoter prediction [10, 11, 12]. This work is highly cross-disciplinary in nature: while the fundamental subject matter is biological and final results of biological interest are obtained, techniques from other fields such as ANN are extensively used. The accurate identification of exonic/intronic regions, donor/acceptor splice sites, and other regions and biological signals (as shown in Fig. 2.6) would result in an ideal gene finding and annotation system.

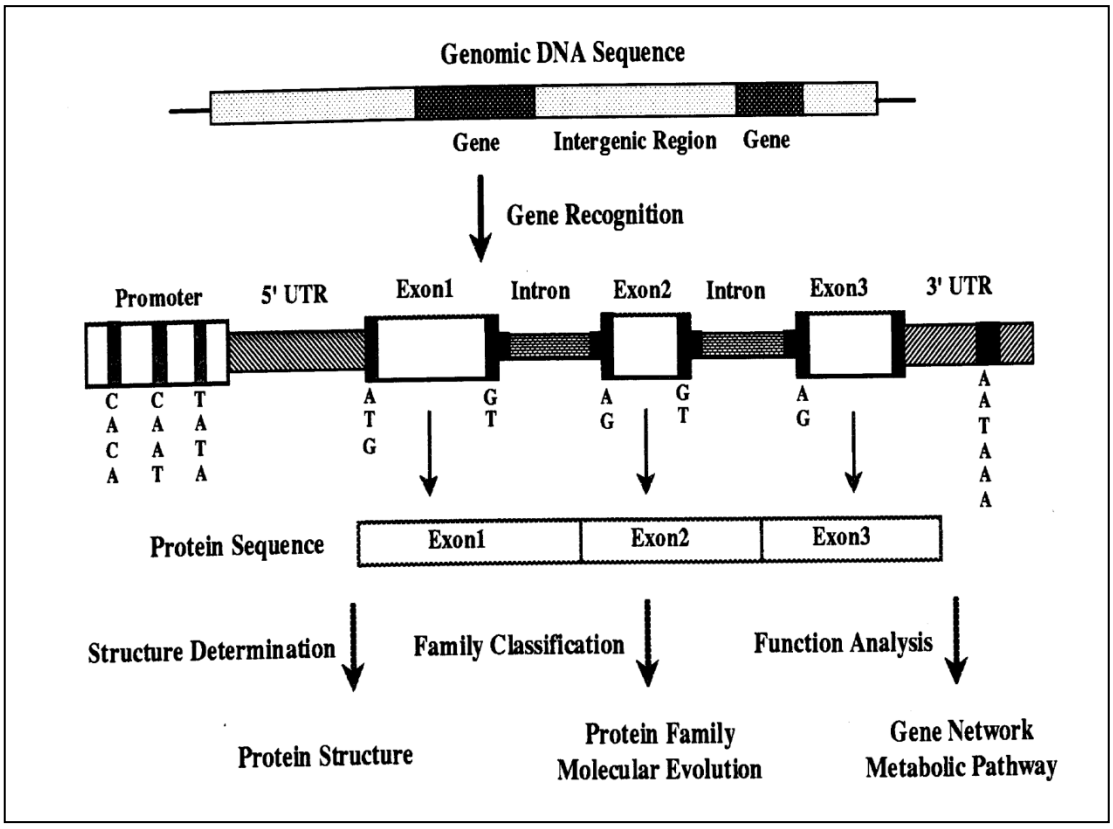

Fig. 1.1 Genome informatics overview (A simplified view of gene structure) [1].

In this work, the role of numerical sequence approach in conjunction with ANNs for human promoter prediction is examined, novel numerical representation selection strategy are proposed and the first investigation of a Digital Promoter Prediction (DigiPromPred) an artificial neural network based system for human promoter region recognition using numerical sequence approach is performed. Most existing DNA symbolic-to-numeric representations map the DNA sequence to more than one numerical sequence, potentially introducing redundancy in representations. The assignment of distinct numbers to each of the four DNA alphabets (i.e., A, C, G, and T) in some representations does not necessarily reflect the mathematical property present in the original DNA sequence. The accuracy of existing ANN based human promoter prediction system is limited, because they rely solely on feature based representation [3-9] and are

not well equipped to capture the complementary properties of promoter and nonpromoter regions (exons, introns, and 3'UTR). There has been limited attempt to apply numerical sequence approach for eukaryotic promoter prediction. As compared to the feature-based approach, the DNA numerical sequence approach can preserve nucleotide positional information and retain all the available sequence information. I carried out an extensive survey on numerical and graphical representations of DNA sequences and found that 2-bit binary, 4-bit binary representations are most suitable for ANN based promoter prediction systems since they fully exploit not only the structural differences of promoters and nonpromoters but also preserve the DNA biological properties to facilitate ANN based human promoter prediction. Thus, I present a Digital Promoter Prediction (DigiPromPred) system for human promoter region recognition using numerical sequence approach. It makes use of two selected DNA numerical representations to preserve the DNA biological properties so that human promoters can be distinguished from non-promoters (such as introns, exons, and 3'UTR). The performance of the proposed system 'DigiPromPred' was primarily evaluated with a small data set and further extensively evaluated with 3-cross validation test using standard datasets such as promoter sequences (30,945) from the DBTSS (Database of transcription start sites) [13]. For non-promoters (30,945), 3'UTR sequences are extracted from the UTRdb database [13], exons and introns are extracted from the Exon-Intron Database (EID) [13]. The predictive accuracy of the proposed system was often evaluated using measures such as sensitivity, specificity, classification error and precision. Furthermore, the DigiPromPred system is evaluated on identifying promoter regions in human chromosome 22 [13]. I tested for the

20 annotated promoters in the chromosome 22 [13]. The experimental results and comparative studies with existing promoter prediction systems are also provided.

It is proposed that the combination of numerical sequence approach in conjunction with ANN based system to discriminate between promoters versus exons, promoters versus introns, and promoters versus 3'UTR can advance the state of the art in promoter prediction, and the resultant system could offer a higher sensitivity and lower false positive than that offered by existing feature based promoter prediction systems.

To reduce the system architecture and computational complexity compared to the existing system, a simple feed forward neural network classifier known as "SDigiPromPred" is proposed to predict promoters in three model organisms using single DNA numerical representation. The SDigiPromPred system is found to be able to predict promoters with a sensitivity of 87%, 87%, 99% while reducing false prediction rate for non-promoter sequences with a specificity of 92%, 94%, 99% for *Human*, *Drosophila*, and *Arabidopsis* sequences respectively with reconfigurable capability compared to existing system. Further, the most suitable numerical representation for *Human*, *Drosophila*, and *Arabidopsis thaliana* promoter prediction is also presented from the experimental results.

1.1 Objectives

The principle objective of this dissertation is to investigate and develop efficient promoter prediction system for accurate prediction of promoters in human genomic sequences and further develop a promoter prediction system with reduced architecture and computational overload to predict promoters in three model organisms. This broad objective may be expressed in terms of number of aims:

- To critically review biological literature, genomic data, and other online resources relevant to promoter prediction in Eukaryotic genomic sequences.
- To carry out extensive survey of the existing numerical and graphical representation of DNA sequences.
- To examine the role of numerical representation selection strategy, and the architecture of the proposed DigiPromPred system in this area.
- To investigate and rigorously evaluate the system performance with a large set of genomic sequences of promoters (30,945) and nonpromoters (exon, intron, 3'UTR each of 30945 sequences) with 3-cross validation test and further evaluate the system for 20 annotated promoters in *human* chromosome 22.
- Comparative studies with state-of-the-art promoter prediction systems.
- To evaluate and analyse the performance of a proposed SDigiPromPred system with reduced architecture and computational overload in three model organisms.

1.2 Organization

The remainder of the dissertation is organized as follows:

Chapter 2 reviews basics of biology, regulation of gene expression and promoters, and provides background about the significance and difficulties in promoter prediction.

Chapter 3 discusses, classifies, analyzes, compares and also evaluates the performance of various DNA numerical representation methods and its variants with DSP techniques like discrete Fourier transform (DFT) to classify human coding and noncoding sequences.

Chapter 4 presents a summary of various DNA graphical representation methods and their applications in envisaging and analyzing long DNA sequences with Digital Signal Processing (DSP) techniques.

Chapter 5 introduces the concepts behind the promoter prediction system and their classification, provides review of existing numerical representation based promoter prediction systems, and the overview of MultiNNProm system (existing).

Chapter 6 presents the numerical representation selection strategy and the architecture of the proposed DigiPromPred system; rigorously evaluate the system performance with a large set of genomic sequences of promoters (30,945) and nonpromoters (exon, intron, 3'UTR each of 30945 sequences) with 3-cross validation test. Evaluate DigiPromPred system with 20 annotated promoters in *human* chromosome 22 and compare the results with state of the art promoter prediction systems.

Chapter 7 describes the numerical representation selection strategy and architecture of another proposed system known as SDigiPromPred with reduced architecture and computational complexity compared to DigiPromPred and further evaluate its performance with three model organisms.

Chapter 8 highlights the achievements of the work and suggests topics for future research.

1.3 Contributions

This research provides original contributions on the use of DNA numerical sequence approach in conjunction with ANNs for human and model organism promoter prediction. The major contributions of this work can be summarized as follows:

- A detailed review and comparison of existing DNA symbolic-to-numeric representations with experimental results based on period-3 property is presented.
- A review of DSP based graphical representation methods is also presented.

The following methods are newly proposed in our human promoter prediction system.

- Presented a numerical sequence approach for recognizing *human* promoters where each individual nucleotide in the DNA sequence is used for promoter identification.
- Introduction of DNA numerical representation selection strategy.
- System has been tested with a large set of genomic sequences of promoters (30,945) and nonpromoters (exon, intron, 3'UTR each of 30,945) and 20 annotated promoters in human chromosome 22.
- Achieved lowest mean and standard deviation with 3-cross validation test.
- Improved sensitivity and reduced false prediction rate.
- Thus our proposed system "DigiPromPred" which is based on numerical representation of DNA sequences is comparable to the state-of-the-art promoter prediction systems that are feature based.

The following features are newly proposed in the promoter prediction system "SDigiPromPred".

- Reduced architecture and computational complexity compared to DigiPrompPred and MultiNNProm systems by using a simple feed forward neural network classifier.

- Promoter sequences are used as positive samples and non-promoters as negative samples.
- SdigiPromPred system architecture is reconfigurable with consistent performance metrics with three model organisms.

CHAPTER II

BIOLOGICAL BACKGROUND AND PROMOTER PREDICTION

In the past few years, the emphasis in molecular biology has shifted from the study of individual genes to the detection of promoters in the entire genome of an organism. As the magnitude of available data grows, detecting these regions by biological experiments is too time-consuming. As a result, the field becomes increasingly dependent on computational methods for the prediction and annotation of promoters. With its many features and capabilities for recognition, generalization and classification, artificial neural network technology is well suited for investigation in this emerging area. This chapter reviews basics of biology, regulation of gene expression and promoters, giving background about significance and difficulties in promoter prediction.

2.1 Cell, Genome and Chromosome

The cell is the structural and functional unit of all living organisms. Living organisms can be classified into two: 1) Prokaryotes and 2) Eukaryotes.

Prokaryotes have their DNA free in the cell, as they lack a nucleus, cell membrane, and many other structures that are seen in eukaryotic cells; in general, they are less developed and simpler. On the other hand, in eukaryotes a nucleus is separated from rest of the cell by a nuclear membrane, and DNA is kept inside the nucleus and their structure is very complex; a comparison of both cells is depicted in Fig. 2.1. For this dissertation,

unless specified, all references to cells are meant to be for eukaryotes, the ones present in humans. As stated, every cell of the human body contains a nucleus, where the DNA is surrounded by a liquid composed of water and other molecules.

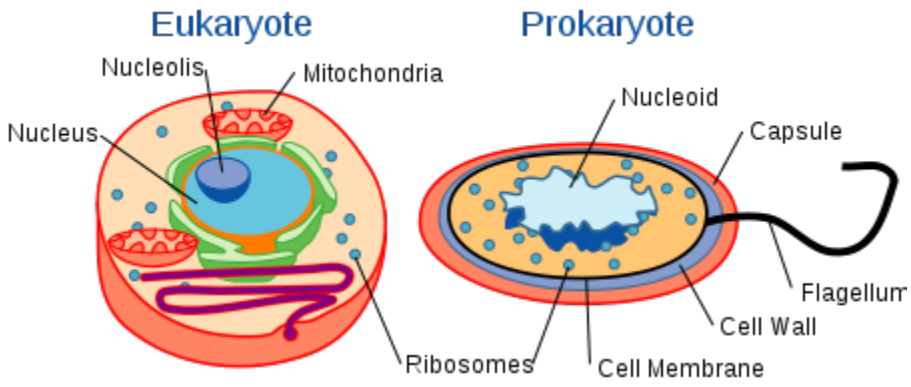

Fig. 2.1 Simplified models for the cells of eukaryotes and prokaryotes **(Image credit:** http://www.soe.uoguelph.ca/webfiles/mleuniss/Cell_Biology.htm).

Prokaryotes generally have just one chromosome, which is sometimes a circular DNA molecule, whereas eukaryotes have two matched sets of chromosomes, one set from each parent. Examples of prokaryotes are bacteria and blue algae. All other organisms including plants, animals, and humans are eukaryotes.

The genome (shown in Fig. 2.2), the total DNA content of an organism, consists of one or more large uninterrupted pieces of chromosomes. A chromosome contains many protein coding genes, regulatory elements, and other intervening sequences. Different organisms may have different number of chromosomes in their genomes. For example, every *Homo sapiens* (Human) cell has 46 chromosomes, whereas in *mice* this number is 40. Table 2.1 gives the genome size, the number of genes, and the number of chromosomes for selected species.

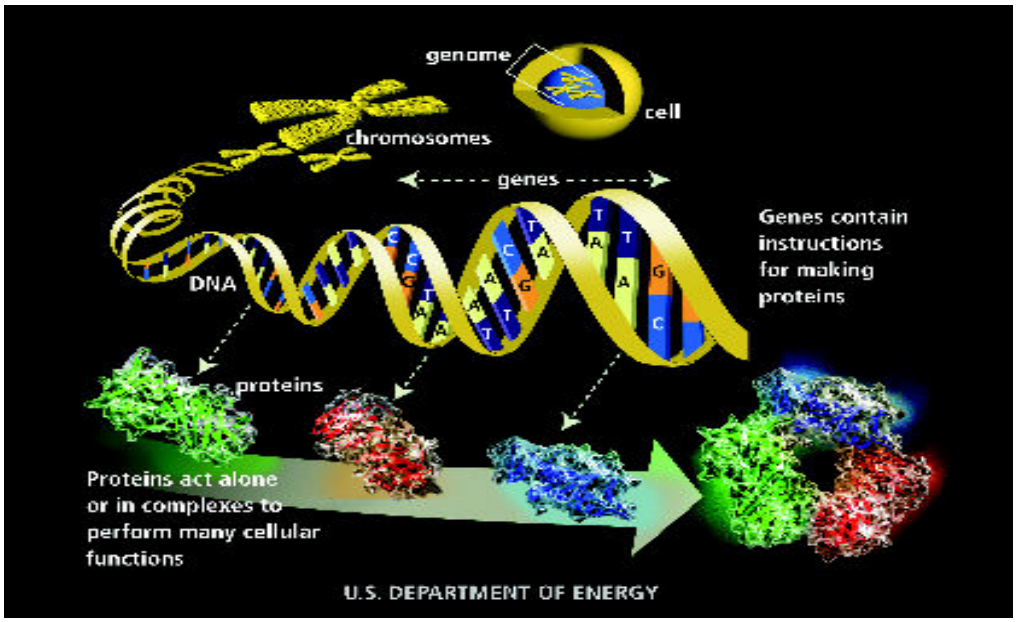

Fig. 2.2 The complete flow from cell to protein (**Image credit:** http://genome.gsc.riken.go.jp/hgmis/project/info.html).

Table 2.1 Genome sizes of certain species (**Table credit**: http://www.ornl.org/sci/techresources/Human_Genome/faq/compgen.shtml)

Organism	Genome size (in base pairs)	Number of genes	Number of Chromosomes (diploid)
Homo sapiens (human)	3.2 billion	~25,000	46
Mus musculus (mouse)	2.6 billion	~25,000	40
Drosophila melanogaster (fruit fly)	137 million	13,000	8
Arabidopsis thaliana (plant)	100 million	25,000	10
Caenorhabditis elegans (round worm)	97 million	19,000	12
Saccharomyces cerevisiae (yeast)	12.1 million	6000	32
Escherichia coli (bacteria)	4.6 million	3200	1
H. influenzae (bacteria)	1.8 million	1700	1

2.2 DNA (Deoxyribonucleic acid)

DNA encodes all the necessary information to run a cell. It can be viewed as the blue print for cell machinery. DNA is made up of linear chains of subunits called nucleotides. Each nucleotide consists of three parts; a nitrogenous base, a five-carbon-atom sugar (deoxyribose) and a phosphate group. The four possible nucleotide bases are A (Adenine), C (Cytosine), G (Guanine), and T (Thymine). However, in a number of viruses (prokaryotes), where ribonucleic acid (RNA) is used as basic building block of their genome, nucleotide Thymine (T) is replaced by Uracil (U). In DNA, individual nucleotides are attached to each other through sugar-phosphate bonds, forming a long one dimensional chain with two distinct ends, the 5' end (upstream), and the 3'end (downstream). Thus, DNA sequence is either shown from 5' to 3' (default is 5 prime end to 3 prime end) or its reverse from 3' to 5'. Therefore, this DNA chain or strand can symbolically be represented by a character string consisting of four alphabet letters A, C, G, and T. The DNA is a double stranded helix (see Fig. 2.3), that holds the following interesting features:

- DNA is double stranded; it twists around its long axis forming a double helix.
- The two strands are antiparallel i.e., each strand has specific orientation and they run in opposite directions.
- Each strand of DNA contains all the necessary information for proper functioning of a cell. This is a key factor in DNA replication, in which the double helix uncoils and two new complementary strands are generated.
- The sequence of DNA is a 1-dimensional string. The convention for the directionality of the chain is to go from the 5' end to the 3' end.

- The purine of one strand is paired with pyrimidine of the other strand i.e., A is paired to T and vice versa, and G is linked to C and vice versa.
- The purine-pyrimidine pairs are being complement to one another. Thus, if the sequence 5'— ACTTAGCTAAGCGG --- 3' occurs on one strand, the other strand must have the sequence 5' --- CCGCTTAGCTAAGT --- 3' i.e.,

  5'— ACTTAGCTAAGCGG --- 3'

  3' --- TGAATCGATTCGCC --- 5'

- The unit of measurement of a single stranded DNA sequence is the total number of nucleotides (nt) and for double stranded DNA sequence it is base pairs (bp).
- The symbols R and Y are used to designate purine (A or G) and pyrimidine (C or T); symbols S and W are used to designate strong (C or G) and weak hydrogen bonding (A or T); symbols K and M are used to designate keto (T or G) and Amino (A or C) respectively for a DNA sequence.

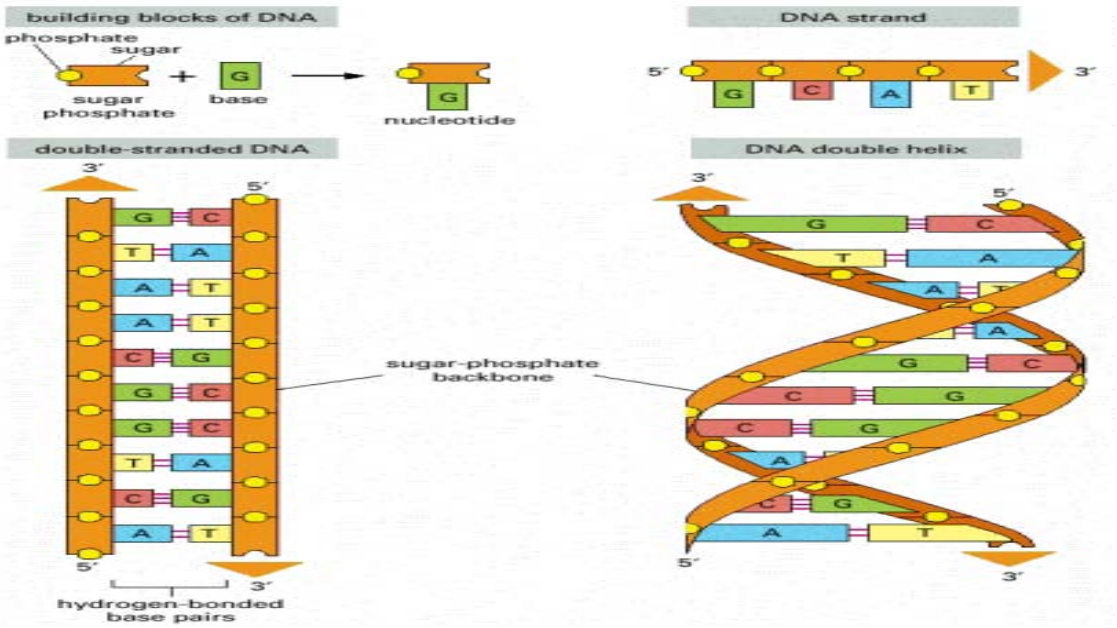

Fig. 2.3 DNA and its building blocks [14].

2.3 Genes

Genes are certain contiguous stretches along DNA which encode information for building proteins. These stretches vary in length and represent different protein coding genes, depending on the organism. As shown in Fig. 2.4, a gene may have two subregions called exons, and introns (prokaryotes do not have introns). The genes of eukaryotic (animals, plants, fungi) organisms can contain non-coding regions called introns that are removed from the messenger RNA (mRNA) in a process known as splicing. The regions that actually encode the gene product, which can be much smaller than the introns, are known as exons. The identification of promoters, start/stop codons, and acceptor (‘ag’)/donor (‘gt’) splice sites (Fig. 2.5) could recognize accurately the start and end points of protein coding gene regions in eukaryotic DNA. Before the synthesis into protein (i.e., chain of amino acids), the DNA sequence is converted into messenger RNA (mRNA: nucleotide T is replaced by U). Triplets of mRNA nucleotides specify each type of amino acid. Each nucleotide triplet is called a codon. The table that maps 64 codons into 20 amino acids and 3 stop signals is known as genetic code, as given in appendix A. In this many to one mapping, the codon ‘AUG’, which codes for methionine, also indicates the start of a gene [15]. There are three possible ways to read a nucleotide sequence in DNA or RNA as a series of non-overlapping triplets, depending upon whether reading starts with the first, second or third base in the sequence, is known as reading frame. For example, the three possible reading frames for nucleotide sequence ‘CAUUGCCAGU’, are [16];

C A U U G C C A G U - -

- C A U U G C C A G U -

- - C A U U G C C A G U

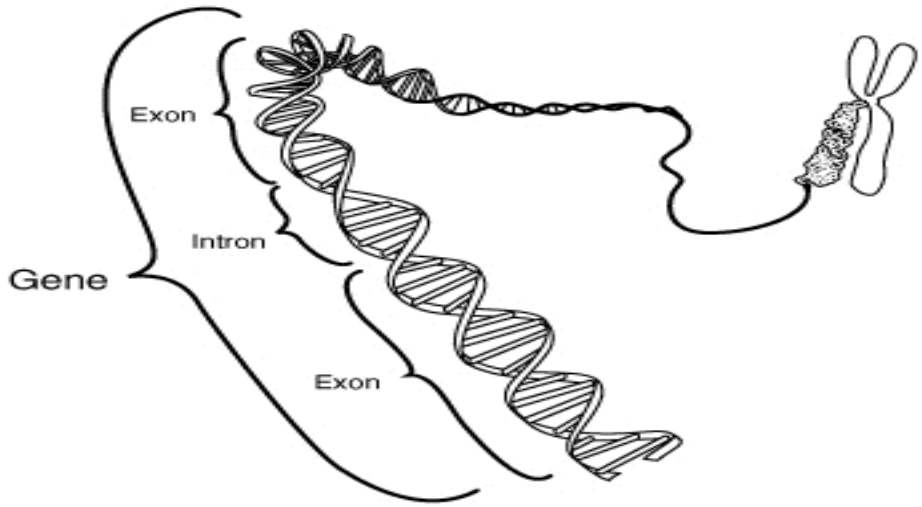

Fig. 2.4 Chromosome consists of genes.

2.4 Regulation of Gene expression and Promoters

Genes in DNA contain information for the production of RNA and proteins inside the cells. Proteins play a vital role in cellular functions. A vast majority of genes are known to produce proteins as their end products. The process by which information from a gene is used for synthesizing proteins in cells is known as gene expression. Gene expression involves transfer of sequential genetic information from DNA to proteins and broadly involves the following stages (Fig. 2.5):

- Transcription, where a gene's DNA sequence is copied into RNA by a polymerase called RNA polymerase (RNAP). This transcription results in a single stranded sequence of primary transcript or pre-mRNA.

- Capping, where primary transcript is capped on the 5’ end, which ensures the stability of the transcript by protecting it from degradation enzymes.
- Poly-adenylation, where a part of 3’end of the primary transcript is replaced by a poly-A tail (a stretch of Adenine bases) for providing stability.
- Splicing, where introns are removed and exons are joined together from the primary transcript to form messenger RNA (mRNA).
- mRNA is transported from nucleus to cytoplasm.
- Translation, where a ribosome produces a protein (made of specific amino acid chain) by using the mRNA template.

The complete process of moving from DNA to proteins is also known as the central dogma of molecular biology.

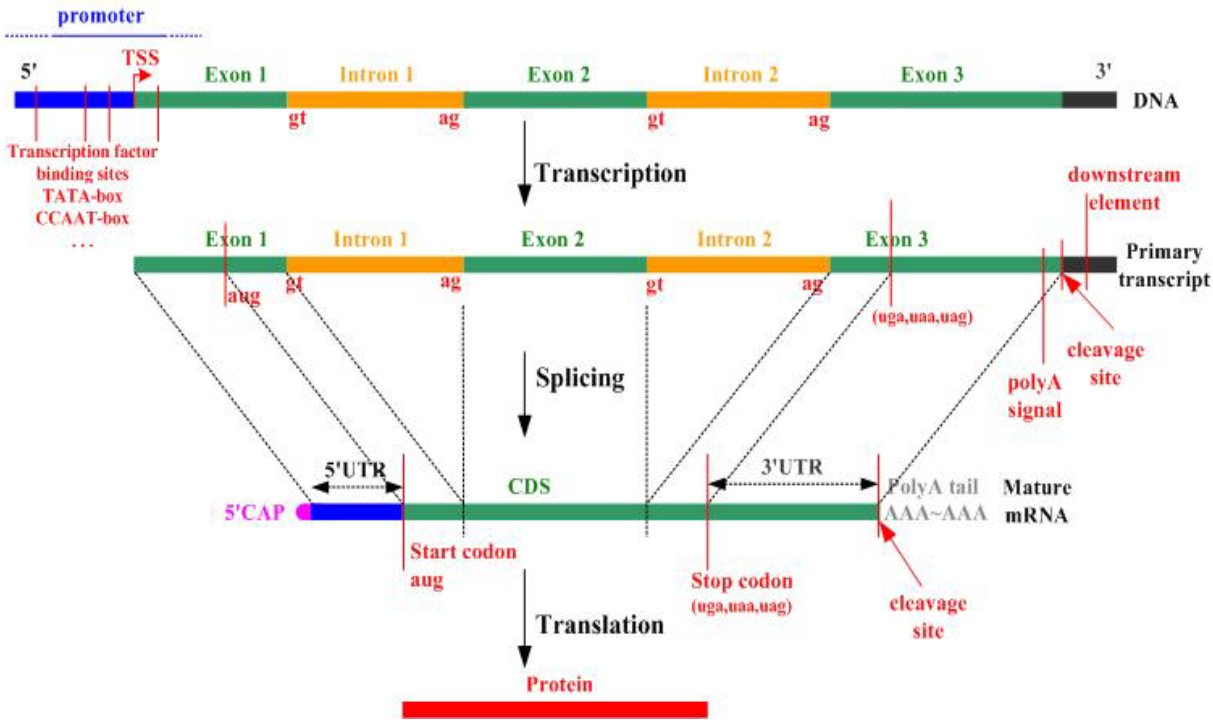

Fig. 2.5 Stages of gene expression in a cell (**Image Credit**: www.carolguze.com/text/442-1-humangenome.shtml).

In the entire process flow of gene expression from transcription to translation (stages shown in Fig. 2.5), transcription is generally believed to be the most important stage. The transcription stage of gene expression involves regulatory DNA regions known as promoters.

Every gene has at least one promoter that facilitates and controls its transcription initiation. This control mechanism occurs through a complex interaction between various transcription factors (TFs for example TATA binding protein) that get attached to their specific transcription factor binding sites (TFBSs for example TATA box) present in the gene's promoter region. A promoter is usually defined as a non-coding region of DNA that covers the transcription start site (TSS) or the 5' end of the genes they regulate. Bulk of promoter region typically lies upstream of the TSS.

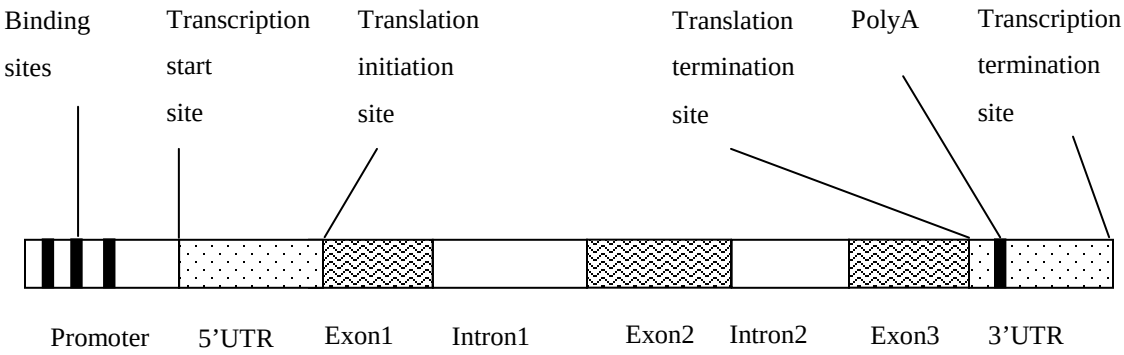

Fig. 2.6 Diagrammatic illustration of a eukaryotic gene (highlighting the locations of the promoter region, TFBs, transcription start site (TSS), 5'untranslated region (5'UTR), exons, introns, and 3'untranslated region (3'UTR)).

Fig. 2.6 shows the structure of a eukaryotic gene sequence and the location of the promoter in the DNA gene sequence. The first box which is of interest indicates the promoter which is close to the transcription start site (TSS).

Reference [17] and [18] provides a detailed survey on the fundamental concept of promoter structure and the biological point of view of eukaryotic promoter prediction. Ref [18] also bought up the idea that, when promoter characteristics are analyzed, physical properties, such as DNA bendability, can be applied to promoter prediction. Ref [18] states that not only promoters to be analyzed, but non-promoters, like exon, intron, and 3'UTR, should be also.

2.5 Significance of Promoter Prediction

Because the gene sequence data are growing exponentially recently, it is important to decode and annotate such data. However, traditional biological experiments are not sufficient. How to design efficient computer algorithms and software to analyze and annotate gene sequences becomes one of the most important issues today.

If we know the position of the promoter, we will know the position of the first exon. With knowing the position of the first exon, we get the starting position of the coding region of the gene, which will be translated into the protein sequence. How to find the promoter of a DNA sequence is one of the critical points in the work of gene sequence analysis.

If we know which segment of a DNA sequence is the promoter sequence, we can use the promoter sequence to regulate the speed of translation from DNA into a protein (for example miniature drug delivery system to control blood clotting; to burn away malignant tumors). Furthermore, the promoter is also useful in genetically modified plants or foods (for example pest resistance, cold tolerance, drought tolerance, salinity tolerance, disease tolerance in plants; improved nutritional content in food products,

vaccines in tomatoes and potatoes for ease of admissibility) [19]. With the similar method, we can also have the protein which causes disease grow more slowly, even destroy it. Through the recombination of DNA sequences with the promoter in transgenic technology, animals may get the gain-of-function or loss-of-function [20].

Presently, only by using biochemical methodology, can researchers locate the exact position of TSS, where the transcription starts. However, the processes are complicated and time consuming. By the promoter prediction method, we may be able to narrow down the promoter regions among massive DNA sequences. A further experiment then can be designed and tested if required. Therefore, much more time and cost will be saved.

2.6 Why is it Difficult to Model Promoters Computationally?

The obstacles in efficient modeling and recognition of promoters are as follows:

- Promoters constitute a very small percentage of the entire genome.
- Length of the promoter may range from a few hundred bases in some genes to thousands of bases in others.
- Properties of promoters are different from the properties of other functional regions (exons, introns, and 3'UTR).
- Promoter sequences are so diverse that it is not yet possible to draw clear generalizations about them and which can be applied universally for all types of promoter recognition.
- Transcription factor binding sites (TFBSs) in promoters are short and not fully conserved in the sequence; they may appear in numerous combinations and order.

Apart from this, the exact location, the orientation, and the mutual distance between the TFBSs may also vary a lot.

- Insufficient and incomplete information about TFs and TFBSs, though several thousands of them have been documented in TRANSFAC database [21]
- Unreliable models based on TFBSs produce high number of false positives on the genome.
- Most of the promoter prediction systems are organism specific.

All these factors have resulted in the inability to produce an efficient computer based promoter systems which can be used for predicting general promoters. However, with an approach focused on modeling the promoter sequences with suitable numerical representation in conjunction with artificial neural networks some of the above problems may be diluted to some extent. This is exactly what has been followed in this dissertation.

CHAPTER III

NUMERICAL REPRESENTATION OF DNA SEQUENCES

Since the completion of the *Human* Genome Program (HGP) [2],[22], deoxyribonucleic acid (DNA) sequence analysis is an important but challenging task. In particular, the identification of genes plays a critical role in gene annotation. DNA sequence analysis can generally be divided into two classes: DNA symbolic representation based analysis and DNA numerical sequence based analysis. In DNA symbolic representation based approach, computational techniques are applied directly on the DNA sequence for analysis. In DNA numerical sequence approach, each individual nucleotide of the DNA sequence is converted to numerical values through a mapping function. As a result, Digital signal processing (DSP) techniques [16],[23]-[25] has become increasingly important in genomic DNA research to reveal genome structures to identify hidden periodicities and features which cannot be revealed by conventional DNA symbolic and graphical representation techniques [26]. In this chapter, I shall discuss, classify, analyze and also evaluate the performance of 35 DNA numerical representation methods and its variants with DSP techniques like discrete Fourier transform (DFT) to classify human coding and noncoding sequences. At the end of this chapter I also discuss and compare various numerical representations and their applications with tables for the ease of use of the readers and some general observations based on the experiments performed is also discussed.

3.1 DNA Numerical Representation Methods

Numerical representation of DNA sequences is necessary to apply a wide range of mathematical tools, including most of signal processing and machine learning techniques. In recent years, a number of schemes have been introduced to map DNA nucleotides into numerical values. Some possible desirable characteristics [43] of a numerical representation include: 1) compact representation; 2) minimum redundancy; 3) Each nucleotide has equal magnitude; 4) complementary structure of nucleotide pairs preserved; 5) Distances between all nucleotide pairs are equal; 6) biological, structural, and statistical information is captured or well modeled in mathematical properties; 7) ability to capture information in three reading frames; 8) representation should not introduce any bias that may give rise to spurious results; 9) Feasible to reconstruct the DNA sequence back from the numerical representation; 10) compatibility with different mathematical analysis tools.

The DNA numerical sequence representation can be divided into three classes: Fixed mapping, DNA physico chemical property based mapping, and statistical property based mapping (summarized in Table 3.1, 3.2, and 3.3). In fixed mapping techniques, the nucleotides of DNA data are transformed into a series of arbitrary numerical sequences. Fixed mapping include the Voss [27]-[29], the tetrahedron [16,30], the 2-bit binary [10,31], the 3-bit binary [32], the 4-bit binary [33], the paired nucleotide [34]-[35], the integer number [36], the real number [37]-[38], the complex number [36],[39]-[40] the pentanary code [41], the quaternion [42]-[44] the 12-letter alphabet [29],[45] and the 18-letter alphabet [46].

In DNA physico chemical property based mapping, biophysical and biochemical properties of DNA biomolecules are used for DNA sequence mapping, which is robust

and is used to search for biological principles and structures in biomolecules. This mapping includes the EIIP [47]-[48], the atomic number [49], the paired numeric [42]-[43], the molecular mass [50]-[51], the paired nucleotide atomic number [49], the DNA walk [52]-[54], the Z-curve [55]-[60], the digital Z-signals [61], the phase specific Z-curve [62], the simple Z [63], the genetic code context [64].

In statistical property based mapping the DNA alphabets are mapped in terms of different properties like the inter-nucleotide distance [65], the binucleotide distance [66], the cumulative categorical periodogram (CCP) [67], the single nucleotide bias probability indicators [68], the correlation function [69], the position count function (PCF) [70], the codon index based on recurrence time [71], the ratio-R [72], the Galois field [73], the complexity [52], and the frequency of nucleotide occurrence [42]-[43]. As compared to the DNA symbolic representation approach, the DNA numerical sequence approach can preserve nucleotide positional information and retain all the available sequence information. Among these three groups, most of the DNA physico chemical property based mapping methods performs better as they are able to reflect the biological properties very well in the numerical domain.

It is well known that protein coding regions of DNA sequences exhibit a period-3 behavior due to codon structure. Identification of the period-3 regions helps in predicting the gene locations, and in fact allows the prediction of specific exons within the genes of eukaryotic cells. Traditionally these regions are identified with the help of discrete Fourier Transform (DFT).

Although good results have been obtained in the recognition of coding and noncoding regions of prokaryote genes, the strengths of the statistical features are not

sufficient to identify exons in humans because of their limited average length. So the classification of coding and noncoding sequences in humans is still a difficult problem in bioinformatics [63].

3.2 Review of Existing DNA numerical Representation Methods

Several authors considered different numerical mapping schemes for gene detection. Reference [37] describes the effects of binary representation; the integer, real, complex representations; and the DNA walk representation for autoregressive modeling and feature analysis of DNA sequences. In [42]-[43] the Voss, the tetrahedron, the real number and its variants, the complex, the quaternion, the EIIP, the paired numeric, the frequency of nucleotide occurrence, and the Z-curve representations are compared using the discrete Fourier transform for period-3 based exon (coding region) prediction. In [74] seven different mapping rules are discussed. In [75] we made an attempt to provide a survey on various numerical representation methods. We classified various numerical representations to two main groups; discussed each representation with a numerical example, its merits and demerits, applications, and also some concluding remarks.

The graphical representation of DNA alphabets is central to digital signal processing based DNA visualization and analysis. Perhaps the most widely used visualization scheme for this purpose is the Fourier spectrum of Voss representation [28], however many other schemes have also been introduced, such as the tetrahedron [16,30], the DNA walk [52], and the Z-curve [55].

Reference [26] describes about various non-DSP based graphical representation techniques to visualize DNA sequence in one-, two-, or three-dimensions. There has

been a limited attempt to provide a survey on various DSP based graphical representation methods.

In [76] we made an attempt to provide a survey on various graphical representations methods. In reference [76], we discussed each graphical representation with a figure, its merits and demerits, applications, and also some concluding remarks.

It is seen that there has been a limited attempt to provide a survey on all the existing DNA numerical representation methods. Also, there has been limited attempt to apply numerical sequence approach to classify human coding and noncoding sequences for the purpose of performance evaluation of the three groups of DNA numerical representation methods.

3.3 Fixed Mapping

In fixed mapping techniques, the nucleotides of DNA data are transformed into a series of arbitrary numerical sequences. Fixed mapping include the Voss, the tetrahedron, the 2-bit binary, the 3-bit binary, the 4-bit binary, the paired nucleotide, the integer number, the real number, the complex number, the pentanary code, the quaternion, the 12-letter alphabet, and the 18-letter alphabet.

3.3.1 Methodology

The Voss representation [27]-[29] maps the nucleotides A, C, G, and T into four binary indicator sequences as A_n, C_n, G_n, and T_n showing the presence with 1 or absence with 0 of the respective nucleotide. In the tetrahedron method [16, 30], the four sequences [A_n, C_n, G_n, T_n] are mapped to the four vertices of a regular tetrahedron which reduces the

number of indicator sequences from four to three but in a manner symmetric to all the four sequences.

The 2-bit binary representation [10,31] maps the nucleotides A, C, G, and T into two-bit binary namely, 00, 11, 10, 01 respectively resulting into a 1-dimensional indicator sequence. The 3-bit binary representation [32] is a 1-dimensional mapping of DNA bases which can be obtained by mapping the nucleotides A, C, G, and T as 100, 010, 001 and 000 respectively. Similarly, in the 4-bit binary encoding [33] the nucleotides A, C, G, and T are mapped as 1000, 0010, 0001 and 0100 respectively resulting into a 1-dimensional indicator sequence.

The paired nucleotide representation [34]-[35] assigns binary values to a set of two letter DNA alphabets. In the first convention all A, T bases are assigned 0 and all C, G bases are assigned 1, In the second convention all C, T bases are assigned 0 and all A, G bases are assigned 1, and in the third convention all G, T bases are assigned 0 and all A, C bases are assigned 1 thus leading to a 1-dimensional indicator sequence of three different conventions.

The integer representation [36] is a one-dimensional (1-D) mapping of DNA bases which can be obtained by mapping numerals {0, 1, 2, 3} to the four nucleotides as: T=0, C=1, A=2, and G=3. In the real number representation [37], [38], A = -1.5, T = 1.5, C = 0.5, and G = -0.5, which bears complementary property and is efficient in finding the complimentary strand of a DNA sequence. The complex representation [10], [36], [39]-[40] reflects the complementary nature of A-T and C-G pairs as A = 1+j, C = -1+j, G = -1-j, and T = 1-j. This results in a 1, 2 or 4 dimensional mapping of DNA bases.

The pentanary code representation [41] is a one-dimensional (1-D) mapping of DNA bases which can be obtained by mapping complex numerals {j, -j, 1, -1, 0} to the four nucleotides as: T=j, C=-j, A=1, G=-1, and unknown nucleotide X=0. In the quaternion representation [42]-[44] of DNA bases, pure quaternions are assigned to each base: A= i+j+k, C= i-j-k, G= -i-j+k, and T= -i+j-k resulting in a 1 or 4 dimensional mapping of DNA bases.

The binary mapping of 12-symbol alphabet representation [29],[45] reflects the nucleotide composition within codons. Coding regions may be described more completely with the 12-symbol alphabet due to the inherent codon bias in exons. Conversely, this codon bias is not common in introns. The phase p of a nucleotide located at a position n where $n \in \{0, 1, 2, 3, 4, 5, \ldots, N\text{-}1\}$ inside a DNA sequence of length N is defined as $p = n \bmod 3$ where $p \in \{0, 1, 2\}$. Each nucleotide of a DNA sequence is substituted by an appropriate symbol from the alphabet: $A_{12}=\{A_0,A_1,A_2,C_0,C_1,C_2,G_0,G_1,G_2,T_0,T_1,T_2\}$ and is represented by a 1-D binary vector of size 1xN. For an example using the A_{12} alphabet, the DNA sequence CGAT of length N = 4 (shown in Table 3.1) is translated into a N-symbol sequence $C_0\ G_1\ A_2\ T_0$ with each symbol represented by a binary vector of length N = 4 in which C_0 means that there is a nucleotide C with p = 0 in the DNA sequence which is assigned a value 1 in position n = 0 as [1, 0, 0, 0]; G_1 means that there is a nucleotide G with p = 1 which is assigned a value 1 in position n = 1 as [0, 1, 0, 0]; A_2 means that there is a nucleotide A with p = 2 which is assigned a value 1 in position n = 2 as [0, 0, 1, 0]; T_0 means that there is a nucleotide T with p = 0 which is assigned a value 1 in position n = 3 as [0, 0, 0, 1]; and the absence of a nucleotide in each of the remaining eight symbols of the A_{12} alphabet is

represented by a zero vector of [0, 0, 0, 0]. This 12-symbol alphabet mapping results in a 12-dimensional (12-D) binary indicator sequence with each dimension represented by a 1-D binary vector of size 1xN. The DFT power spectrum of the DNA sequence CGAT is obtained by summing the absolute squared of the DFT of each of the four non-zero 1x4 binary vectors C_0, G_1, A_2, T_0. Alternatively, this mapping can yield a 1-D binary indicator sequence of size 1x12N.

The 18-symbol alphabet representation [46] is the extension of the A_{12} representation that takes into account the non-uniform distribution of stop codons along the three phases $p \in \{0, 1, 2\}$. For the 18-symbol alphabet, the DNA nucleotides and the stop codons on both strands of DNA are substituted by symbols from the alphabet A_{18}={A_0,A_1,A_2,C_0,C_1,C_2,G_0,G_1,G_2,T_0,T_1,T_2,S_0,S_1,S_2,S_0',S_1',S_2'}. The symbols S_0, S_1, and S_2 represent any of the stop codons TAA, TGA, TAG in a given DNA strand, and S_0', S_1', and S_2' represent any of the corresponding stop codons TTA, TCA, CTA on its reverse DNA strand. For a DNA sequence CGAT of length N = 4 containing no stop codon, the sequence is converted using the A_{18} alphabet representation as $C_0G_1A_2T_0$. Consider another DNA sequence TGACGCTA of length N = 8 containing 2 stop codons, the sequence is converted using the A_{18} alphabet representation as $S_0T_0G_1A_2C_0G_1S_2'C_2T_0A_1$ by searching stop codons three consecutive nucleotides at a time after obtaining its 12-symbol alphabet representation as follow. Step 1: Based on the 12-symbol alphabet representation, we obtain $T_0G_1A_2C_0G_1C_2T_0A_1$. Step 2: TGA codon emerges with phase 0 is a stop codon, yielding representation S_0T_0. Step 2: GAC codon emerges with phase 1 is a not stop codon, yielding representation $S_0T_0G_1$. Step 3: ACG codon emerges with phase 2 is not a stop codon, yield representation $S_0T_0G_1A_2$. Step 4: CGC emerges with phase 0

is not a stop codon, yield representation $S_0T_0G_1A_2C_0$. Step 5: GCT codon emerges with phase 1 is not a stop codon, yielding representation $S_0T_0G_1A_2C_0G_1$. Step 5: CTA codon emerges with phase 2 is a stop codon on its reverse DNA strand, yielding the complete representation $S_0T_0G_1A_2C_0G_1S_2'C_2T_0A_1$. The presence (or absence) of a nucleotide in a position n is represented by 1 (or 0) in a 1xN vector. Therefore, the symbol S_0 = [1, 0, 0, 0, 0, 0, 0, 0], the symbol T_0 = [1, 0, 0, 0, 0, 0, 1, 0], and etc. The absence of a nucleotide in each of the remaining ten symbols of the A_{18} alphabet is represented by a zero vector of [0, 0, 0, 0, 0, 0, 0, 0]. This 18-symbol alphabet mapping results in a 18-dimensional (18-D) binary indicator sequence with each dimension represented by a 1xN vector. The DFT power spectrum of the DNA sequence TGACGCTA is obtained by summing the absolute squared of the DFT of each of the ten non-zero vectors with each vector of dimension 1xN. Alternatively, this mapping can yield a 1-D binary indicator sequence of size 1x18N.

3.3.2 Merits and Demerits

Numerical representation of a DNA sequence when it is being used in conjunction with DSP techniques can identify hidden periodicities, nucleotide distributions, and features which cannot be revealed easily by conventional methods such as DNA symbolic and graphical representations. A summary of the merits and demerits of the fixed mapping methods is shown in Table 3.1.

Each of the DNA numerical representations in fixed mapping offers different properties, and maps a DNA sequence into one to eighteen numerical sequences. The Voss binary indicator representation of a DNA sequence does not predefine any mathematical relationship among the bases, but only indicates the frequencies of the bases [64]. Voss or

the tetrahedron maps a DNA sequence onto four or three numerical sequence, potentially introducing different redundancy in each individual representation [43].

The 2-bit binary representation features linearly dependent set vectors, where the Hamming distance among the vectors is different [31]. The 3-bit binary representation is suitable for inductive inference [32]. The 4-bit binary representation is a unitary coding matrix with identical Hamming distance among each vector [31]. Thus the 2-bit, 3-bit, and the 4-bit binary representation are mostly applied for neural networks based systems that require extensive training [10],[31],[32],[33].

The paired nucleotide representation comprises of three different conventions that describe the distribution of the paired nucleotide {A,G}; {C,G}; {A,C} separately along the DNA sequence. In principle, the results obtained with each of these representations are independent of each other because they refer to different aspects of the DNA chain, all of them containing relevant information [34,35]. The integer representation implies a structure on the nucleotides such as purine (A, G) > pyrimidine (C, T) [42]. The real number representation bears complementary property [42], [43]. However, the assignment of a real number to each of the four bases does not necessarily reflect the structure present in a DNA sequence [42], [43].

The complex representation reflects some of the complementary features of the nucleotides in its mathematical representation [42], [43]. One of the disadvantages of periodicity transform (which it shares with the spectral methods) is its symbol bias that is inherent in the mapping of DNA symbols to complex numbers [44]. In the pentanary code representation the additional symbol X is introduced to deal the nucleotide ambiguity (that arises due to sequence estimation problems) in a numerical manner [41].

The quaternion representation results in a periodicity transform that is symbol balanced and that detects all repetitive structures [44]. Quaternion representation is limited to specific mathematical analysis tools. For example, a discrete quaternion Fourier transform (DQFT) based spectral analysis is required to detect certain DNA patterns [43]. The 12-letter alphabet representation captures the differential base composition at each codon position in a DNA sequence [45]. The 18-letter alphabet representation captures the differential base composition in codons and the differential stop codon composition along the three phases in both DNA strands [46].

3.3.3 Applications

The Voss representation has shown to be useful for locating periodicities and for locating potential gene sequences characterized by simple nonuniform codon usage. The power spectrum of the Voss's binary indicator sequence reveals a peak at frequency 0.33 (period-3) for coding region and shows no peak for noncoding region in *S.Cerevisiae* chromosome #3 [28]. The Frequency-domain analysis of DNA sequences (with the Voss representation) is a powerful tool for specifically identifying gene coding regions in DNA sequences of *C-elegans* chromosome #33 [16,28]**.**

Studies [28, 64, 67] indicate that one of the drawbacks of power spectrum analysis of Voss representation is that it is not able to detect period-3 feature in some *viral* and *bacterial* organisms and it produces the same Fourier spectra for two different DNA sequences which is not suitable for DNA sequence comparison. The Voss and the tetrahedron representations are two equivalent representations when being used in power spectrum analysis.

The 2-bit binary representation is applied for a neural network based multi-classifier system for gene identification in DNA sequences [10], the 3-bit binary representation is employed for splice junction recognition on gene sequences by BRAIN learning algorithm [32], and the 4-bit binary representation is applied for neural network optimization for *E.coli* promoter prediction [33]. With paired nucleotide representation wavelets are used to smooth G+C profiles to locate characteristic patterns in genome sequences and secondly, a wavelet scalogram is used as a measure for sequence profile comparison in bacterial genomes [34]-[35].

Integer and real representation is applied in autoregressive modeling and feature analysis of DNA sequences [37]. Integer and real representation is also applied to identify protein coding regions using the DFT based spectral content measure [42,43]. Arbitrarily assigned integer and real number representations may introduce some mathematical property which does not exist in a base sequence. Hence their DSP applications are limited suggesting that these integer and real mappings need to be used carefully for a given application [43].

Various complex representation based analysis tools are phase analysis, 2D (complex) and 3D (vector) sequence path analysis and stem diagram [39,40]. The Complex representation is applied in conjunction with digital signal processing (DSP) techniques for analyzing long range correlations in DNA sequences [41,52]. Pentanary complex representation resulted in improved results for the wavelet transform of the 3-dimensional DNA walk for *HUMDPI* sequence [41] when compared to using quaternary complex representation. Genomic researchers are encouraged to use pentanary complex representation widely.

Quaternion based representation can be analyzed only in conjunction with the DQFT to detect certain DNA patterns. It was conjectured [42]-[44] that the quaternion approach could improve DNA pattern detection via the discrete quaternionic Fourier transform (DQFT). DFT power spectrum using 12-letter alphabet representation produces a stronger spectral component for *bactereophage phi-X174* [29] when compared with Voss representation for the same DNA sequence. The 12-letter alphabet is applied to find borders between coding and noncoding DNA regions by an entropic segmentation method for *R. prowazekii, E.coli, M.jannaschii* DNA sequences, whose results are more accurate than those obtained with moving window technique [45].

The recursive segmentation with Jensen-Shannon divergence with 18-letter alphabet achieves maximum accuracies for the whole genomes of two bacteria, *Rickettsia prowazekii* and *Borrelia burgdorferi* compared to Jensen-Renyi divergence with 12-letter alphabet for the detection of borders between coding and noncoding regions [46].

Table 3.1 Fixed Mapping Numerical Representation of DNA Sequences, Merits, Demerits, and Applications (1: Voss; 2: Tetrahedron; 3: 2-Bit Binary; 4: 3-Bit Binary; 5: 4-Bit Binary; 6: Paired Nucleotide; 7: Integer; 8: Real; 9: Complex; 10: Pentanary Code; 11: Quaternion; 12: 12-Letter Alphabet; 13: 18-Letter Alphabet)

	Representation	**S(n) = [CGAT]**	**I**	**Merits**	**Demerits**	**Applications**
1	$X_n = 1$ for S(n) = X $X_n = 0$ for S(n) ≠ X X_n applies to any of C_n, G_n, A_n, T_n	C_n = [1, 0, 0, 0] G_n = [0, 1, 0, 0] A_n = [0, 0, 1, 0] T_n = [0, 0, 0, 1]	4	Efficient spectral detector of base distribution and periodicity features; offering numerical and graphical visualization [16,28,29].	Redundancy [43]; linearly dependent set of representation.	Identification of protein coding regions [16,28].
2	$x_r(n)=\frac{\sqrt{2}}{3}[2T_n-C_n-G_n]$ $x_g(n)=\frac{\sqrt{6}}{3}[C_n-G_n]$ $x_b(n)=\frac{1}{3}[3A_n-T_n-C_n-G_n]$	$x_r(n)=\frac{\sqrt{2}}{3}$[-1,-1,0,2] $x_g(n)=\frac{\sqrt{6}}{3}$[1,-1,0,0] $x_b(n)=\frac{1}{3}$[-1,-1,3,-1]	3	Periodicity detection [16].	Reduced redundancy [16].	Power spectrum analysis [16].
3	A=00,C=11, G=10, T=01	[11,10,00,01]	1		Linearly dependent vectors with different Hamming distances [33]; Training is	Gene identification with a neural network based multi-classifier

	Representation	S(n) = [CGAT]	I	Merits	Demerits	Applications
					required [31].	system [31].
4	A=100,C=010, G=001,T=000	[010,001,100,000]	1	Suitable for inductive inference [32].	Training is required [32].	Splice junction recognition with neural network [32].
5	A=1000,C=0010, G=0001,T=0100	[0010,0001, 1000,0100]	1	Unitary coding matrix with identical hamming distances [33].	Training is required. [33]	Promoter prediction with neural network [33].
6	A,G=1; C,T=0 C,G=1; A,T=0 G,T=1; A,C=0	[0110] [1100] [0101]	1	Describes the distribution of the paired nucleotide along the sequence [34, 35].		To locate characteristic patterns, sequence profile comparison with wavelets [35].
7	A = 2, C = 1, G = 3, T = 0	[1, 3, 2, 0]	1	Simple integer representation.	(A, G) > (C, T) ; introducing mathematical properties not present in DNA sequence [42].	Autoregressive modeling and feature analysis of DNA sequences [37].; Identifying protein coding regions using the DFT based spectral content measure [42],[43].
8	A = -1.5, C = 0.5, G = -0.5, T = 1.5	[0.5, -0.5, -1.5, 1.5]	1	A-T and C-G are complement. [42],[43].	Introducing mathematical properties not present in DNA sequence [42],[43].	Autoregressive modeling and feature analysis of DNA sequences [16]; Identifying protein coding regions using the DFT based spectral content measure [42],[43].
9	A = 1+j, C = -1-j, G = -1+j, T = 1-j	[-1-j, -1+j, 1+j, 1-j]	1	A-T and C-G are complex conjugate; reflecting complementary feature of nucleotides [42],[43].	Introducing base bias in time domain analysis [44].	Phase analysis, 2D (complex) and 3D (vector) sequence path analysis and stem diagram [39],[40].
		C_n = [-1-j, 0, 0, 0] G_n = [0, -1+j, 0, 0] A_n = [0, 0, 1+j, 0] T_n = [0, 0, 0, 1-j]	4			
10	A = 1, C = -j, G = -1, T = j, Unknown nucleotide X=0	[-j, -1, 1, j]	1	Deals effectively with sequencing estimation problems [41].		Wavelet transform of the 3-dimensional DNA walk [41].
11	A = i+j+k, C = i-j-k, G = -i-j+k, T = -i+j-k	[i-j-k, -i-j+k, i+j+k, -i+j-k]	1	Overcoming base bias [44].	Working with DQFT only [43].	Identifying protein coding regions using the DFT based spectral content measure [42], [43].
		C_n = [i-j-k, 0, 0, 0] G_n = [0, -i-j+k, 0, 0] A_n = [0, 0, i+j+k, 0] T_n = [0, 0, 0, -i+j-k]	4			
12	$A_{12} = \{A_0, A_1, A_2, C_0, C_1, C_2, G_0, G_1, G_2, T_0, T_1, T_2\}$	A_0 = [0, 0, 0, 0] ; A_1 = [0, 0, 0, 0] A_2 = [0, 0, 1, 0] ; C_0 = [1, 0, 0, 0] C_1 = [0, 0, 0, 0] ; C_2 = [0, 0, 0, 0] G_0 = [0, 0, 0, 0] ; G_1 = [0, 1, 0, 0] G_2 = [0, 0, 0, 0] ; T_0 = [0, 0, 0, 1]	12	Captures differential base composition at each codon position [45].		Produces a stronger spectral component when compared with Voss mapping [29]; finding borders between coding and noncoding DNA regions [45].

	Representation	S(n) = [CGAT]	I	Merits	Demerits	Applications
		T_1 = [0, 0, 0, 0] ; T_2 = [0, 0, 0, 0]				
13	$A_{18} = \{A_0, A_1, A_2, C_0, C_1, C_2, G_0, G_1, G_2, T_0, T_1, T_2, S_0, S_1, S_2, S'_0, S'_1, S'_2\}$	A_0 = [0, 0, 0, 0] ; A_1 = [0, 0, 0, 0] A_2 = [0, 0, 1, 0] ; C_0 = [1, 0, 0, 0] C_1 = [0, 0, 0, 0] ; C_2 = [0, 0, 0, 0] G_0 = [0, 0, 0, 0] ; G_1 = [0, 1, 0, 0] G_2 = [0, 0, 0, 0] ; T_0 = [0, 0, 0, 1] T_1 = [0, 0, 0, 0] ; T_2 = [0, 0, 0, 0] S_0 = [0, 0, 0, 0] ; S_1 = [0, 0, 0, 0] S_2 = [0, 0, 0, 0] ; S'_0 = [0, 0, 0, 0] S'_1 = [0, 0, 0, 0] ; S'_2 = [0, 0, 0, 0]	18	Captures (i) differential base composition in codons, and (ii) the differential stop codon composition along three phases in both DNA strands [46].		Achieves maximum accuracies compared to 12-letter alphabet for the detection of borders between coding and noncoding regions [46].

3.4 DNA Physico Chemical Property based Mapping

In this type of mapping, biophysical and biochemical properties of DNA biomolecules are used for DNA sequence mapping, which is robust and is used to search for biological principles and structures in biomolecules. This mapping includes the EIIP, the atomic number, the paired numeric, the molecular mass, the paired nucleotide atomic number, the DNA walk, the Z-curve, the digital Z-signals, the phase specific Z-curve, the simple Z, the genetic code context.

3.4.1 Methodology

EIIP [47] represents the distribution of the free electrons' energies along a DNA sequence. A single EIIP indicator sequence [48] is formed by substituting the EIIP of the nucleotides A=0.1260, C=0.1340, G=0.0806, and T=0.1335 in a DNA sequence. A single

atomic number indicator sequence [49] is formed by assigning the atomic number in each nucleotide as: A=70, C=58, G=78 and T=66 in a DNA sequence.

In the paired numeric representation [42]-[43], nucleotides (A-T, C-G) are to be paired in a complementary manner and values of +1 and -1 are to be used respectively to denote A-T and C-G nucleotide pairs. It can be represented as one or two indicator sequences.

The molecular mass representation [50]-[51] is a 1-dimensional indicator sequence formed by mapping the molecular mass of the nucleotides A=134, C=110, G=150, and T=125 respectively in a DNA sequence.

In paired nucleotide atomic number representation [49] the paired nucleotides are assigned with the atomic numbers A,G=62 and C,T=42 respectively resulting in a 1-dimensional indicator sequence.

The DNA-Walk model [52]-[54] shows a graph of a DNA sequence in which a step is taken upwards (+1) if the nucleotide is pyrimidine (C or T) or downwards (-1) if it is purine (A or G). The graph continues to move upwards and downwards as the sequence progresses in a cumulative manner, with its base number represented along the x-axis. Another variant of DNA walk is the complex DNA walk where in the nucleotides are assigned complex numbers as: A=1; G=-1; C=-j; and T=j respectively. The DNA walk can be used as a tool to visualize changes in nucleotide composition, base pair patterns, and evolution along a DNA sequence.

The Z-curve [55]-[60] is a 3-D curve that provides a unique representation for visualization and analysis of a DNA sequence. The three components of the Z-curve, $\{x_n, y_n, z_n\}$, represent three independent nucleotide distributions that completely describe a

DNA sequence. The components x_n, y_n, z_n display respectively the distributions of purine versus pyrimidine (R versus Y), amino versus keto (M versus K), and strong H-bond versus weak H-bond (S versus W) bases along the sequence.

Digital Z-signals [61] decomposes the DNA sequence into three series of digital signals, based on Z-curves. These three series of digital signals, Δx_n, Δy_n and Δz_n can only have the values of 1 or -1. Δx_n is equal to 1 when the n^{th} base is A or G (purine), or -1 when the n^{th} base is C or T (pyrimidine); Δy_n is equal to 1 when the n^{th} base is A or C (amino-type) or -1 when the n^{th} base is G or T (keto-type). Similarly, Δz_n is equal to 1 when the n^{th} base is A or T (weak hydrogen bond), or -1 when the n^{th} base is G or C (strong hydrogen bond).

Phase-specific Z-curves [62] describe the distribution of bases at first, second and third codon positions, respectively resulting in 9-dimensional (9-D) feature representation. The curve for the DNA sequence with bases at positions 0, 3, 6, ……, forms a phase-specific curve called phase-1 Z curve. Similarly, the Z curves with bases at positions 1, 4, 7, …., and 2, 5, 8, …., are called the phase-2 and phase-3 Z curves, respectively. Thus, the phase-1, phase-2, and phase-3 Z curves describe the distributions of bases at first, second and third codon positions respectively. For each phase-specific Z curve there are three components, as for the ordinary Z curve. The three components of the phase-1 Z curve are denoted by x_0, y_0, z_0, respectively, and $x_1, y_1, z_1, x_2, y_2, z_2$ are defined similarly.

The simple Z (SZ) representation [63] is obtained by performing the max operation on the 9-dimensional phase-specific Z-curves resulting in reduction of the size of SZ features to one-third of the phase specific Z-curves.

$x_n = \max\{x_0, x_1, x_2\}, y_n = \max\{y_0, y_1, y_2\}, z_n = \max\{z_0, z_1, z_2\}$ where $x_0, x_1, x_2, y_0, y_1, y_2, z_0, z_1, z_2$ are the values obtained from the phase-specific Z-curves and *n* is the position of the nucleotide in the DNA sequence.

Genetic code context (GCC) representation [64] incorporates the composition and distribution of the amino acid information in three reading frames. In this method, each consecutive codon from the three reading frames in a DNA sequence is converted to an amino acid and each amino acid in turn is represented by a unique complex number, of which the real parts and imaginary parts are from the hydrophobicity properties and residue volumes of the amino acids, respectively. This results in a single dimension indicator sequence in amino acid domain. For the example sequence CGAT. It has two consecutive codons; CGA and GAT, these two codons are converted to amino acids R and D (from Fig. 4 of reference [64]) and this two amino acids are further converted to numerical values by substituting the values from Table 1 of reference [64].

3.4.2 Merits and Demerits

Each of the DNA numerical representations in DNA physico chemical property based mapping reflects the DNA biological properties, and maps a DNA sequence into one to twelve numerical sequences.

The EIIP representation identifies protein coding regions in some genomes where the Voss representation fails to detect them and it also reduces the computational overhead by 75% [48]. There are a number of genes where both the Voss and the EIIP representation fails to detect protein coding regions [48].

The atomic number representation and paired nucleotide atomic number representation [49] is a new and recent mapping; further exploration is required to reveal its potentials.

The paired numeric representation incorporates a useful DNA structural property and is characterized by reduced complexity [42],[43]. The resultant one-sequence or two-sequence DNA representation offers reduction in DFT processing compared to the Voss, the Tetrahedron, and the Z-curve representations [42],[43]. The molecular mass representation [50] needs to be explored by genomic researchers.

The 3-D DNA walk based on complex representation provides useful information such as long range correlation information, sequence periodicities, and changes in nucleotide composition [52]. This technique is suitable only for DNA sequences of a few hundreds of base pairs, and for lengthy sequences the DNA walk tends to become too complicated since there is too much information for extracting anything useful [52].

For the Z-curve, the compositional patterns of genomic sequences can be quickly recognized in a perceivable form. Studies [55]-[60] indicate that the Z-curve representation is more robust and computationally efficient for DNA sequence analysis as compared to the Voss representation. The x_n, y_n, and z_n components of the Z-curve are independent [55] and contain all the information in a DNA sequence, and mapping between a DNA sequence and the Z-curve is bidirectional [55, 64]. Last but not the least, each of the x_n, y_n, and z_n components of the Z-curve generates a discrete signal that has a clear biological interpretation [55].

The Digital Z-signal has clear biological meaning that describes the distribution of the bases of purine/pyrimdine, amino/keto, and strong/weak hydrogen bonds along the sequence [61].

The phase-specific Z-curves capture information at three codon positions and offer good recognition rate using the Fisher discriminate algorithm [62]. The simple Z (SZ) representation will ensure that the SZ features capture information at the correct reading frame and maintain a good recognition rate using fewer features [63].

Genetic code context generate different Fourier spectrum for different sequences, unlike Voss representation that produce same Fourier spectrum for two different sequences [64]. GCC incorporates the amino acid information such as composition and its distribution in three reading frames [64].

3.4.3 Applications

The power spectrum of EIIP indicator sequence with a sliding Kaiser window is a computationally better tool for identifying gene coding regions in Burset and Guigó and HMR 195 datasets with certain limitations [48].

The atomic number representation is applied to study the nucleotide fluctuations in radiation resistance-repair genes in *Deinococcus radiodurans* (DR) and *E-coli* [49].

The paired numeric representation provides an improved accuracy in identifying protein coding regions with DFT-based spectral content measure over Voss, tetrahedron, integer, real, complex, quaternion, EIIP, inter-nucleotide distance, frequency of nucleotide occurrence, and Z-curve representations using GENSCAN data set [42]-[43].

Table 3.2 DNA Physico Chemical Property based Numerical Representation of DNA Sequences, Merits, Demerits, and Applications (14: EIIP; 15: Atomic Number; 16: Paired Numeric; 17: Molecular Mass; 18: Paired Nucleotide Atomic Number; 19: DNA Walk; 20: Z-Curve; 21: Digital Z-Signals; 22: Phase Specific Z-Curve; 23: Simple Z-Curve; 24: Genetic Code Context)

	Representation	S(n) = [CGAT]	I	Merits	Demerits	Applications
14	A= 0.1260, C= 0.1340,	[0.1340, 0.0806, 0.1260, 0.1335]	1	Reflecting DNA physico chemical	Failing to detect coding	Identifying protein coding regions with

	Representation	S(n) = [CGAT]	I	Merits	Demerits	Applications
	G= 0.0806, T= 0.1335	C_n = [0.1340, 0, 0, 0] G_n = [0, 0.0806, 0, 0] A_n = [0, 0, 0.1260, 0] T_n = [0, 0, 0, 0.1335]	4	property [47]; reducing computational overhead [48]; improving gene discrimination capability [48].	region in some genomes [48].	sliding Kaiser window.
15	A = 70, C = 58, G = 78, T = 66	[58, 78, 70, 66]	1	Reflecting DNA physico chemical property [49].	Requiring further exploration.	Nucleotide fluctuations in genes [49].
		C_n = [58, 0, 0, 0] G_n = [0, 78, 0, 0] A_n = [0, 0, 70, 0] T_n = [0, 0, 0, 66]	4			
16	A or T = 1, C or G = -1	P_{1n} = [-1, -1, 1, 1]	1	Reflecting DNA structural property[42],[43]; reduced complexity; reduced DFT processing [42],[43]; improved coding region identification accuracy over other methods [42],[42].	Requiring further exploration [42],[43].	Identifying protein coding regions using the DFT based spectral content measure [42],[43].
		P_{2n} = [-1, -1, 0, 0] & [0, 0, 1, 1]	2			
17	A=134,C=110, G=150, T=125	[110,150,134,125]	1	Reflecting DNA physico chemical property [50].	Requiring further exploration.	A standardized molecular weight is applied in identification of a new motif in gene data with Kohonen self-organizing map [51].
		C_n = [110, 0, 0, 0] G_n = [0, 150, 0, 0] A_n = [0, 0, 134, 0] T_n = [0, 0, 0, 125]	4			
18	A,G=62; C,T=42	P_{1n} = [42, 62, 62, 42]	1	Reflecting DNA physico chemical property [49].	Requiring further exploration.	Fractal dimension analysis of nucleotide fluctuations in DNA sequences[49].
		P_{2n} = [42, 0, 0, 42] & [0, 62, 62, 0]	2			
19	C or T = 1, A or G = -1	[1, -1, -1, 1]	1	Providing long range correlation information [52]; sequence periodicities [52]; changes in nucleotide composition [52]; offering numerical and graphical visualization [52].	Not suitable for lengthy (> 1000 bases) sequences [52].	Provides a graphical representation [52]; distinguishes between coding and noncoding sequences with quantitative measurement [52]; locates periodicities or nucleotide structures [52].
	A=1, G=-1, C=-j, T=j	[-j, -1, 1, j]	1			
20	$x_n=(A_n+G_n)-(C_n+T_n) \equiv R_n-Y_n$ $y_n=(A_n+C_n)-(G_n+T_n) \equiv M_n-K_n$ $z_n=(A_n+T_n)-(C_n+G_n) \equiv W_n-S_n$	x_n =[-1, 0, 1, 0] y_n =[1, 0, 1, 0] z_n =[-1,-2,-1, 0]	3	Clear biological interpretation [55]; independent x_n, y_n, z_n components [55,64]; bidirectional mapping; reduced computation; superior to sliding window technique; offering numerical and graphical visualization [55-60].		Gene identification and protein; genomic islands detection, comparative genomics, distribution of nucleotide composition; identifying replication origins; wavelet denoising technique for DNA sequences has been studied for isochore, coding region, and gene detections [55-60].

	Representation	S(n) = [CGAT]	I	Merits	Demerits	Applications
21	A,G=1; C,T=-1 A,C=1; G,T=-1 A,T=1; C,G=-1	$\triangle x_n = [-1,\ 1,\ 1,\ -1]$ $\triangle y_n = [1,\ -1,\ 1,\ -1]$ $\triangle z_n = [-1,\ -1,\ 1,\ 1]$	3	Clear biological meaning [61].		Detects small length coding regions with lengthen-shuffle Fourier transform algorithm [61].
22	$x_i = (A_i + G_i) - (C_i + T_i)$ $y_i = (A_i + C_i) - (G_i + T_i)$ $z_i = (A_i + T_i) - (C_i + G_i)$ $x_i, y_i, z_i \in [-1,1]$, $i = 0,1,2$	$x_0 = [-1, 0, 0, -1]$ $x_1 = [0, 1, 0, 0]$ $x_2 = [0, 0, 1, 0]$ $y_0 = [1, 0, 0, -1]$ $y_1 = [0, -1, 0, 0]$ $y_2 = [0, 0, 1, 0]$ $z_0 = [-1, 0, 0, 1]$ $z_1 = [0, -1, 0, 0]$ $z_2 = [0, 0, 1, 0]$	9	Capture information at three codon positions; good recognition rate [62].	Higher number of features (9) [62].	Classify the coding from non-coding regions by Fisher discriminate analysis [62].
23	$\left\{ \begin{array}{l} \max_i[(A_i + G_i) - (C_i + T_i)] \\ \max_i[(A_i + C_i) - (G_i + T_i)] \\ \max_i[(A_i + T_i) - (G_i + C_i)] \end{array} \right\}$ $(i = 1,2,3)$.	$x_n = [0, 0, 1, 0]$ $y_n = [1, 0, 1, 0]$ $z_n = [0, 0, 1, 1]$	3	Capture information at the correct reading frame; good recognition rate using fewer features (3) [63].	Applied for short length sequences (140 nt) [63].	Recognition of short human exons and introns using the K-nearest-neighbor (KNN) classifier [63].
24	Amino acid CGA =R=0.60+181.2j Amino acid GAT =D=0.46+110.8j	[0.60+181.2j, 0.46+110.8j]	1	Unique numerical and as well spectral representation; incorporates amino acid composition and distribution; includes the nucleotide information for all the three reading frames [64].	Requiring further exploration [64].	Period-3 property was successfully investigated in DNA sequences [64].

The standardized molecular weight is applied in identification of a new motif in *human insulin receptor* gene data with Kohonen self-organizing map [51]. The paired nucleotide atomic number representation has been applied to study the fractal dimension analysis of nucleotide fluctuations in *human* and *chimpanzee* DNA sequences [49].

The DNA walk provides a graphical representation for each gene and permits the degree of correlation in the base pair (GC versus AT) sequence to be directly visualized [52]. DNA walk technique also distinguishes between coding and noncoding sequences with quantitative measurement. This has been applied to *viral*, *bacterial*, *yeast* and *mammalian* DNA sequences [41, 52-54]. Locating periodicities or nucleotide structure may be found

using complex DNA walk representation [52]. Wavelet analysis is applied to extract structural information (corresponding to sequence periodicity) from the complex DNA walk in the transform domain [52]. Wavelet transforms also allows measures to quantify correlations based on DNA walk [54].

The Z-curve was employed in gene identification and protein coding measurement of DNA [56]-[57]. The distribution of the C+G content in the *human* genome has been studied by using a windowless technique derived from the Z-curve method [57]. The Z-curve's detected distribution of the C+G content along genomic sequences (such as *budding yeast* and *vibro cholerae* genomes [56]) is generally superior to the widely used sliding-window technique. The Z-curve has been used for horizontally transferred genomic islands detection, comparative genomics, studying the distribution of nucleotide composition [58], and identifying replication origins of *archaeal* genomes with 3-D Z-curve and 2-D Z-curve based on CG and AT disparity [58]. A Z-curve based wavelet denoising technique for DNA sequences has been studied in [59]-[60] for isochore, coding region, and gene detections.

Digital Z-signal representation in conjunction with lengthen-shuffle Fourier transform algorithm has been successful in detecting period-3 property in short length coding regions in Fickett and Tang benchmark data sets [61].

Phase-specific Z-curves forms a 9-feature vector which helps to classify the coding from non-coding regions in the *yeast* genome at better than 95% accuracy by Fisher discriminate analysis [62].

SZ representation is applied for the recognition of short *human* exons and introns using the K-nearest-neighbor (KNN) classifier [63].

The period-3 property was successfully investigated in DNA sequences of *myeloid zinc finger* protein 1 splice variants ZNF42 gene of *Human* genome using the GCC mapping method [64]. GCC method has more potential in gene finding and DNA sequence classification and function prediction [64].

3.5 Statistical Property based Mapping[1]

In this mapping the DNA alphabets are mapped in terms of different properties like the nucleotide distance, the categorical data, the nucleotide bias in terms of the coding and noncoding statistics, the correlation function, the position count function, the codon index based on recurrence time, the ratio-R, the Galious field, the multinomial characteristics of nucleotides, and the frequency of nucleotide occurrences. [1]Note: The name "Statistical Property based Mapping" was suggested by Dr. H. K. Kwan to replace "Other Mapping Methods".

3.5.1 Methodology

In the inter-nucleotide distance representation [65] each base symbol is replaced by a number k which is the base distance between the next similar base in the DNA sequence. In case a similar base is not found then the sequence value of that base is the length of the remaining base in the DNA sequence. This is represented as one dimensional indicator sequence.

In the binucleotide distance representation [66] every base 'A' is replaced by a number k which is the distance to the next base 'T', every base 'T' by the distance to the next base 'A', every base 'C' by the distance to the next base 'G' and every base 'G' by the

distance to the next base ‘C’. In case a similar base is not found then the sequence value of that base is the length of the remaining base in the DNA sequence.

The cumulative categorical periodogram (CCP) representation [67] is based on the concept of categorical periodograms. A categorical periodogram is a numeric sequence with the n^{th} element of the sequence indicating the number of occurrences of cycles with period *n* in it. The period of the cycle is defined as the number of intervening events plus one. CCP measures the existence of pairs of identical elements at a distance of k base pairs. In a DNA sequence CGAT, as there are 4 categorical elements, cycles of various periods between A & A, C & C, G & G, and T & T emerge that are quantified in CCP for the complete DNA sequence.

The single nucleotide bias probability indicator [68] is the ratio of the normalized frequencies of nucleotides A, C, G, and T in the coding and noncoding regions of the dataset constructed by Burset and Guigo [68]. That is, if $f_e(A)$, $f_e(C)$, $f_e(G)$, $f_e(T)$ are the frequencies of A, C, G, and T normalized to the length of exons and $f_i(A)$, $f_i(C)$, $f_i(G)$, $f_i(T)$ are the corresponding ones for introns, of the same genome, the (nucleotide) relative frequency indicator $f_r(X)= f_e(X)/ f_i(X)$, for X=A, C, G, or T. This can replace the ‘ones’ in the binary indicator sequences in the original technique of filtering. The resulting sequence is renamed as single nucleotide bias probability indicator sequence in 1-dimension.

The Correlation function [69] compares each base in a DNA sequence to its various neighbors along the sequence, scoring 1 when the two bases are identical and 0 otherwise and then summing them along the complete DNA sequence and this process is repeated from the first base to the last base in the DNA sequence.

The position count function (PCF) [70] measures the number of times the nucleotides A, C, G, and T appear in the three positions within a codon along a DNA sequence.

The codon index based on recurrence time [71] represents a genomic DNA sequence hierarchically by quantifying repeating patterns in a genomic sequence so as to properly characterize the sequence period-3 feature and its enthropy. The DNA sequence is denoted by $S=b_1b_2b_3\ldots\ldots\ b_N$, where N is the sequence length and each b_i (i=1,,……, N) is a nucleotide base. Next, group consecutive nucleotide bases of window size *w* and call that a word of size *w*. using maximal overlapping sliding window W_i has n=N-*w*+1 such words.. Two words in W_i are considered equal if all their corresponding bases match and its recurrence time T(i) for a position i along the W_i sequence is the distance between the two matched words, if no match exists in the sequence such a situation is represented as T(i)=-1. This process is carried out for all the grouped words in W_i. Discarding all -1 terms from the T(i) sequence, one gets a recurrence time T(i) series with nontrivial terms.

The ratio-R representation [72] is the ratio of the count of bases (C or T) to count of bases (A or G) in an interval of window size defined by the user and repeat the process for the complete DNA sequence.

A single Galois indicator sequence GF (4) [73] is formed by assigning the numerical values to the nucleotides A=0, C=1, G=3, and T=2 in a DNA sequence. The complexity representation [52] of a DNA sequence is a quantification based on multinomial characteristics of the nucleotides. Given a window of length L, the local complexity state

Table 3.3 Statistical Property Based Mapping of DNA Sequences, Merits, Demerits, and Applications (25: Inter-Nucleotide Distance; 26: Binucleotide Distance; 27: CCP; 28: Single Nucleotide Probability Indicators; 29: Correlation Function; 30: PCF; 31: Codon Index based on Recurrence Time; 32: Ratio-R; 33: Galois Field; 34: Complexity; 35: Frequency of Nucleotide Occurrence)

	Representation	S(n) = [CGAT]	I	Merits	Demerits	Applications
25	C=4-1=3 G=4-2=2 A=4-3=1 T=4-4=0	[3, 2, 1, 0]	1	Based on distance measure [65].	Does not conserve the biological information [65]; pattern matching is better [65].	Reveals the existence of discriminatory spectral envelope in the coding region and for some in promoter regions of coding sequences [65].
26	C=2-1=1 G=4-2=2 A=4-3=1 T=4-4=0	[1, 2, 1, 0]	1	Based on distance measure [66].	Does not conserve the biological information [66]; requiring further exploration [66].	Reveals the existence of discriminatory spectral envelope in the coding region and for some in promoter regions of coding sequences [66].
27	Cycle count $A \leftrightarrow A$ = [0 0 0 0] Cycle count $C \leftrightarrow C$ = [0 0 0 0] Cycle count $G \leftrightarrow G$ = [0 0 0 0] Cycle count $T \leftrightarrow T$ = [0 0 0 0] CCP = [0 0 0 0]	[0, 0, 0, 0]	1	Unique representation and computationally efficient spectral detector of period-3 property compared to Voss representation [67].	Failed to detect period-3 in some prokaryotic genes [67].	Identification of period-3in gene coding regions [67].
28	A=0.19, G=0.20, C=0.27, T=0.36	[0.27, 0.20, 0.19, 0.36]	1	Incorporates genome statistics [68].	Model dependent [68].	Improves the discrimination capability of existing gene finding techniques [68].
		C_n = [0.27, 0, 0, 0] G_n = [0, 0.20, 0, 0] A_n = [0, 0, 0.19, 0] T_n = [0, 0, 0, 0.36]	4			
29	$S(0) = \sum_{i=1}^{3} g_{i,i+1} = 0$ $S(1) = \sum_{i=1}^{2} g_{i,i+2} = 0$ $S(2) = \sum_{i=1}^{1} g_{i,i+3} = 0$	[0, 0, 0]	1	A visualization tool for rapidly scanning the sequences for various regular patterns with Fourier and wavelet transforms [69].		Displays regular patterns in DNA sequences [69].
30	$C_w^A(s) = \sum_{i=0}^{\lfloor \frac{N-1}{w} \rfloor} A[wi+s]$ $C_w^C(s) = \sum_{i=0}^{\lfloor \frac{N-1}{w} \rfloor} C[wi+s]$ $C_w^G(s) = \sum_{i=0}^{\lfloor \frac{N-1}{w} \rfloor} G[wi+s]$ $C_w^T(s) = \sum_{i=0}^{\lfloor \frac{N-1}{w} \rfloor} T[wi+s]$ for $(0 \le s < w)$	$C_3^A(0) = 0;\ C_3^A(1) = 0;$ $C_3^A(2) = 1;\ C_3^C(0) = 1;$ $C_3^C(1) = 0;\ C_3^C(2) = 0;$ $C_3^G(0) = 0;\ C_3^G(1) = 1;$ $C_3^G(2) = 0;\ C_3^T(0) = 1;$ $C_3^T(1) = 0;\ C_3^T(2) = 0$	12	Provides an automated DFT based approach in predicting coding regions [70]; computationally efficient and faster than STFT-based algorithms [70].	Function of window size [70]; cannot predict the actual boundaries between coding and noncoding regions [70]; inability to detect very small coding regions [70].	Identification of protein coding regions [70].
31	$W_i = CGA, GAT$ $T(i) = 0, -1$ $T(i) = 0$	[0]	1	Identifies dominant features first before compressing [71]; it is about 7% more accurate than Fourier transform	Slightly lower accuracy than Fourier transform for C.elegans and human genome in characterizing the period-3 property [71].	Identify protein-coding regions [71].

	Representation	S(n) = [CGAT]	I	Merits	Demerits	Applications
				method for yeast genome in characterizing the period-3 property [71]; it is largely species independent [71]; works well on short DNA sequences [71].		
32	R(1)=0.5; R(2)=0.5	[0.5, 0.5]	1	Based on ratio of nucleotides [72].	Function of interval size [72].	Study of long range correlation [72].
33	A=0; C=1; G=3; T=2	[1, 3, 0, 2]	1	Symbolic Galois field operations [73].	Requiring further exploration [73].	DNA sequence analysis [73].
34	A=0.79; C=0.60; G=0.35; T=0.00	[0.60, 0.35, 0.79,0.0]	1	Suitable method for visualizing different complexity domains [52].	Not suitable for length (>1000 bases) sequences [52].	To display regions exhibiting periodicity [52].
35	A=0.23326,C=0.28142, G=0.28179, T=0.20354.	[0.28142, 0.28179, 0.23326, 0.20354]	1	Improved results over all existing schemes, with less processing for the purpose of identifying protein coding regions [42],[43].	Data driven [42], [43]. Less superior than paired nucleotide representation [42],[43].	Identifying protein coding regions using the DFT based spectral content measure [42],[43].
		C_n = [0.2812,0, 0, 0] G_n=[0,0.28179, 0, 0] A_n=[0,0, 0.23326, 0] T_n= [0, 0, 0, 0.20354]	4			

vector for the nucleotide sequence is defined as n = [n_1, n_2, n_3, n_4]. Such that $n_1 \geq n_2 \geq n_3 \geq n_4$ are all nonnegative integers that satisfy $\sum_i n_i = L$. The number n_1 represents the number of occurrences of the most frequent nucleotide a_1, n_2 the number of the second most frequent nucleotide a_2 and so on where $a_i \in \{A,C,G,T\}$. The multinomial coefficient is given by

$$\Omega = \frac{L!}{n_1!n_2!n_3!n_4!} \tag{3.1}$$

and it is the total number of distinguishable arrangements of n_1 outcomes of a_1, n_2 outcomes of a_2 and so forth, over a sequence window length L. A measure of complexity is given by

$$K = \frac{1}{L} \ln \Omega \tag{3.2}$$

This process is repeated by moving the window along the length of the entire DNA sequence.

In the frequency of nucleotide occurrence representation [42]-[43], the nucleotides are represented by their frequency of occurrence of A, C, G, and T in exons of GENSCAN data set, allowing either single-or four-sequence representation of DNA.

3.5.2 Merits and Demerits

The inter-nucleotide and binucleotide signals are based on distance measurement between nucleotides along the DNA sequence. This is obviously not an information conserving representation. The pattern match for identifying promoter regions seems to be superior to the inter-nucleotide distance based spectral analysis at the moment. The inter-nucleotide and binucleotide representation require more fine tuning before it is ready for applications [65],[66].

CCP is a unique representation of the DNA sequence. It is observed that the spectral signatures in CCP are functionally equivalent to the established period-3 peak in the spectrum of Voss representation. CCP being a single sequence is computationally better compared to Voss representation [67].

By incorporating the single nucleotide probability indicator in the existing filtering techniques (model dependent) gave improved results in locating exons. It is also observed that in many cases the results are not very much appreciable. So incorporating other coding measures into DSP techniques may be explored for different DNA sequences [68].

The correlation function is a visualization tool for rapidly scanning the sequences for various regular patterns with Fourier and wavelet transforms [69].

PCF provides an automated DFT based approach in predicting coding regions. It is computationally efficient and faster than STFT-based algorithms. The performance of PCF based on DFT algorithm is efficient in detecting coding regions whose lengths are comparable to the window size (351). However, when the length of the coding regions is smaller than 150 nucleotides, the PCF based DFT algorithm, does not perform well and the prediction accuracy is relatively inaccurate [70].

The probability distribution of codon index based on recurrence time is an effective method to capture a DNA sequence's period-3 property. It identifies dominant features first before compressing; it is about 7% more accurate than Fourier transform method for yeast genome in characterizing the period-3 property, and slightly lower for *C.elegans* and *human* genomes. This strongly suggests that the method is largely species independent and it also works well on short DNA sequences [71].

The ratio-R is based on ratio of nucleotides, and is a function of interval size [72]. Galois field is based on symbolic Galois field operations [73]. Genomic researchers are encouraged to explore this representation.

Complexity representation is a suitable method for visualizing different complexity domains but not suitable for length (>1000 bases) sequences [52]. Frequency of nucleotide occurrence (data driven) provides improved results over all existing schemes but less than the paired numeric representation, with less processing for the purpose of identifying protein coding regions [42]-[43].

3.5.3 Applications

The inter-nucleotide and binucleotide signals are a novel way of digital signal representation of genomic data that reveals the existence of discriminatory spectral envelope (with DFT) in the coding region and for some in promoter regions of coding sequences in the Burset and Guigo datasets [65]-[66]. These methods require more fine tuning for which more works need to be done, before it is ready for application [65]-[66].

The spectral analysis of the CCP signals of DNA show prominent peaks which are exactly coinciding with the well known period-3 peaks in the spectrum of Voss indicator sequence of DNA. Spectral analysis studies have been conducted on CCP of 10 genes from the data set of Burset and Guigo [67]. By incorporating the single nucleotide probability indicator in the existing filtering techniques gave improved results in locating exons in most of the Burset and Guigo data set [68].

The Correlation function procedure has been applied in conjunction with Fourier and wavelet transforms to sequences from the *human* chromosome 22, to *nef* genes from various HIV clones and to *myosin* heavy chain DNA [69]. The position count function in conjunction with Bartlett window based DFT was applied for identification of protein coding regions in chromosome III of *C.elegans* [70].

The probability distribution of codon index based on recurrence time method is applied to identify protein-coding regions in chromosome 3 of *C.elegans*, chromosome 19 and 22 of *human* genome, and 16 yeast chromosomes [71].

Ratio-R representation has been applied exclusively to study long range correlation in chicken, *C.elegans*, *adenovirus*, and *human* DNA sequences [72]. Galois field enables powerful tools for DNA analysis that can be explored further by genome researchers [73]. Complexity representation is applied to *Helicobacter pylori*, strain J 99 bacteria sequence. Frequency of nucleotide is applied for exon detection in the GENSCAN data set of human genomic sequences using the DFT-based spectral content measure [42]-[43].

3.6 DSP based Features for Coding and Noncoding Region Classification

It has been observed by researchers that base sequences in the protein coding regions of DNA molecules exhibit a period-3 behavior due to codon structure involved in the translation of base sequences into amino acids [28],[77]. For eukaryotes (cells with nucleus, example *humans*) this periodicity has mostly been observed within the exons (coding sub regions inside the genes [14]) and not within the introns (noncoding sub regions in the genes). The period-3 behavior of exons has been widely used to identify these regions using DSP-methods, following conversion to numerical sequences.

The discrete Fourier transform (DFT), the most commonly used method for spectrum analysis of a finite-length numerical sequence $x[n]$ of length N, is defined as [78]:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j(2\pi nk/N)}, \qquad 0 \le k \le N-1 \tag{3.3}$$

Equation 1 can be used to calculate DFTs for numerical sequences representing DNA sequences, for example each of the four binary indicator sequences (i.e., $X_A[k], X_C[k], X_G[k], and\ X_T[k]$). The periodicity of 3 in exon regions of a DNA sequence suggests that the DFT coefficient corresponding to $k = N/3$ (where N is chosen to be a multiple of 3) in each DFT sequence should be large [16]. Note that the calculation of the DFT at a single frequency ($k = N/3$) is sufficient. The spectral content (SC) measure [43] combines the individual DFTs (i.e., $X_A[k], X_C[k], X_G[k], and\ X_T[k]$) to obtain a total Fourier magnitude spectrum of the DNA sequence, as follows:

$$S[k] = \sum_{m} \left| X_m[k] \right|^2, \qquad \text{m}=\{\text{A,C,G,T}\} \tag{3.4}$$

The GeneScan program [28], based on the SC measure, computes the signal-to-noise ratio of the peak at $k = N/3$ as $P = S[N/3] / \hat{S}$, where $\hat{S}$ represents the average of the total spectral content $S[k]$. Regions having $P \ge 4$ are assumed to be protein coding (exon) regions else considered as noncoding (introns) regions.

3.7 Performance Evaluation based on Matlab Simulation

1500 *human* exons and 1500 introns with length less than 140 bp (140 bp was chosen, since the average length of exons of vertebrate gene is 137 bp) were extracted from the database (http://bit.uq.edu.au/altExtron/). The exons are not frame-specific. Although introns in humans can be very long, short introns were selected to avoid introducing any bias in recognition due to length [79].

In this chapter, all the numerical representation methods are compared for the purpose of classifying the coding (1500 sequences) and noncoding regions (1500 sequences) in *human* dataset [79] as shown in Fig. 3.1, using the discrete Fourier transform (DFT) based on period-3 property using Matlab toolbox 7.5.0 (R2007b) as mentioned in appendix B. The classification criteria is the extracted period-3 value in Fourier domain divided by the sequence length (less than 140 bp); if this is equal or greater than 4 it is classified as coding otherwise it is a noncoding sequence. Reference [28] uses threshold value of 4 as a discriminator between coding and noncoding sequences based on survey of a large number of coding and noncoding sequences from a variety of organisms. In our experiments I also used this threshold value of 4, over different numerical representations and its variants. Precision is used for the evaluation of the classification results which is defined as the ratio of sum of correctly classified exons and introns to the total number of sequences presented.

For 4-D Voss, 3-D Tetrahedron, 4-D Complex, 4-D Quaternion,12-D 12-letter alphabet, 18-D 18-letter alphabet, 4-D EIIP, 4-D Atomic number, 2-D paired numeric, 4-D Molecular mass, 2-D paired nucleotide atomic number, 3-D Z-curve, 3-D Digital Z-signals, 9-D phase specific Z-curve, 3-D Simple Z-curve, 4-D Single nucleotide probability indicators, and 4-D Frequency of nucleotide occurrence, I have adopted sum of power spectra on their respective dimension values. For 4-D position count function (PCF) representation the period-3 DFT spectrum is computed by parsing the DNA sequences in codons and counting the number of different nucleotides at each of three phases. The approach is of O(N) and is computationally efficient than the fast Fourier transform (FFT) [70].

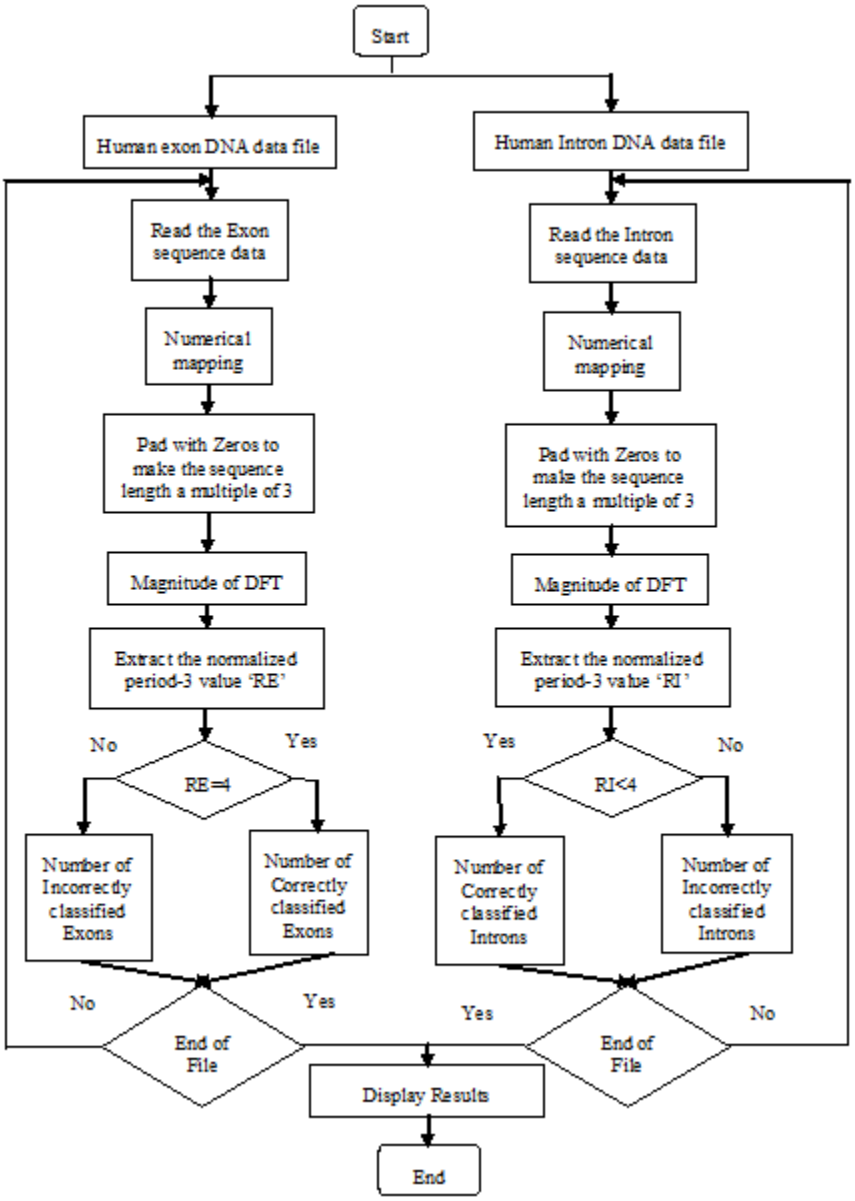

Fig. 3.1 Flowchart for *human* dataset classification with DFT using various numerical represeresentation methods.

The simulation results are shown in Table 3.4, 3.5, 3.6. Note '*' indicates that in [29],[45] the 12-letter alphabet representation have an inherent assumption of period-3 in both coding and noncoding regions. In principle, the results obtained with each of these

numerical representations are independent of each other, because they refer to the different aspects of the DNA molecule, all of them containing relevant information.

From Table 3.4, under fixed mapping methods, 4-D quaternion representation is the best with classification accuracy of 70.9% because the Quaternion representation results in a periodicity transform that is symbol balanced [42], that detects period-3 in short length DNA sequences with discrete quaternion Fourier transform (DQFT), and from Table 3.5, under DNA physicochemical property based representation, phase specific Z-curve is the best with classification accuracy of 73.5% since the phase-specific Z-curve capture information at three codon positions respectively resulting in 9-dimensonal representation [62] that detects period-3 property in short length DNA sequences, and from Table 3.6, under statistical property based mapping, PCF representation is the best with classification accuracy of 61.2% because the position count function (PCF) [70] measures the number of times the nucleotides A, C, G, and T appear in the three positions within a codon along a DNA sequence and gives a direct measure of period-3 property in short length DNA sequences.

Overall from the 3 major groups of numerical representation methods, phase specific Z-curve is the best with classification accuracy of 73.5% for short length human dataset. The possible reasons for under performance of other numerical representation methods could be: incapable to detect period-3 property in short length DNA sequences; cannot capture information at three codon positions; contains redundancy; presence of bias; the biological, structural, and statistical information is not captured or not well modeled in mathematical properties.

Table 3.4 Fixed Mapping Numerical Representation of DNA Sequences Simulation Results (1: Voss; 2: Tetrahedron; 3: 2-Bit Binary; 4: 3-Bit Binary; 5: 4-Bit Binary; 6: Paired Nucleotide; 7: Integer; 8: Real; 9: Complex; 10: Pentanary Code; 11: Quaternion; 12: 12-Letter Alphabet; 13: 18-Letter Alphabet; Pre: Precision (%); I: Number of Indicator Sequences)

	Classification of Human exons (1500)		**Classification of Human Introns (1500)**		**Pre**	**I**
	Correct	**Incorrect**	**Correct**	**Incorrect**		
1	55	1445	1500	0	51.8	4
2	156	1344	1500	0	55.2	3
3	581	919	1454	46	67.8	1
4	409	1091	1363	137	59.1	1
5	1251	249	761	739	67.1	1
6	41	1459	1499	1	51.3	1
7	291	1209	1461	39	58.4	1
8	169	1331	1484	16	55.1	1
9	569	931	1313	187	62.7	1
	447	1053	1483	17	64.3	4
10	396	1104	1453	47	61.6	1
11	900	600	1087	41.3	66.2	1
	737	763	1390	110	**70.9**	4
12*	1500	0	0	1500	50.0	12
13	1500	0	0	1500	50.0	18
Mean					59.43	
Standard deviation					6.99	

Table 3.5 DNA Physico Chemical Property based Numerical Representation of DNA Sequences Simulation Results (14: EIIP; 15: Atomic Number; 16: Paired Numeric; 17: Molecular Mass; 18: Paired Nucleotide Atomic Number; 19: DNA Walk; 20: Z-Curve; 21: Digital Z-Signals; 22: Phase Specific Z-Curve; 23: Simple Z-Curve; 24: Genetic Code Context; Pre: Precision (%); I: Number of Indicator Sequences)

	Classification of Human exons (1500)		**Classification of Human Introns (1500)**		**Pre**	**I**
	Correct	**Incorrect**	**Correct**	**Incorrect**		
14	0	1500	1500	0	50.0	1
	0	1500	1500	0	50.0	4

15	1451	49	72	1428	50.8	1
	1500	0	0	1500	50.0	4
16	531	969	1471	29	66.7	1
	160	1340	1499	1	55.3	2
17	1493	7	21	1479	50.5	1
	1500	0	0	1500	50.0	4
18	1470	30	42	1458	50.4	1
	1500	0	0	1500	50.0	2
19	62	1438	1372	128	47.8	1*
	155	1345	1253	247	46.9	1 #
20	638	862	530	970	38.9	3
21	942	558	1210	290	71.7	3
22	705	795	1500	0	73.5	9
23	84	1416	1477	23	52.0	3
24	1500	0	3	1497	50.1	1
Mean					53.21	
Standard deviation					9.00	

*: 1-D DNA walk; #: 1-D Complex DNA walk

Table 3.6 Statistical Property Based Mapping of DNA Sequences Simulation Results (25: Inter-Nucleotide Distance; 26: Binucleotide Distance; 27: CCP; 28: Single Nucleotide Probability Indicators; 29: Correlation Function; 30: PCF; 31: Codon Index based on Recurrence Time; 32: Ratio-R; 33: Galois Field; 34: Complexity; 35: Frequency of Nucleotide Occurrence; Pre: Precision (%); I: Number of Indicator Sequences)

	Classification of Human exons (1500)		**Classification of Human Introns (1500)**		**Pre**	**I**
	Correct	**Incorrect**	**Correct**	**Incorrect**		
25	1161	339	311	1189	49.1	1
26	1262	238	149	1351	47.0	1
27	229	1271	430	1070	22.0	1
28	2	1498	1500	0	50.1	1
	263	1237	1494	6	58.6	4

29	0	1500	1500	0	50.0	1
30	336	1164	1500	0	**61.2**	4
31	1459	41	52	1448	50.4	1
32	81	1419	224	1276	10.2	1
33	293	1207	1500	0	59.8	1
34	0	1500	1500	0	50.0	1
35	0	1500	1500	0	50.0	1
	0	1500	1500	0	50.0	4
Mean					46.80	
Standard deviation					14.55	

From these experiments, I have examined the influence of the choice of numerical representation on the quality of classification in the *human* data set. The experimental results show that the selection of the appropriate numerical representation can greatly influence the classification results since it depends on how well the numerical representation capture the biological information or modeled in mathematical properties. In the beginning of this chapter in section 6.1 I have stated that there are three groups of numerical mapping namely, fixed mapping, and DNA physico chemical property based mapping, and statistical property based mapping. My hypothesis is that the numerical representation should be selected based on the DNA physico chemical properties of the DNA for a specific classification problem since this representation reflects the biological properties very well in the numerical domain compared to the other two groups. The obtained classification results confirm this hypothesis: the numerical representation with the best classification results among the three groups is the phase specific Z-curve which belongs to the DNA physicochemical property based mapping. Hence, the DNA physico

chemical property based mapping is the most stable among the three groups of numerical mapping methods for this experimental setup.

3.8 Comparison of Numerical Representation Methods

Primarily, fixed mapping representation methods, such as the Voss or the tetrahedron maps a DNA sequence onto four or three numerical sequence, potentially introducing different redundancy in each individual representation. Studies [16], [28] indicate that the Voss representation is an efficient representation among the fixed mapping methods for spectral analysis of DNA sequences.

The 2-bit, 3-bit, and 4-bit binary do not reflect the biological property so well and these mappings are found to be exclusively suitable for neural network based classifiers for DNA sequence analysis. The paired nucleotide representation reflects the distribution of the nucleotides and has been successful in detecting structures with Fourier transform and wavelet analysis.

The arbitrary assignment of integer and real number to DNA nucleotides does not necessarily reflect the structure present in the original DNA sequence. The concept of pentanary code could be extended to any numerical representation depending on the requirement. The quaternion approach requires further exploration.

The physico chemical property based mapping techniques such as the EIIP mappings, the atomic number, the paired numeric, the DNA walk, and the Z-curve, in which each exploits the structural difference of protein coding and noncoding regions to facilitate DSP-based gene and exon predictions.

Further, the digital Z-signals, the SZ, and the GCC reflects the DNA physico chemical property very well, it could be explored further with wavelet transform or time-frequency analysis for identification of protein coding regions. All these methods are robust, independent, less redundant, and have biological interpretation which can be regarded as useful representations for DNA sequence analysis.

Under statistical property based mapping the correlation function is a useful tool to visualize various different periodicities in a DNA sequence, the single nucleotide probability indicator and the frequency of nucleotide occurrence is limited to standard data sets thus tending to be a model dependent method. PCF provides an automated approach for identification of protein coding regions with certain limitation in terms of window size.

Various other representations like the inter-nucleotide, the binucleotide, the CCP, the codon index based on recurrence time, the ratio-R, the complexity, the Galois field, are based on some mathematical or physical property and the results of these representations are not that appreciable. These methods require more fine tuning for which more works need to be done, before it is ready to apply for various organisms.

From numerical and as well graphical representation point of view, studies [55]-[60] and as well our experimental results indicate that the phase specific Z-curve a variant of Z-curve representation is more robust and computationally efficient for DNA sequence analysis as compared to the Voss representation and other representation. The x_n, y_n, and z_n components of the Z-curve are independent and contain all the information in a DNA sequence, and mapping between a DNA sequence and the Z-curve is bidirectional. On the other hand, the four indicator sequences of the Voss representation form a linearly

dependent set which add up to a sequence of ones; and the DNA walk is not bidirectional. Last but not the least, each of the x_n, y_n, and z_n components of the Z-curve generates a discrete signal that has a clear biological interpretation.

3.9 Discussion

In this chapter, I have examined the influence of the selection of about 35 DNA numerical representations (that I have across in my survey) on the quality of classification of coding and noncoding regions in the human data set with DFT technique based on period-3 property. The quality of classification depends on many different choices such as (1) the type of mapping of DNA data into a numerical domain, (2) the selection of the threshold for classification, (3) the DSP technique applied to extract period-3 property, (4) the data set selected. These classification parameters are usually selected by the user ad hoc.

Which one of the numerical representation techniques is to be used in association with DSP depends on a particular application. The user can choose the numerical representation that (1) reflects the desired property for example period-3 property; (2) that has no bias; (3) that captures information at three codon positions; (4) there is no redundancy involved; (5) the biological, structural, and statistical information is captured or well modeled in mathematical properties; (6) complementary structure of nucleotide pairs preserved; (7) possibility to reconstruct DNA sequence from the representation,; (8) access to different mathematical analysis tools; (9) a combination of these properties.

There are several areas, however, where the numerical representation techniques need considerable development before they can become common tools of the biologist's

profession. Numerical representation techniques lack precision in identifying specific sequence motifs, they cannot be directly adapted to distinguish between reading frames and are not yet quantitative enough to precisely characterize gene sequences. Some of these problems are being tackled by several groups [16, 43] and hopefully a more complete set of analytical tools will be available in the near future.

At the current stage, the numerical representation methods complement DNA letter series representations and together present a set of techniques that can be expected to help solve in a large manner the mounting problem of sensibly viewing and analyzing the rapidly growing sequence databases.

CHAPTER IV

GRAPHICAL REPRESENTATION OF DNA SEQUENCES

The sequencing of genomic data requires methods to allow this data to be visualized and analyzed. With the emergence of genomic signal processing, graphical representation techniques play a key role in applying digital signal processing techniques like Fourier transforms, and more recently, wavelet transform for visualizing DNA sequence. The choice of the graphical representation technique for a DNA sequence affects how well its biological properties can be reflected in the graphical domain for visualization and analysis of the characteristics of special regions of interest within the DNA sequence. This chapter presents a summary of various DNA graphical representation methods and their applications in envisaging and analyzing long DNA sequences. A discussion on the comparative merits and demerits of the various methods and observations has also been included.

4.1 Graphical Representation Methods

Graphical representation approach has been attracting attention in genomic DNA research and has become increasingly important in highlighting local and global base dominances, to identify repetitive sequences, differentiate between coding and non-coding regions which cannot be revealed accurately by conventional DNA symbolic and non-DSP based graphical representation techniques [26]. Fig. 4.1 illustrates the steps of the genomic signal processing (GSP) approach wherein a DNA sequence is converted to a numerical sequence before digital signal processing (DSP) is applied [16,23,24,41] for visualization

and analysis. The graphical representation of DNA alphabets is central to digital signal processing based DNA visualization and analysis. Perhaps the most widely used visualization scheme for this purpose is the Fourier spectrum of Voss representation [28], however many other schemes have also been introduced, such as the tetrahedron [16,30], the DNA walk [52], and the Z-curve [55].

Reference [26] describes about the non-DSP based graphical representation techniques to visualize DNA sequence in one-, two-, or three-dimensions. These techniques are complex, computationally intensive, require high ended computer graphics for visualization and are not yet quantitative enough to precisely characterize gene sequences. There has been a limited attempt [26] to provide a survey on various graphical representation methods. In this chapter the focus will be on various DNA graphical representation methods and their signal processing applications.

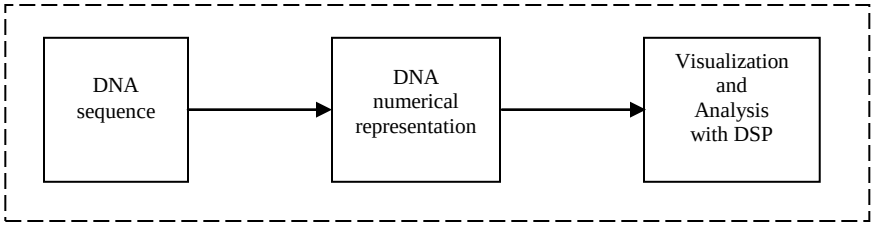

Fig. 4.1 Steps of genomic signal processing.

In order to apply graphical representation techniques, the alphabets of a DNA sequence need to be mapped onto their corresponding numerical values for visualization and analysis with DSP methods.

The Voss representation [28] is a binary representation, which maps the nucleotides A, C, G, and T into four binary indicator sequences as $U_A[n]$, $U_C[n]$, $U_G[n]$, and $U_T[n]$ showing the presence with 1 or absence with 0 of the respective nucleotides. The major signal in

coding regions of genomic sequences is a three-base periodicity in biological domain. Fourier transform approach is applied to visualize this periodicity i.e. the three-base periodicity in the nucleotide arrangement is evidenced as a sharp peak at the frequency f=1/3 in the Fourier spectrum based on Voss representation. Studies [16,28,29] indicate that the spectral analysis of Voss representation is one of the efficient graphical representations to recognize coding regions in genomic DNA sequences**.**

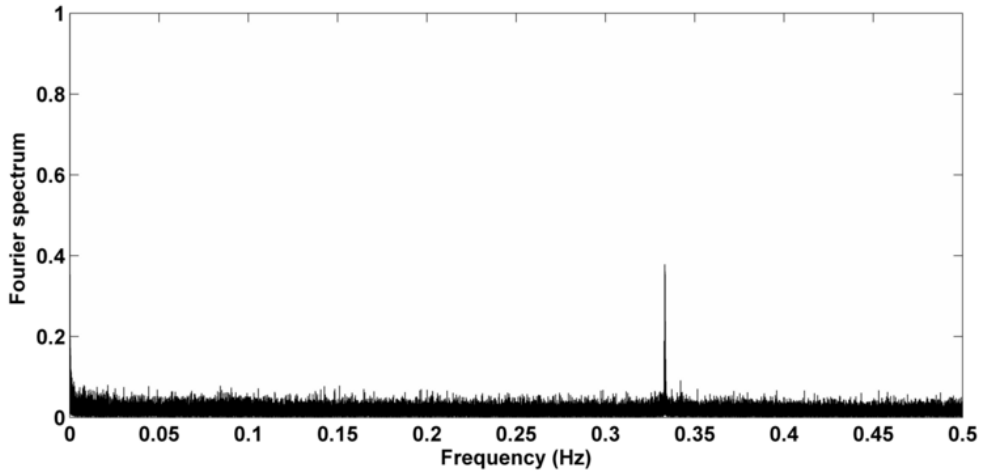

Fig. 4.2 DFT power spectrum (Voss representation) demonstrating a peak for the coding region of *S. Cerevisiae* chromosome #3.

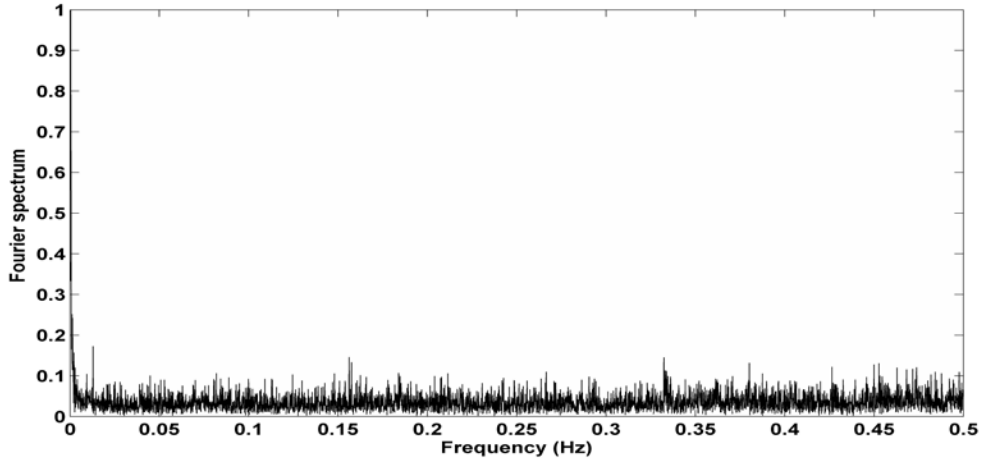

Fig. 4.3 DFT power spectrum (Voss representation) demonstrating no significant peak for the non-coding region of *S. Cerevisiae* chromosome #3.

The power spectrum of the Voss's binary indicator sequence reveals a peak at frequency 0.33 (period-3) for coding region (Fig. 4.2) and shows no peak for noncoding region (Fig. 4.3).

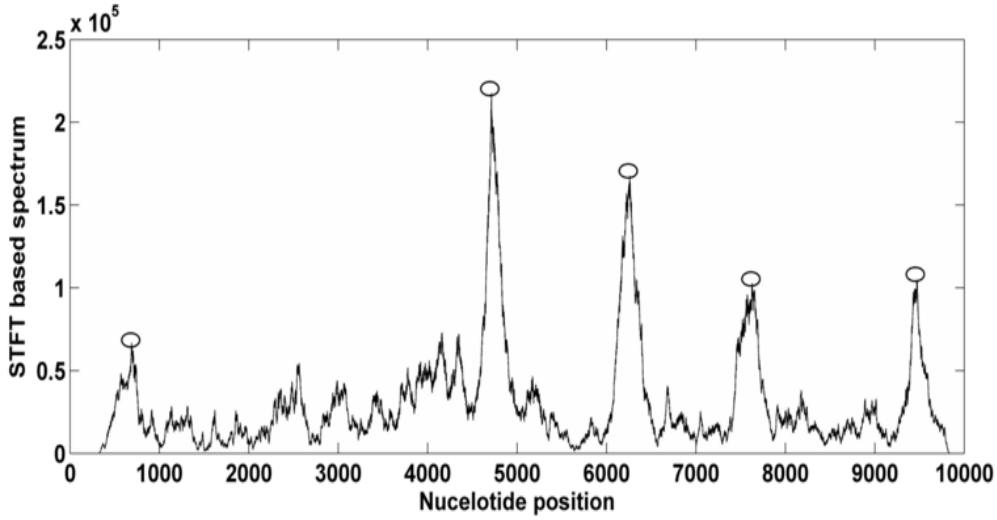

Fig. 4.4 Squared magnitude of the STFT (Voss representation) for gene F56F11.4 in *C-elegans* chromosome #3 (The circles indicate the locations of coding regions).

The Frequency-domain analysis of DNA sequences (with the Voss representation) is a powerful tool for specifically identifying gene coding regions in DNA sequences**.** This coding measure has been utilized in the program 'Genescan' [28] to evaluate the period-3 component of the Fourier power spectrum of the binary indicator sequences. It operates on a sliding window (also known as STFT) to search for a peak with a strength (against the average strength of the power spectrum in the region) surpasses a threshold which indicates the presence of exons as shown in Fig. 4.4. Thus the Fourier spectrum of the Voss mapping is widely utilized to detect and visualize the periodicity feature of coding regions, and the STFT of Voss mapping to identify and display graphically the probable protein coding regions of a sequence. The tetrahedron method [16,30] reduces the number of indicator sequences from four to three in a manner symmetric to all four

components reflecting the same property of the Voss. Spectrally it is identical to Voss representation.

The DNA-Walk model [52,53] shows a graph of the DNA sequence in which a step is taken upwards (+1) if the nucleotide is pyrimidine (C or T) or downwards (-1) if it is purine (A or G). The graph continues to move upwards and downwards as the sequence progresses in a cumulative manner, with its base number represented along the x-axis. The DNA walk can be used as a tool to visualize changes in nucleotide composition (Fig. 4.5), to observe base pair patterns (Fig. 4.6), and sequence evolution along a DNA sequence.

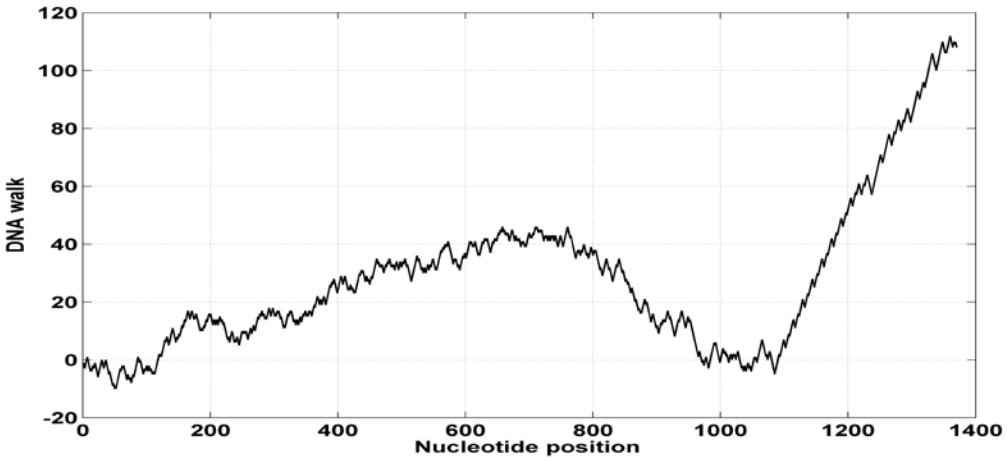

Fig. 4.5 One-dimensional DNA walk (This illustrates the relative content of purine and pyrimidine residues within the noncoding region of *Helicobacter pylori* strain J99 bacteria).

The DNA walk (Fig. 4.5) provides a graphical representation for each gene and permits the degree of correlation in the base pair (GC versus AT) sequence to be directly visualized. DNA walk technique also distinguishes between coding and noncoding sequences with quantitative measurement. This has been applied to *viral*, *bacterial*, *yeast* and *mammalian* DNA sequences [41,53,54], [80-83]. Locating periodicities or nucleotide structure may be found using complex DNA walk representation. Also, by projecting this

walk down on the xy-plane reveals another visualization (staircase-like behavior of the walk that corresponds to a repeat sequence that occurs eight times) shown in Fig. 4.6. Wavelet analysis is applied to extract structural information (corresponding to sequence periodicity) from the complex DNA walk in the transform domain [52] as shown in Fig. 4.7 (the dark conical structure at the end is the same periodic staircase shown in Fig. 4.6). Wavelet transforms also allows measures to quantify correlations based on DNA walk [54].

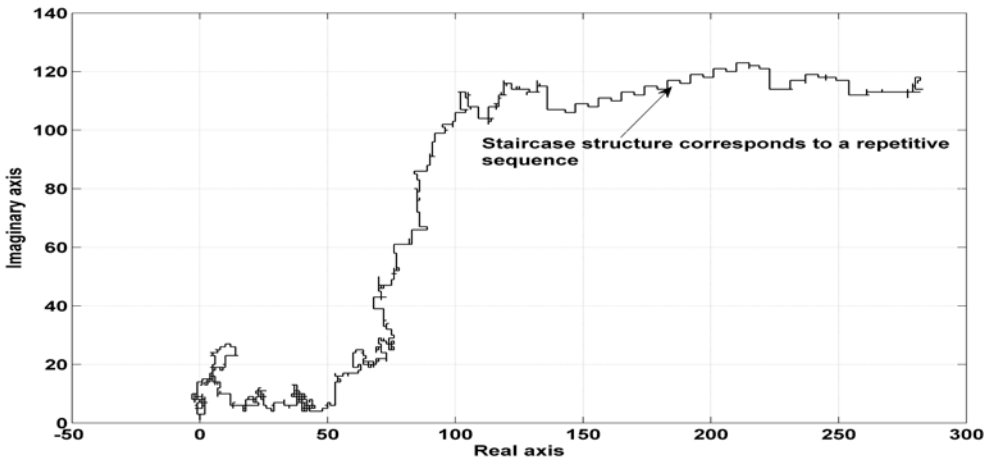

Fig. 4.6 Two-dimensional DNA walk projection (This illustrates the repetitive sequence structure in the noncoding region of *Helicobacter pylori* strain J99 bacteria).

The Z-curve [55] is a 3-dimensional curve (Fig. 4.8) that provides a unique representation for visualization and analysis of a DNA sequence. The three components of the Z-curve (Fig. 4.7**)** x_n, y_n, z_n represent three independent nucleotide distributions that completely describe the DNA sequence under study. The components x_n, y_n, z_n display the distributions of purine versus pyrimidine (R versus Y), amino versus keto (M versus K), and strong H-bond versus weak H-bond (S versus W) bases along the sequence,

respectively. The compositional patterns of genomic sequences can be quickly recognized in a perceivable form with Z-curve.

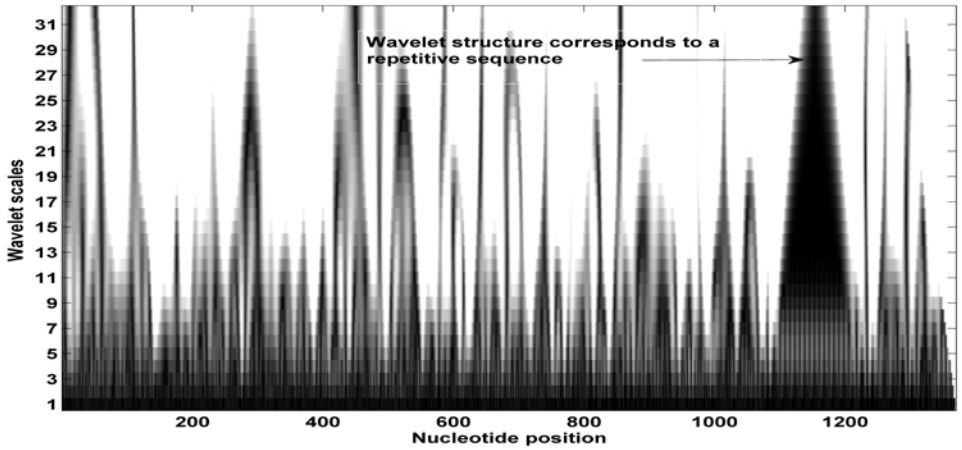

Fig. 4.7 Wavelet transform analysis for complex-valued DNA walk for the noncoding region of *Helicobacter pylori* strain J99 bacteria (here the repetitive structures are highlighted more clearly compared to the analysis of one-dimensional walk in Fig. 5 in ref [52]).

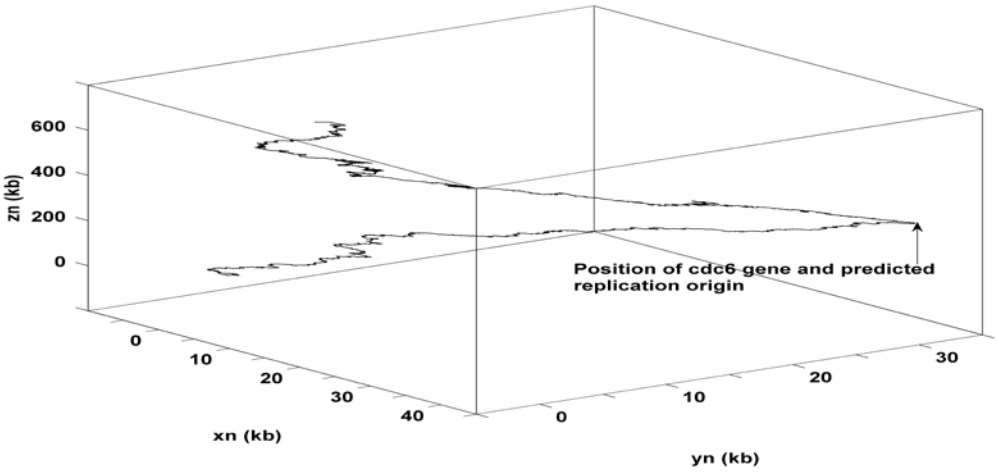

Fig. 4.8 Z-curve for *M.mazei* genome.

In particular, the distribution of the G+C content along genomic sequences (such as budding yeast and vibro cholerae genomes [56]) is generally superior to the widely used sliding-window technique.

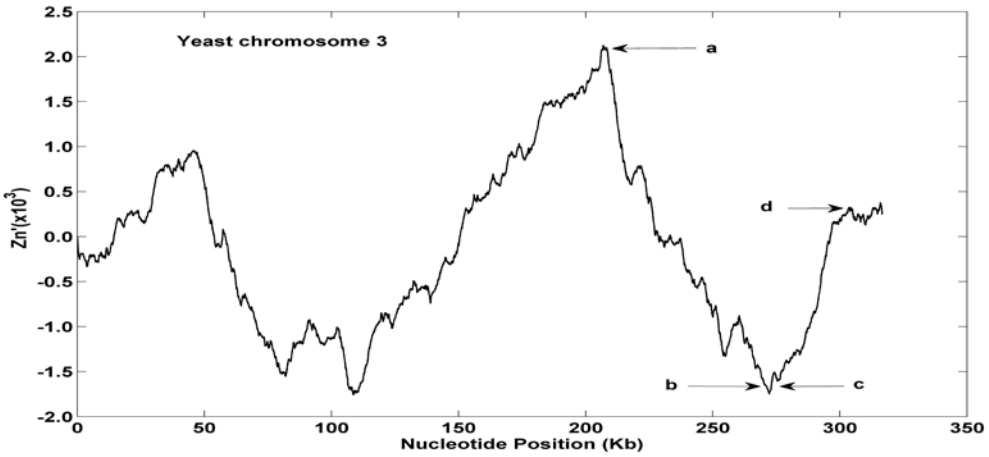

Fig. 4.9 The Z-curves for chromosome III of the *S.cerevisiae* genome (The arrow a and b indicte the begin and end of a G+C –rich region, while c and d indicate the begin and end of a A+T region, respectively).

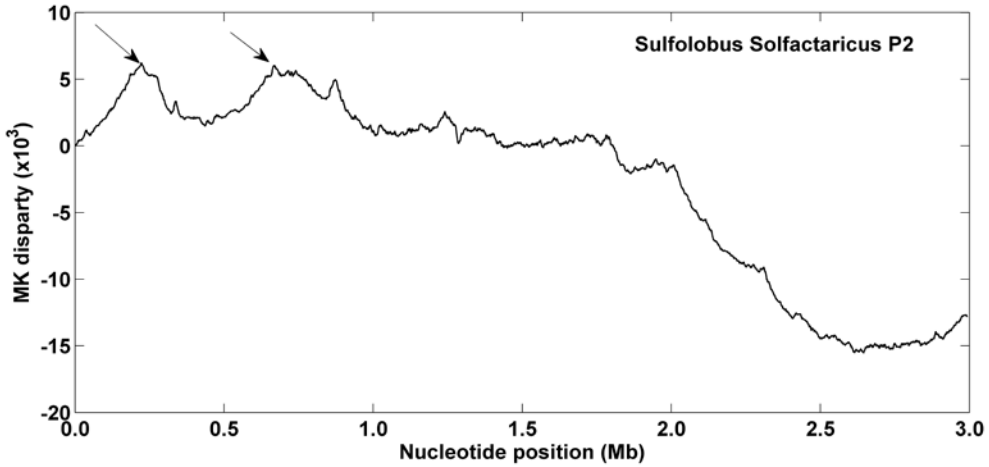

Fig. 4.10 The MK disparity curve for the *Sulfolobus solfataricus* P2 genome (The arrows indicate two replication origins, oriC1 and oriC2, are close to genes cdc6-1 and cdc6-3, respectively).

The distribution of the G+C content in the human genome has been studied by using a windowless technique derived from the Z-curve method [57] shown in Fig.4.9. There is a strong correlation between the G+C content and gene density. The higher gene density regions for any chromosome are almost all situated at the GC-isochores with highest G+C content. Another application of the Z-curve is in identifying replication origins of

archeal genomes with 3-D Z-curve (Fig. 4.8) and 2-D Z-curve based on GC and MK disparity [58] shown in Fig. 4.10. A Z-curve based wavelet denoising technique for DNA sequences has also been reported in [59-60].

4.2 Comparisons and Analyses

The Fourier power spectrum analysis of the DNA sequence and the results rely on mapping the nucleotides to binary (Voss) indicator sequences. This treatment has shown to be useful for locating periodicities and for locating potential gene sequences characterized by simple nonuniform codon usage. The Voss and the tetrahedron representations are equivalent representation when used for power spectra analysis. Studies [28,64,67] indicate that one of the drawbacks of power spectrum analysis of Voss representation is that it is not able detect period-3 feature in some *viral* and *bacterial* organisms and it produces the same Fourier spectra for two different DNA sequences which is not suitable for DNA sequence comparison.

DNA walk is a novel and relatively easy to implement technique. The DNA walk based on complex representation provides useful information such as long range correlation information, sequence periodicities, and changes in nucleotide composition. Further, the wavelet transform of complex DNA walk readily identifies sequence structures related to sequence periodicities. Thus, the DNA walk technique allows direct visualization of sequence evolution in the base domain, as well as in a transform domain unlike the power spectrum of Voss representation whose visualizations are in the transformed domain only. The DNA Walk graphical representation technique is able to identify and study sequence behavior over short stretches of DNA sequences ranging over a few hundreds

of nucleotides, and for lengthy sequences the DNA walk tends to become too complicated since there is too much information for extracting anything useful.

Z-curve is an innovative method of mapping the DNA sequence into a 3-dimensional folding curve. This representation is one-to-one correspondence, i.e., there exists a unique Z-curve for a given DNA sequence, or there exists a unique DNA sequence for a given Z-curve. The Z-curves generally have rich spatial folding structures which reflect the symmetry, the periodicity, the local motif, and the global features of the distribution of bases of the DNA sequences. The Z-curve may have any resolution and is suitable for visualizing and analyzing the DNA sequence with any length. The Z-curve generally gives the observer a more intuitive impression than the conventional letter series representation of the DNA sequence does.

4.3 Observations

The Fourier transform of Voss representation or Tetrahedron can be used as a preliminary spectral indicator of periodicity in a DNA sequence.

Small scale features, which range over a few hundreds of nucleotides, can be visualized effortlessly using different forms of DNA walks and periodic behavior can be readily illuminated using frequency domain transforms such as the wavelet transform based on complex DNA walk. Thus, the complex DNA walk is suitable for such purposes and has been visualized here over a sample sequence. However, for lengthy sequences, DNA walk visualization tends to become complicated.

The Z-curve is a robust, independent, less redundant, and has clear biological interpretation which can be regarded as a useful visualization technique for DNA sequence analysis of any length and with any resolution

Among all these methods, it can be concluded that the Z-curve is more robust and efficient representation for DNA sequence analysis compared to Fourier spectrum of Voss representation and the DNA walk. The Z-curve contains all the information in the DNA sequence; mapping from DNA sequence to Z-curve is one to one i.e., each can be reconstructed given the other. One advantage of the Z-curve is that each of the three Z-curves generated a digital signal that has a clear biological interpretation. A second advantage of the Z-curve map is that the x, y, and z components are independent unlike the voss representation where the four elementary sequences form a linearly dependent set since they add up to a sequence of ones. Last but not the least, the computational cost of the DNA spectrum using the Z-curve mapping is much lower than in the voss representation case.

CHAPTER V

EXISTING PROMOTER PREDICTION SYSTEM

This chapter briefly introduces the concepts behind the promoter prediction system and their classification. This chapter also discusses about the state-of-the-art of existing numerical representation based promoter prediction system and this chapter ends with a description of existing numerical representation based promoter prediction system.

5.1 Promoter Prediction Systems

Deciphering human genome sequences to find genes and their regulatory network is an important but challenging task. In particular, the identification of promoter region plays a critical role in gene annotation as this region essentially controls the biological activation of a gene [84-86]. Detecting these regions by biological experiments is too time-consuming. As a result, a number of computational methods have been developed in the past few years such as Eponine [3], FirstEF [4], PromoterInspector [5], DPF [6], DragonGSF [7,8] and PCA-HPR [9].

The promoter prediction tools can generally be divided into two classes: feature-based approach and DNA numerical representation based approach. In feature-based approach, discriminative features such as CpG islands, TATA and CAAT boxes are extracted from promoter sequences to differentiate themselves from non-promoters. In prokaryotes such as E. coli, the promoter consists of two conserved short sequences at -10 and -35 positions upstream from the transcription start site (TSS). The sequence at -10 is called

the Pribnow box or the -10 element. It usually consists of six nucleotides TATAAT. The other sequence at -35 (the -35 element) consists of six nucleotides TTGACA. Unlike prokaryotic promoters, eukaryotic (e.g., human) promoters are so diverse that it is not yet possible to draw any clear generalizations about them. Some have a consensus sequence called the TATA box located about 30 bp upstream from the TSS (commonly referred to as -30). The consensus sequence is TATAAA, but there is a lot of variation. In addition, many eukaryotic promoters have no recognizable TATA box. In such cases, the transcriptional initiation site is usually less precisely defined which could start occurring at several different locations. Eukaryotic TATA box has sequence similarity to the prokaryotic -10 box (TATAAT), but the TATA box is located substantially further upstream. A second upstream site that is often encountered in eukaryotic promoters is the CCAAT box**,** commonly referred to as the "cat box". It has a consensus sequence of GGCCAATCT, again with substantial variation, particularly at the ends. Typically, it is located near the -75 position relative to the transcriptional start site. CpG islands are found around gene starts in approximately half of mammalian promoters and are estimated to be associated with about 60% of human. Due to the dissimilar and diverse nature of eukaryotic promoters, none of these features alone can characterize all promoters. Among the existing feature-based promoter prediction systems [3-9], the system in [9] used the most informative and discriminative 6-words based on the KL divergence and position specific information as features in characterizing promoters. It performs well with a sensitivity of 63.8% for *human* chromosome 22. However, it still has an unacceptable high false positive with specificity of 50.3%. In other words, about 50% of non-promoter sequences are incorrectly classified as promoters in this system [9].

In DNA numerical representation based approach, each individual nucleotide of the promoter and non-promoter sequences is converted to numerical values through a mapping function. A combination of such numerical sequences is then used to distinguish promoters from non-promoters [10]. Typical examples of numerical representation based approaches for promoter predictions are: binary representation [92], 2-bit binary [10,75,87,92], binary representation based on DNA physicochemical properties [92], 4-bit binary [10,75,87,91,92], and integer representation [90]. There has been limited attempt to apply numerical representation based approach for eukaryotic promoter prediction. As compared to the feature-based approach, the DNA numerical representation based approach can preserve nucleotide positional information and retain all the available sequence information.

5.2 Review of Existing Numerical Representation based Promoter Prediction Systems

Several authors considered different numerical mapping schemes for promoter prediction. In [88], feed forward neural networks were used to predict the exact location of the TSS in *human* using 4-bit binary representation. However, only 44.6% recognition rate for true sites and 4.9% false positive rate were achieved. Reference [89] used the orthogonal vectors of 4-bit binary numbers to represent A, C, G, and T and neural networks for classification. It achieved a correlation coefficient of 0.63 only.

For prokaryotic promoters, it was shown experimentally that 4-bit binary DNA representation is superior to 2-bit binary DNA encoding [87]. However, the biological reasons for choosing the 2-bit and 4-bit binary DNA representation were not discussed.

Using the 4-bit binary representation, a classification accuracy of 89.5% and an error rate of 10.37% were achieved for *E. coli* promoter recognition [91]. Reference [93] used a purine-pyrimidine encoding scheme and reported an average precision of 90.9% for *E.coli* promoters. Reference [10] combined four representations, namely 2-bit binary, 4-bit binary and variants of integer representation of a DNA sequence for *E-coli* promoter prediction. A high sensitivity of 97.48% and a specificity of 98.83% were reported, though the biological significance of different numerical representations was not discussed.

In [92], eleven binary mapping rules were analyzed for promoter recognition using support vector machines. They achieved a sensitivity of 99.89%, specificity of 99.75%, for *drosophila* classification and a sensitivity of 99.75%, specificity of 99.89% for *human* promoter classification. Despite the high performance for eukaryotic promoters, the system did not consider UTR sequences and thus the performance has not yet been evaluated for promoter prediction at chromosome level.

In [90], a hybrid approach was suggested where both the numerical sequence and promoter features were considered as input features to a cascade AdaBoost algorithm. It achieved a sensitivity of 68.5% and specificity of 68.6% in *vertebrate* (eukaryotic) promoters. However, it only considered one type of integer mapping for nucleotide symbols and there is no evaluation of its effectiveness in retaining biological properties.

We can see that a good performance has been achieved in prokaryotic promoter recognition, but not for eukaryotic promoter. Also, there was limited attempt to apply numerical sequence approach for eukaryotic (e.g., *human*) promoter prediction. It thus

motivated us to explore the influence of DNA sequence numerical mapping on the quality of classification for human promoter prediction.

5.3 Overview of MultiNNProm System (Existing)

In a DNA numerical sequence approach for promoter recognition, the choice of numerical representation used to map nucleotide symbols to numerical values is important. Different numerical representations preserve different biological properties embedded in a DNA nucleotide sequence [75,76]. Fig. 5.1 shows the block diagram of an existing system for identifying E. coli promoter sequences called "MultiNNProm" [10]. In [10], a DNA sequence comprising nucleotides {A, C, G, T} is converted to four numerical sequences using the following mapping functions:

E1 (integer): A = -2; T = -1; C = 1; G = 2.

E2 (integer): A = -1; T = -1; C = 1; G = 1.

C2 (2-bit binary): A = 00; T = 01; C = 11; G = 10.

C4 (4-bit binary): A = 1000; T = 0100; C = 0010; G = 0001.

These four numerical sequences serve to retain different biological properties and they are fed to four neural networks separately so that each neural network is trained to extract promoter properties embedded in that particular numerical sequence. Outputs from neural networks are then passed onto a probability builder function to assign probabilities as to whether a tested sequence is an *E. coli* promoter or not. Finally, the outputs of the probability functions are combined through an aggregation function with weights determined by a genetic algorithm (shown in Fig. 6.3) [10]. In this way, promoter

properties extracted from these four numerical representations are used simultaneously to identify promoters.

The four neural networks have been trained on the training data set consisting of *E. coli* promoters and exons. In the testing data set, the four neural networks have different performance. In particular, the sensitivities for NNE1, NNE2, NNC2 and NNC4 are respectively 88.68%, 87.42%, 90.57% and 94.97%. The specificities are 87.72%, 88.30%, 91.81%, 96.49% respectively. Each encoding method retains different biological properties of the sequence and thus gives slightly different performance. After combining the results from these four neural networks, the "MultiNNProm" system outperforms other prediction systems for *E. coli* promoters with sensitivity of 97.5% and specificity of 98.8%.

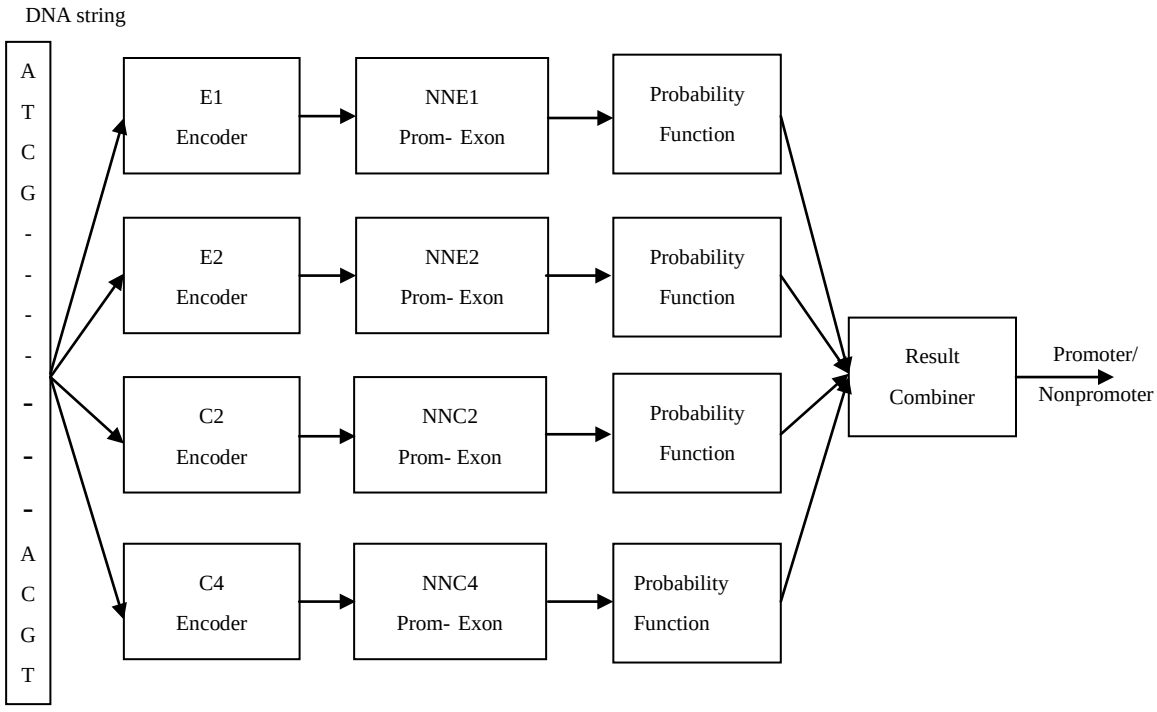

Fig. 5.1 MultiNNProm for *E.coli* promoter prediction.

It is well known that eukaryotic promoters have more diverse structure than prokaryotic promoters and thus it is more difficult to characterize eukaryotic promoters than prokaryotic promoters. The performance of the “MultiNNProm” on eukaryotic promoter identification remains unknown and requires further investigation. Besides, a genomic sequence comprises introns and 3’UTR in addition to promoters and exons. Unlike “MultiNNProm”, state-of-the-art promoter prediction systems always contain three classifiers to distinguish 1) promoters from exons; 2) promoters from introns; and 3) promoters from 3’UTR. Thus “MultiNNProm” should be extended to have classifiers for distinguishing promoters from introns and 3’UTR in addition to exons.

CHAPTER VI

THE PROPOSED DIGIPROMPRED SYSTEM

In this chapter the numerical representation selection strategy and the architecture of the proposed DigiPromPred system are discussed. At the end of the chapter, the experimental results and comparative studies with existing promoter prediction systems are provided.

6.1 Numerical Representation Selection Strategy and the Architecture of the Proposed DigiPromPred System

In the "MultiNNProm" system, a single DNA sequence is represented by using four different numerical representations to retain different biological properties embedded in the sequence. With the use of three classifiers for 1) promoters versus exons; 2) promoters versus introns and 3) promoters versus 3'UTR, altogether 4x3=12 numerical representations are generated which result in the use of 12 neural networks. The resultant system would be computationally intensive. An alternative way to manage this computation problem is to find the most suitable numerical representations out of these four representations {E1, E2, C2, C4} for each of the three classifiers. In this way, only one numerical representation is used to capture sequence characteristics for a particular classifier, thus requiring only three neural networks.

The four representations {E1, E2, C2, C4} use different mapping functions for nucleotides so that different biological properties are preserved in these representations. E1 and E2 use integers while C2 and C4 use binary numbers. The E1 representation

introduces unequal distance among the nucleotides, i.e., G>C and T>A. This can introduce undesirable characteristics to the resultant numerical sequences and thus affect features that are retained in the promoter sequences. The E2 representation assigns same value to some nucleotides, i.e., A = T = -1, C = G = 1. Thus the difference between A and T (also C and G) are ignored by the E2 representation. Previous studies have shown that the C2 representation is effective for gene (i.e., exons) identification [75]. Thus C2 appears to be a good candidate for retaining reliable features for distinguishing promoters from exons. C4 uses a unity coding approach for {A, T, G, C}. The binary numbers are orthonormal and have identical hamming distance to each other. Since our proposed system is a three classifier system, to choose the 3 most suitable numerical representation out of four representation (E1,E2,C2,C4), I need to evaluate the system performance 4^3=64 times, which is time consuming and computationally expensive. So I decided first to evaluate the performance of promoter versus exon and promoter versus intron only and this requires 2^4=16 experiments and the figure of merit to select the most suitable numerical representation was the optimum value of classification error. I found that C2 and C4 was most suitable for promoter versus exon and promoter versus introns classifier respectively that achieved a lowest classification error of 13.21% as shown in Table 6.1 and the associated flowchart is shown in Fig. 6.2 and Fig. 6.3. Now I am left with the task to find the most suitable numerical representation for promoter versus 3'UTR classifier. For this I consider the total three classifier system as I have already chosen C2 and C4 for promoter versus exon and promoter versus introns classifier to choose the most suitable numerical representation out of four (E1,E2,C2,C4) for promoter versus 3'UTR I need to evaluate the system performance 4^1=4 times and the outcome of this is C4 is the most

suitable numerical representation for promoter versus 3'UTR with lowest classification error of 9.12% as shown in Table 6.2. Thus, these experimental results are in agreement with the studies that C2 and C4 representations are widely used with neural network based prediction systems [10, 92].

Fig. 6.1 shows the block diagram of the "DigiPromPred" system in which feed-forward neural networks are employed. Each neural network has three layers: input layer, hidden layer and output layer, and is trained using a supervised backpropagation algorithm. Neural networks are used as they can capture high-order relationships among nucleotides in a DNA sequence. Due to the use of different numerical sequences in representing a DNA sequence, the three neural networks have slightly different configurations, which are summarized in Table 6.3 and the associated flowchart is shown in Fig. 6.4.

The networks are trained to produce an output of '-1' if the input sequence is found to be a non-promoter, and an output of '1' if the sequence is deemed to be a promoter. The outputs of the individual neural networks are then passed through a probability function [10] which is defined as follows. If the output of the neural network is 1 or more than 1, it is likely that the given sequence is a promoter and a probability of 0.9999 is assigned. If the output of the neural network is -1 or less than -1, it is unlikely that the given sequence is a promoter and a probability of 0.001 is assigned. The probabilities are assigned due to the use of the logarithm function within the result combiner. The three outputs of the probability function are then aggregated through a weighted sum in the result combiner. The weights are obtained by first initializing a random population of weights. Then they are refined using a genetic algorithm (shown in Fig. 6.3) with the classification accuracy of the combined classifier as the fitness function [10].

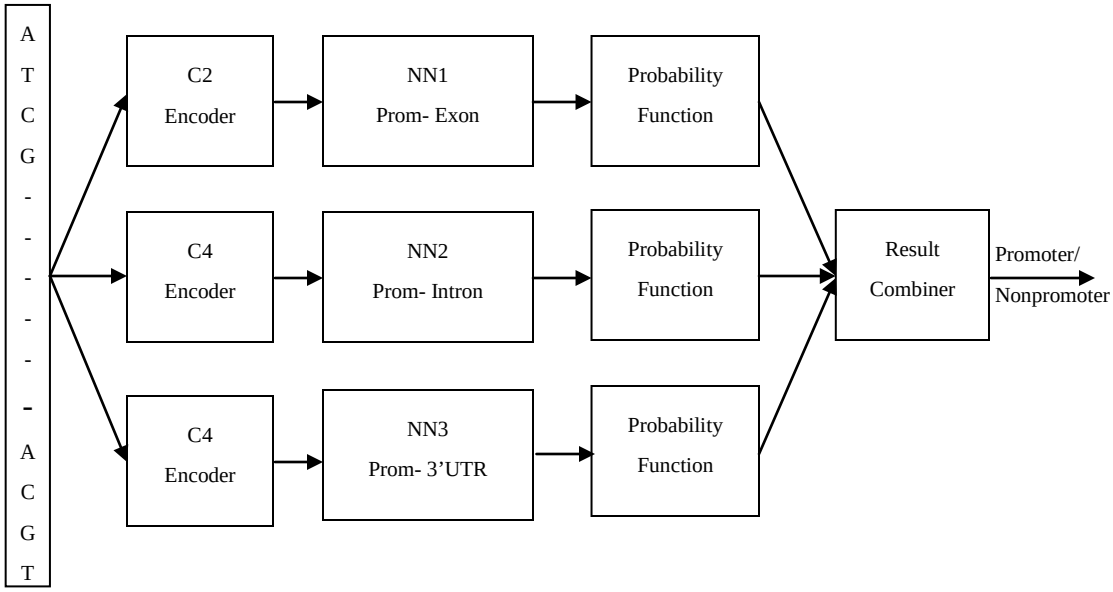

Fig. 6.1 DigiPromPred: a *human* promoter prediction system

Table 6.1 Performance Evaluation of a Two Neural Network System with Various Numerical Representations (Sen: Sensitivity (%); Spe: Specificity (%); Er: Classification error (%); Pre: Precision (%))

Numerical Representation		**Sen**	**Spe**	**Er**	**Pre**
E1	E1	80.00	74.24	23.58	76.41
E1	E2	74.34	69.27	28.93	71.07
E1	C2	57.48	98.44	34.28	65.72
E1	C4	91.60	80.90	15.09	84.90
E2	E1	80.00	74.24	23.58	76.41
E2	E2	82.57	72.73	23.90	76.10
E2	C2	80.15	77.54	21.38	78.62
E2	C4	88.49	77.07	18.87	81.13
C2	E1	71.64	97.43	18.87	81.13
C2	E2	71.01	100.00	18.87	81.13
C2	C2	69.67	100.00	20.12	79.87
C2	C4	88.89	85.24	13.21	86.79
C4	E1	85.94	80.53	17.29	82.70
C4	E2	88.09	81.25	16.04	83.96
C4	C2	89.43	81.02	15.72	84.28
C4	C4	90.98	81.63	14.78	85.22

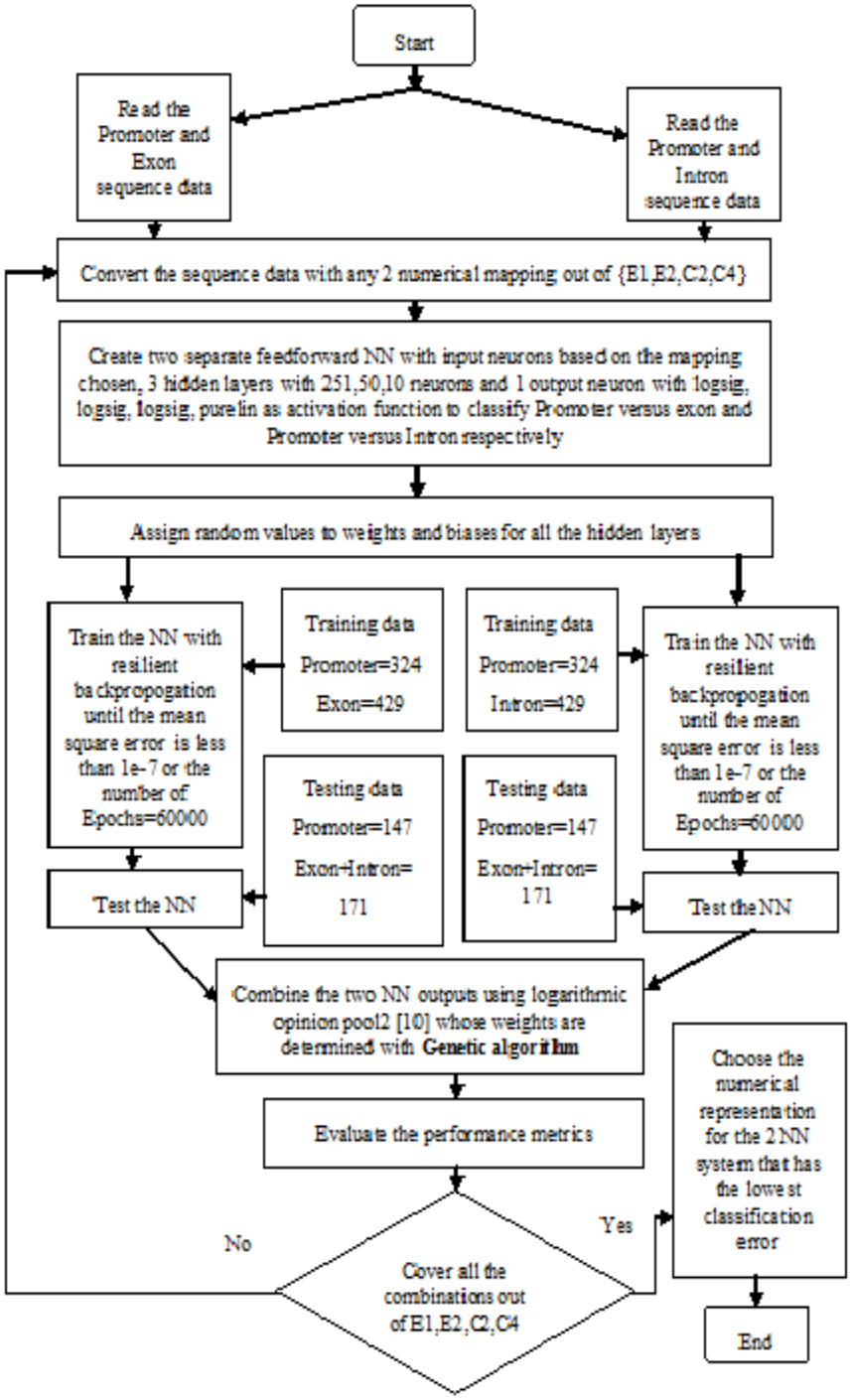

Fig. 6.2 Flowchart for performance evaluation of a two NN system with various numerical representations.

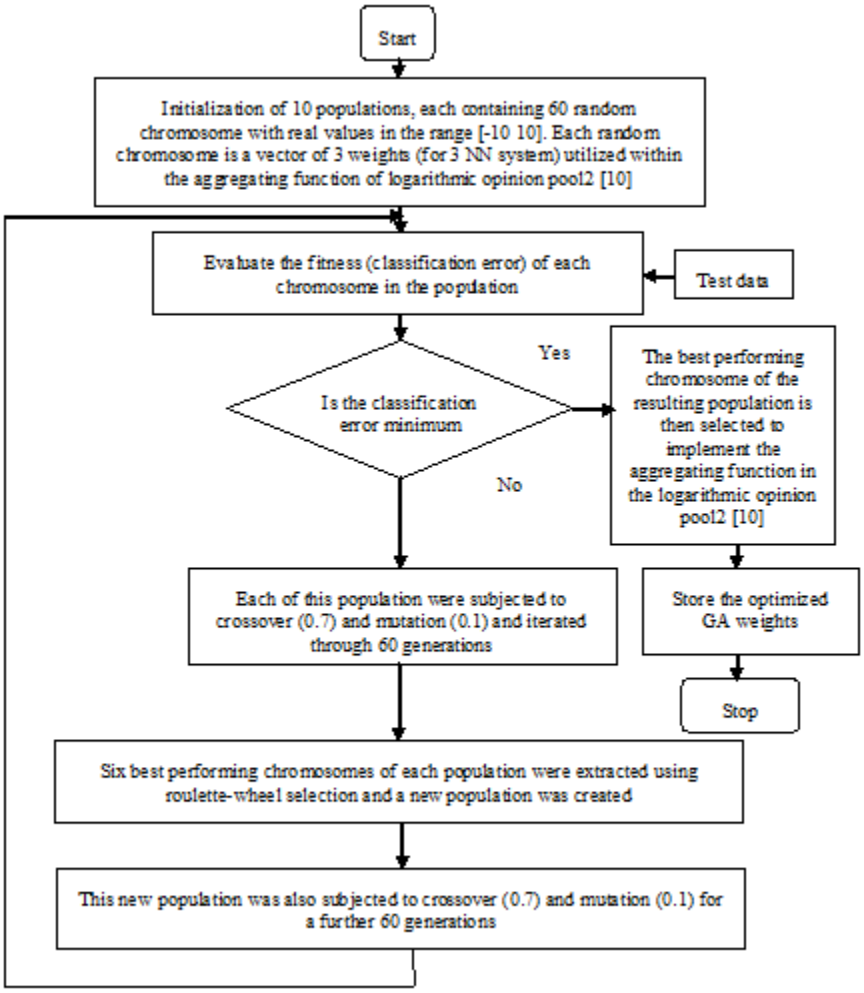

Fig. 6.3 Flowchart for determining aggregating coefficients using Genetic algorithm.

Table 6.2 Performance Evaluation of the Total Proposed System "DigiPromPred" with Various Numerical Representations (Sen: Sensitivity (%); Spe: Specificity (%); Er: Classification error (%); Pre: Precision (%))

Numerical Representation			Sen	Spe	Er	Pre
C2	C4	E1	81.56	99.28	10.69	89.31
C2	C4	E2	82.86	98.60	10.06	89.94
C2	C4	C2	84.61	97.31	9.43	90.57
C2	C4	C4	83.52	100.00	9.12	90.88

Table 6.3 Neural Network Configurations

Neural Network	Encoding method	Number of input neurons	Number of neurons in three hidden layers	Activation function
NN1	C2	502	251:50:10	Logsig: Logsig: Logsig: Purelin
NN2	C4	1004	251:50:10	Logsig: Logsig: Logsig: Purelin
NN3	C4	1004	251:50:10	Logsig: Logsig: Logsig: Purelin

6.2 Experimental Results

The DigiPromPred system was evaluated on 471 *human* promoters sequences [-200, +50] bp around the TSSs from the DBTSS (Database of transcription start sites) [13]. For non-promoters, 600 3'UTR sequences with 251 bp in length were extracted from the UTRdb (Untranslated region database) [13], 600 exons and 600 introns with 251 bp in length were extracted from the Exon-Intron Database (EID) [13] as mentioned in appendix C. These data were divided into two sets: training and testing sequences. The training sequences consist of 324 promoter sequences, 429 exons, 429 introns and 429 3'UTR sequences. The remaining sequences are the testing sequences. These publicly available databases provide the best combination of coverage, quality and reliability [13]. Each neural network performed perfectly on the training data set and displayed a recognition rate of 100%. Results for the testing sequences were then used to calculate the sensitivity, specificity, classification error and precision. They are found to be 83.52%, 100%, 9.12%, 90.88%, respectively.

Both statistical analysis and chromosome level testing were performed to investigate the effectiveness of the DigiPromPred system in identifying human promoter sequences. In statistical analysis, the 3-cross validation test was performed to test the stability of the DigiPromPred system. In chromosome level testing, the DigiPromPred system was applied to the *human* chromosome 22 to identify the 20 experimentally annotated promoters.

6.3 Evaluation of Results

The predictive accuracy of a promoter prediction program is often evaluated using measures such as sensitivity and specificity. True positives (TP) are the promoters that are predicted as promoters. True negatives (TN) are the non-promoters that are predicted as non-promoters. False positives (FP) are the promoters that are predicted as non-promoters. False negatives (FN) are the non-promoters that are predicted as promoters. Let the numbers of TP, TN, FP and FN be denoted as respectively N_{TP}, $N_{TN`}$, N_{FP}, N_{FN} and the total of positive (promoter) sequences and negative (non-promoter) sequences as N_{PO}, N_{NG}. The sensitivity S_n is defined as the proportion of promoter that is correctly predicted as promoter, i.e.,

$$S_n = \left(1 - \frac{N_{FP}}{N_{NG}}\right) \tag{6.1}$$

The specificity S_p is defined as the proportion of promoter sequences out of all the positive sequences that can be predicted, i.e.,

$$S_p = \left(\frac{N_{TP}}{N_{PO}} \right) \tag{6.2}$$

Positive predictive value (*PPV*) is a measure of the capability of identifying the positive cases, i.e.,

$$PPV = \left(\frac{N_{TP}}{N_{TG} + N_{FP}} \right) \tag{6.3}$$

Precision P is the ratio of the test sequences that are correctly classified (TP + TN) and the total number of testing sequences (N), i.e.,

$$P = \frac{C}{N} \tag{6.4}$$

Classification error CE is the percentage of erroneous identifications with respect to the total number of testing sequences, i.e.,

$$CE = \left(\frac{N_{FP} + N_{FN}}{N} \right) * 100 \tag{6.5}$$

Finally, *F-measure* reflects the balance between sensitivity and *PPV*. It is defined as,

$$F - measure = \left(\frac{2 * S_n * PPV}{S_n + PPV} \right) * 100 \tag{6.6}$$

In general, the higher is the *F-measure*, the better is the performance.

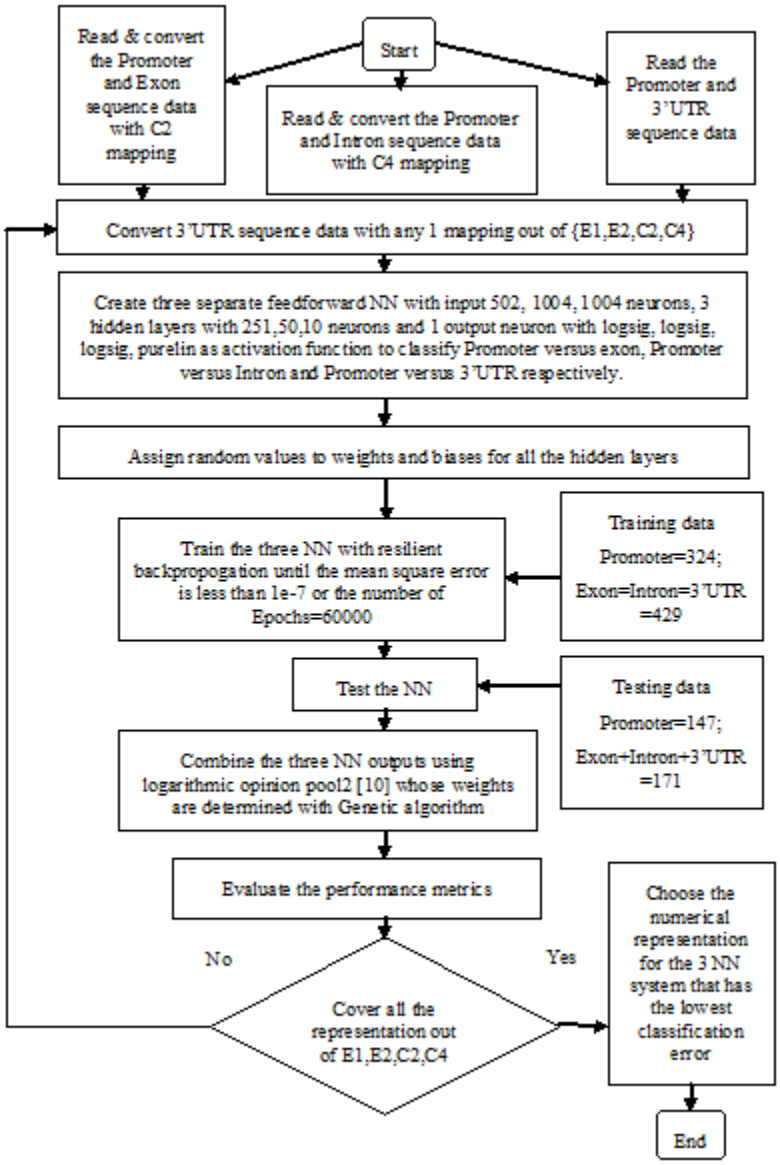

Fig. 6.4 Flowchart for performance evaluation of the total proposed system "DigiPromPred" with various numerical representations.

6.4 3-Cross Validation Test

Cross validation is a kind of statistical process which is used to estimate how accurate the DigiPromPred system performs in practice and its stability. One round of cross validation involves partitioning data into complementary subsets in which the analysis is made on one subset (called the testing set) while training is done on other subsets (called the training set). To reduce variability, multiple rounds of cross validation are performed using different partitions, and the validation results are averaged over all rounds. In 3-cross validation, the total data set is split into three disjoint sets. From these sets, three different training sets are built by joining two of the sets in turn, while the third set is used as a testing set.

Altogether 30,945 human promoter sequence segments from -200 to +50 relative to TSS from the DBTSS were considered. For non-promoters, I used 30,945 exon and 30,945 intron sequences with 251 bp (base pair) in length from the EID database, and 30,945 3'UTR sequences with 251 bp in length from the UTRdb database [13]. The prediction accuracies were analyzed using sensitivity, specificity, classification errors, and precision. The three cross validation results are summarized in Table 6.4 with the last two rows showing the mean and the standard deviation for each of the performance measures and the associated flowchart is shown in Fig. 6.5.

The DigiPromPred system gives a mean sensitivity of about 91%. Thus a large proportion of promoter sequences can be correctly detected. The mean specificity of 90% means that only 10% of non-promoter sequences are incorrectly classified as promoters. This represents an improvement over existing human promoter detection tools. The classification error is defined as portion of erroneous identifications (including both false positive and false negative) out of all the testing sequences. The DigiPromPred

system gives a low value of 9.5% only. The precision, which is defined as the portion of all correct identifications out of all the sequences, is 90.5%. This implies that both

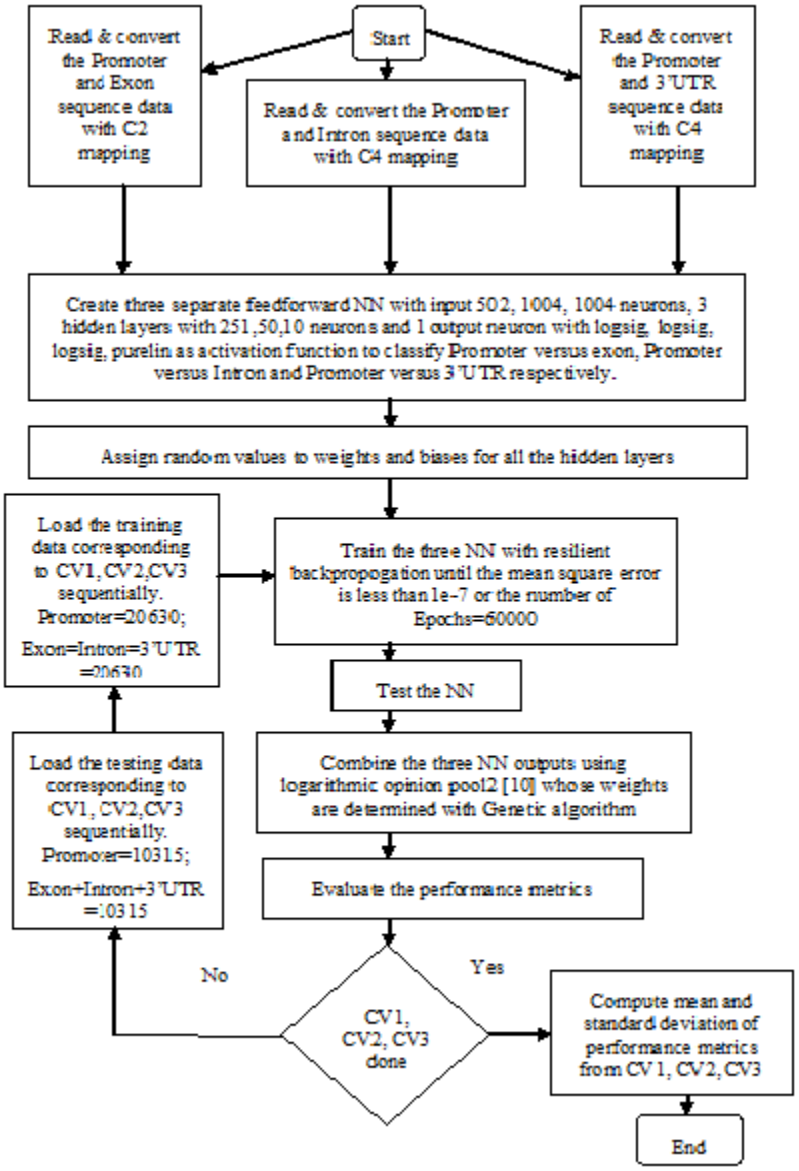

Fig. 6.5 Flowchart for performance evaluation of 3-Cross validation test.

promoters and non-promoters can be correctly detected. The prediction accuracy analysis results shown in Table 6.4 indicate that the DigiPromPred system is effective in

identifying *human* promoter sequences while at the same time reduces the false prediction rate for non-promoter sequences.

Table 6.4 Performance Evaluation of 3-Cross Validation Test (Sen: Sensitivity (%); Spe: Specificity (%); Er: Classification error (%); Pre: Precision (%))

Label	Sen	Spe	Er	Pre
Cross Validation 1	91.2	88.3	10.3	89.7
Cross Validation 2	100.0	96.7	1.7	98.3
Cross Validation 3	81.1	86.4	16.5	83.6
Mean	90.8	90.4	9.5	90.5
Standard deviation	9.4	5.5	7.4	7.4

6.5 Human Chromosome -22 Test

The DigiPromPred system was evaluated on identifying promoter regions in *human* chromosome 22 [13]. I tested for the 20 experimentally annotated promoters in the chromosome 22 [13]. A window of size 251 bp [13] was used to move over the sequence at a step size of 70 bp. If one or more predictions fall in the region [-2000,+2000] relative to the annotated TSS location, the respective gene is counted as a true positive (TP). If the annotated TSS is missed by this count, it represents a false negative (FN). All predictions that fall on the annotated part of the gene within the region [+2001, EndofTheGene] are counted as false positives (FPs). If there are no predictions in the region [+2001, EndofTheGene], it represents a true negative (TN).

The performance of the DigiPromPred system is compared with other state-of-the-art promoter prediction systems including Promoter 2.0 [13], Promoter Scan 1.7 [13], NNPP [13], TSSG [13], TSSW [13], FPROM [94]. The results of the comparative studies are summarized in Table 6.5 and the associated flowchart is shown in Fig. 6.6. The second

row shows the total number of true positive (TP), i.e., the number of promoters that are correctly classified as promoters.

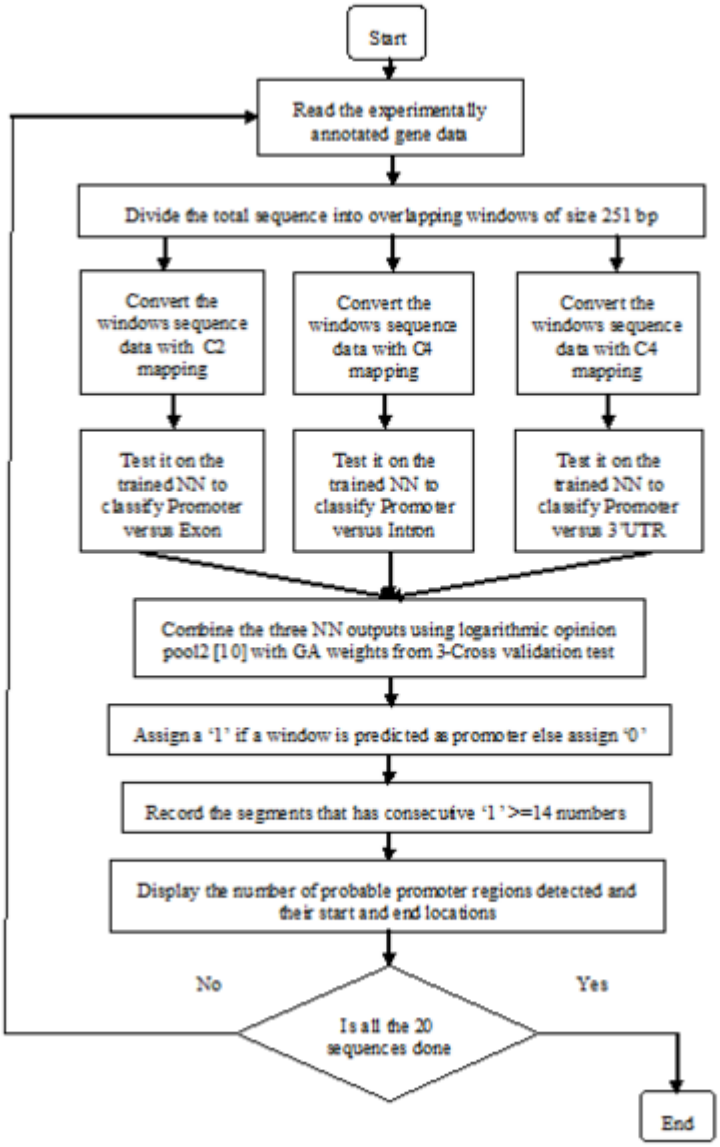

Fig. 6.6 Flowchart for performance evaluation of the *Human* chromosome 22 level test.

As there are 20 annotated promoters, the Promoter 2.0 system, and the NNPP system can correctly detect all the 20 promoters and our proposed system detects 16 out of 20

promoters. The third row shows the total number of false negative (FN), i.e., the number of promoters that are not detected by the prediction tools. The fourth row shows the average number of false positive (FP) over the 20 annotated promoters. As the false positive represents promoters that are wrongly classified as non-promoters, this number should be small. Our proposed system "DigiPromPred", which is a numerical representation based system, gives the smallest number of FP=2 as compared to state-of-the-art promoter prediction systems that are feature based. Fig. 6.7 shows the comparative study of the proposed 'DigiPromPred' with other state-of-the-art promoter prediction systems. A reliable prediction system should have FP as small as possible. From Table 6.5, it can be inferred that the DigiPromPred system provides the highest prediction accuracy and the number of FP is small as compared to the Promoter 2.0, and the NNPP system. The results indicate that the DigiPromPred system achieves a good balance between prediction accuracy and reliability. In terms of the *F-measure*, our system performs better than other systems.

Table 6.5 Comparison of Seven Prediction Systems for Experimentally Annotated Promoters on *Human* Chromosome 22 (1: Promoter 2.0 [21]; 2: Promoter Scan 1.7 [21]; 3: NNPP [21]; 4: TSSG [21]; 5: TSSW [21]; 6: FPROM [22]; 7: DigiPromPred (Proposed System).

Label	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Total number of TP	20	11	20	17	18	17	***16***
Total number of FN	0	9	0	3	2	3	***4***
Average number of FP	20	3	31	5	5	5	***2***
Total number of TN	2	7	1	4	2	3	***8***
Sensitivity	100	55	100	85	90	85	***80***
Specificity	9	70	3	44	29	38	***80***
PPV	50	79	39	77	78	77	***89***
F-measure	67	65	56	81	84	81	***84***
Precision	52	60	40	72	74	71	***80***
Classification Error	48	40	60	28	26	29	***20***

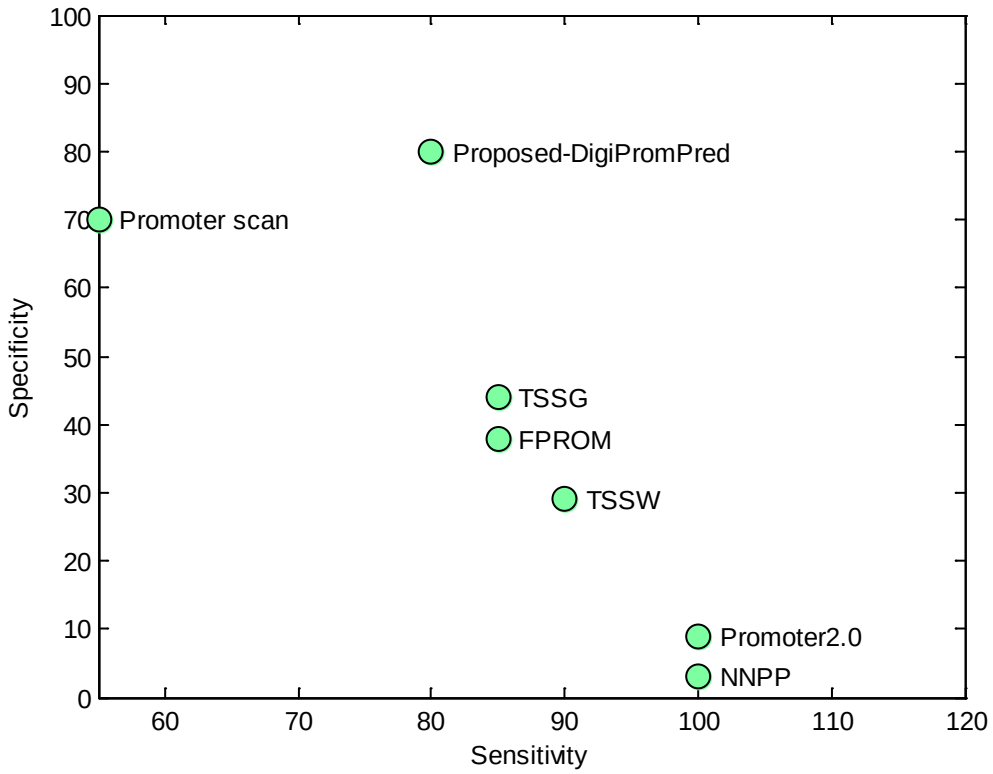

Fig. 6.7 Comparison of the proposed "DigiPromPred' with other state-of-the-art promoter prediction systems

CHAPTER VII

A SIMPLE PROMOTER PREDICTION SYSTEM

In this chapter a simple feed forward neural network classifier known as "SDigiPromPred" is proposed to predict promoters in three model organisms with reduced architecture and computational complexity. The numerical representation selection strategy and the architecture of the proposed SDigiPromPred system are discussed. At the end of the chapter, the experimental results are evaluated.

7.1 The Proposed SDigiPromPred System

In the "MultiNNProm" system, a single DNA sequence is represented by using four different numerical representations to retain different biological properties embedded in the sequence with four different neural networks shown in Fig. 5.1 and extensively discussed in chapter 5. With the use of three classifiers for 1) promoters versus exons; 2) promoters versus introns and 3) promoters versus 3'UTR, the architecture and computational complexity proposed in "DigiPromPred" is reduced by 75.0% compared to MultiNNProm system shown in Fig. 6.1 and extensively discussed in chapter 6. In order to further reduce the architecture and computational complexity (Appendix E) of DigiPromPred system further, I propose a single neural network based system using only one DNA numerical representation titled "SDigiPromPred" compared to DigiPromPred and MultiNNprom systems.

The proposed new solution "SDigiPromPred" uses a feed-forward neural network, which has three layers: input layer, hidden layer and output layer, and trained using a supervised

backpropagation algorithm as shown in Fig. 7.1 to predict promoters in three model organisms like *Human*, *Drosophila*, and *Arabidopsis thaliana* using numerical representation based approach. Due to the use of different numerical sequences corresponding to *human*, *Drosophila*, and *Arabidopsis thaliana*, in representing a DNA sequence, the neural network have slightly different configurations, which are summarized in Table 7.1.

The network is trained to produce an output of '-1' if the input sequence is found to be a non-promoter, and an output of '1' if the sequence is deemed to be a promoter. The output of the neural network is then passed through a probability function [10] which is defined as follows. If the output of the neural network is 1 or more than 1, it is likely that the given sequence is a promoter and a probability of 0.9999 is assigned. If the output of the neural network is -1 or less than -1, it is unlikely that the given sequence is a promoter and a probability of 0.001 is assigned. The probabilities are assigned due to the use of the logarithm function within the result scaling. The output of the probability function is then scaled by a weight in the result scaling. The weight is obtained by first initializing a random population of weight. They are refined using a genetic algorithm with the classification accuracy of the total system as the fitness function [10].

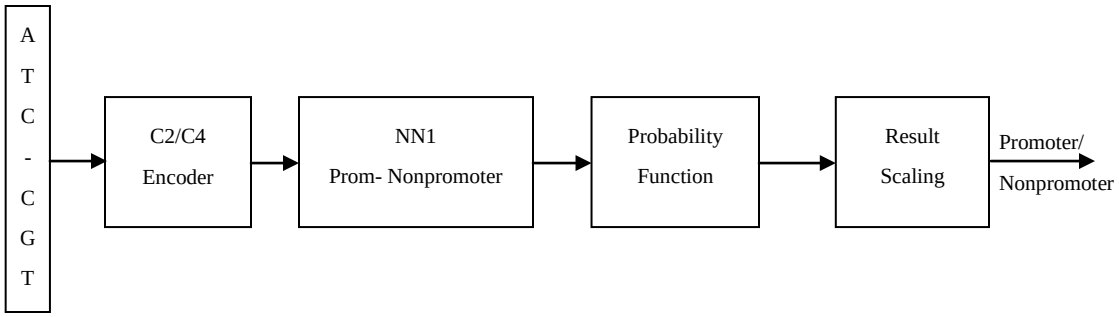

Fig. 7.1 SDigiPromPred: a *human* promoter prediction system with C2 encoding, a *Drosophila* promoter prediction system with C2 encoding, and a *Arabidopsis thaliana* promoter prediction system with C4 encoding respectively.

Table 7.1 Neural Network Configurations for SDigiPromPred System

Dataset for Neural Network NN1	Encoding method	Number of input neurons	Number of neurons in three hidden layers	Activation function
Human	C2	600	300:50:10	Logsig: Logsig: Logsig: Purelin
Drosophila	C2	600	300:50:10	Logsig: Logsig: Logsig: Purelin
Arabidopsis thaliana	C4	1004	251:50:10	Logsig: Logsig: Logsig: Purelin

7.2 Numerical Representation Selection Strategy of the Proposed SDigiPromPred System

In the "MultiNNProm" system, a single DNA sequence is represented by using four different numerical representations to retain different biological properties embedded in the sequence. To find the most suitable numerical representations out of these four representations {E1, E2, C2, C4} for our proposed system SDigiPromPred for three model organisms, I carried out additional experiments whose results are tabulated in Table 7.2, 7.3, and 7.4. The best representation chosen based on the lowest classification error for the three model organism *human*, *Drosophila*, *Arabidopsis thaliana* are C2, C2, and C4 respectively.

7.3 Experimental Results

A feed forward neural network based system as shown in Fig. 7.1 using different numerical mappings E1, E2, C2, and C4 respectively were trained and tested with Matlab R2007b for *Human*, *Drosophila*, and *Arabidopsis* thaliana datasets respectively as mentioned in appendix D. The details of the data sets used, and the associated

performance metrics and the most suitable numerical representation for each model organism are discussed below.

7.3.1 Case study with *Human* dataset

The SDigiPromPred system was evaluated on 565 *human* promoters sequences [-250, +50] bp around the TSSs from the Berkeley human database [92]. For non-promoters, 680 exons sequences with 300 bp in length were extracted from the Berkeley human database [92], and 853 introns with 300 bp in length were extracted from the same source [92]. These data were divided into three sets: training the neural network, training the total system to get the weight using genetic algorithm, and testing the total system with this optimized weight used in the result scaling. The training set 1 consist of 138 promoter sequences, 178 exons, and 212 introns. The training set 2 consists of 138 promoter sequences, 178 exons, and 212 introns. The remaining sequences are the testing sequences. The difficulty with this dataset is that promoter sequences include both introns and exons (from +1 bp to +50 bp), thus classification of promoters is much more difficult than pure introns or exons [92]. Furthermore, the datasets are imbalanced: there are 26.1% promoters against 73.9% nonpromoters in the *human* training dataset1 and dataset2, and 27.7% promoters against 72.3% nonpromoters in the *human* dataset3 for testing. Classification results of the test sequences (dataset3) are used to calculate the sensitivity, specificity, classification error and precision (Table 7.2) for E1, E2, C2, and C4 numerical representation respectively. From the Table 7.2, it is found that C2 encoding is the most suitable numerical representation for *Human* data set as it has the lowest classification error of 9.50% compared to E1, E2, and C4 numerical representations. The associated flowchart is shown in Fig. 7.2.

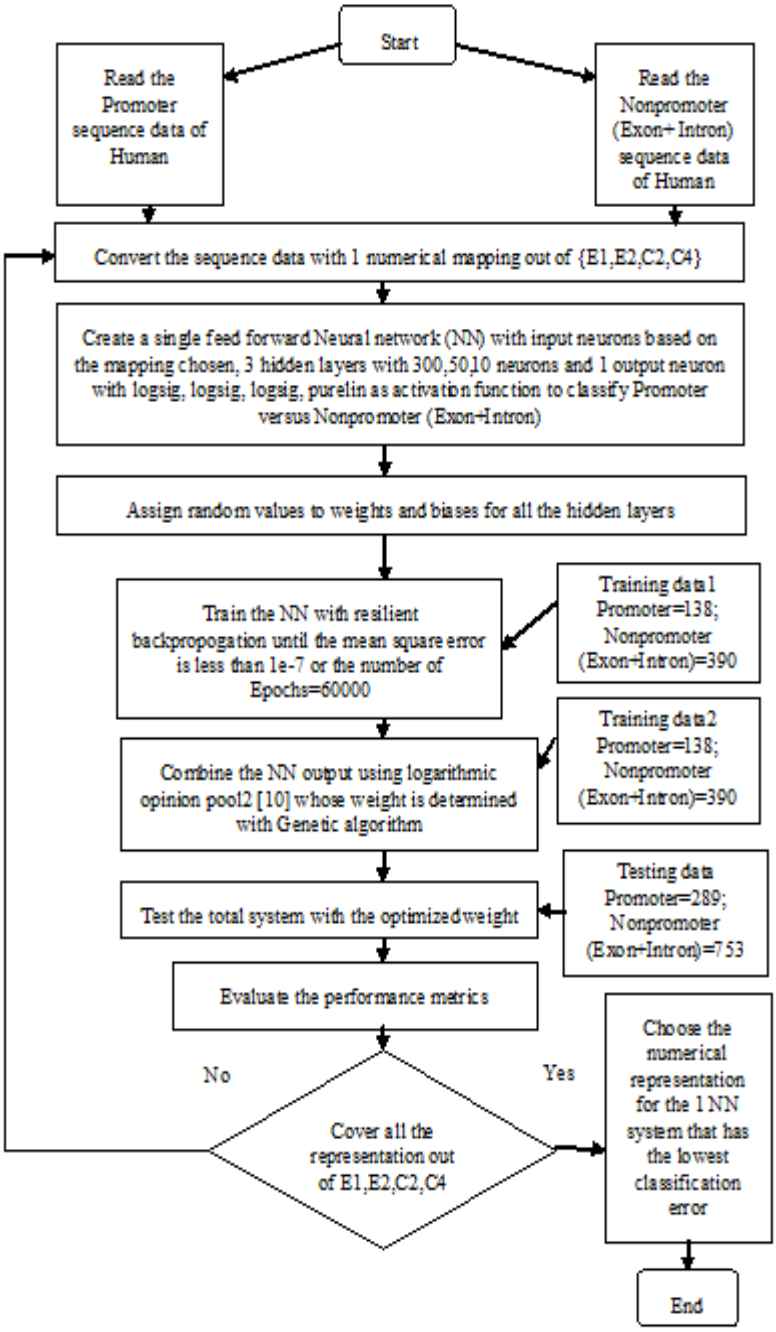

Fig. 7.2 Flowchart for performance evaluation of SDigiPromPred with various numerical representations for *Human* data set.

Table 7.2 Performance Evaluation of SDigiPromPred System with Various Numerical Representations for *Human* dataset (Sen: Sensitivity (%); Spe: Specificity (%); Er: Classification error (%); Pre: Precision (%))

Numerical Representation	Sen	Spe	Er	Pre
E1	77.45	90.09	13.24	86.76
E2	77.19	90.89	12.86	87.14
C2	**86.54**	**91.82**	**9.50**	**90.50**
C4	51.71	87.82	26.39	73.61

7.3.2 Case study with *Drosophila* dataset

The SDigiPromPred system was evaluated on 1842 *Drosophila* promoters sequences [-250, +50] bp around the TSSs from the Berkeley Drosophila database [92]. For non-promoters, 2859 exons sequences with 300 bp in length were extracted from the Berkeley Drosophila database [92], and 1799 introns with 300 bp in length were extracted from the same source [92]. These data were divided into three sets: training the neural network, training the total system to get the weight using genetic algorithm, testing the total system with this optimized weight used in the result scaling. The training set 1 consist of 429 promoter sequences, 709 exons, and 419 introns. The training set 2 consists of 429 promoter sequences, 709 exons, and 419 introns. The remaining sequences are the testing sequences. The difficulty with this dataset is that promoter sequences include both introns and exons (from +1 bp to +50 bp), thus classification of promoters is much more difficult than pure introns or exons [92]. Furthermore, the datasets are imbalanced: there are 27.6% promoters against 72.4% nonpromoters in the *Drosophila* training dataset1 and dataset2, and 29.1% promoters against 70.9% nonpromoters in the human dataset3 for testing. Classification results of the test sequences (dataset3) are used to calculate the sensitivity, specificity, classification error and precision (Table 7.3) for E1, E2, C2, and

C4 numerical representation respectively. From Table 7.3, it is found C2 encoding is the most suitable numerical representation for Drosophila data set as it has the lowest classification error of 8.24% compared to E1, E2, and C4 numerical representation. The associated flowchart is shown in Fig. 7.3.

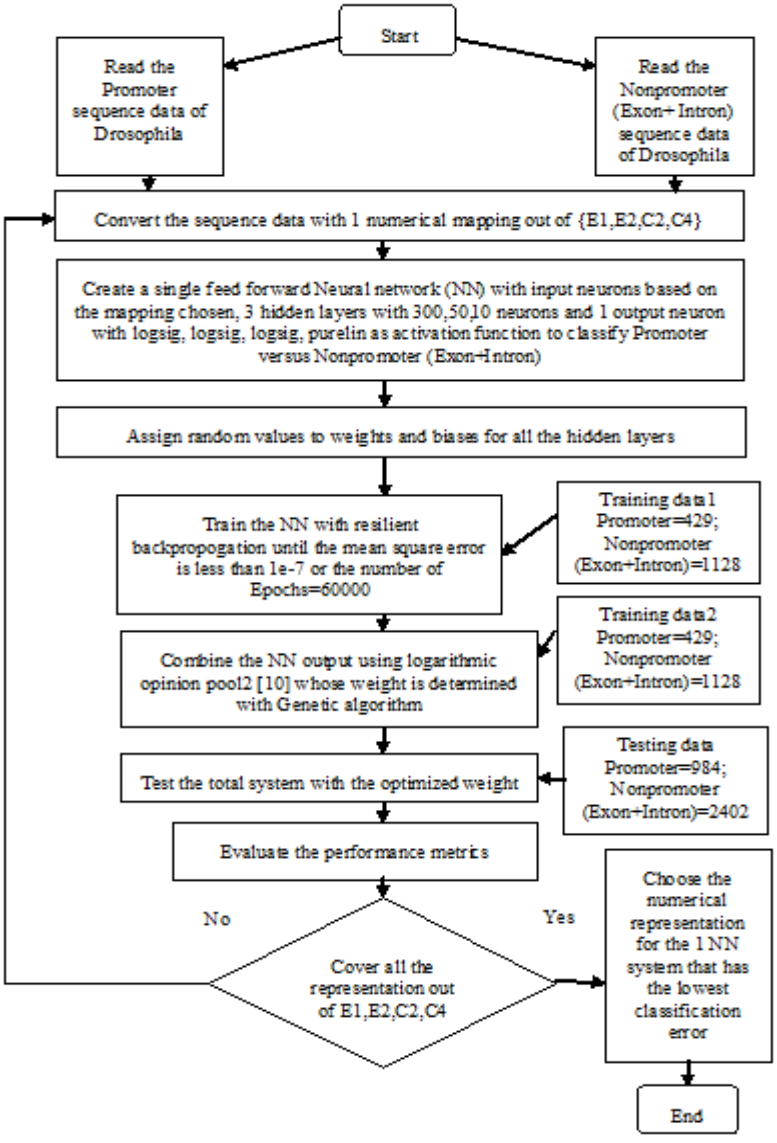

Fig. 7.3 Flowchart for performance evaluation of SDigiPromPred with various numerical representations for *Drosophila* data set.

Table 7.3 Performance Evaluation of SDigiPromPred System with Various Numerical Representations for *Drosophila* dataset (Sen: Sensitivity (%); Spe: Specificity (%); Er: Classification error (%); Pre: Precision (%))

Numerical Representation	Sen	Spe	Er	Pre
E1	74.37	87.77	15.86	84.14
E2	74.73	87.93	15.65	84.35
C2	**87.14**	**93.56**	**8.24**	**91.76**
C4	40.00	76.79	36.03	63.97

Table 7.4 Performance Evaluation of SDigiPromPred System with Various Numerical Representations for *Arabidopsis thaliana* dataset (Sen: Sensitivity (%); Spe: Specificity (%); Er: Classification error (%); Pre: Precision (%))

Numerical Representation	Sen	Spe	Er	Pre
E1	92.47	96.88	4.33	95.67
E2	82.69	95.20	8.47	91.53
C2	96.55	98.19	2.26	97.74
C4	**99.32**	**99.48**	**0.56**	**99.44**

7.3.3 Case study with *Arabidopsis thaliana* dataset

The SDigiPromPred system was evaluated on 305 *Arabidopsis thaliana* promoters sequences [-200, +50] bp around the TSSs from the TAIR Arabidopsis Thaliana [95]. For non-promoters, 291 exons sequences with 251 bp in length were extracted from the TAIR Arabidopsis Thaliana [95], and 541 introns with 251 bp in length were extracted from the same source [95]. These data were divided into three sets: training the neural network, training the total system to get the weight using genetic algorithm, testing the total system with this optimized weight used in the result scaling. The training set 1 consist of 79

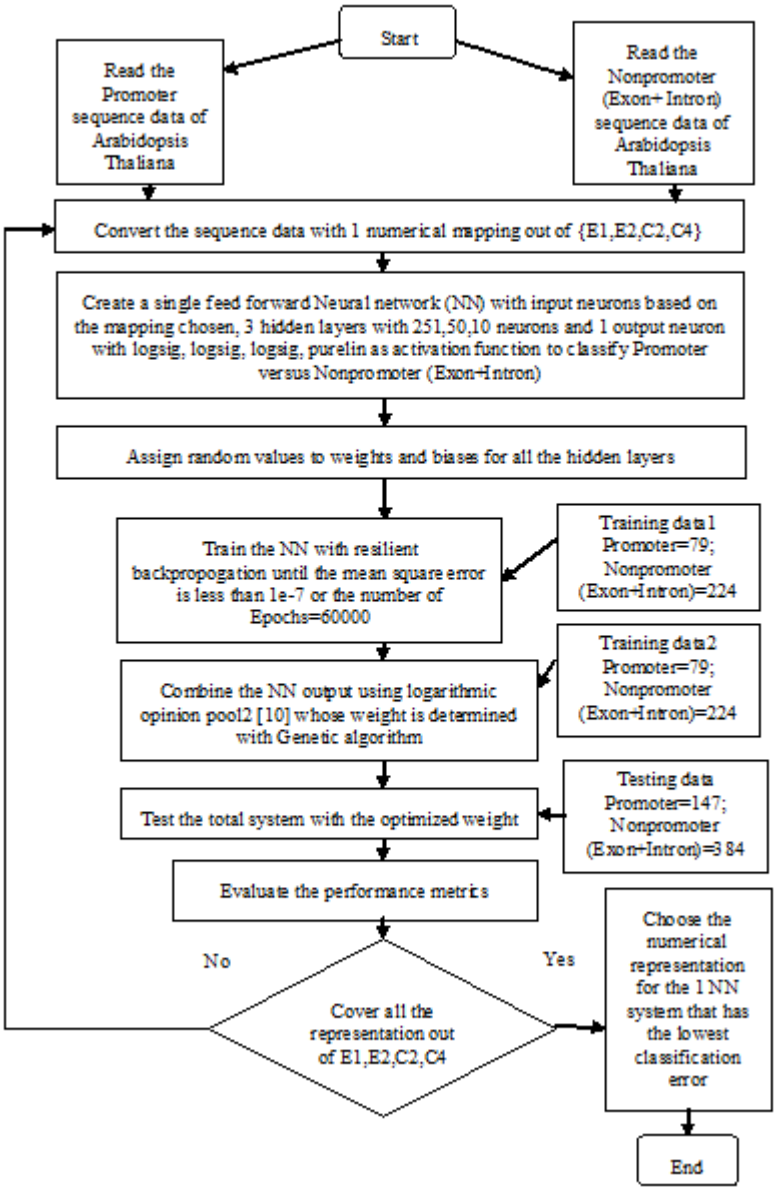

Fig. 7.4 Flowchart for performance evaluation of SDigiPromPred with various numerical representations for *Arabidopsis thaliana* data set.

promoter sequences, 94 exons, and 130 introns. The training set 2 consists of 79 promoter sequences, 94 exons, and 130 introns. The remaining sequences are the testing sequences. The difficulty with this dataset is the limited number of promoter sequences available.

Furthermore, the datasets are imbalanced: there are 26.1% promoters against 73.9% nonpromoters in the TAIR Arabidopsis Thaliana training dataset1 and dataset2, and 27.7% promoters against 72.3% nonpromoters in the TAIR *Arabidopsis Thaliana* test dataset. Classification results of the test sequences (dataset3) are used to calculate the sensitivity, specificity, classification error and precision (Table 7.4) for E1, E2, C2, and C4 numerical representation respectively. From the Table 7.4, it is found that C4 encoding is the most suitable numerical representation for *Arabidopsis thaliana* data set as it has the lowest classification error of 0.56% compared to E1, E2, and C2 numerical representations. The associated flowchart is shown in Fig. 7.4.

7.4 Evaluation of Results

The classification results for the different encoding methods are presented in Table 7.2 (for the *Human* dataset), Table 7.3 (for the *Drosophila* dataset), and Table 7.4 (for *Arabidopsis thaliana* dataset).

The best classification results for the *human* and *drosophila* dataset is obtained using 2-bit binary and for Arabidopsis thaliana dataset using 4-bit binary representation respectively. The reason for this is that the *drosophila* promoter sequences are characterized by the repeating occurrences of the TATA box (TATAA or TATAAA subsequences) or the *Pribnow* box (TATAAT subsequences), thus the 2-bit binary encoding, where A and T nucleotides are mapped using different binary values can extract these repeating sequences effectively leading to best results [92].

DPE (*downstream promoter elemen*t) which is a distinct 7-nucleotide sequence (A/G)G(A/T)CGTG [92] occur in *human* promoter sequences frequently, thus the best

classification results are obtained, when C and G nucleotides are represented using 2-bit binary mapping, because they separates A/T and C/G nucleotides efficiently [92].

Arabidopsis thaliana sequences are also characterized by the occurrence of the TATA box (TATAAA subsequence), Y patch (TCTCTC subsequence), REG (regulatory element group) with AGGCCC subsequence, and YR rule (Y: C or T; R: A or G) [96], thus the best results is achieved using the 4-bit binary encoding because this mapping is a unitary coding matrix with identical hamming distance between the nucleotide encodings and is capable to detect all these sub sequences efficiently.

CHAPTER VIII

CONCLUSION AND FUTURE WORK

The primary aim of this research was the development of DigiPromPred, a numerical representation based human promoter prediction system and also the development of SDigiPromPred, for promoter prediction in three model organisms with reduced architecture and computational overload. In this chapter, I summarize our main results and contributions.

8.1 Conclusion

Eukaryotic promoters are well known to be diverse in nature and are difficult to characterize. This makes most existing feature-based promoter prediction systems suffer from low specificity with a large number of false positives, i.e., non-promoters are incorrectly identified as promoters. In this dissertation, I have presented a numerical representation based approach for recognizing human promoters where each individual nucleotide in the DNA sequence is used in the promoter identification. A DNA sequence contains exons, introns, promoters and 3'UTR with different regional characteristics. By using appropriate numerical representation for different regions in DNA sequences, regional properties can be characterized more effectively. The system has been tested with a large set of genomic sequences of promoters (30,945) and nonpromoters (exon, intron, 3'UTR each of 30945) and 20 annotated promoters in *human* chromosome 22. Chromosome level experimental results show that the DigiPromPred system is effective in identifying human promoter sequences with sensitivity, *PPV*, and *F-measure* that are

found to be 80%, 89%, 84%, respectively while reducing the false prediction rate for non-promoter sequences with a specificity of 80%. Thus the proposed “DigiPromPred” system which is based on numerical representation of DNA sequences outperforms the state-of-the-art promoter prediction systems that are feature based.

Compared to several feature based human promoter prediction systems, the study clearly demonstrates better performance of DigiPromPred on the analyzed human data set. Apart from this, the DigiPromPred system has several advantages compared to the existing feature based system for human promoter prediction. These have been highlighted below along with several new concepts that have been introduced in this study.

First study where a detailed review and comparison of existing DNA symbolic-to-numeric representations is presented. Tables 3.1, 3.2, and 3.3 are provided for reader’s quick reference with numerical examples of all the 35 DNA numerical representation methods and their variants.

It is the first time to evaluate the performance of about 35 DNA numerical representation methods and its variants with discrete Fourier transform (DFT) based on period-3 property to classify short length human coding and noncoding sequences.

- First study and development of DigiPromPred a human promoter predictions system totally based on numerical mapping of DNA sequences.
- Introduction of DNA numerical representation selection strategy
- Prediction of promoter in a DNA sequence
- Performance evaluation of DigiPromPred with 3-cross validation test with large data sets

- Identification of promoter regions in large genomic sequences, viz, 20 annotated promoters in *human* chromosome 22.
- Achieved lowest mean and standard deviation with 3-cross validation test.
- Improved sensitivity and reduced false prediction rate.
- First study and development of SDigiPromPred system with reduced architecture and computational complexity compared (Appendix E) to DigiPromPred.
- Performance evaluation and analysis of SDigiPromPred system in three model organisms

The proposed system DigiPromPred which is based on numerical representation of DNA sequences outperforms the state-of-the-art promoter prediction systems that are feature based and

I believe that DNA research community will find DigiPromPred a useful complement to the existing set of promoter analysis tools and SDigiPromPred system as an alternate approach for promoter prediction compared to systems that are feature based.

8.2 Future Work

The future work is to evaluate the suitability of 35 different numerical representations and its variants presented in this dissertation for prokaryotic and eukaryotic promoter prediction using DigiPromPred system.

To extract regional features from the numerical representations so as to improve the efficiency of the DigiPromPred system for chromosome level testing with improved sensitivity and specificity.

Also, the implementation of DigiPromPred on the web to facilitate research community for testing their data sets for human promoter prediction.

Finally, the feasibility and evaluation of SDigiPromPred for promoter prediction for 3-cross validation test and chromosome level test and comparison with DigiPromPred and other state of the art promoter prediction systems.

Appendix A

The Genetic Code: Codons to Amino Acids Mapping

The mRNA codons (i.e., triplets of available four types of mRNA bases A, C, G, and U) in exon regions encode 20 amino acids and 3 stop signals, as listed in Table A.1:

Table A.1. The genetic code: codons to amino acid mapping [15]

First Position (5' end)	Second position				Third position (3'end)
	G	A	C	U	
G	Gly	Glu	Ala	Val	G
	Gly	Glu	Ala	Val	A
	Gly	Asp	Ala	Val	C
	Gly	Asp	Ala	Val	U
A	Arg	Lys	Thr	Met	G
	Arg	Lys	Thr	Ile	A
	Ser	Asn	Thr	Ile	C
	Ser	Asn	Thr	Ile	U
C	Arg	Gln	Pro	Leu	G
	Arg	Gln	Pro	Leu	A
	Arg	His	Pro	Leu	C
	Arg	His	Pro	Leu	U
U	Trp	STOP	Ser	Leu	G
	STOP	STOP	Ser	Leu	A
	Cys	Tyr	Ser	Phe	C
	Cys	Tyr	Ser	Phe	U

Three-letter code and name of 20 amino acids commonly found in proteins are: Ala = Alanine, Cys = Cysteine, Asp = Aspartic Acid, Glu = Glutamic Acid, Phe = Glycine, His = Histidine, Ile = Isoleucine, Lys = Leucine, Met = Methionibe, Asn = Asparagine, Pro = Proline, Gln = Glutamine, Arg = Arginine, Ser = Serine, Thr = Threonine, Val = Valine, Trp = Tryptophan, Tyr = Tyrosine.

Appendix B

Simulation Platform Specifications

Hardware Specifications:

- Processor : Intel (R) Pentium (R) D CPU 3.20 GHz 3.20 GHz
- Memory : 1.00 GB
- System type : 32-bit Operating System
- Operating System : Microsoft windows Vista Enterprise

Software Specifications:

- Matlab : Version 7.5.0.342 (R2007b)
- Toolbox : Bioinformatics & Neural Network

Appendix C

Dataset [79] used for Classification of coding and noncoding regions in *human* sequence for various numerical representation based on period-3 with DFT

S.No.	Title	Human Dataset File Name	Number of Sequences	Average Sequence length (bp)
1	Coding	H_exonsout_1500_1.txt	1500	140
2	Noncoding	H_intronsout_1500_1.txt	1500	140

Dataset [13] used for 2 NN, 3 NN and 3-Cross validation test

S.No.	Title	Human Dataset File Name	Number of Sequences	Sequence length (bp)
1	Promoter	promoter.txt	30946	251
2	Nonpromoter	exon.txt	30946	251
3	Nonpromoter	intron.txt	30946	251
4	Nonpromoter	utr3.txt	30946	251

Dataset [13] used for Chromosome level test

S.No.	Human Chromosome 22 File Name	Sequence length (bp)
1	chr22_193_3.txt	10036
2	chr22_193_4.txt	48920
3	chr22_193_6.txt	27557
4	chr22_193_9.txt	81776
5	chr22_193_10.txt	32465
6	chr22_193_11.txt	92507
7	chr22_193_12.txt	15548
8	chr22_193_13.txt	22940
9	chr22_193_14.txt	29407
10	chr22_193_15.txt	22461

S.No.	Human Chromosome 22 File Name	Sequence length (bp)
1	chr22_193_17.txt	4460
2	chr22_193_18.txt	11847
3	chr22_193_19.txt	7823
4	chr22_193_20.txt	26115
5	chr22_193_21.txt	4109
6	chr22_193_24.txt	5563
7	chr22_193_25.txt	42720
8	chr22_193_26.txt	12311
9	chr22_193_28.txt	28891
10	chr22_193_30.txt	29222

Appendix D

Dataset [96] used for Classification of promoter and nonpromoter (exon+intron) sequences in *Arabidopsis thaliana* model organism

S.No.	Title	Human Dataset File Name	Number of Sequences	Sequence length (bp)
1	Promoter	promoter.txt	305	251
2	Nonpromoter	exon.txt	291	251
3	Nonpromoter	Intron.txt	541	251

Dataset [92] used for Classification of promoter and nonpromoter (exon+intron) sequences in *Drosophila* model organism

S.No.	Title	Human Dataset File Name	Number of Sequences	Sequence length (bp)
1	Promoter	promoter.txt	984	300
2	Nonpromoter	exon.txt	1441	300
3	Nonpromoter	Intron.txt	961	300

Dataset [92] used for Classification of promoter and nonpromoter (exon+intron) sequences in *Human* model organism

S.No.	Title	Human Dataset File Name	Number of Sequences	Sequence length (bp)
1	Promoter	promoter.txt	289	300
2	Nonpromoter	exon.txt	324	300
3	Nonpromoter	Intron.txt	429	300

Appendix E

Computational complexity of promoter prediction systems

S.No.	Promoter prediction system	Number of Neural Networks used	Training time	Testing time	Data set	Numerical Representation used
1	SDigiPromPred	1	4.24 mins.	1.36 min	*Human*	C2
2	DigiPromPred	3	6.10 hours	5.20 min	*Human*	C2,C4,C4
3	MultiNNProm	4	10.12 hours	7.37 min	*E. Coli*	E1,E2,C2,C4
4	MultiNNProm-extended	12	27.36 hours	7.57 min	*Human*	E1,E2,C2,C4; E1,E2,C2,C4; E1,E2,C2,C4

References

[1] C. H. Wu and J. W. McLarty, *Neural Networks and Genome Informatics*. 1st Ed. New York: Elsevier Science, 2000.

[2] R. J. Robbins, B. David, and S. Jay, "Informatics and the Human Genome Project," *IEEE Engineering in Medicine and Biology Magazine*, vol. 14, pp. 694-701, Nov.-Dec. 1995.

[3] T. A. Down, and T. J. P. Hubbard, "Computational detection and location of transcription start sites in mammalian genomic DNA," *Genome Research*, vol. 12, pp. 458-461, Mar. 2002.

[4] R. V. Davuluri, I. Grosse, M. Q. Zhang, "Computational identification of promoters and first exons in the human genome," *Nature Genetics*, vol. 29, pp. 412-417, Dec. 2001.

[5] M. Scherf, A. Klingenhoff, T.Werner, "Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach," *J. of Molecular Biology*, vol. 297, pp. 599-606, Mar. 2000.

[6] V. B. Bajic, S. H. Seah, A. Chong, S. P. T. Krishnan, J. L. Y. Koh, V. Brusic, "Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates," *J. of Molecular Graphics and Modelling*, vol. 21, pp. 323-332, Mar. 2003.

[7] V. B. Bajic, and S. H. Seah, "Dragon Gene Start Finder: An advanced system for finding approximate locations of the start of gene transcriptional units," *Genome Research*, vol. 13, pp. 1923-1929, Aug. 2003.

[8] V. B. Bajic, S. L. Tan, Y. Suzuki, S. Sugano, "Promoter prediction analysis on the whole human genome," *Nature Biotechnology*, vol. 22, pp. 1467-1473, Nov. 2004.

[9] X. Li, J. Zeng, H. Yan, "PCA-HPR: A principle component analysis model for human promoter recognition," *Bioinformation*, vol. 2, pp. 373-378, Jun. 2008.

[10] R. Ranawana, V. Palade, "A neural network based multi-classifier system for gene identification," *Neural Comput & Applic.*, vol. 14, pp. 122-131, Jul. 2005.

[11] Q. Ma, J. Wang, D. Shasha, C. Wu, "DNA sequence classification via an expectation maximization algorithm and neural networks: A case study," *IEEE Transactions on Systems, Man, and Cybernetics, part C: Applications and Reviews, Special Issue on Knowledge Management*, vol. 31, pp. 468-475, Nov. 2001.

[12] I. Mahadevan, I. Ghosh, "Analysis of E.coli promoter structures using neural networks," *Nucl. Acids Res.*, vol. 22, pp. 2158-2165, Jun. 1994.

[13] J. Zeng, X.-Y. Zhao, X.-Q. Cao, H. Yan, "SCS: signal, context and structure features for genome-wide human promoter recognition," I*EEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 7, pp. 550-562, Jul.-Sep. 2010.

[14] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, *Essential Cell Biology*, 1st Ed. New York: Garland Publishing, 1998.

[15] C. Setubal and J. Meidanis, *Introduction to Computational Molecular Biology*, 1st Ed. Boston: PWS Publishing, 1997.

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

[17] A. G. Pedersen, P. Baldi, Y. Chauvin, S. Brunak, "The biology of Eukaryotic promoter prediction--a review," *Computers and Chemistry*, vol. 23, pp. 191-207, Jun. 1999.

[18] T. Werner, "Models for prediction and recognition of Eukaryotic promoters," *Mammalian Genome*, vol. 10, pp. 168-175, Feb. 1999.

[19] T. Reichhardt, "Will souped up salmon sink or swim?" *Nature*, vol. 406, pp. 10-12, Jul. 2000.

[20] P. Rorth, K. Szabo, A. Bailey, T. Laverty, J. Rehm, G. M. Rubin, K. Weigmann, M. Milan, V. Benes, W. Ansorge, S. M. Cohen, "Systematic gain-of-function genetics in Drosophila," *Development*, vol. 125, pp. 1049-1057, Mar. 1998.

[21] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloss, S. Land, B. Lewicki-Patapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, E. Wingender, "TRANSFAC: Transcriptional regulation, from patterns to profiles," *Nucleic Acids Res.*, vol. 31, pp. 374-378, Jan. 2003.

[22] R. J. Robbins, "Challenges in the Human Genome Project," *IEEE Engineering in Biology and Medicine*, vol. 11, pp. 25-34, Mar. 1992.

[23] P. P. Vaidyanathan, "Genomics and Proteomics: A Signal Processor's Tour," *IEEE Circuits and Systems Magazine*, vol. 4, pp. 6-29, Fourth quarter 2004.

[24] E. A. Cheever, D. B. Searls, W. Karunaratne, and G. C. Overton, "Using signal processing techniques for DNA sequence comparison," in *Proc. of Fifteenth Annual Northeast Bioengineering Conference*, Boston, USA, Mar. 1989, pp. 173-174.

[25] J. P. Mena-Chalco, H. Carrer, Y. Zana, R. M. Cesar Jr., "Identification of protein coding regions using the modified Gabor-wavelet transform," I*EEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, pp. 198-207, Apr.–Jun. 2008.

[26] A. Roy, C. Raychaudhury, and A. Nandy, "Novel techniques of graphical representation and analysis of DNA sequences---A review," *Journal of Biosciences*, vol. 23, pp. 55-71, Mar. 1998.

[27] Richard F. Voss, "Evolution of Long-range Fractal Correlations and 1/*f* noise in DNA base sequences," *Physical Review Leters*, vol. 68, pp. 3805-3808, June 1992.

[28] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *Computer Applications in the Biosciences (CABIOS)*, vol. 13, pp. 263-270, Jun. 1997.

[29] J. A. Berger, S. K. Mitra, and J. Astola, "Power spectrum analysis for DNA sequences," in *Proc. of Seventh International Symposium on Signal Processing and its Application*s, Paris, France, vol. 2, Jul. 2003, pp. 29-32.

[30] B. D. Silverman and R. Linker, "A measure of DNA periodicity," *J. Theor. Biol.*, vol. 118, pp. 295-300, Feb. 1986.

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

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

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

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

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

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

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

[38] J. Zhao, X. W. Yang, J. P. Li, and Y. Y. Tang, “DNA sequence classification based on wavelet packet analysis,” in *Proc. of the Second International Conf. on Wavelet Analysis and its Applications: Lecture Notes in Computer Science*, vol. 2251, pp. 424-429, Jan. 2001.

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

[40] P. D. Cristea, “Representation and analysis of DNA sequences,” in *Genomic signal processing and statistics: EURASIP Book Series in Signal Processing and Communications*, vol. 2, E. R. Dougherty et al, Eds. New York: Hindawi, 2005 , pp. 15-66.

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

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

[43] Akhtar, M., J. Epps, E. Ambikairajah, "Signal processing in sequence analysis: advances in eukaryotic gene prediction," *IEEE Journal of Selected Topics in Signal Processing*, vol .2, pp. 310-321, Jun. 2008.

[44] A. K. Brodzik, O. Peters, "Symbol-balanced Quaternionic periodicity transform for latent pattern detection in DNA sequences," *IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing (ICASSP '05)*, Philadelphia, USA, vol. 5, pp. 373-376, Mar. 2005.

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

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

[47] I. Cosic, “Macromolecular Bioactivity: Is it resonant interaction between macromolecules? Theory and Applications,” *IEEE Transactions on Biomedical Engg.*, vol. 41, pp. 1101-1114, Dec. 1994.

[48] Achuthsankar S. Nair and Sreenadhan S. Pillai, “A coding measure scheme employing electron-ion interaction pseudo potential (EIIP),” *Bioinformation*, vol. 1, pp. 197-202, Oct. 2006.

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

[50] H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, Z. D. Goldberger, S, Havlin, S. M. Ossadnik, C. –K. Peng, M. Simmons, “Statistical mechanics in biology: how ubiquitous are long-range correlations?” *Physica A*, vol. 205, pp. 214-253, Apr. 1994.

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

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

[53] C. K. Peng, S.V. Buldyrev, S. Havlin, M. Simmons, H. E. Stanley, and A. L. Goldberger, "Mosaic organization of DNA nucleotides," *Physical Review E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics*, vol. 49, pp. 1685-1689, Feb. 1994.

[54] A. Arneodo, E. Bacry, P. V. Graves, J. F. Muzy, "Characterizing long-range correlations in DNA sequences from wavelet analysis," *Physical Review Letters*, vol. 74, pp. 3293-3296, Apr. 1995.

[55] R. Zhang and C. T. Zhang, "Z curves, An Intuitive Tool, for Visualizing and Analyzing the DNA sequences," *J. BioMol. Struct. Dyn.*, vol. 11, pp. 767-782, Feb. 1994.

[56] C. T. Zhang, J. Wang, and R. Zhang, "A novel method to calculate the G+C Content of Genomic DNA sequences," *J. BioMol. Struct. Dyn.*, vol. 19, pp. 333-341, Oct. 2001.

[57] C. T. Zhang and R. Zhang, "An isochore map of the human genome based on the Z curve method," *Gene*, vol. 317, pp. 127-135, Oct. 2003.

[58] R. Zhang and C. T. Zhang, "Identification of replication origins in archaeal genomes based on the Z-curve method," *Archaea*, vol. 1, pp. 335-346, May 2005.

[59] B. Y. M. Kwan, J. Y. Y. Kwan, H. K. Kwan, R. Atwal, and O. T. Shen, "Wavelet analysis of the genome of the model plant *Arabidopsis thaliana*," in *Proc. of TENCON*, Hong Kong, China, Nov. 2006, pp. 1-4.

[60] H. K. Kwan, R. Atwal, and B. Y. M. Kwan, "Wavelet analysis of DNA sequences," in *Proc. of Int. Conf. on Communications, Circuits and Sytems*, Xiamen, China, May 2008, pp. 917-921.

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

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

[63] Y. Wu, and A. W.-C. Liew, “Classification of short human exons and introns based on statistical features,” *Physical Review E Stat. Nonlin soft Matter Phys.*, vol. 67:, pp. 061916.1-061916.7, Jun. 2003.

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

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

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

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

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

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

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

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

[72] Dan Larhammar and C. A. C. Dreismann, “Biological origins of long-range correlations and compositional variations in DNA,” *Nucleic Acids Research*, vol. 21, pp. 5167-5170, Nov. 1993.

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

[74] R. Nini, Q. Lijun, “Study of numerical mapping methods for DNA sequences,” *Journal of Biomedical Engineering-Chengdu*, vol. 22, pp. 681-685, Aug. 2005 [article in Chinese].

[75] Hon Keung Kwan and Swarna Bai Arniker, "Numerical representation of DNA sequences," in *Proc. Of IEEE Inter. Conf. on Electro/Information Technology*, Windsor, Ontario, Canada, June 2009, pp. 307-310.

[76] Swarna Bai Arniker and Hon Keung Kwan, "Graphical representation of DNA sequences," in *Proc. of IEEE Inter. Conf. on Electro/Information Technology*, Windsor, Ontario, Canada, June 2009, pp. 311-314.

[77] E. N. Trifonov, J. L. Sussman, "The pitch of chromotin DNA is reflected in its nucleotide sequence," *Proc. of the Natl. Acad. Sci., USA*, vol. 77, pp. 3816-3820, Jul. 1980.

[78] S. K. Mitra, *Digital Signal Processing: A Computer-Based Approach*, 2nd Ed. Singapore: McGraw-Hill, 2002.

[79] A. W. C. Liew, Y. Wu, H. Yan, M. Yang, "Effective statistical features for coding and non-coding DNA sequence classification for yeast, C.elegans and human," *Int. J. of Bioinformatics Res. and Applications*, vol. 1, pp. 181-201, Aug. 2005.

[80] A. Arneodo, Y. D. Carafa, E. Bacry, P. V. Graves, J. F. Muzy, C. Thermes, "Wavelet based fractal analysis of DNA sequences," *Physica D*, vol. 96, pp. 291-320, Sep. 1996.

[81] A. Arneodo, Y. D. Carafa, B. Audit, E. Bacry, J. F. Muzy, C. Thermes, "What can we learn with wavelets about DNA sequences?," *Physica A*, vol. 249, pp. 439-448, Jan. 1998.

[82] B. Audit, C. Theremes, C. Vaillant, Y. D. Carafa, J. F. Muzy, A. Arneodo, "Long-Range Correlations in Genomic DNA: a signature of the nucleosomal structure," *Physical Review Letters*, vol. 86, pp. 2471-2474, Mar. 2001.

[83] A. D. Haimovich B. Byrne, R. Ramaswamy, W. J. Welsch, "Wavelet analysis of DNA walks," *Journal of Computational Biology*, vol. 13, pp. 1289-1298, Sep. 2006.

[84] W. W. Wasserman, A. Sandelin, "Applied bioinformatics for the identification of regulatory elements," *Nat. Genet.*, vol. 5, pp. 276-287, Apr. 2004.

[85] U. Ohler, G.Liao, H.Niemann, G.M. Rubin, "Computational analysis of core promoters in the Drosophila genome," *Genome Biol.*, vol. 3, pp. research0087.1-0087.12, Dec. 2002.

[86] T. Werner, "Models for prediction and recognition of eukaryotic promoters," *Mamm. Genome*, vol. 10, pp. 168-175, Feb. 1999.

[87] F. Zhang, M. D. Kuo, A. Brunkhors, "E.coli promoter prediction using feed-forward neural networks," in *Proc. of the 28th IEEE EMBS Annual International Conference*, New York, USA, Aug.-Sept. 2006, pp. 2025-2027.

[88] N. L. Larsen, J. Engelbrecht, S. Brunak, "Analysis of eukaryotic promoter sequences reveals a systematically occuring CT-signal," *Nucleic Acids Research*, vol. 23, pp. 1223-1230, Apr. 1995.

[89] S. Knudsen, "Promoter 2.0: for the recognition of Pol II promoter sequences," *Bioinformatics*, vol. 15, pp. 356-361, May 1999.

[90] X. Xie, S. Wu, K.-M. Lam, H. Yan, "PromoterExplorer: an effective promoter identification method based on the AdaBoost algorithm," *Bioinformatics*, vol. 22, pp. 2722-2728, Sep. 2006.

[91] P. C. Conilione, D. Wang, "Effect of non-target examples on E.coli promoters recognition using neural networks," in *Proc. Of IEEE Inter. Joint Conf. on Neural Networks*, Montreal, Canada, vol. 1, Jul.-Aug. 2005, pp. 310-315.

[92] R. Damasevicius, "Analysis of binary feature mapping rules for promoter recognition in imbalanced DNA sequence datasets using support vector machine," in *4th Inter. IEEE Conf. on Intelligent Systems*, Varna, Bulgaria, Sep. 2008, pp. 11-20-11-25.

[93] C.-J. Lin, C.-C. Peng, C.-Y. Lee, "Prediction of RNA polymerase binding sites using purine-pyrimidine encoding and hybrid learning methods," *International Journal of Applied Science and Engineering*, vol. 2, pp. 177-188, Jul. 2004.

[94] J. Zeng, S. Zhu, H. Yan, "Towards accurate human promoter recognition: a review of currently used sequence features and classification methods," *Briefings in Bioinformatics*, vol. 10, pp. 498-508, Sep. 2009.

[95] Y. Zuo, Q. Li, "Predicting plant pol-II promoter based on subsequence increment of overlap content diversity," in *2nd International Conference on Biomedical Engineering and Informatics*, Tianjin, China, Oct. 2009, pp. 1-5.

[96] Y. Y. Yamamota, H. Ichida, et al, "Identification of plant promoter constituents by analysis of local distribution of short sequences," *BMC Genomics*, vol. 8, pp. 67-89, Mar. 2007.

VITA AUCTORIS

Swarna Bai Arniker was born in Hyderabad, Andhra Pradesh state, India. She has received her M.Sc. (Applied Electronics) degree in first class with **Distinction** from Osmania University, Hyderabad, India in 1990, and M.Tech. (Micro-Electronics & VLSI Design) from Indian Institute of Technology (IIT), Kanpur, India in 2001 on full scholarship from D.R.D.O., Government of India. In the last nearly thirteen years' R&D experience, She worked as a Senior Scientist RCI, Government of India, Hyderabad, India. Now, she is a Ph.D. Candidate at the Department of Electrical and Computer Engineering, University of Windsor, Ontario, Canada on study leave from RCI, Government of India.