TECHNIQUES FOR ANALYSIS OF GENOMIC SEQUENCE DATA

A THESIS

Submitted in partial fulfilment of the
requirements for the award of the degree
of
DOCTOR OF PHILOSOPHY
in
ELECTRONICS AND COMPUTER ENGINEERING

by

RAVI GUPTA

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE-247 667 (INDIA)

APRIL, 2007

INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled TECHNIQUES FOR ANALYSIS OF GENOMIC SEQUENCE DATA in partial fulfillment of the requirement for the award of the Degree of Doctor of Philosophy and submitted in the Department of **Electronics and Computer Engineering** of the Indian Institute of Technology Roorkee, Roorkee is an authentic record of my own work carried out during a period from July 2004 to April 2007 under the supervision of **Dr. Ankush Mittal**, Associate Professor and **Dr. Kuldip Singh**, Professor, **Department of Electronics and Computer Engineering** of Indian Institute of Technology Roorkee, Roorkee.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other Institute.

(RAVI GUPTA)

This is to certify that the above statement made by the candidate is correct to the best of our knowledge.

(Dr. Ankush Mittal)
Associate Professor
Dept. of E&CE

(Dr. Kuldip Singh)
Professor
Dept. of E&CE

Date:

The Ph.D. Viva-Voce Examination of **Ravi Gupta**, Research Scholar, has been held on ....Nov. 10, 2007..

Signature of Supervisor(s) **Signature of External Examiners**

Abstract

Over the past few decades advances in genomic technologies have led to an explosive growth in the biological information generated by the scientific community. There are over 65 billion nucleotides from more than 61 million individual sequences in GenBank as on September 2006. With the enormous amount of genomic and proteomic data available in the public domain, it is becoming increasingly important to be able to analyze the data and interpret the results to decipher the connections between the genomic data and the biological functionality of living cells and organisms.

Mapping the symbolic data into one or more numerical sequences opens the possibility of applying signal processing techniques, especially digital signal processing (DSP) for solving highly relevant problems of biological sequence analysis. Genomic signal processing (GSP) is a quickly evolving interdisciplinary field that blends bioscience, medicine and signal processing. GSP offers several robust and computationally efficient tools like discrete Fourier transform (DFT), digital filters, discrete wavelet transform and several other tools for obtaining solutions to biological problems. In several biological problems, application of signal processing techniques forms the foundation of data analysis.

The goal of the current research work is to apply digital signal processing (DSP) concepts for solving important problems related to sequence analysis.

The thesis combines the advantages of DSP with pattern recognition technique for identification and classification of sequence patterns. GSP techniques have been successful especially for identification of hidden structures but have not made much impact in sequence identification and sequence classification problems. Furthermore, there exist very few signal processing techniques for protein sequence analysis. In this thesis a broad methodology for analysis of DNA and protein sequences is proposed. The aim of the thesis is not to replace the existing techniques but to provide complementary approaches, to explore novel applications of signal processing in bioinformatics, to devise simple and efficient algorithms, to provide novel biological features and to apply machine learning algorithms for improving the analysis capability of GSP algorithms. In chapter two of the thesis, a brief review of the existing techniques and their limitations for the problems that were taken up for the current research work is presented.

Chapter three of the thesis presents a signal processing technique for identification of exact and inexact tandem repeat patterns in DNA sequences. It is well known that tandem repeats in telomeres play important role in cancer and are linked to over a dozen major neurodegenerative genetic disorders in humans. Short tandem repeats are used for DNA fingerprinting. Despite their importance, locating and characterizing these repeats within anonymous DNA sequences remain a challenge. In past, signal processing (SP) algorithms based on DFT and short periodicity transform (PT) techniques have been applied for identifying tandem repeats. Periodicity transform based approach is computationally expensive and inaccurate for inexact tandem repeat identification, especially where it occurs due to insertion and deletion operations in DNA sequences. Furthermore, both DFT and PT techniques for the case of inexact repeats cannot clearly ascertain whether a pattern is due to period 'P' or its multiple. The proposed algorithm applies a novel periodicity measure based on orthogonal exactly periodic subspace decomposition (EPSD) technique. The algorithm is based on the concept of identifying local periods in the input signal and is robust in identifying inexact and hidden repeat patterns which otherwise are very difficult to detect. The EPSD measure also resolves the problems that were present in previous signal processing based approaches. The time complexity of the algorithm is $O(NL_w \log_2 L_w)$, where N is the length of the DNA sequence and L_w is the window length for identifying repeats. To demonstrate the capabilities of the algorithm, experiments were performed on artificially generated DNA sequences and actual DNA sequences covering both exact and inexact repeats.

Chapter four of the thesis addresses the problem of identifying exact and inexact inverted repeats present in DNA sequences. Fast correlation and periodicity measure based algorithms are presented in this chapter for identifying both exact and inexact IRs. Inverted repeats (IRs) are widespread in both prokaryotic and eukaryotic genomes, and have been associated with a large number of possible functions. Identification of inverted repeats and especially inexact inverted repeats in a DNA sequence has remained one of the challenging problems in DNA sequence analysis. Most of the existing methods for inverted repeat identification are either very difficult to handle, as they require a large number of input parameters or are inefficient in identifying inexact inverted repeats. Also, till date no signal processing algorithm exists for identifying IRs. The algorithms require the user to input only two easily understood parameters: maximum inverted repeat size and minimum length of contiguous repeat. This makes IRs identification job easier for the users, especially for biologists. The algorithm is evaluated by performing experiment on biological dataset download from NCBI website. The obtained results are compared with standard tool available online and the results show the effectiveness of the proposed approach. In Chapter five the correlation based approach is extended for RNA secondary structure prediction problem after modification in merging algorithm for IR detection.

Chapter six of the thesis deals with the problem of identifying protein coding patterns and presents a novel pattern recognition framework based on wavelet variance features for identifying protein coding DNA patterns. The identification task is a very challenging because there is no specific criterion based on which every coding and non-coding pattern can be identified. Currently, the most accurate identification techniques are based on linear/slope model of Z-curve components. However, the linear model provides a poor approximation for highly non-linear Z-curve components. In addition, the slope based techniques ignore the local statistical information present in DNA sequences which are important for identifying small coding patterns. The existing signal processing methods are based on only period-3 feature of protein coding region. In the proposed approach a wavelet based time series analysis technique has been applied for extracting coding feature from Z-curve components. Till now, wavelets have never been applied in identification of coding patterns. Also, pattern recognition approach has not been explored for identification of coding DNA sequences. The wavelet coefficient provides both local and global information contents of DNA sequences. The proposed approach provided a 10-fold cross-validation accuracy of more than 93% on recall patterns of *Yeast* genome. Furthermore, a combined feature vector (i.e., slope and wavelet variance features) based SVM classification is also proposed. The combined feature vector provided a 10-fold cross-validation accuracy of 96% for recall patterns of *Yeast* genome and more than 96% recall pattern accuracy for the *E. coli* genome.

Chapter seven of the thesis presents the development of a novel feature vector for efficient identification of G-protein-coupled receptors (GPCRs), GPCRs families, subfamilies and sub-subfamilies using SVM. GPCRs are one of the largest groups of proteins in vertebrate species. Their classification and functional annotation are very important in present medical and pharmaceutical research because GPCRs play key roles in many diseases. The large dimension of feature vector for the existing popular SVM based technique (SVMpred) makes the classification task quite expensive in terms of computational and memory used. The proposed feature vector is based on wavelet variance of seven important physicochemical properties of amino acids. Furthermore, the dimension of the proposed feature vector is also reduced to 35. This helps in building faster and memory efficient classifier which can be implemented on any normal desktop computer. The technique classifies GPCRs and non-GPCRs using a 5-fold cross-validation with accuracy and Matthews correlation coefficient (MCC) of 99.9% and 0.998 respectively. The technique is further able to detect major classes or families, subfamilies and sub-subfamilies of GPCRs with a total accuracy of 97.63%, 96.64% and 93.38% respectively. In addition, the technique classifies the human GPCRs with accuracy and MCC of 99.88% and 0.998 respectively. Finally in chapter eight, the contributions made in the thesis are summarized and scope of future work is outlined.

Acknowledgements

First and foremost, I would like to thank my supervisor Dr. Ankush Mittal who has been a source of inspiration for me. His invaluable comments during the whole work with this dissertation and the encouraging way in which he guided me to a deeper level of understanding of problems has enriched my learning abilities. He was always there to meet and brainstorm my ideas that helped me to think through my problems. He taught me how to write academic papers and had confidence in me when I doubted myself. I also thank him for providing facilities for my research work.

My special thanks to my co-supervisor, Prof. Kuldip Singh and my research committee members Prof. R. C. Joshi, Dr. Manoj Misra and Dr. (Mrs.) Sunita Gakkhar for their valuable advice.

I am grateful to my parents who in spite of several hardships supported my education, provided me unconditional support and encouragement to pursue my interests and believed in me. I also thank my brother who always encouraged me to study and sacrificed so many things for me. I am grateful to my sister and brother-in-law for their love, motivation and guidance towards my studies.

I am also grateful to many teachers in the past, especially Dr. M. K. Das (University of Delhi) who advised me to pursue research work after my post-graduation. Dr. M. K. Das is a modest, soft spoken person and an excellent teacher. His encouragement and moral support for students is truly commendable.

I am indebted to my cousin, Deepender Kumar who has played a significant role in my life. He has guided me from the days of school to my post-graduation level. My special thanks to Vikash Ranjan and Anjit Choudhury for the inspiration they have provided to me.

Thanks to my friends, Basudeb, Rohit, Rishi, Bikas Sahoo, Zahid, Rajesh Roshan Dash, Navin, Rakesh Roshan Dash, J. S. Chadchan and Praveen for their motivation, for healthy and interesting discussions and for helping me at several difficult moments. I thank Rohit for introducing me the concepts of machine learning. I am also grateful to other colleagues of my department for providing a good working atmosphere.

I am grateful to my special friends Arijit, Sameer and Vipul for their help, motivation, and moral support before and during my research work. Vipul helped me in deciding my area of research for my dissertation and also provided some books from IIT Bombay library for my research work. I owe a lot to all three of them. I am also grateful to members of Dhakkans group – Roopali, Anushree, Mitika, Bhavna and Anju for their support. I also thank parents of Arijit and Vipul who encouraged me to do research. I am also grateful to my friends of Bersarai.

I am grateful to Basudeb, Jayashree and Palli Mishra for their motivation, love and delicious food in the campus.

I am also very grateful to everyone who has read parts of the manuscript, especially Rohit and Basudeb.

I am also thankful to my department for providing computing and other important facilities for my research work. I am also grateful to the Indian Institute of Technology Roorkee for providing access to several research journals.

Last but not the least, I thank Ministry of Human Resource Development (MHRD), Govt. of India for providing me financial support during my dissertation work.

Ravi Gupta

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The entire genetic information of any living organism is written as linear information within DNA sequences and is coded by four different nucleotides. DNA molecules serve as backup for genetic information for the whole organism. The particular well defined fragments of this information, so-called coding sequences, are then translated, using complex molecular mechanisms, into other linear information then contained with protein sequences and coded with 20 different amino acids.

One of the significant genomic achievements in recent times has been the development of fast methods for sequencing genes and proteins. This has enabled the creation of large databases which can be processed by considering sequences of nucleic acids (DNA, RNA) and amino acids (proteins) as strings of characters. Today there exists over 65 billion nucleotides from more than 61 million individual sequences in GenBank [1](see Figure 1.1). On an average, these databases double in size every 18 months. Analysis and interpretation of the extremely large size of biomolecular sequences are some of the most exciting challenges faced by the scientific community.

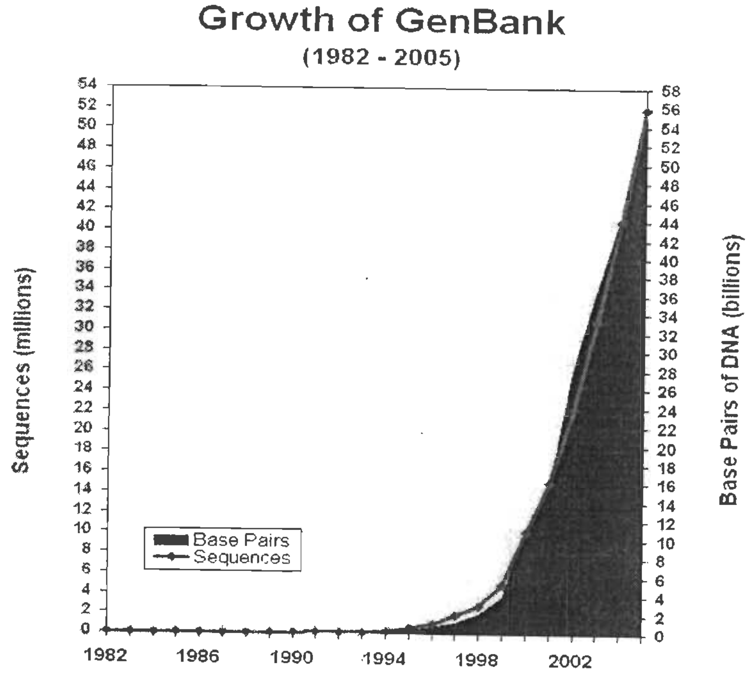


Figure 1.1: Growth of entries and nucleotides in GenBank from 1982–2005.

Signal processing techniques offer a great promise in analyzing genomic data. Genomic information is expressed digitally in nature much like as information is encoded digitally in computers as strings of zeros and ones. For example, DNA sequences are encoded by four nitrogenated bases: Adenine, Thymine, Cytosine, and Guanine. Similarly, protein molecules are encoded by twenty types of amino acids. Both DNA and protein molecules can be mathematically represented by character strings. The character string can be properly mapped into one or more numerical sequences, and in this way signal processing techniques provide a set of novel and useful tools for solving highly relevant problems of genomics.

Signal processing has played an important role in the area of sequence analysis (DNA and proteins) and DNA microarray analysis. The sequence analysis techniques have been applied to reveal some hidden structures, for sequence comparison and classification, to distinguish coding from non-coding regions in DNA sequences, and in genotype to phenotype mapping and several other problems. DNA microarray analysis technique helps in the monitoring of gene expression for tens of thousands of gene simultaneously. It has found applications in the medical and biological fields such as gene discovery, disease diagnosis, drug discovery, and toxicological researches. For the current study the signal processing techniques for sequence analysis is explored.

Digital signal processing (DSP) techniques offer more efficient ways to identify regions of the DNA exhibiting periodic behaviour. In [2, 3, 4], digital filters are applied to extract the period-3 component (the protein-coding regions of DNA demonstrate a period-3 performance due to codon structures. A codon is a sequence of three adjacent nucleotides constituting the genetic code that determines the insertion of a specific amino acid in a polypeptide chain during protein synthesis or the signal to stop/start protein synthesis). In [5, 6, 7], sliding window based discrete Fourier transform (DFT) technique was proposed for identification of coding DNA sequences. DFT based algorithms are also used to identify tandem repeat patterns from a DNA sequence [8, 9]. In [10], DSP techniques for predicting potential promoter are proposed.

Wavelet analysis provides a useful SP approach for the visual description of inherent structure underlying DNA sequences. In [11], cross-correlation of wavelet coefficients is used for protein sequence comparison. Lio and Vannucci [12], applied discrete wavelet transform (DWT) to find pathogenicity islands and gene mutation events in genome data. Wavelet is also used [13] to search the DNA sequence construction rules. The salient spots in the final two-dimensional (2-D) analysis results revealed significant features in the DNA sequence. Their results demonstrated that while the non-coding sequences showed spectra similar to those from random sequences, coding sequences revealed specific periodicities of variable length and a common periodicity of three. In [14], the authors discuss the use of the continuous wavelet transform (CWT) and the resonant recognition model (RRM) to predict the location of oncogene protein active sites and to gain insight into their structures and functions. Arneodo *et al.* [15] used wavelet transform modulus maxima (WTMM) to analyze the fractal scaling properties in DNA sequences. They demonstrated the existence of long-range correlation in genes containing introns and non-coding regions, and also quantified that correlation. They also found that the fluctuations in the DNA walk profiles were homogeneous with Gaussian statistics. This result reveals useful information about the role of introns and non-coding intergenic regions in the non-equilibrium dynamic process that produced DNA sequences.

1.1 Research Problems

The aim of the present research work is to contribute towards furnishing novel signal processing measures and features for analysis of DNA, RNA and protein sequences. Furthermore, supervised machine learning technique is proposed for identification of DNA and protein sequence patterns based on features extracted using signal processing techniques. The challenges inherent in the development of measures and features applying signal processing include:

- Forming a close relationship between sequence analysis issues and signal processing problems.
- Providing a suitable mapping technique for converting symbolic DNA, RNA and protein sequences into numeric sequences.

- Selection of an appropriate signal processing tool either for calculation of measures or for extraction of features from mapped biomolecular sequence data.
- Dealing with huge genomic data in an efficient and effective manner.

The sequence analysis problems that were taken up for the current research work are:

- Detection of exact and inexact tandem from DNA sequences.
- Identification of inverted repetitive patterns from DNA sequences.
- Prediction of RNA secondary structure from its primary sequence.
- Identification of protein coding DNA sequence patterns.
- Recognition of G-coupled protein receptor protein sequences and classification of GPCRs into its family, subfamilies and sub-subfamilies.

1.2 Framework of the Research

A general strategy was followed for solving the selected research problems. The steps are as follows:

1. Understand the issues related to the problem and identify the limitations of the exiting solutions.
2. Visualize the problems in terms of signal processing and pattern recognition task.
3. Identify the DSP and machine intelligence techniques for calculating measures or features.

4. Propose a suitable mapping technique for DNA, RNA or protein sequences.

5. Present a technique that implements DSP and pattern recognition concepts and solves the issues related to the problem.

6. Evaluate the performance of the proposed technique on actual biological datasets and compare the results with the existing techniques.

The complete framework of our research work is presented in Figure 1.2. The various tasks performed for solving our research problems are: acquisition of datasets, mapping of genomics and proteomics datasets into numeric sequences, processing of mapped datasets using DSP techniques and calculation of measures or features, interpretation of measures and annotation of biological data, application of machine intelligence algorithms to extracted features and identification of pattern class. Depending on the task to be performed the complete framework is divided into different modules.

The proposed methodology for the sequence analysis problems of the thesis can be considered as 3-steps procedure: mapping, processing and analysis/classification. The data acquisition module is a pre-processing step which helps in constructing datasets for analysis purpose. The mapping module helps in converting a symbolic sequence (input) into numeric sequences (output). A mapping function is selected depending upon the formulations of the problem. An arbitrary selection of mapping function may lead to incorrect or no result. After obtaining a numeric sequence, an appropriate signal processing technique is applied to the input data and the desired information is extracted either in the form of measures or feature vector. The third step consists of two modules: annotation and classification. The tandem repeat identification, inverted repeat identification and RNA secondary structure prediction problem move through annotation module. However, the pattern recognition research problems: identification of protein coding regions and GPCRs identification and classification move through the classification modules. The second step, i.e., application of signal processing on mapped data acts as a feature extraction module from pattern recognition point of view. The feature vectors are processed using a classification algorithm for identification of patterns.

1.3 Organization of the Thesis

The thesis is organized as follows. Chapter 2 presents a brief introduction to signal processing tools and an overview of signal processing contributions to several sequence analysis problems. A novel signal processing measure for identification of tandem and other hidden patterns is presented in chapter 3. Chapter 4 presents a correlation based framework for identification of inverted repeat. Chapter 5 presents a correlation based framework for prediction of RNA secondary structure from a primary RNA sequence. A pattern recognition approach for identification of coding and non-coding patterns is proposed in chapter 6. Chapter 7 presents an efficient wavelet based features extraction and SVM classification technique for GPCRs recognition and identification of GPCR families and subfamilies. Chapter 8 concludes our thesis by presenting our contribution and future research directions.

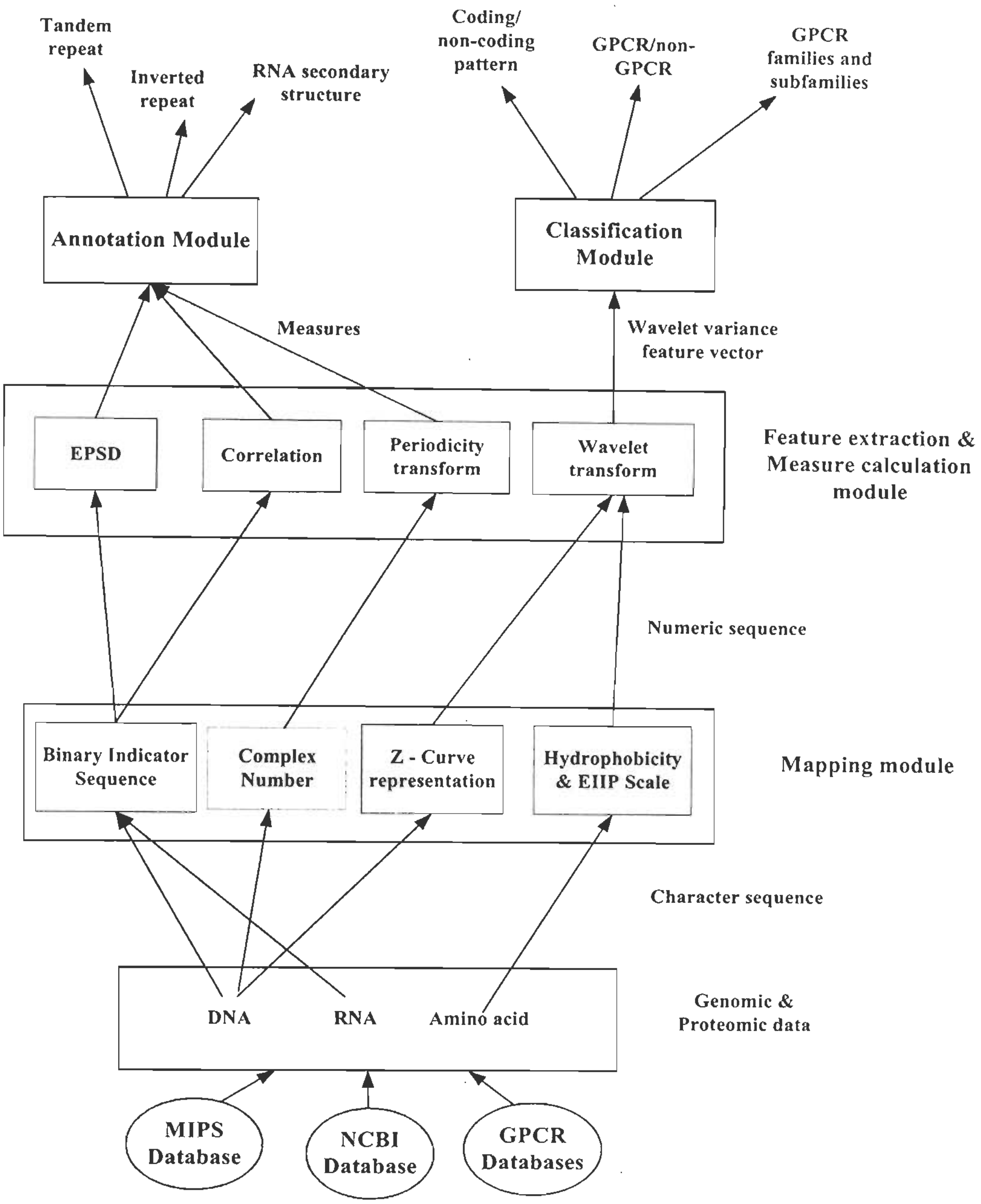


Figure 1.2: Framework of the proposed research work.

Chapter 2

Review and Preliminaries

With the enormous amount of genomic and proteomic data available in the public domain, it is becoming increasingly important to be able to analyze the data and interpret the results in a biologically meaningful manner. The need goes far beyond database management, which is still essential for the organization and easy access to the huge quantity of data, to the necessity of deciphering the connections between the genomic data and the biological functionality of living cells and organisms. This chapter presents a brief review on contributions of signal processing (SP) techniques for sequence analysis problems of bioinformatics.

2.1 SP Tools for Sequence Analysis

In this section, a brief introduction to some transformation techniques of signal processing useful for sequence analysis is provided.

2.1.1 Transforms

A transform is a special type of function. To understand transforms, one must first understand the concept of a function. A function is a relationship between two sets (called the domain and codomain), and this relationship must satisfy two conditions. First, every element in the domain must correspond to some element in the codomain. Second, no two elements in the codomain can correspond to the same element in the domain. These two requirements can be combined by saying that for each element in the domain there corresponds exactly one element in the codomain. Transforms are special type of functions whose domain and the codomain contains frequency functions.

2.1.2 Discrete Fourier transform

Let $\{a_t\} = \{a_t : t = 0, \ldots, N-1\}$ is a sequence of N real or complex valued variables. The discrete Fourier transform (DFT) is the sequence $\{A_k\}$ of N variable given by

$$A_k \equiv \sum_{t=0}^{N-1} a_t \exp^{-j2\pi tk/N}, \quad k = 0, \ldots, N-1. \tag{2.1}$$

where A_k is associated with frequency $f_k \equiv k/N$.

$\{a_t\}$ can reconstructed from its DFT $\{A_k\}$ using the following equation

$$a_t = \frac{1}{N} \sum_{k=0}^{N-1} A_k \exp^{j2\pi tk/N}, \quad t = 0, \ldots, N-1. \tag{2.2}$$

Since we can reconstruct $\{a_t\}$ from its DFT, these two sequences $\{a_t\}$ and $\{A_k\}$ can be considered representations of common mathematical entity.

2.1.3 Discrete wavelet transform

Wavelets are becoming increasingly popular in different areas of applied and theoretical science. Data compression, signal processing, turbulence, geophysics, statistics, numerical analysis and bioinformatics are only few examples from a long list of disciplines in which wavelet have been successfully employed.

Discrete wavelet transform (DWT) is an orthonormal transform. Let X_0, X_1, ..., X_{N-1} represent a time series of N real-valued variable. If $\{W_n : n = 0, \ldots, N-1\}$ represent the DWT coefficients, then we can write $\mathbf{W} = W\mathbf{X}$, where $\mathbf{W}$ is a column vector of length $N = 2^J$ whose nth element is the nth DWT coefficient W_n, and W is an NxN real-valued matrix defining the DWT and satisfying $W^T W = I_N$ (the condition that the length of $\mathbf{X}$ be a power of two is restrictive). Orthonormality of DWT implies that $\mathbf{X} = W^T\mathbf{W}$ and $\| \mathbf{W}\|^2 = \| \mathbf{X}\|^2$. Hence W_n^2 represents the contribution to the energy attributable to the DWT coefficient with index n.

In DFT coefficients are associated with frequencies, the nth wavelet coefficients W_n is associated with a particular set of times. The elements of the vector $\mathbf{W}$ can be decomposed into $J+1$ subvectors, where $J = \log_2 N$. The first J subvectors are denoted by $\mathbf{W}_j$, $j = 1, \ldots, J$, and the jth such subvector contains all of the DWT coefficients for scale τ_j. Note that $\mathbf{W}_j$ is a column vector with $N/2^j$ elements. The final subvector is denoted as $\mathbf{V}_J$ and contains just the scaling coefficient W_{N-1}. The wavelet transform can be written as

$$\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_J, \mathbf{V}_J]^T \tag{2.3}$$

The energy preservation for DWT can be written as:

$$\| \mathbf{X}\|^2 = \| \mathbf{W}\|^2 = \sum_{j=1}^{J} \| \mathbf{W}_j\|^2 + \| \mathbf{V}_J\|^2 \tag{2.4}$$

Let $D_j \equiv \mathcal{W}_j^T \mathbf{W}_j$ for $j = 1, \ldots, J$, which is an N dimensional column vector whose elements are associated with changes in $\mathbf{X}$ at scale τ_j; i.e., $\mathbf{W}_j = \mathcal{W}_j \mathbf{W}_j$ represents the portion of the analysis $\mathbf{W} = \mathcal{W}\mathbf{X}$ attributable to scale τ_j, while $\mathcal{W}_j^T \mathbf{W}_j$ is the portion of the synthesis $\mathbf{X} = \mathcal{W}^T \mathbf{W}$ attributable to scale τ_j. Let $S_J \equiv \mathcal{V}_J^T \mathbf{V}_J$, which has all of its elements equal to the sample mean $\overline{X}$. The input time series $\mathbf{X}$ can be written as follow:

$$\mathbf{X} = \sum_{j=1}^{J} D_j + S_J \tag{2.5}$$

which defines a *multiresolution analysis* (MRA) of $\mathbf{X}$, i.e., we express the series $\mathbf{X}$ as the sum of a constant vector S_J and J other vectors D_j, $j = 1, \ldots, J$, each of which contains a time series related to variations in $\mathbf{X}$ at a certain scale. And D_j is referred as the jth level *wavelet detail.*

2.2 SP Techniques for Sequence Analysis

In this section, signal processing techniques for sequence analysis issues in bioinformatics are briefly discussed. The list is not exhaustive but covers several key problems of bioinformatics.

2.2.1 Identification of protein-coding regions

It has been noticed that protein-coding regions (exons) in gene have a period-3 component because of coding biases in the translation of codons into amino acids. This observation can be traced back to the 1980 work of Trifonov and Sussman [16]. The period-3 property is absent outside exons, and hence can be exploited to locate exons. Automatic identification of protein coding regions in genomic DNA sequences is a fundamental step in the computational annotation of genes and is one of the central issues in bioinformatics.

Digital signal processing (DSP) techniques for identification of coding regions (exons) in DNA sequences include the application of the DFT on overlapping windows [5, 6, 7] and the application of bandpass digital filters that are centered at $2\pi/3$ [2, 3, 4]. The output of the bandpass digital filter at $2\pi/3$ can be thought of as one measure of the DNA spectral content at frequency $2\pi/3$. A computer program GeneScan, based on concept of DFT for period-3 identification was developed by Tiwari *et al.* [7]. It locates coding open reading frames and exonic regions in genomic sequences. In [17], a new measure based on the arguments of DFT is presented and is very useful in locating short genes and exons. A fast DFT based gene prediction algorithm is proposed by Datta and Asif in [18] for identifying protein coding regions. The authors also provide theoretical justification for the period-3 property by defining a new parameter referred to as the position count function (PCF), which measures the number of times different nucleotides appear in the three phases with a DNA codon.

2.2.2 Identification of non-coding regions

The period-3 property indicates a strong short-term correlation in the coding regions but there is also a long-range correlation exhibited by DNA sequences both in gene and intergenetic regions. One of the earliest papers to point this out appeared in *Nature* in 1992 [19]. The study made is based on a concept of DNA *walk* [19, 20]. Later studies by other authors [21, 22, 23, 24, 25] examined correlations over much longer regions which contained many genes. Long range correlations have been found both in coding and noncoding regions. According to Fourier transform theory, long range correlation implies that the Fourier transform has $1/f$-behavior in low frequency region [26]. In [27], Voss demonstrated that the power spectrum has power-law of $1/f^{\beta}$ behavior for the human Cytomegalovirus strain AD169. Later studies have indicated that such long range correlation is valid even further, extending to several millions of bases [28]. Li [29] has written a comprehensive review paper on this topic, and has also observed that the $1/f$ behavior in natural phenomenon can be traced to the so-called duplication-mutation model [30].

2.2.3 Identification of hidden structures

Predicting and detecting the underlying structural patterns accurately in a DNA sequence is a difficult problem for researchers. Traditional techniques for structure detection were based on calculating average base composition in a DNA sequence for a fixed window size. However, by applying multiresolution analysis concept of wavelet transform, the problem of deciding window size is resolved. The wavelet transform allows efficient extraction of basic components at different scales. In [12], discrete wavelet transform (DWT) is applied to find pathogenicity islands and gene mutation events in genome data. DWT is used to smooth G+C profiles to locate characteristics patterns in a genome sequences, and a wavelet scalogram is used for sequence profile comparison. In [31], a wavelet change-point (WCP) technique is used to predict the location and topology of transmembrane helix (HTM) in a primary amino acid sequences. Wavelet is applied to decompose the propensity profile based on frequency of the residues in HTM segments, into wavelet coefficients. Later on, a data dependent threshold is used to select wavelet coefficients that detect abrupt changes in the profile. The reported result was comparable with other methods, such as HMM and NN architectures which are computationally much more complex than WCP technique. In [32], a non-decimated wavelet transform and wavelet variance, correlation scale-by-scale decompositions, based approach is applied to determine the location of HTM and

G+C regions in genome sequences. Another method based on wavelet transform has been proposed in [33]. The proposed approach combines wavelet multiresolution analysis and the cumulative GC profile to precisely identify the boundaries of isochores in the human genome. A novel approach based on IIR lowpass filters for detecting CpG islands in a genomic sequence is presented in [34]. In this research, a Markov chain model as elaborated in [35] is coupled with IIR lowpass filters to identify CpG islands. The proposed approach is very simple and capable of identifying CpG islands efficiently at a low computational expense.

2.2.4 Sequence similarity

Protein sequence comparison and alignment techniques are one of the most important and widely used methods for protein sequence analysis. The aim of protein comparison and alignment is to find the similarities or differences between two or more protein sequences. These comparative techniques have provided new insight into the structure-function relationships of the active site of a protein. Many algorithms such as BLAST [36, 37], PSI-BLAST [38] and FASTA [39] have been developed based on character based similarity, though differing in approaches. The concept of similarity for those approaches only means how many identical pairs of amino acids exist for the query sequence and the subject sequence. These algorithms fail to extract subsequences that are not identical in characters but show similarities in their physicochemical properties, tertiary structure, resonance recognition model (RRM) spectra and biological functions [40].

Signal processing provides a non character based approach to establish similarity between protein sequences. In [11], wavelet analysis is used to extract characteristic bands from protein sequences. In this work, the sequence-scale analysis with wavelet gives a multiresolution similarity comparison between protein sequences. This similarity expands the traditional sequence similarity concept which take into account only the local pair-wise amino acid and disregards the information contained in coarser spatial resolution. In addition, this wavelet-based approach does not require the complex sequence alignment processing for sequences. In [41], a technique based on spectral similarity is proposed to compare subsequences of amino acids that show similarly. This approach finds a similarity score between sequences based on any given attribute, like hydrophobicity of amino acid on the basis of spectral information after partial conversion to the frequency domain. Other than sequence comparison, sequence classification is also a major problem in DNA sequence analysis.

A Fourier transform-based support vector machine approach is presented in [42] for predicting and classifying GPCRs subfamilies. This method couples fast Fourier transform (FFT) with SVM on the basis of the hydrophobicity profile of the amino acid sequences. A wavelet packet (WP) technique is applied in [43] for DNA sequence classification, i.e., to classify exons (a segment of DNA that is transcribed to RNA and specifies a portion of a protein) and introns (noncoding subregions in genes). The wavelet coefficients is later on used as a criteria for sequence classification.

2.2.5 Identification of tandem repeats

SP solutions to tandem repeat pattern identification problem include the application of discrete Fourier transform (DFT) [8, 9] and the application of short-time periodicity transform (STPT) [44]. In [8], DFT is used as a pre-processing tool for identifying the significant periodic regions through a sliding window analysis, and then an exact search method is used for finding the repetitive units. In [9], instead of a product spectrum, a sum spectrum is proposed as a measure for identifying repeats. The product spectrum is especially sensitive to the presence of inexact repeats. A STPT based approach for finding tandem repeats in DNA sequence is presented in [44]. Details of SP algorithms for tandem repeat identification are discussed in chapter 3.

2.2.6 Protein function prediction

Prediction of protein function from its 3-D structure is a big problem in bioinformatics. A relational map between protein structure and function is considered as the third genetic code (the relationship between amino acid sequences and the 3-D structure of proteins is thought to be as the second genetic code). If biologists could predict the action of a protein by looking at its 3-D structure, they would have an increased chance of designing more effective drugs. Furthermore, if they could solve the sequence-structure problem, they would be able to make those structural modification exactly. Unfortunately, this still remains a dream for now, as we do not really know how to interpret protein functions, both in theory and in practice.

In [14], researchers have investigated the oncogene functional group using digital signal analysis methods, in particular Fourier transform and continuous wavelet transform (CWT). They incorporated the continuous wavelet transform (CWT) into the RRM [40] to predict the active sites for a chosen protein example. The RRM is a novel physico-mathematical approach established to analyze the interaction between a protein and its target. The RRM assumes that the specificity of protein interaction is governed by the resonant electromagnetic energy transfer at the specific frequency for each interaction. RRM was also known as ISM (information spectrum method). ISM main interest is in giving each amino acid a set of electron-ion interaction potentials and comparing the frequency one obtains from a chain of these amino acids. This is mostly an information-domain approach that assumes that proteins can only interact if they share a peak in this frequency space through which energy can be transferred. RRM now includes not only resonant energy, which happens at very small distances, but also longer range interactions. One of the main applications of this model is to predict the location of a protein's biological active site(s) using DSP. The results provided a new insight into the structure-function relationships of the analyzed oncogene protein family.

2.2.7 Visualization of sequences

Data visualization is a key challenge in bioinformatics, especially so with the increasing number of complex genomes that are currently being sequenced. One motivation among many, that requires such analysis techniques deals with studying the fractal behavior of DNA sequences. In [45, 46], DNA walk representation and Gauss-wavelet-based method is used for visualization and analysis of DNA sequences. DNA walk representation allows one to graphically visualize how DNA sequences evolve. Gaussian-wavelet-based analysis is used for locating periodicity and extracting structural information from a complex-valued DNA walk. In [47], a correlation function is proposed to compare each base in a DNA sequence to its various neighbours and which is subsequently processed by Fourier and wavelet transform (Walsh-Hadamard).

Chapter 3

Identification of Exact and Inexact Tandem Repeat Patterns

The identification and analysis of repetitive patterns are active areas of biological and computational research. Tandem repeats in telomeres play a role in cancer and hyper-variable tri-nucleotide tandem repeats are linked to over a dozen major neurodegenerative genetic disorders. Despite their importance, locating and characterizing these repeats within anonymous DNA sequences remain challenging tasks. The difficulty is due to presence of imperfect and complex repeat patterns.

In this chapter, an application of signal processing technique for identification of exact and inexact tandem repeat patterns in DNA sequences is presented. The motivation for developing a signal processing approach for the current problem comes from similarity between period detection problem in signal processing and tandem repeat identification problem. The algorithm proposed in this chapter applies a novel periodicity measure based on orthogonal exactly periodic subspace decomposition technique. The measure resolves the problems like whether the repeat pattern is of period P or its multiple (i.e., $2P$, $3P$ and so on) and several other problems that are present in previous signal processing based approaches. The time complexity of the algorithm is $O(NL_w \log L_w)$, where N is the length of DNA sequence and L_w is the window length for identifying repeats. To demonstrate the capabilities of the algorithm experiments were performed on pseudo DNA, i.e., artificially generated and actual DNA sequences covering both exact and inexact repeats.

3.1 Introduction

A direct or tandem repeat is the same pattern recurring on the same strand in the same nucleotide order, e.g., TGAC recurs as TGAC. Tandem repeats play significant structural and functional roles in DNA. They occur in abundance in structural areas such as telomeres, centromeres and histone binding regions [48]. They also play a regulatory role near genes and perhaps even within genes. Over a dozen of major human degenerative diseases [49, 50, 51, 52] are associated with dramatic increase in the number of copies of a trinucleotide pattern and are listed in Table 3.1. In afflicted individuals, the repeat number has been amplified from the normal range of tens of copies to hundred or thousands, resulting in the disease. It has been suggested that the repeats themselves produce unusual physical structures in the DNA causing polymerase slippage and the resulting amplification. Cancer is also correlated to regions containing tandem repeats [48]. Short tandem repeats are used as a convenient tool for genetic profiling of individuals [53]. Thus, identification and analysis of repetitive DNA are active areas of biological and computational research.

The main objectives of tandem repeat pattern identification algorithms are to identify its periodicity, its pattern structure, its location and its copy number.

The algorithmic challenges for identification problem are: lack of prior knowledge regarding the composition of the repeat pattern and presence of inexact and hidden repeats. Inexact repeats, formed due to mutations of exact repeat, are thought to be representation of historical events associated with sequence. Thus, it is important for any repetitive pattern identification algorithm to identify inexact in addition to exact repeat structures in a DNA sequence.

In past, several algorithms and measures based on heuristic, combinatorial, dynamic programming [54, 55, 56, 57, 58, 59] and SP approaches [8, 9, 44] have been proposed for finding tandem repeat structure in DNA sequences. In this chapter, a novel signal processing (SP) approach for identifying exact and inexact tandem repeats in DNA sequences is presented. An exactly periodic subspace decomposition (EPSD) [60] based measure for identifying repeats is proposed. EPSD technique, unlike the Fourier transform, is obtained by taking projection onto exactly periodic orthogonal multidimensional subspaces. By having subspaces of dimensions larger than one, the exactly periodic subspace (EPS) [60] can better capture, in one coefficient, the periodic energy than the Fourier transform. Hence, the new measure is more sensitive than previous techniques for identifying repeats. In addition to identification of exact repeats, the proposed measure is useful in identifying inexact and other hidden repeat patterns unannotated by GenBank database. The EPSD based approach also helps in identifying whether a particular pattern is due to period P or its multiple. Thus the ambiguity that is present in [8, 9, 44] is taken care by the proposed algorithm.

In section 3.3.2, an analogy between tandem repeat and periodicity in a signal is discussed. As with other signal processing approaches, the proposed approach strictly deals with numeric sequences. In section 3.3.1, a numeric mapping for

Table 3.1: Human diseases caused by expansion of simple DNA repeats.

Disease	Gene	Pattern	Repeat number [51]	
			Normal	Affected
Fragile X syndrome	FMR1	CGG	< 50	> 200
Fragile X-E mental retardation	FMR2	CCG	< 35	> 200
Myotonic dystrophy	DMPK	CTG	< 35	> 50
Spinocerebellar ataxia type 8	SCA8	CTG	< 40	> 110
Friedreich's ataxia	X25	GAA	< 35	> 100
Spinobulbar muscular atrophy	AR	CAG	< 40	> 40
Huntington's disease	IT15	CAG	< 40	> 40
Dentatorubralpallidoluysian atrophy	DRPLA	CTG	< 35	> 50
Spinocerebellar ataxia type 1	SCA1	CAG	< 40	> 40
Spinocerebellar ataxia type 2	SCA2	CAG	< 30	> 35
Spinocerebellar ataxia type 3	SCA3	CAG	< 40	> 40
Spinocerebellar ataxia type 6	CACNA1A	CAG	< 20	> 20
Spinocerebellar ataxia type 7	SCA7	CAG	< 20	> 40

nucleotides (DNA bases) is presented. The need for an alternate signal approach other than Fourier [8] and STPT [44] methods for identifying inexact repeats is explained in section 3.3.2. The theory of EPSD technique and its modification for the current problem is discussed in section 3.3.3. Further, the proposed tandem repeat algorithm is given in section 3.4. Experiments were performed on pseudo DNA sequences and several actual DNA sequences to verify the effectiveness of the novel approach and are given in section 3.5.

3.2 Literature Review

This section presents a brief review on existing methods for identification of tandem repeats.

3.2.1 Non-signal processing approaches

Benson [54, 61] proposed a heuristic program (i.e., Tandem repeat finder) for finding tandem repeats. The algorithm does not require specification of pattern or pattern size and is based on a stochastic model of tandem repeats, rather than some minimal alignment score. The program has detection and analysis components. The detection component uses a set of statistically based criteria to find candidate tandem repeats. The analysis component attempts to produce an alignment for each candidate and if successful, gathers a number of statistics about the alignment (such as percent, identity, percent indels) and the nucleotide sequence (such as composition, entropy measure). The stand-alone version of tandem repeat finder (TRF) is available free of charge for non-commercial purpose at *http://tandem.bu.edu/trf/trf.download.html* and an online version of TRF can be accessed at (*http://tandem.bu.edu/trf/trf.submit.options.html*).

Kurtz *et al.* [62, 63, 55] proposed *REPuter* program family for identification of tandem. It implements an efficient and compact implementation of suffix trees in order to locate exact repeats in linear space and time. These exact repeats are then used as seeds for identifying inexact repeats by allowing mismatches, insertions, and deletions. The program is not heuristic and guarantees to find all inexact repeats as specified by the parameters. Output size is controlled via parameters for minimum length and maximum error. Output is sorted by significance score (E-values) calculated according to the distance model used. The stand-alone version of *REPuter* is available free of charge for non-commercial purpose at *http://www.genomes.de/download.html*. An online version of *REPuter* providing some basic functionality can be used on the Bielefeld Bioinformatics web server (*http://bibiserv.techfak.uni-bielefeld.de/reputer/submission.html*).

Kolpakov *et al.* in [56] proposed a combined combinatorial/heuristic paradigm algorithmic approach for finding approximate tandem repeats and a software program, named *mreps*. The program finds all approximate tandem arrays (under the Hamming distance model) that verify a certain combinatorial definition. The program is divided into two main stages. In the first stage, the program applies an efficient combinatorial algorithm for finding all repetitive structures of a certain kind in a given sequence. These structures serve as 'raw material' for the second stage, which applies to them an heuristic treatment consisting of several steps to obtain relevant repeats. The *mreps* program is open-source software and can be freely downloaded or queried through a web-based interface at *http://bioinfo.lifl.fr/mreps/*.

An algorithm to identify complex repeat pattern is proposed by Hauth and Joseph in [59, 64]. The algorithm involves both locating and characterizing repeat regions. It is divided into three major tasks: (1) isolate a tandem repeat by determining its period and its approximate sequence location, (2) determine the pattern associated with a region period and (3) characterize the region using the pattern. The proposed technique analyze k-length substrings in a DNA sequence by finding recurring distances between identical substring and is similar to Benson [54]. However, instead of using a statistical model for locating interesting periods, a simple and accurate filter technique is applied to determine repeat patterns.

Apart from the above discussed algorithms, there exists several other methods for identifying repeats. EQUICKTANDEM available in the EMBOSS package [65] is a simple statistically-based algorithm that identifies tandem repeat structures in DNA for each pattern size up to given bound. RepeatMasker (*http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker*) is a program which uses a database of known sequences and implements a string-matching algorithm to find copies of those repeats in a new sequence. A clustering method for analysis of the repetitive structure of genomic sequences is described in [66]. In [67], TROLL program for finding exact tandemly repeated copies of *priori* specified patterns is presented.

3.2.2 Signal processing approaches

SP based algorithms for identifying tandem repeats have their own advantages because of its sensitivity towards detection of inexact repeats and application of faster signal processing tool like DFT. These algorithms also provide an easy solution to biologist or non-computer experts. Unlike the non-SP approaches which require a number of error tolerances parameters like match, edit distance, Hamming distance and several other parameters which are very difficult to understand for any normal user, the SP based algorithms require mainly one parameter which acts as a threshold for identifying repeats.

In [44], authors have presented a novel signal processing algorithm (periodicity explorer (PE)) for identification of tandem repeats. The algorithm is based on the short-time periodicity transform (STPT) and is adapted from the periodicity transform [68] to provide a more localized measure of periodic content. The periodicity transform provides a method for detection of periodicities in finite-duration sequences. This transform decomposes a finite-duration sequence into a sum of periodic sequences where the decomposition is accomplished by projecting the sequence onto a set of periodic subspaces. These subspaces are not orthogonal leading to decompositions that are not unique. The PE algorithm is composed of three main components. The first component is the mapping of the nucleotide bases {A, C, T, G} into numerical values. The second is the primary processing component in which STPT is applied to the DNA sequence and periodogram is obtained. The last component consists of a search for repeat sequences and a repeat characterization based on the periodogram.

The PE algorithm has several shortcomings. The nucleotide mapping in [44] was taken as follows: A$= 1 + j$, C$= -1 + j$, G$= -1 - j$, and T$= 1 - j$, where $j = \sqrt{-1}$. Let the two DNA sequences be ACATACAC and ACAGACAC. The projection of the DNA sequences onto the periodic subspace P_2 (where P is the set of all periodic sequences) is given by $\{(1 + j), (-0.5 + 0.5j), (1 + j), (-0.5+0.5j), (1+j), (-0.5+0.5j), (1+j), (-0.5+0.5j)\}$ and $\{(1+j), (-1+0.5j), (1+j), (-1+0.5j), (1+j), (-1+0.5j), (1+j), (-1+0.5j)\}$ respectively. And the periodogram coefficient values obtained for projection of the DNA sequence on P_2 subspace are 0.75 and 0.895 respectively. By comparing the two DNA sequences, it is observed that even though the two DNA sequences have equal degree of 2-periodic component (differ just by one symbol from becoming ETR), the projection of DNA sequences are different and also the periodogram coefficient obtained are different. This shows that the periodogram coefficient cannot act a good estimator for measuring periodicity.

Sharma *et al.* [8] presented a Fourier transform (FT) based method (i.e., Spectral repeat finder (SRF)) for locating and identifying repetitive DNA and its constituent units. The method first identifies the length of the potential repeat unit present in a given DNA sequence by evaluating the power spectrum. Subsequently, the sequence is scanned at particular individual frequencies to locate the approximate region(s) where the repeat units are present. Potential seed patterns from these regions are then used to identify repeats through an exact method. The SRF algorithm is summarized in the following steps:

Step 1: Take a DNA sequence of length N as an input and map the DNA sequence into numbers using binary indicator mapping technique.

Step 2: Compute the power spectrum, $S(f)$, and the average power spectra $\overline{S}$ for the mapped DNA sequence.

Step 3: Identify all the peaks with $S(f_i)/\overline{S} > T$ (the threshold). After identifying the peaks, calculate the period of the repeat pattern, $p_i = 1/f_i$.

Step 4: For each identified peaks, compute $P_m(j) = S(f_i)/\overline{S}$ in a sliding window of length l centered on position j in the sequence. Regions containing a repeat of length p_i can be identified directly as those where $P_m(j)$ is greater than threshold (T).

Step 5: Since both the period of repeat pattern p_i, and its location are known, an exact method is used to identify the repeat units.

The power spectrum gets crowded when the length of the input DNA sequence is high and therefore assessing the significance of peak in the power spectra becomes difficult. Hence, the long DNA sequences (>15Kb) are divided into overlapping segments of length 10 Kb; each of these segments is analyzed individually for the presence of repetitive units. The time complexity of SRF algorithm is $O(n^2)$, where n is the length of the sequence. The threshold value (T) is fixed to 4 and the choice of window length and slide length is made heuristically.

In [9], a product spectrum of Fourier nucleotide subsequences is presented to detect hidden periodicity. The algorithm is summarized in the following steps:

Step 1: Convert the DNA sequence into four nucleotide subsequence $x_A[n]$, $x_C[n]$, $x_G[n]$, $x_T[n]$ using binary indicator mapping technique.

Step 2: Remove the mean value from mapped sequences and then take a normalized Fourier transform.

$$m_\alpha = \frac{1}{N}\sum_{n=0}^{N-1} x_\alpha[n] \tag{3.1}$$

where N is the length of the DNA sequence

$$S_\alpha(f) = \frac{1}{N}\sum_{n=0}^{N-1}(x_\alpha[n] - m_\alpha)e^{-j2\pi fn} \tag{3.2}$$

for $0 \leq f \leq 0.5$ and $\alpha \in \{\text{A, C, G, T}\}$.

Step 3: Calculate the Fourier product spectrum

$$S(f) = \prod_{\alpha\in\{\text{A, C, G, T}\}} (|S_\alpha(f)| + c) \tag{3.3}$$

where c is a small positive constant. The constant c is to prevent the nulling of $S(f)$ if a particular character is not part of the repeat pattern.

Step 4: The beginning and end location of the repeat of the tandem repeat region is identified by selecting a threshold value and using the Fourier product spectra (equation 3.3).

However, the existing SP approaches for tandem repeat identification have several disadvantages. All three existing approaches [8, 44, 9] cannot ascertain whether a repeat is due to period P or its multiple, i.e., $2P$, $3P$ and so on. For example, from Figure 3.1 one cannot tell whether the repeat period is of 21 or its multiple, i.e., 42, 63 and so on. In SRF [8], the power spectrum gets crowded when the length of the input DNA sequence is high and hence it becomes difficult to identify tandem repeats, especially inexact repeats. The PE algorithm is designed to be executed separately for every period because the periodicity transform provides non-orthogonal decomposition of the signal. This means that the run time of the PE algorithm is $O(NWP_{max})$ where N is the length of analyzed DNA sequence, W is the window size and P_{max} is the maximum period. Further, PE algorithm does not work well in identifying inexact repeats which occur due to insertions or deletions.

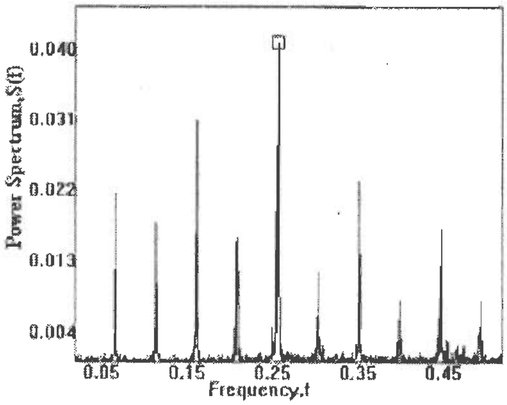


Figure 3.1: Power spectra provided by SRF program for an inexact repeat DNA sequence of period 21.

3.3 Mathematical Formulation of Tandem Repeat Pattern Identification

The standard representation of genomic information by sequences of nucleotide symbols in DNA, RNA or amino acids limits the processing of genomic information to pattern matching and statistical analysis. Providing mathematical representation to symbolic DNA sequences opens the possibility to apply signal processing techniques for the analysis of genomic data [5] and reveals features of genomes that would be difficult to obtain by using standard statistical and pattern matching techniques. The arbitrary assignment of a number to each symbol would impose a mathematical structure not present in the original data. Thus, a nucleotide mapping should be chosen such that it preserves the biological features and does not introduce any artifact into the mapped signal. For the proposed algorithm, a binary indicator sequence [7] representation for the DNA sequence is selected. This mapping helps in formulating the tandem repeat identification problem analogous to period detection in signal processing.

3.3.1 Numerical representation of DNA sequences

Consider a DNA sequence $S[n] = s_1 s_2 \dots s_L$ of length L, consisting of a sequence of a series of four nucleotides symbols A, C, G, T. The binary indicator sequences are obtained as follows:

$$S_\Omega[n] = \begin{cases} 1, & \text{if } S[n] = \Omega \text{ where } \Omega \in \sum(=\{\text{A,C,G,T}\}) \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

For example, the binary indicator sequences for GGCATACACAGACACGCC are given in Table 3.2.

Table 3.2: Binary indicator sequences for a random DNA sequence.

	G	G	C	A	T	A	C	A	C	A	G	A	C	A	C	G	C	C
S_A	0	0	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0	0
S_C	0	0	1	0	0	0	1	0	1	0	0	0	1	0	1	0	1	1
S_G	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
S_T	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

3.3.2 Definitions of different repeats in DNA sequences

Definition 1: A subsequence $S'[n] = s_i s_{i+1} \dots s_{i+l-1}$ of $S[n]$ is an exact tandem repeat (ETR) of period 'p' and repeat pattern $\alpha = r_1 r_2 \dots r_p$ (where 'i' is the starting position and 'l' is the length of ETR), if the following conditions are satisfied:

1. $\lfloor l/p \rfloor \geq 2$, where $\lfloor l/p \rfloor$ is the count for pattern (α), i.e., number of times α has occurred in subsequence $S'[n]$. The count of repeat pattern (α) should at least be equal to two.

2. $\Lambda = \{r_1 r_2 \dots r_p\}$, where $\Lambda \subseteq \Sigma$ & $|\Lambda| \geq 1$.

3. $S_\Delta[n]$ is p-periodic, $\forall \Delta \in \Lambda$, where $i \leq n \leq i + l$.

For example, if $S[n]$ = GGCATACT**ACGACGACG**CCG, then $S'[n]$ = **ACGACGACG**, $i = 9$, $p = 3$, $l = 9$, $\lfloor l/p \rfloor = 3$, $\alpha =$**ACG**, and $S_A[n]$, $S_C[n]$, $S_G[n]$ are 3-periodic sequence.

Definition 2: A subsequence $S''[n] = s_i s_{i+1} \dots s_{i+l-1}$ of $S[n]$ is an inexact tandem repeat (InTR) of period 'p' and repeat pattern $\alpha = r_1 r_2 \dots r_p$ (where 'i' is the starting position and 'l' is the length of InTR), if the following conditions are satisfied:

1. $\lfloor l/p \rfloor \geq 2$.

2. $\Lambda = \{r_1 r_2 \ldots r_p\}$, where $\Lambda \subseteq \Sigma$ & $|\Lambda| \geq 1$.

3. $S_\Delta[n]$ is non-periodic, for at least one $\Delta \in \Lambda$, where $i \leq n \leq i + l$.

4. $\forall \Delta \in \Lambda$, p-period measure of $S_\Delta[n] \geq$ *Threshold*.

For example, if $S[n]$ = GGC<u>ATACACAGACAC</u>GCCGGCG, then $S''[n]$ = **ATACACAGACAC**, $i = 4$, $p = 2$, $l = 12$, α =**AC**, $\Lambda \equiv$ {A,C} and $S_A[n]$ is 2-periodic sequence (not necessarily exact).

From the above formulations it is observed that the repeat identification in DNA is analogous to period detection in signals. Thus, the knowledge of periodicity in the binary signals (i.e., $S_\Omega[n]$) helps in identifying tandem repeats present in DNA sequences. Therefore, the main objective of SP algorithm for tandem repeat identification problem is to develop a good measure for identifying periods in the binary signals.

3.3.3 Exactly periodic subspace decomposition

The exactly periodic subspace decomposition (EPSD) technique was proposed by Muresan & Parks in [60]. The EPSD technique generates orthogonal subspaces that correspond to periods ranging from 1 upto the maximum expected sub-period of the input signal S. The energy of the expected sub-periods are obtained by taking orthogonal projections of S onto these different orthogonal subspaces. The key idea behind the EPSD technique is the concept of exactly periodic signals (EPS). The definition of exactly periodic signal is given as:

Definition 3: A signal S is of exactly period P if S is in Φ_P (where Φ_P is the subspace of the signal of period P) and the projection of S onto subspace $\Phi_{P'}$ for all $P' < P$ (where $\Phi_{P'}$ is the subspace of signal of period P') [60].

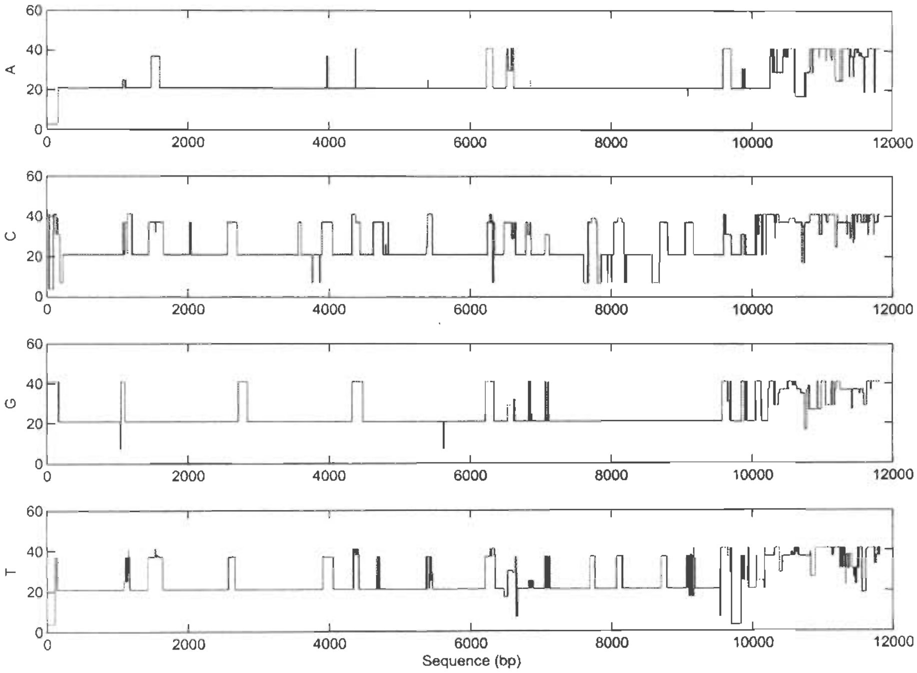


Figure 3.2: Output period provided by EPSD technique for an inexact repeat DNA sequence of period 21.

Thus, a signal of exactly period P is not exactly period $2P$, $3P$, etc. although it continues to be of period $2P$, $3P$, etc. Also, not every periodic signal is exactly periodic, but every exactly periodic signal is periodic. This property of EPSD technique helps in removing the problem with existing signal processing techniques. For example, by looking at the output period provided by EPSD technique as shown in Figure 3.2 one can easily say that the repeat present in the DNA sequence is of period 21. However, this is not possible for the SRF program whose spectral coefficient was previously shown in Figure 3.1. Some of the important properties of the EPSD technique are:

1. The EPSD technique completely decomposes the input signal $S \in \mathbf{R}^n$ into exactly periodic orthogonal components corresponding to each of the exactly periodic signals of n and all possible factors of n.

2. Unlike the STPT [44], the decomposition of the EPSD technique is unique. Thus, the input signal can be uniquely decomposed on the orthogonal subspaces.

3. The EPSD of signal is achieved by taking projections onto exactly periodic orthogonal multidimensional subspaces of periods that divides n, whereas the discrete Fourier transform is obtained by taking orthogonal projections onto one-dimensional (1-D) complex exponentials $e^{j(2\pi/N)k}$ with frequencies (k/N), $k = 0, \ldots, N - 1$. The EPS is spanned by a collection of Fourier exponentials, which is dictated by the period. Thus, by having spaces of dimensions larger than one, EPS can capture the periodic energy in one coefficient better than the Fourier transform.

The EPSD technique, proposed in [60], was applied to identify periodic signal by considering the entire input signal, i.e., it provides information about the periods that are present in complete input data sequence. However, in tandem repeat identification problem, even though the core objective is to identify periods in DNA sequences, there is one major difference. Instead of looking for periods that are present in entire input DNA sequence, local periodic information is considered because most of the tandem repeats that are present in the DNA sequences are localized to small portion of the complete genome. In addition, the tandem repeats form only small fraction of total genome. Thus, the main objective of tandem repeat identification program is to provide the localized periodic information. The EPSD technique is adapted for the current problem to provide a measure for localized periodic information that is present in DNA sequences.

Instead of analyzing the complete input DNA sequence in one go, we divide the DNA sequence into a set of subsequences defined by a point-wise multiplication of the original DNA sequence by a stationary window. The EPSD technique is then applied to the resulting subsequences. Let the window be represented by W_i of length L_w and beginning at ith element, where

$$W_i[n] = \begin{cases} 1, & n = i, i+1, \ldots, i + L_w - 1 \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

The localized portion of the sequence S, S_{W_i} is defined as

$$S_{W,i} = S[n] \cdot W_i[n] \tag{3.6}$$

3.4 Tandem Repeat Detection Algorithm

The objectives of the proposed algorithm are to identify the position, period and the length of repeat patterns in DNA sequences. For identifying repeats, the symbolic DNA sequences are first mapped into four digital signals and then EPSD mathematical tool is applied. Later on, repeat coefficient measure is calculated for each window and the potential repetitive patterns are reported depending on the value of input parameters provided by the user. The algorithm is designed to identify tandem repeats from 2-period to maximum period (P_{max}) provided by the user within an observation window of size L_w. The complete repeat detection process is divided into three major phases. The proposed algorithm is described

Table 3.3: Algorithm for calculating repeat coefficients.

```
1. Accept window size ($L_w$), Maximum period ($P_{max}$).
2. for $i$=1 to $N + L_w - 1$ do // $N$ is the length of DNA sequence.
3.     $S_{W,i}[n] = S_{W,i}[n] - \overline{S}_{W,i}[n]$, where $\overline{S}_{W,i}[n] = \mathrm{MEAN}(S_{W,i}[n])$.
4.     $\alpha_{W,i}[1,\ldots,P_{max}] = \mathrm{EPSD}(S_{W.i}[n], P_{max})$.
5.     $\pi_{W,i}[1,\ldots,P_{max}] = \frac{\|\alpha_{W,i}[1,\ldots,P_{max}]\|^2}{\|S_{W,i}[n]\|^2}$.
6.     OUTPUT($p_i$,$\pi_{W,i}[p_i]$), where $\pi_{W,i}[p_i]$ =max($\pi_{W,i}[1],\ldots,\pi_{W,i}[P_{max}]$.
```

below.

Nucleotide mapping of a DNA sequence $S[n]$ into nucleotide subsequences — The nucleotide mapping procedure was discussed in the section 3.3.1. In this phase, four binary subsequences ($S_A[n]$, $S_C[n]$, $S_G[n]$ and $S_T[n]$) are obtained using (equation 3.4) that act as input signals for the proposed algorithm.

Calculation of tandem repeat coefficient for subsequences — For identifying the position of the tandem repeats in DNA sequences a sliding window based approach is used. The algorithm for calculating period with maximum energy for the input DNA sequence of length N and input parameters (P_{max}, L_w) is provided in Table 3.3, where the value of P_{max} can vary from 2 to $L_{w/2}$. The prior knowledge of maximum repeat pattern size restricts the search to pattern size P_{max}. However, if the user does not have prior knowledge then the value of P_{max}

can be fixed to $L_{w/2}$. In step 3 of the algorithm, the DC component (i.e., period 1) is removed from the input signal. This step helps in removing the repeats that occur due to single base repeat pattern, for instance repeat like AAAAA in DNA sequence ACGACAAAAACAACG, because the repeat pattern of period 1 is of no interest. In step 4, the energy of the input signal is decomposed on the subspaces from 2 to P_{max} using EPSD technique. The energies of the subspaces are stored in the array $\alpha_{W,i}$. The array $\pi_{W,i}$ which is calculated in step 5, measures the fraction of power of the periodic subspaces from 2 to P_{max}. The value $\pi_{W,i}$ acts as an indicator for identifying the local periodicity of the input sequence and is said to as *tandem repeat coefficient.* Finally in step 6, a tuple $< p, \pi_{W,i}[p] >$ for each window is obtained where p is the periodic subspace that has maximum fraction of power in the subsequence for the window positioned at i. The algorithm presented in Table 3.3, unlike the PE algorithm, needs just a single scan for identifying the period ($\leq P_{max}$) of repeat patterns in the input DNA sequence. This step is performed on all four binary subsequences obtained from the previous step.

Identification and characterization repeat from binary subsequences — In this phase, utilizing the value of threshold parameter (τ) and tuple $< p, \pi_{W,i}[p] >$ calculated in the previous phase, the repeats in all four binary subsequences are identified first. A repeat is represented by tuple $< \Omega, i, l, p >$, where $\Omega \in$ {A, C, G, T}, i is the starting position of the repeat (position of the window), l is the length of the repeat, and p is the period of repeat. A repeat satisfies the following conditions:

- $\pi_{W,i}, \pi_{W,i+1}, \ldots, \pi_{W,i+l-1} \geq \tau$

- $p_i = p_{i+1} = \cdots = p_{i+l-1} = p$

After the repeats in each subsequences are identified, all four subsequences are processed together and classified into ETR, and InTR based on the definitions provided in section 3.3.2.

3.5 Experimental Results

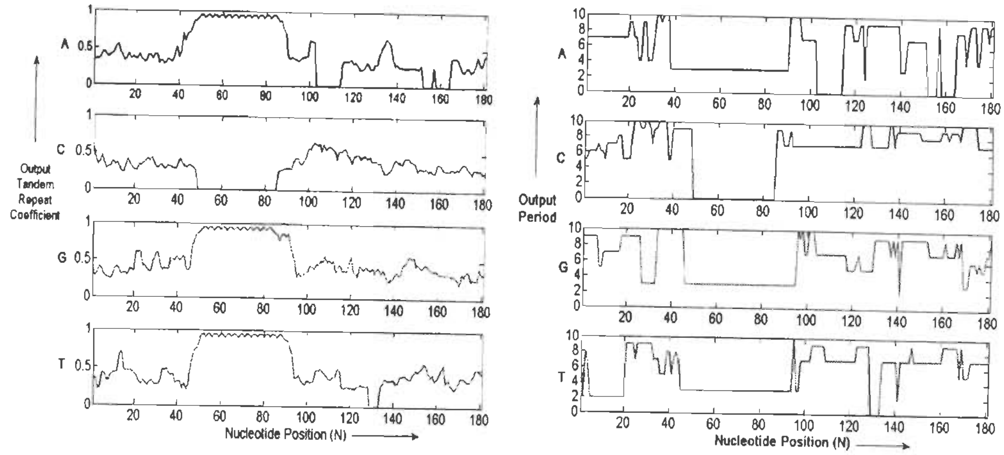
In this section, experiments are performed on two categories of datasets to show the effectiveness of the proposed algorithm. First, the algorithm is used to analyzes pseudo DNA sequences. The second datasets included actual DNA sequences available at GenBank database. The datasets were selected such that the experiment covers exact, inexact (complex, dispersed, and hidden) repeat patterns. Results obtained from other tandem repeat identification algorithm when applied to the DNA sequences considered for analysis is also provided in this section. The proposed algorithm was implemented in C++ for Microsoft Windows ® platform.

A. Pseudo DNA sequence testings: In this category the datasets were created by adding a tandem repeat pattern in a random DNA sequence. The datasets includes both exact and inexact tandem repeats. For inexact tandem repeats test cases with *substitution*, *insertion* and *deletion* operations in the initial exact tandem repeat were taken up for the simulation study. For each test case in (Figures 3.3– 3.12), three sub-figures have been provided. The first figure shows the test DNA sequence where the substituted nucleotides are underlined, inserted nucleotides are circled and the deleted nucleotides are shown by rectangular box. The symbol * in the result represents any of the four nucleotides. The second and third figures provide information about the tandem repeat coefficient and output period respectively, obtained for the test DNA sequence. In Table 3.4, a summary of tests performed on pseudo DNA sequences is provided.

Test 1: An exact tandem repeat pattern. Repeat position: 51–100.

1 TTAACTGTCC	11 GAGTCGGAAT	21 CCATCTCTGA	31 GTCACCCAAG	41 AAGCTGCCCT
51 **ATGATGATGA**	61 **TGATGATGAT**	71 **GATGATGATG**	81 **ATGATGATGA**	91 **TGATGATGAT**
101 GATCCTGCAG	111 GCTGTGGGCG	121 GTGGGCCTGG	131 GACAGGCAGC	141 TACGGGCCCG
151 AGTGTGACTG	161 GTGGGCGCTG	171 GGTGTATTCG	181 CCTATGAAAT	191 GTTCTATGGG

(a) DNA sequence



(b) Tandem repeat coefficient

(c) Output period

Figure 3.3: Exact tandem repeat.

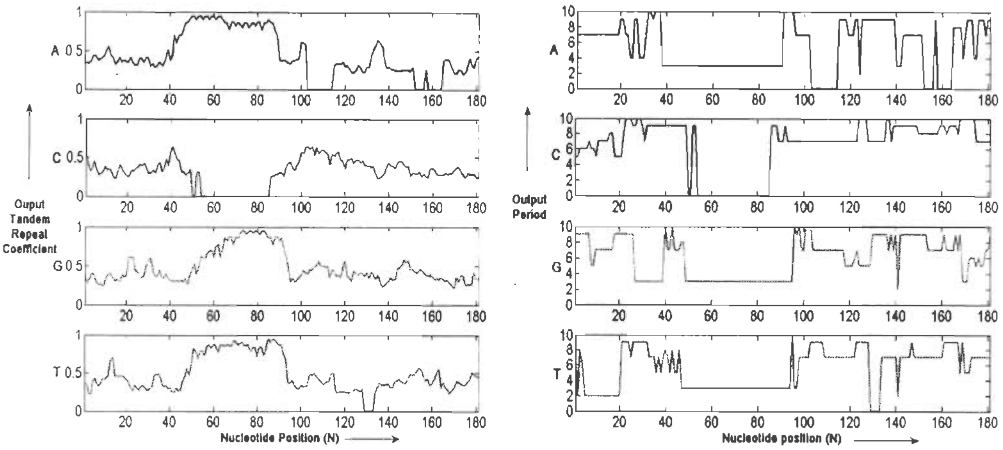
Input parameters: Maximum period=10, Window length=20, Threshold=0.7.

Result: Tandem repeat pattern=**ATG**, Output period=3, Start position=48, Length=61.

Test 2: Tandem repeat with 10% *substitutions*, (at positions 53, 56, 64, 71 and 84), Location: 51–100.

1	11	21	31	41
TTAACTGTCC	GAGTCGGAAT	CCATCTCTGA	GTCACCCAAG	AAGCTGCCCT
51	61	71	81	91
ATCATTATGA	**TGAGGATGAT**	**CATGATGATG**	**ATGTTGATGA**	**TGATGATGAT**
101	111	121	131	141
GATCCTGCAG	GCTGTGGGCG	GTGGGCCTGG	GACAGGCAGC	TACGGGCCCG
151	161	171	181	191
AGTGTGACTG	GTGGGCGCTG	GGTGTATTCG	CCTATGAAAT	GTTCTATGGG

(a) DNA sequence



(b) Tandem repeat coefficient

(c) Output period

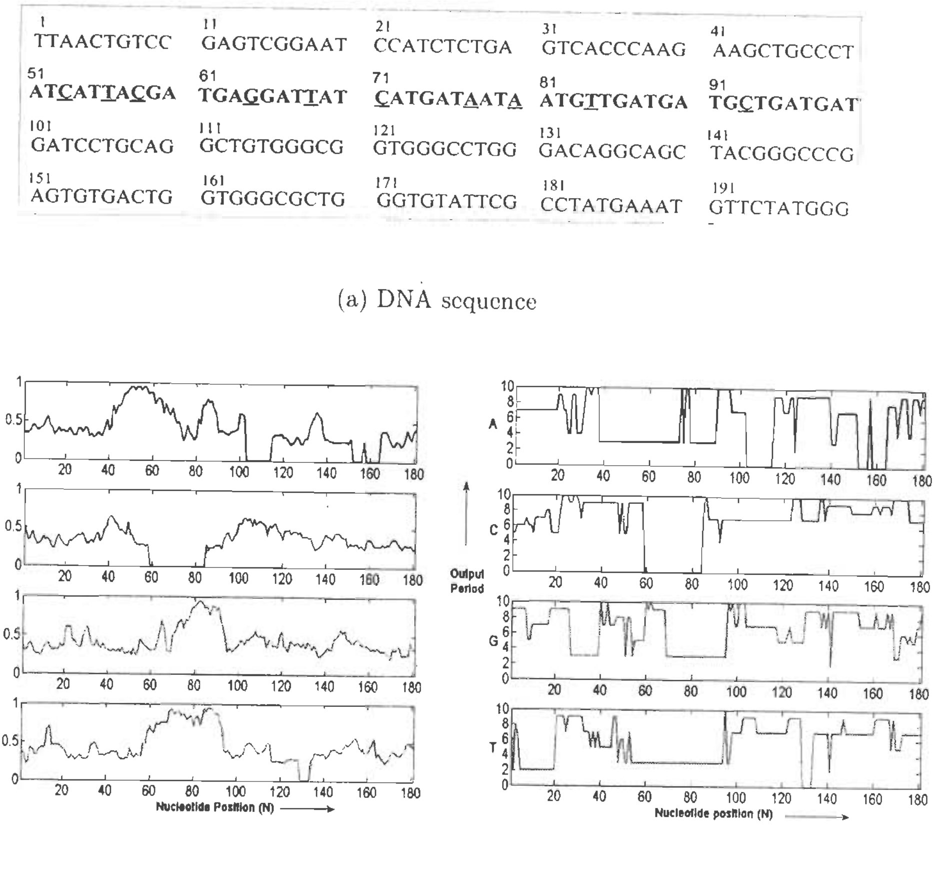
Figure 3.4: Tandem repeat with 10% substitutions.

Input: Maximum period=10, Window length=20, Threshold=0.9.

Result: Tandem repeat pattern= **ATG**, Output period= 3, Start position= 61, Length= 48.

Test 3: Tandem repeat with 20% *substitutions*,(at positions 53, 56, 58, 64, 68, 71, 77, 80, 84 and 93). Location: 51–100.

Input: Maximum period=10, Window length=20, Threshold=0.7.



1 TTAACTGTCC	11 GAGTCGGAAT	21 CCATCTCTGA	31 GTCACCCAAG	41 AAGCTGCCCT
51 **ATCATTACGA**	61 **TGAGGATTAT**	71 **CATGATAATA**	81 **ATGTTGATGA**	91 **TGCTGATGAT**
101 GATCCTGCAG	111 GCTGTGGGCG	121 GTGGGCCTGG	131 GACAGGCAGC	141 TACGGGCCCG
151 AGTGTGACTG	161 GTGGGCGCTG	171 GGTGTATTCG	181 CCTATGAAAT	191 GTTCTATGGG

(a) DNA sequence

(b) Tandem repeat coefficient

(c) Output period

Figure 3.5: Tandem repeat with 20% substitutions.

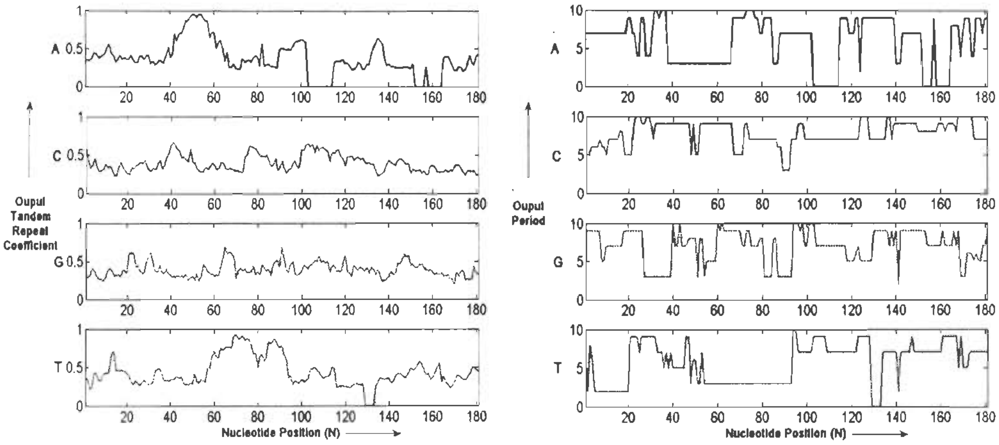
Result: Tandem repeat pattern= **ATG**, Output period=3, Start position=78, Length=30.

Test 4: Tandem repeat with 30% *substitutions*, (at positions 53, 56, 58, 64, 68, 71, 77, 80, 84, 86, 90, 93, 97 and 99). Location: 51–100.

Input: Maximum period=10, Window length=20, Threshold=0.7.

1	11	21	31	41
TTAACTGTCC	GAGTCGGAAT	CCATCTCTGA	GTCACCCAAG	AAGCTGCCCT
51	61	71	81	91
ATCATTACGA	**TGAGGATTAT**	**CATGCTAATA**	**ATGTTCATGG**	**TGCTGAGGCT**
101	111	121	131	141
GATCCTGCAG	GCTGTGGGCG	GTGGGCCTGG	GACAGGCAGC	TACGGGCCCG
151	161	171	181	191
AGTGTGACTG	GTGGGCGCTG	GGTGTATTCG	CCTATGAAAT	GTTCTATGGG

(a) DNA sequence



(b) Tandem repeat coefficient

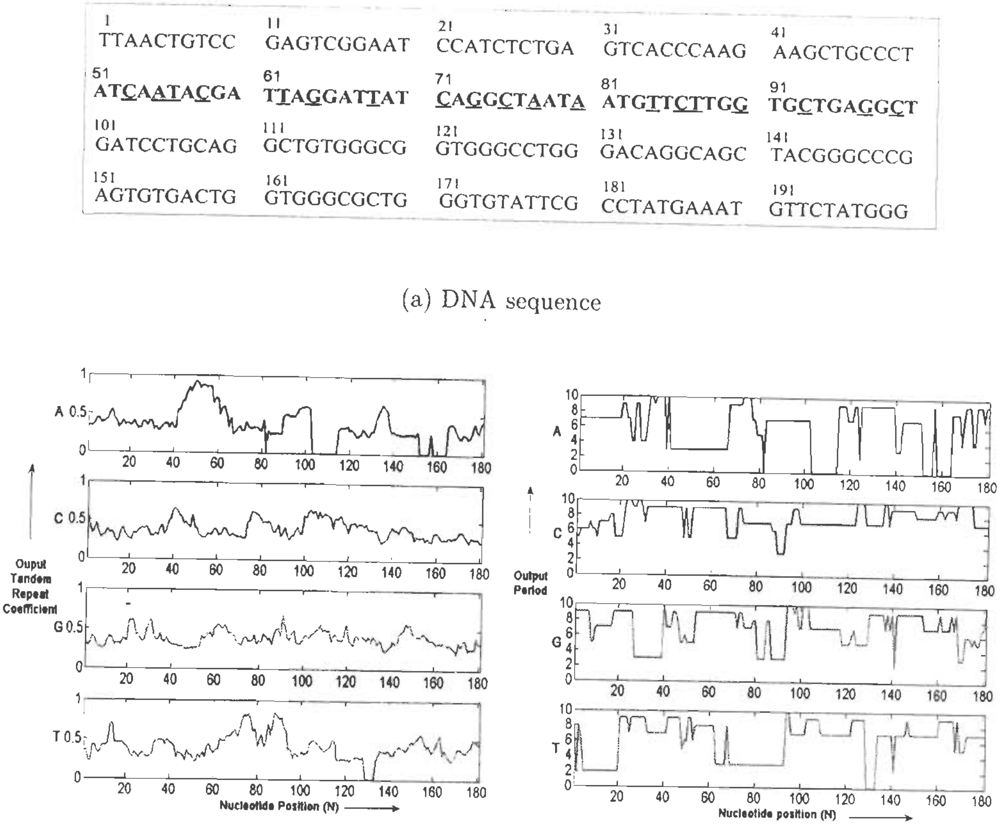
(c) Output period

Figure 3.6: Tandem repeat with 30% substitutions.

Result:

- Tandem repeat pattern = **T*A**, Output period = 3, Start position = 60, Length = 21.
- Repeat pattern = **T****, Output period = 3, Start position = 81, Length = 31.

Test 5: Tandem repeat with 40% *substitutions*,(at positions 53, 55, 56, 58, 62, 64, 68, 71, 73, 75, 77, 78, 80, 84, 86, 87, 90, 93, 97 and 99), Location: 51–100. Input: Maximum period=10, Window length=20, Threshold=0.7.



1 TTAACTGTCC	11 GAGTCGGAAT	21 CCATCTCTGA	31 GTCACCCAAG	41 AAGCTGCCCT
51 **ATCAATACGA**	61 **TTAGGATTAT**	71 **CAGGCTAATA**	81 **ATGTTCTTGG**	91 **TGCTGAGGCT**
101 GATCCTGCAG	111 GCTGTGGGCG	121 GTGGGCCTGG	131 GACAGGCAGC	141 TACGGGCCCG
151 AGTGTGACTG	161 GTGGGCGCTG	171 GGTGTATTCG	181 CCTATGAAAT	191 GTTCTATGGG

(a) DNA sequence

(b) Tandem repeat coefficient

(c) Output period

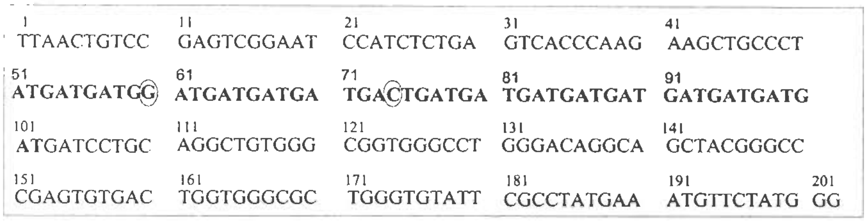
Figure 3.7: Tandem repeat with 40% substitutions.

Result:

- Periodic repeat pattern = **A****, Output period = 3, Start position = 46, Length = 27.
- Periodic repeat pattern = **T****, Output period = 3, Start position = 81, Length = 31.

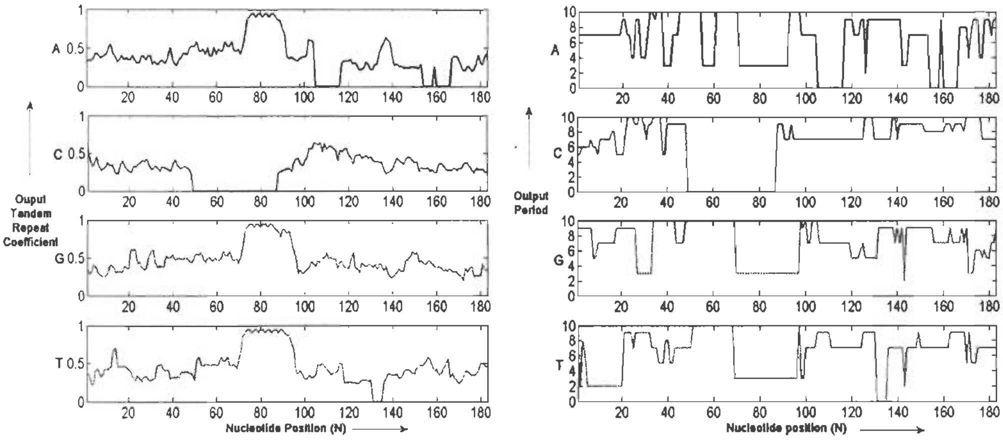
Test 6: Tandem repeat with 2 *insertions* after position 59 and 72 in input DNA sequence. Location of repeat: 51–104.

Input: Maximum period=10, Window length=20, Threshold=0.7.



```
1           11          21          31          41
TTAACTGTCC  GAGTCGGAAT  CCATCTCTGA  GTCACCCAAG  AAGCTGCCCT
51          61          71          81          91
ATGATGATGG  ATGATGATGA  TGACTGATGA  TGATGATGAT  GATGATGATG
101         111         121         131         141
ATGATCCTGC  AGGCTGTGGG  CGGTGGGCCT  GGGACAGGCA  GCTACGGGCC
151         161         171         181         191         201
CGAGTGTGAC  TGGTGGGCGC  TGGGTGTATT  CGCCTATGAA  ATGTTCTATG  GG
```

(a) DNA sequence



(b) Tandem repeat coefficient

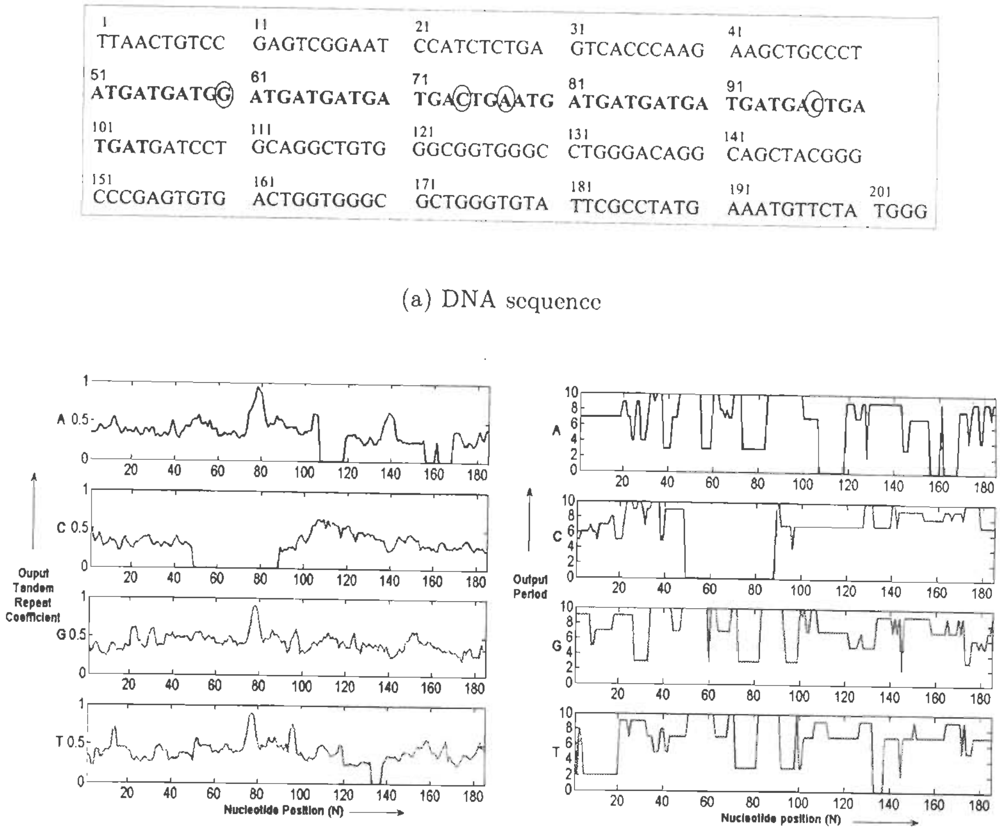
(c) Output period

Figure 3.8: Tandem repeat with 2 insertions.

Result: Tandem repeat pattern= **ATG**, Output Period=3, Start position= 74, Length= 37.

Test 7: Tandem repeat with 4 *insertions* after position 59, 72, 74 and 93 in input DNA sequence. Location of repeat: 51–104.

Input: Maximum period=10, Window length=20, Threshold=0.7.



```
1           11          21          31          41
TTAACTGTCC  GAGTCGGAAT  CCATCTCTGA  GTCACCCAAG  AAGCTGCCCT
51          61          71          81          91
ATGATGATGG  ATGATGATGA  TGACTGAATG  ATGATGATGA  TGATGACTGA
101         111         121         131         141
TGATGATCCT  GCAGGCTGTG  GGCGGTGGGC  CTGGGACAGG  CAGCTACGGG
151         161         171         181         191        201
CCCGAGTGTG  ACTGGTGGGC  GCTGGGTGTA  TTCGCCTATG  AAATGTTCTA TGGG
```

(a) DNA sequence

(b) Tandem repeat coefficient

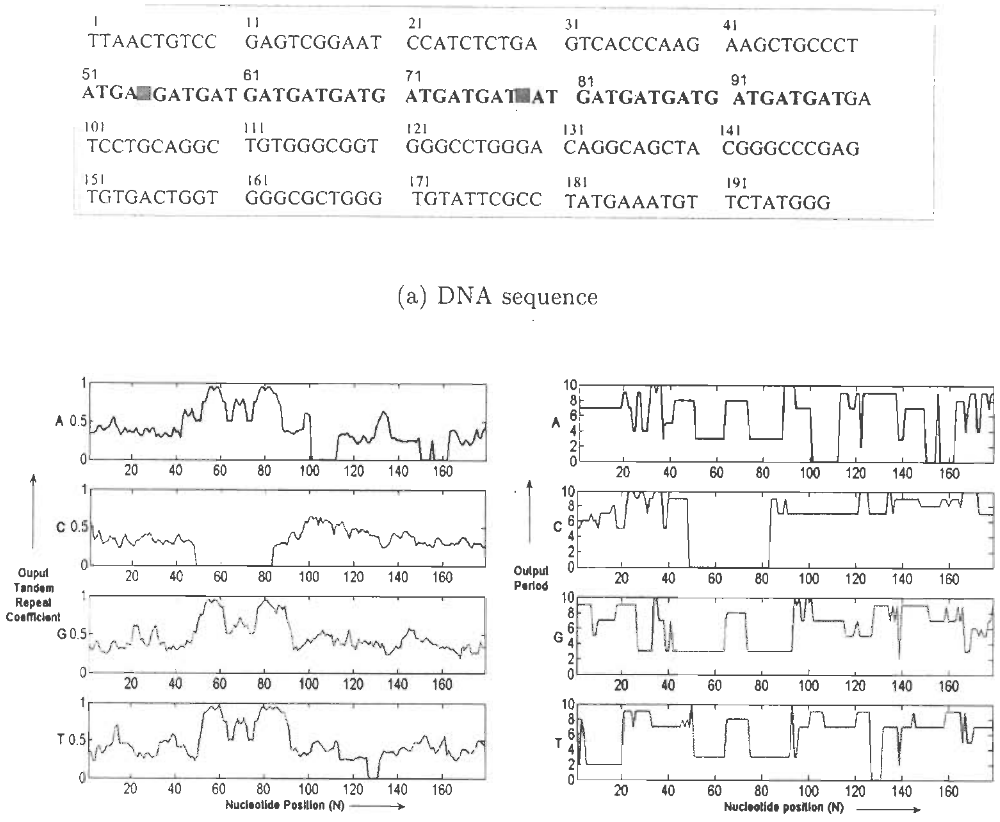
(c) Output period

Figure 3.9: Tandem repeat with 4 insertions.

Result: Tandem repeat pattern= **TGA**, Output Period=3, Start position=77, Length=23.

Test 8: Tandem repeat with 2 *deletions* after position 54 and 79. Location of repeat: 51–98.

Input: Maximum period=10, Window length=20, Threshold=0.7.



```
1           11          21          31          41
TTAACTGTCC  GAGTCGGAAT  CCATCTCTGA  GTCACCCAAG  AAGCTGCCCT
51          61          71          81          91
ATGA■GATGAT GATGATGATG  ATGATGAT■AT GATGATGATG  ATGATGATGA
101         111         121         131         141
TCCTGCAGGC  TGTGGGCGGT  GGGCCTGGGA  CAGGCAGCTA  CGGGCCCGAG
151         161         171         181         191
TGTGACTGGT  GGGCGCTGGG  TGTATTCGCC  TATGAAATGT  TCTATGGG
```

(a) DNA sequence

(b) Tandem repeat coefficient

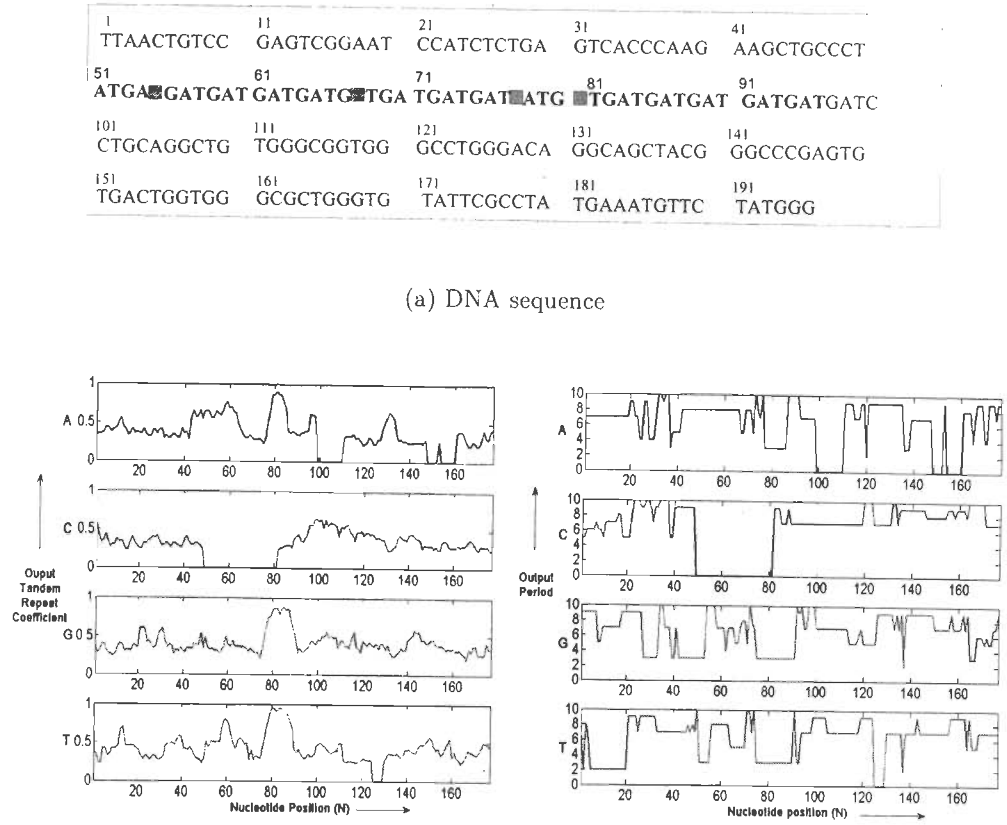
(c) Output period

Figure 3.10: Tandem repeat with 2 deletions.

Result: Tandem repeat pattern=**TGA**, Output Period=3, Start position=53, Length=52.

Test 9: Tandem Repeat with 4 *deletions* after position 54, 68, 79 and 83. Location of repeat: 51–96.

Input: Maximum period=10, Window length=20, Threshold=0.7.



```
1           11          21          31          41
TTAACTGTCC  GAGTCGGAAT  CCATCTCTGA  GTCACCCAAG  AAGCTGCCCT
51          61          71          81          91
ATGA■GATGAT GATGATG■TGA TGATGAT■ATG ■TGATGATGAT GATGATGATC
101         111         121         131         141
CTGCAGGCTG  TGGGCGGTGG  GCCTGGGACA  GGCAGCTACG  GGCCCGAGTG
151         161         171         181         191
TGACTGGTGG  GCGCTGGGTG  TATTCGCCTA  TGAAATGTTC  TATGGG
```

(a) DNA sequence

(b) Tandem repeat coefficient

(c) Output period

Figure 3.11: Tandem repeat with 4 deletions.

Result: Tandem repeat pattern= **TGA**, Output Period=3, Start position=79, Length=26.

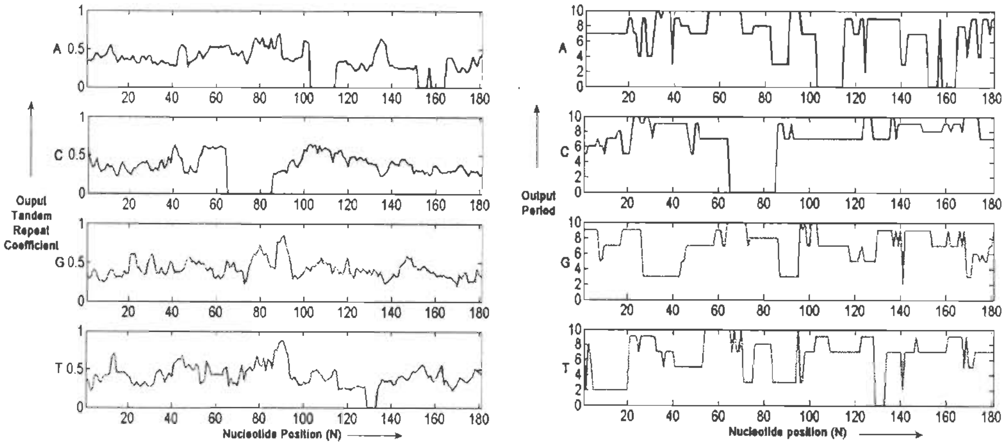
Test 10: Tandem repeat with 10% *substitutions*, 2 *insertions* and 2 *deletions*.

Location of repeat: 51–100.

Input: Maximum period=10, Window length=20, Threshold=0.7.

```
1            11           21           31           41
TTAACTGTCC   GAGTCGGAAT   CCATCTCTGA   GTCACCCAAG   AAGCTGCCCT
51           61           71           81           91
ATCATT▨TGAT  GAGCGATGAT   CATGAATGAT   GATGTTGAT■A  TGATGATGAT
101          111          121          131          141
CTGCAGGCTG   TGGGCGGTGG   GCCTGGGACA   GGCAGCTACG   GGCCCGAGTG
151          161          171          181          191
TGACTGGTGG   GCGCTGGGTG   TATTCGCCTA   TGAAATGTTC   TATGGG
```

(a) DNA sequence



(b) Tandem repeat coefficient

(c) Output period

Figure 3.12: Tandem repeat with 10% substitutions, 2 insertions and 2 deletions.

Result: Periodic Repeat pattern= **TG***, Output Period=3, Start position=88, Length=25.

Table 3.4: Summary of test performed on 10 pseudo DNA sequences.

Repeat pattern	**Operation**	**Detected repeat pattern**
Exact tandem repeat pattern ATG from 51-100	None	Repeat pattern= ATG, Output period= 3, Position= 48-108.
Tandem repeat pattern ATG from 51-100 with 20% substitutions	Substitutions at positions 53, 56, 58, 64, 68, 71, 77, 80, 84 and 93	Repeat pattern=ATG, Output period=3, Position=78-107.
Tandem repeat pattern ATG from 51-100 with 30% substitutions	Substitutions at positions 53, 56, 58, 64, 68, 71, 77, 80, 84, 86, 90, 93, 97 and 99	(a) Repeat pattern = T*A, Output period = 3, Position = 60 - 80. (b) Repeat pattern = T**, Output period = 3, Position = 81 - 110.
Tandem repeat pattern ATG from 51-100 with 40% substitutions	Substitutions at positions 53, 55, 56, 58, 62, 64, 68, 71, 73, 75, 77, 78, 80, 84, 86, 87, 90, 93, 97 and 99	(a) Repeat pattern = T*A, Output period = 3, Position = 60 - 80. (b) Repeat pattern = T**, Output period = 3, Position = 81 - 110.

Continued on next page

Table 3.4 – continued from previous page

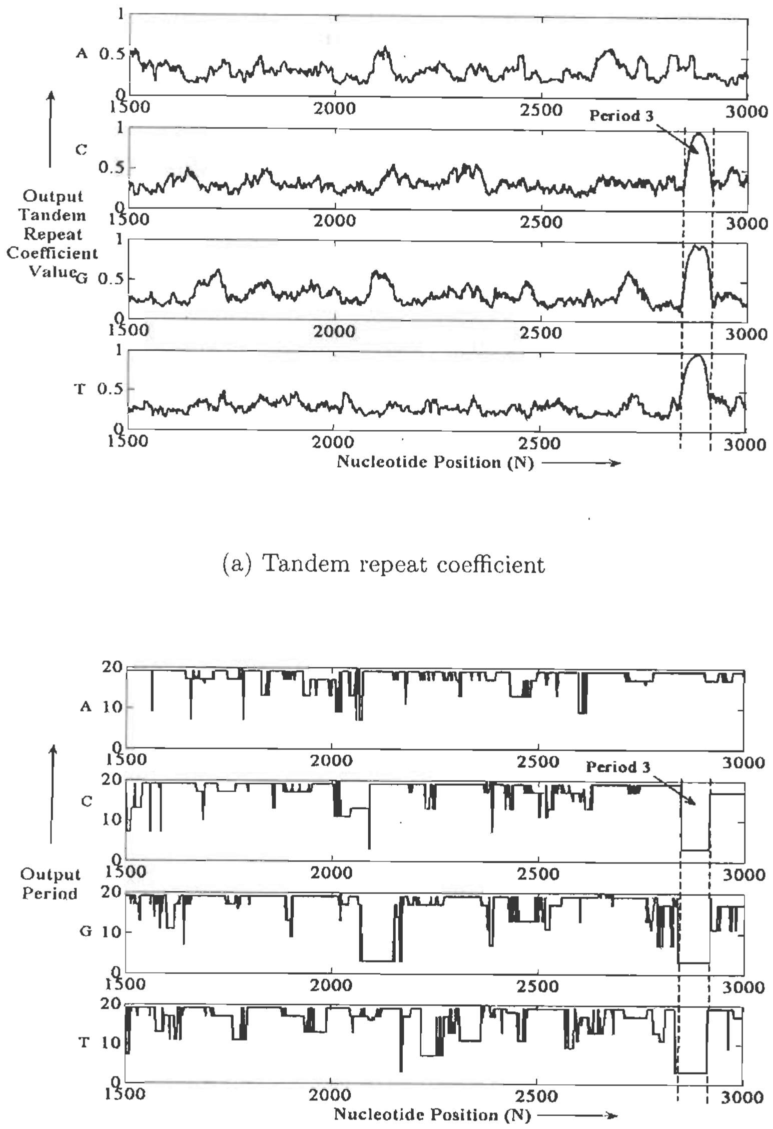
Repeat pattern	Operation	Detected repeat pattern
Tandem repeat ATG with 2 Insertions. Location of repeat: 51-102	Insertions after position 49 and 72	Repeat pattern= ATG, Output Period=3, Position= 74-110.
Tandem repeat ATG with 4 Insertions. Location of repeat: 51-104	Insertions after position 59, 72, 74 and 93	Repeat pattern= TGA, Output Period= 3, Position= 77-99.
Tandem Repeat with 2 Deletions. Location of repeat: 51-98	Deletions after position 54 and 79	Repeat pattern= TGA, Output Period=3, Position= 53-104.
Tandem Repeat with 4 Deletions. Location of repeat: 51-96	Deletions after position 54, 58, 79 and 83	Repeat pattern= TGA, Output Period=3, Position= 79 - 104.
Exact Tandem Repeat with 10% substitutions, 2 Insertions and 2 Deletions	Location of repeat: 51-100, 10% substitutions, 2 Insertions and 2 Deletions	Repeat pattern= TG*, Output Period=3, Position= 88-112.

B. Actual DNA sequence testings

Test 1: Myotonic dystrophy disease, the most common muscular dystrophy in Humans, is caused by an expansion of the CTG repeat located in the 3′-UTR (untranslated region) of dystrophia myotonica protein kinase (DMPK) gene [69]. The 3′-UTR region is present after a coding region in a DNA sequence. For a normal person the repeat number of CTG is less than 35 and for a person suffering from myotonic dystrophy the CTG count is above 50 [51]. This dataset consists of DNA sequence (GenBank accession number: XM_027572, length = 3436 base pairs (bp)) of Homo sapiens DMPK gene sequenced under NCBI annotation project.

The DNA sequence was tested with input parameters for window size (L_w)=40 and maximum period (P_{max})=10 and threshold (τ)=0.95. The tandem repeat coefficients obtained for subsequences $S_A[n]$, $S_C[n]$, $S_G[n]$, $S_T[n]$ are shown in Figure 3.13(a) and in Figure 3.13(b) the output periods obtained for the subsequences are provided. The subsequences $S_C[n]$, $S_G[n]$ and $S_T[n]$ have repeat coefficient values greater than 0.95 from 2876 to 2967 and the corresponding output period is 3 (shown in Figure 3.13(b)). An exact tri-nucleotide tandem repeat pattern CTG of repeat length 62 (repeat number $\approx$ 21), beginning at 2890 was identified in the DNA sequence. The protein coding sequence for Human DMPK gene is from 779-2668 bp. The identified tandem repeat lies after 2668 bp in DMPK gene sequence; this confirms the presence of CTG repeat in 3′-UTR of Human DMPK. Apart from exact tandem repeats, weak patterns of period 3 were also identified for nucleotides C (beginning at 1864, length of 21) and G (beginning at 2114, length of 63).

Experiments were also conducted using TRF 4.0 [61]and PE [44] for a max-



(a) Tandem repeat coefficient

(b) Output period

Figure 3.13: (GenBank Accession: XM_027572, length=3436 base pair (bp)) with input parameters (window length=80 and maximum period=20).

Table 3.5: Repeat patterns identified in HSVDJSAT DNA sequence.

Program	Consensus period	Repeat region
Our Algorithm	$2^{a,c}$	825-865
	$9^{a,c}$,$10^{a,c}$, $19^{b,d}$, $49^{b,d}$	1177-1545
Hauth program	9, 10, 19, 37, 38, 48	1197-1538
TRF 4.0[e]	2^{c}	826-856
	10^{c}	1199-1539
	19^{d}	1190-1539
	49^{d}	1195-1539

[a]Maximum period size (P_{max})≤10, [b]Maximum period size(P_{max})>10

[c]Simple tandem repeat, [d]Multi-period tandem repeat

[e]Alignment parameter(match,mismatch,indel)=(2,7,7), Min. score=30 and P_{max}=50

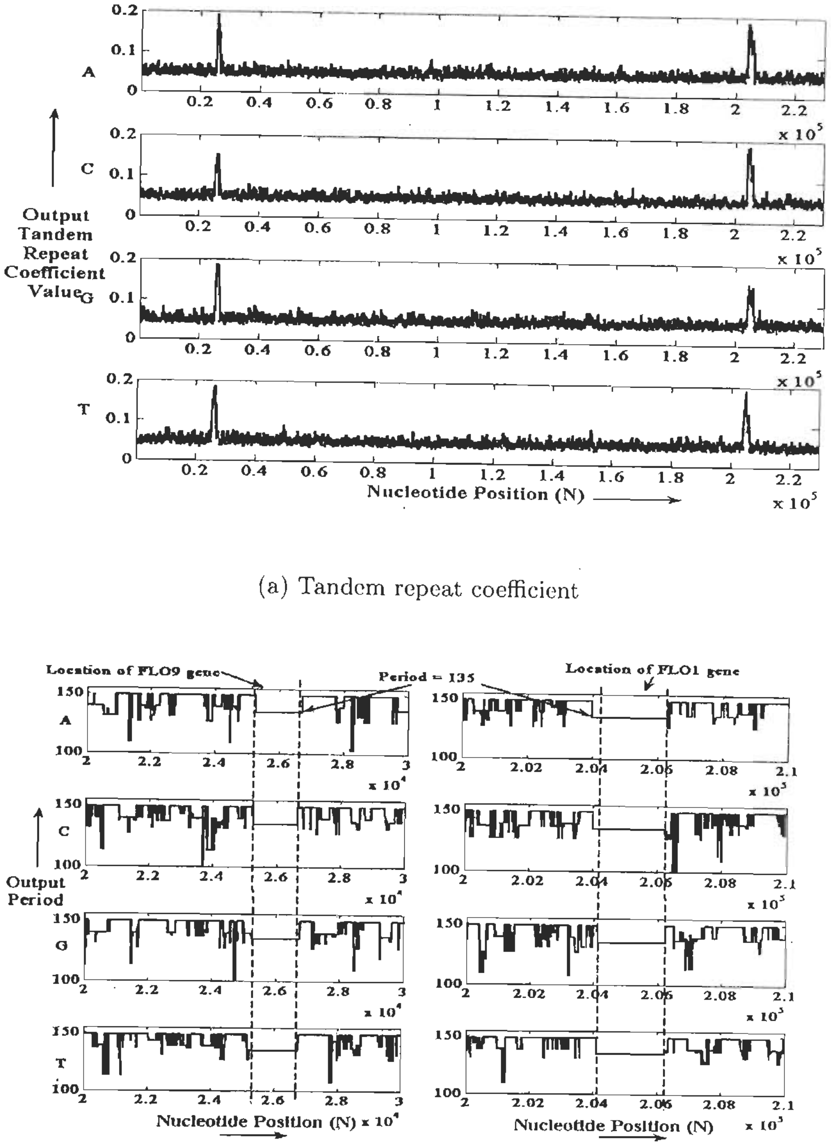
imum period size equal to 10. TRF 4.0 with default input parameters provides output consisting of tandem repeat of pattern TGC starting at 2890 and repeat length 62. The PE program provided output pattern of period 3 (TGC), period 6 (TGCTGC) and period 9 (TGCTGCTGC).

TEST 2: The analysis of Homo sapiens, GeneBank Locus: HSVDJSAT of length 1985 bp is provided in this example. This DNA sequence consists of simple and multi-period tandem repeat patterns. Periods of size 2, 9, 10, 19 and 48 were identified in the DNA sequence. The details of the identified repeats are provided in Table 3.5. The consensus tandem repeat patterns of size 2, 19, and 49 reported by the algorithm are: AC, CTGGGAGAGGCTGGGATTG, CTGGGAGAG-GCTGGGAGAG, GAGGCTGGGAGAGGCTGGGAGAG*CTGGGAGAGGCT G*GATTGCTGGGA (where * represents any of the four nucleotides i.e. A, C, G or T). Tests were also performed using tandem repeat finder (TRF) 4.0 [54, 61] and Hauth program [59] for identifying repeats. In [64], Hauth reported the 49th period as period of 48 and missed the simple repeat pattern of period 2. The TRF 4.0 program missed the tandem repeat pattern of period size 9.

TEST 3: The complete chromosome I sequence contains two flocculation genes (FLO1 and FLO9), one at each end of the chromosome, that contain a tandem repeat region having similar 135 bp pattern [70]. The GeneBank details of the DNA sequence and genes (FLO1 and FLO9) are as follows:
Locus: NC_001133, Total base pairs: 230,208
Organism: *Saccharomyces cerevisiae* (baker's yeast)
Gene: FLO1, Region in DNA sequence: 24,001 - 27,969
Gene: FLO9, Region in DNA sequence: 203,394 - 208,007

The DNA sequence is processed by the algorithm with input parameters, window size (L_w) = 600 and maximum period (P_{max}) = 150. The outputs (i.e. repeat coefficients and maximum period) of the algorithm for the nucleotide subsequences are provided in Figure 3.14(a) and Figure 3.14(b). Two sharp peaks are present in Figure 3.14(a). These peaks are due to presence of strong tandem repeats in the DNA sequence at these positions. The first peak starts at 25324 and last for 1842 bp. The maximum period for this region as shown in Figure 3.14(b) is 135. This tandem repeat region lies in gene FPO9. The second peak starts at 204207 and last for 2466 bp. This region also have maximum period of 135 bp. However, the total number of copies for this tandem repeat is higher than the previous one. The result confirms the presence of strong tandem repeats which are present in FLO1 and FLO9 genes of *Saccharomyces cerevisiae*, chromosome I.

TEST 4: The analysis of *Homo sapiens* collagen gene, GenBank accession number: NM_001847 of length 6574 bp containing weak tandem repeat pattern



(a) Tandem repeat coefficient

(b) Output period

Figure 3.14: (GenBank Accession: NM_001133, length=230,208 bp) with input parameters (window length=600 and maximum period=150).

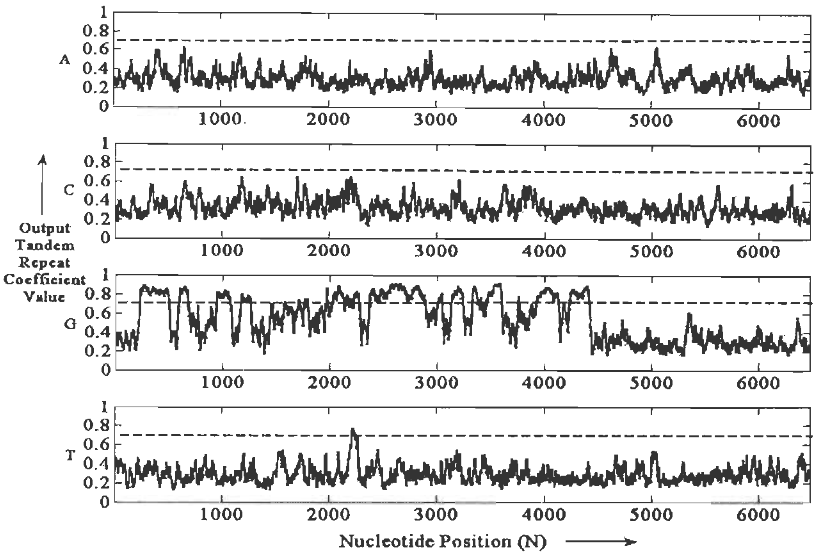


Figure 3.15: Tandem repeat coefficient value of subsequences $S_A[n]$, $S_C[n]$, $S_G[n]$, $S_T[n]$ for DNA sequence (GenBank Accession: NM_001847, length=6574 bp) with input parameters (window length=100 and maximum period=20).

is provided in this example. The tandem repeat coefficients obtained for subsequences $S_A[n]$, $S_C[n]$, $S_G[n]$, $S_T[n]$ for window size (L_w) = 100 and maximum period (P_{max}) = 20 are shown in Figure 3.15. In the figure, subsequence $S_G[n]$ has significant repeat coefficient value from 250 to 4400, while for subsequence $S_T[n]$, the repeat coefficient is above threshold=0.7 from 2233 to 2326. However, for other subsequences i.e., $S_A[n]$ and $S_C[n]$ the value of repeat coefficient lies between 0.4 and 0.6. This shows the presence of repetitive pattern involving nucleotide G and T.

Tests were also performed using PE and TRF program. PE program gave tandem repeat of period 9 and multiple of 9 (i.e., 18, 27 and so on). This is due to problem with the PE algorithm because it cannot distinguish whether a repeat is of period p or it's multiple. However, this problem did not appear in our algorithm because of unique decomposition property of EPSD technique. The TRF program provided two tandem repeat region of period 9 starting at 963 and 1404. Both PE and TRF fail to inform the user regarding hidden periodicity of nucleotide G. This has happened because the TRF and PE programs are designed only to detect tandem repeat and not hidden periodicity of individual nucleotides in DNA sequences.

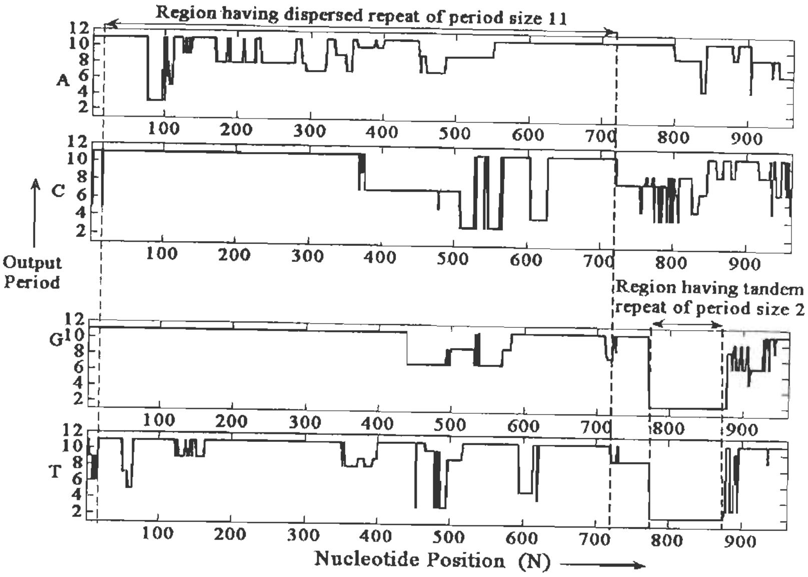


Figure 3.16: Output period of subsequences $S_A[n]$, $S_C[n]$, $S_G[n]$, $S_T[n]$ for DNA sequence (GenBank Accession: M65145) with input parameters (window length=110 and maximum period=11).

TEST 5: A Human microsatellite repeat (GenBank Accession: M65145) is taken up for analysis in this dataset. Figure 3.16 shows the periods identified in the DNA sequence. From the figure, it is clear that the DNA sequence contains two repeat regions of period 2 and 11. The dinucleotide repeats of pattern TG occur between positions 780 and 933 bp (GenBank annotation is between 860 and 900 bp). And the period 11 repeats are located between 92 and 781 bp (unannotated by GenBank). The analysis of the period 11 repeat region of the DNA sequence reveals the dispersed (hidden repeat) copy of the period 11 i.e., TGACTTTGGGG. The TRF program was unable to detect the period 11 repeats in the DNA sequence with default input parameters. This shows the sensitiveness of the proposed algorithm in identifying and locating dispersed or hidden periodic patterns.

3.6 Conclusion

In this chapter, a signal processing approach for tandem repeat identification problem in a DNA sequence is presented. A novel measure based on adapted exactly periodic subspace decomposition technique is proposed in this chapter. The algorithm offers an easy solution to biologist or non-computer experts and is a complementary technique to existing non-signal approaches for identifying tandem and other hidden repeats in a DNA sequence. Based on the concept of local periodicity in a signal as discussed in section 3.3.3, the algorithm has the potential to identify and locate exact, inexact tandem repeat and other hidden and complex repeat pattern unannotated by GenBank databases. The algorithm resolves the problems like whether the repeat pattern is of period P or its multiple (i.e., $2P$, $3P$ and so on) and other issues related to detection of inexact tandem repeats that were present in previous signal processing based approaches.

Experimental results obtained from pseudo (random) DNA sequences and actual DNA sequences, and comparison with other algorithms show the effectiveness of our algorithm. The proposed algorithm is able to identify the inexact repeats that were missed by TRF algorithm due to two major advantages of the proposed approach. First, all four bases were analyzed separately so even if one base out of four bases is repeating in a portion of DNA sequence the repeat can be easily identified. Secondly, the periodicity measure is calculated by taking projection onto exactly periodic multidimensional subspaces.

Chapter 4

Novel Approaches for Identification of Inverted Repeats

The detection of inverted repeat (IR) structure is important in biology because it has been associated with large biological function. The existing tools for inverted repeat identification are too complex for non-computer experts such as biologists. These tools require a number of input parameters before starting a search for IR detection. Also, there does not exist any signal processing based algorithm for IR detection in a DNA or RNA sequence. The goal of this chapter to investigate a signal processing based approach for IR identification. Additionally, the aim is to develop an algorithm which require few input parameters and easier to use.

First, a periodogram measure for identification of exact IR and inexact IR (due to substitution) is presented. Later on, a correlation based approach is provided to identify all type of inexact IRs present in a DNA sequence. This method applies FFT algorithm for faster calculation of correlation based measure for IR identification. The algorithm operates in two different stages. In the first stage it identifies exact inverted repeats present in the DNA sequence. In the second stage the exact repeats are merged to identify inexact IRs. Experiments were performed on actual datasets downloaded from standard databases and results demonstrate the effectiveness of the algorithms.

4.1 Introduction

An inverted repeat is a DNA or RNA sequence that becomes a palindrome if each character in one half of the sequence is changed to its complement character (in DNA, A–T, C–G are complements; in RNA A–U, C–G). For example, ATGCATGCAT is an inverted repeat. Inverted repeats (IRs) are widespread in both prokaryotic and eukaryotic genomes [71, 72, 73], and have been associated with a large number of possible functions. IRs have been implicated in the regulation of initiation of DNA replication in plasmids, bacteria, eukaryotic viruses and mammals [74]. Restriction enzyme cutting sites are interesting examples of IRs. For example, the restriction enzyme EcoRI recognizes the inverted repeat GAATTC and cuts between the G and the adjoining A (the substring TTC when reversed and complemented GAA). A detail report about the roles of IR in human diseases is presented in [75].

Thus, it is important to detect the inverted repeat structure in a DNA sequence. A major difficulty in identification of repeats arises from the fact that the inverted repeats present in DNA sequence can be either exact or inexact, and are of unspecified length. The detection of exact inverted repeat is simple and can be achieved in linear time but the detecting an inexact inverted repeat has proven to be challenging task. One way to detect inverted repeats in a sequence is done using suffix technique [76]. This technique transforms the inverted repeat detection problem to finding longest common extension subsequence and solve exact or inexact repeat with fixed number of mismatches in linear time. However, the technique becomes both complex and inefficient for finding inexact inverted repeat without any prior knowledge of mismatches due to substitution or insertion/deletion of nucleotides.

Another technique for detecting approximate inverted repeats in nucleotide sequences is inverted repeat finder (IRF) [77]. The methodology for detecting inverted repeat using IRF is similar to the tandem repeat finder [54]. Tandem repeat finder (TRF) is a statistically based heuristic algorithm. The general approach of TRF is similar to BLAST algorithm. The program detects candidate IRs by finding short, exact, reverse complement matches of 4-7 nucleotides (k-tuples) between non-overlapping fragments of a sequence. A "center" position is defined for each k-tuple match. Inverted repeat finder (IRF) detects "clusters"of k-tuple matches having the same or nearly the same center and falling within a small interval of sequence. Candidate inverted repeats are confirmed (aligned and extended) or rejected by computing Smith-Waterman style similarity alignment. IRF run against a genome sequence using parameters match, mismatch, indel, and minimum score. Major drawback of this technique is the requirement of input parameters listed above. A user has to do many trial sessions specifying different set of values for these parameters in order to get a good match.

Einverted available at *http://bioweb.pasteur.fr/seqanal/interfaces/einverted.html* is another program which is used for finding inverted repeats in a DNA sequence. The algorithm is based on dynamic programming methodology. It works by finding alignments between the sequence and its reverse complement that exceeds a threshold score. Gaps and mismatches are assigned penalty (negative score). Matches are assigned a positive score. The score is calculated by summing the values of each match, the penalties of each mismatch and the large penalty of any gaps. Any region whose score exceeds the threshold value are reported.

In this chapter, algorithms for identifying inverted repeats based on efficient signal processing techniques are presented. The aim of this work is to provide easier techniques for identifying inverted repeats and to explore the possibility of applying signal processing tools for inverted repeat problem. Also, the aim of this work is to provide a complementary and easier approach for identifying inverted repeats.

4.2 Periodicity Transform

The periodicity transform [68] offers a technique which helps in detecting periodicities in a given sequence. This transform decomposes finite-duration sequences into a sum of periodic sequences by projecting it onto a set of periodic subspaces. The periodic subspaces are not orthogonal and hence the decomposition may not be unique.

4.2.1 Periodic subspaces

A sequence of real numbers $S(k)$, is said to be p-periodic if there is an integer p such that $S(k+p) = S(k)$ for all integers k. Let P_p be the set of all p-periodic sequences, and P be the set all periodic sequences. Consider a sequence $S \subset P$ containing N elements. This can be considered to be a single period of N elements, *i.e.*, $S \in P_N \subset P$, and the goal is to locate smaller periodicities within S. In order to locate smaller periodicities within S, the sequences must be projected on subspaces P_p for $p < N$. When S is close to some periodic subspace

P_p , then there exists a p-periodic sequence S_p which is closest to S. This S_p is said to be the ideal choice for decomposing S. For every period p and time shift s, we can define the sequence $\delta_p^s(j)$ for all integers j such that:

$$\delta_p^s(j) = \begin{cases} 1 & \text{if } (j-s) \bmod p = 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

The sequences δ_p^s for $s = 0, 1, \ldots, p-1$ are called p-periodic basis vectors since they form a basis for P_p. An inner product can be defined on the periodic subspaces as a $P \times P \rightarrow \Re$ function given by:

$$\langle S1, S2 \rangle = \lim_{k\to\infty} \frac{1}{2k+1} \sum_{i=-k}^{k} S1(i)S2(i) \tag{4.2}$$

for arbitrary elements $S1$ and $S2$ in P. If $S1 \in P_{p_1}$ and $S2 \in P_{p_2}$, then the product sequence $S1(i)S2(i) \in P_{p_1p_2}$ is P_1P_2-periodic, and the inner product satisfying the basic properties of inner product is given as:

$$\langle S1, S2 \rangle = \frac{1}{p_1p_2} \sum_{i=0}^{p_1p_2-1} S1(i)S2(i) \tag{4.3}$$

A sequence S is said to be orthogonal to the subspace P_p if $\langle S, S_p \rangle = 0, \ \forall \ S_p \in P_p$. Any two subspaces are orthogonal if every periodic basis vector in a subspace is orthogonal to every vector in the other subspace. However, it is important to note that the periodic subspaces P_p are not orthogonal to each other.

4.2.2 Projection onto periodic subspaces

Consider an arbitrary sequence $S \in P$. Then, a minimizing vector in P_p is defined as $S_p^* \in P_p$ such that:

$$||S - S_p^*|| \ \leq \ ||S - S_p|| \tag{4.4}$$

$\forall S_p \in P_p$. Thus, S_p^* is the p-periodic sequence that is closest to the original sequence S. The projection theorem [68] states that S_p^* can be characterized as an orthogonal projection of S onto P_p. Applying the projection theorem, $S_p^* \in P_p$ can be expressed as a linear combination of the periodic basis vectors δ_p^s (equation 4.1) as:

$$S_p^* = \alpha_0\delta_p^0 + \alpha_1\delta_p^1 + \quad \cdots \quad + \alpha_{p-1}\delta_p^{p-1} \tag{4.5}$$

where the unique minimizing vector is the orthogonal projection of S on P_p. Hence, $S - S_p^*$ is orthogonal to each of the periodic basis vectors δ_p^s for $s = 0, 1, \ldots, p-1$, *i.e.*,

$$\langle S - S_p^* \rangle = \langle S - \alpha_0\delta_p^0 - \alpha_1\delta_p^1 \cdots \alpha_{p-1}\delta_p^{p-1},\ \delta_p^s \rangle \ = \ 0 \tag{4.6}$$

After simplification, the coefficients α_s are obtained as:

$$\alpha_s = \ p\ \langle S, \delta_p^s \rangle \tag{4.7}$$

Since $S \in P$, it is periodic with some period N. From the definition of inner product in (equation 4.3), α_s are given by:

$$\alpha_s = p\frac{1}{pN}\sum_{j=0}^{pN-1} S(i)\delta_p^s(i) \tag{4.8}$$

However, $\delta_p^s(i)$ is zero except when $(s - i) \bmod p = 0$, and therefore, it simplifies to

$$\alpha_s = \frac{1}{N}\sum_{k=0}^{N-1} S(s + kp) \tag{4.9}$$

And if N/p is integer then this reduces to

$$\alpha_s = \frac{1}{N/p}\sum_{k=0}^{N/p-1} S(s + kp) \tag{4.10}$$

4.2.3 Periodicity measure for IR identification

The inverted repeat measure is based on the idea that if a given rearranged DNA sequence (where the second half of the DNA is complemented and reversed as shown in Figure 4.1) is found to be $\lfloor L/2 \rfloor$ periodic, where L is the length of the DNA sequence, then this shows that the DNA sequence is an inverted repeat. Let S be the given DNA sequence and let S' be the rearranged DNA sequence of S. And let $T_{\lfloor L/2 \rfloor}$ be the projected sequence of S' on $\lfloor L/2 \rfloor$ subspace, then the measure for inverted repeat is given by the following periodogram coefficient.

$$\lambda = \frac{||S'||^2}{||T_{\lfloor L/2 \rfloor}||^2} \tag{4.11}$$

where $||S'||^2$ and $||T_{\lfloor L/2 \rfloor}||^2$ are squared norm of sequences S' and $T_{\lfloor L/2 \rfloor}$. The value of λ is always less than or equal to unity. For the case of exact inverted repeat the value of λ is equal to unity. In section 4.4.1, an algorithm based on λ is presented to identify inverted repeats.

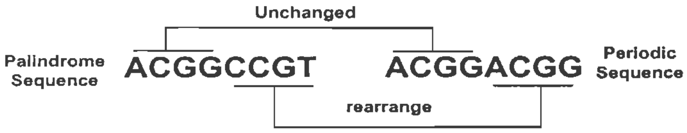


Figure 4.1: Palindrome sequence and period-2 sequence.

4.3 Correlation Function

Correlation is a mathematical tool to quantify the degree of interdependence of one data upon another. In other words, it is used to establish the similarity between one set of data and another. The process of correlation occupies a significant place in signal processing. Applications of correlation are found in image processing for robotic vision or remote sensing by satellite in which data from different images are compared, in detection and identification of noise, and in many other fields, such as, climatology. The correlation function can also be used for discovering weak periodic signals in time series data [78].

The correlation for N-point data is given as:

$$\rho_{f_1,f_2}[k] = \frac{1}{N}\sum_{n=1}^{N} f_1[n]f_2[n+k] \quad \text{where } k = 0, 1, \ldots, N-1. \tag{4.12}$$

where f_1 and f_2 are two functions for which the correlation is to be calculated. When $f_1[n] = f_2[n]$ then correlation is said to be auto-correlation and when $f_1[n] \neq f_2[n]$ then it is said to be cross-correlation.

The complexity of the correlation operation is $O(N^2)$ which is quite expensive, especially when dealing with very large sequences. The correlation computation may be speeded up by exploiting the correlation theorem, usually stated as

$$\rho_{F_1,F_2}[k] = \frac{1}{N}F^{-1}(F_1[n]F_2[n]) \quad \text{where } k = 0, 1, \ldots, N-1. \tag{4.13}$$

where $F_1[n]$ and $F_2[n]$ are Fourier transforms (FT) of $f_1[n]$ and $f_2[n]$ respectively and is the inverse Discrete Fourier transform (IDFT). This approach requires computation of two discrete Fourier transforms (DFTs) and one inverse DFT, each of which is most easily executed using the FFT algorithm in $O(N \log_2 N)$. If the sequence is sufficiently large, it is faster to use this FFT technique than to calculate the correlation directly.

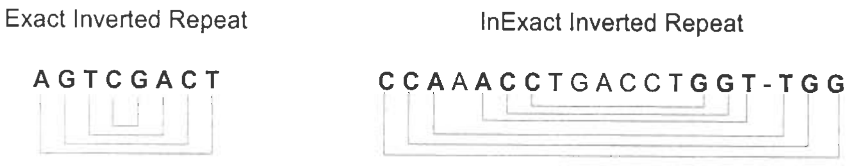


Figure 4.2: An exact and inexact inverted repeat.

4.3.1 Correlation measure for IR identification

Detection of an inverted repeat of a fixed size and fixed number of mismatches (due to substitution or insertion/deletion) is easy. However, detecting an inverted repeat without *apriori* knowledge of its length and number of mismatches is really a challenging task. An example of exact an inexact inverted repeat is shown in Figure 4.2. The two major objectives of any inverted repeat identification algorithm are: *length* and *location* or *position* of inverted repeats in a given sequence. In this section, a measure based on correlation function for identification of length and position of inverted repeats is presented.

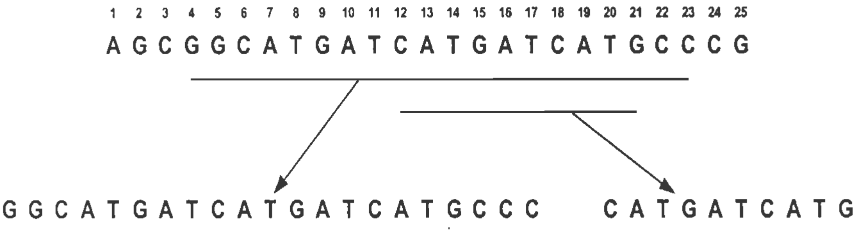
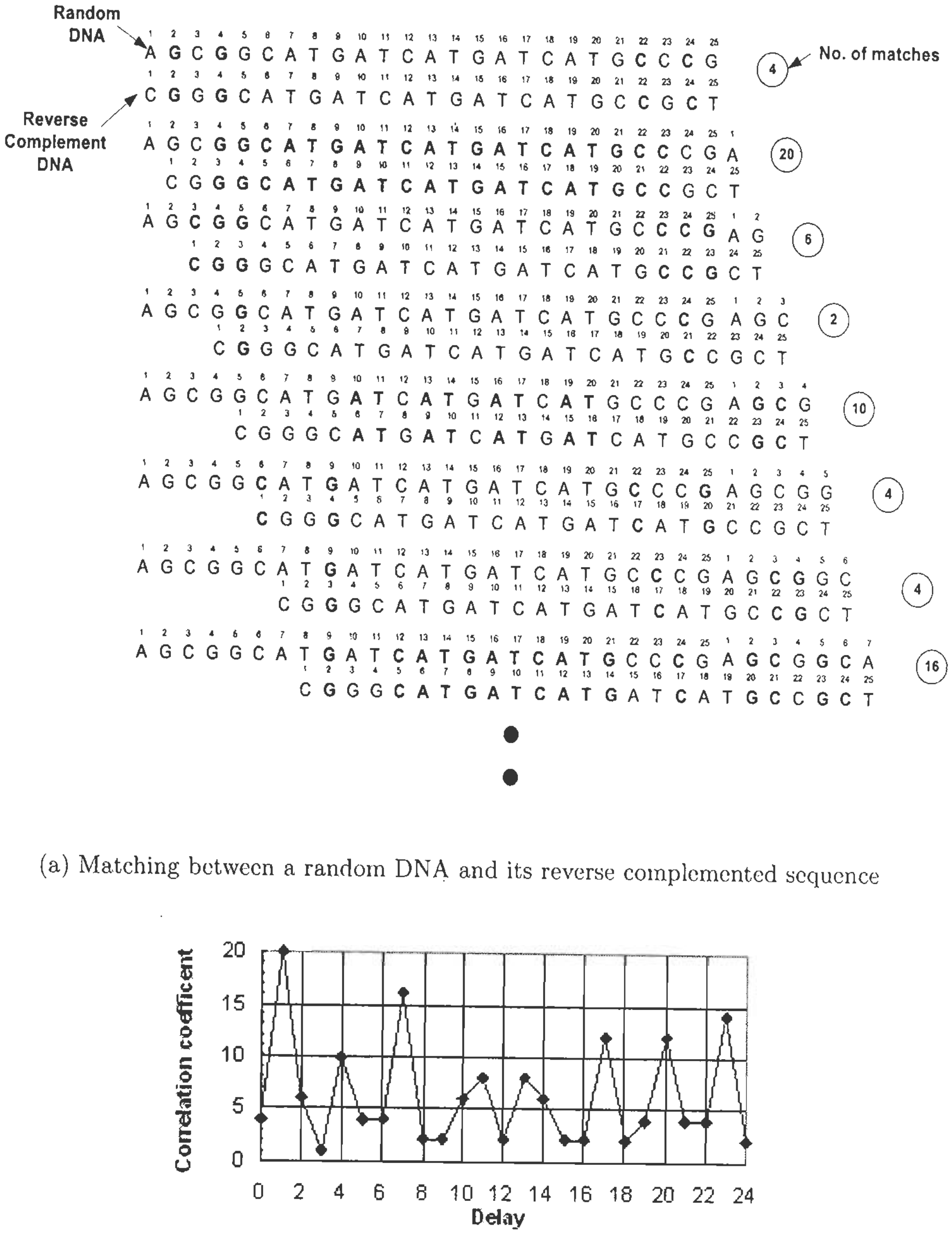


Figure 4.3: A random DNA sequence and inverted repeats.

Consider a random DNA sequence AGCGGCATGATCATGATCATGCCCG. The two main inverted repeats present in the random sequence are shown in Figure 4.3. The reverse complemented of the random DNA sequence is given by CGGGCATGATCATGATCATGCCGCT, where the DNA sequence is first



(a) Matching between a random DNA and its reverse complemented sequence

(b) Correlation coefficients graph

Figure 4.4: Correlation between a DNA and its reverse complemented sequence.

reversed and then the nucleotides are complemented, i.e., A→T, T→A, C→G and G→C. The number of matches that are obtained between the random DNA and its reverse complemented sequence (where the position of the DNA sequence is fixed and the reverse complemented is shifted by one position towards right) are shown in Figure 4.4(a). This matching process between the two DNA sequences is also called as cross correlation. The collection of cross correlation coefficients (or coefficients having number of matches between two sequences) between the random DNA and its reverse complemented sequences are shown in Figure 4.4(b). Note that the two maximum peaks in the graph occurs at delay (or shifting) = 1 and 7. The magnitude of peak in the graph is actually related to the length of IR and the delay or (shifting) is related to the start position of inverted repeat. The two maximum peak present in Figure 4.4(b) provide information about the major IR present in the DNA sequence and are shown in Figure 4.3. Hence, using correlation of a DNA sequence and its reverse complemented sequence, information about the position and length of the repeat, i.e., number of matches can be easy obtained.

A primary step before performing DFFT based correlation of DNA sequences is to assign some numeric values to the nucleotides. An arbitrary assignment of number to the nucleotides would not give a correct correlation measure between the DNA sequences. For example, let A=1, C=2, G=3, T=4, and 3 sequences be $S1$ = AACC, $S2$ = AACG, $S3$ = AACT. By observation, the correlation between $S1$ and $S2$ is equal to the correlation between $S1$ and $S3$, since they are off by one element. However, using the given numeric mapping of the sequence, the correlation coefficient between $S1$ and $S2$ is 0.9, $S1$ and $S3$ is 0.82. So, in order to obtain a correlation measure for which the correlation coefficient is dependent on the sequences but not on the numeric mapping, a binary assignment of the

DNA sequence is used [29, 7]. Consider a DNA sequence S of length L given by,

$$S = s_1 s_2 \dots s_{L-1} s_L \tag{4.14}$$

And, nucleotide mapping for the DNA sequence is given by,

$$S_\Omega[n] = \begin{cases} 1, & \text{if } S[n] = \Omega \quad \text{where } \Omega \in \sum (= \{\text{A,C,G,T,U}\}) \\ 0, & \text{otherwise} \end{cases} \tag{4.15}$$

In this way, the original DNA sequence is decomposed into four binary sequences. The decomposition of the DNA sequence into binary sequences helps in obtaining correlation coefficients which provides the actual level of correlation between the two DNA sequences. Let 'I' represent the reverse complemented sequence of S, then the correlation between the two sequences is calculated as follows:

$$\rho_{S,I}[k] = \sum\nolimits_\Omega \rho_{S_\Omega, I_\Omega}[k] \qquad 0 \leq k \leq L-1 \tag{4.16}$$

$$\rho_{S_\Omega, I_\Omega}[k] = \sum\nolimits_{i=0}^{L-1} S_\Omega[i] I_\Omega[i+k] \qquad \text{where } I_\Omega[i] = I_\Omega[i+L] \tag{4.17}$$

where $\rho_{S_\Omega, I_\Omega}[k]$ is an array of correlation coefficients. The magnitude of correlation coefficient provide information about the number of matches and k provides information about the location of IR. The proposed algorithm in the section 4.4.2 for inverted repeat identification uses correlation coefficients as a measure for identification of inverted repeats in a given DNA sequence.

4.4 Inverted Repeat Identification Algorithms

4.4.1 Algorithm 1

In this algorithm, a periodogram coefficient based measure as defined in section 4.2.3 is applied for identification of exact and inexact repeat due to substitution. The algorithm is divided into three major stages and is given as follows:

Preprocessing: The nucleotides of S are assigned a complex value. This step is very important because signal processing techniques deal only with real or complex values. Hence, mapping symbols to numeric values is a necessary step before the signal processing techniques can be applied. The mapping used in this paper is as follows:

$$\mathrm{A} \rightarrow 1+j, \quad \mathrm{T} \rightarrow -1-j, \quad \mathrm{C} \rightarrow 1-j, \quad \mathrm{G} \rightarrow -1+j \tag{4.18}$$

The nucleotides are mapped to numbers of same magnitude because of the algorithm requirement. The mapping is independent of the order in which the numbers are assigned to nucleotides. If S is of odd length then the central character from S is removed and length of S is decreased by 1. Now, the rearrangement of S is done as shown in Figure 4.1 and S' represents the rearranged sequence.

Calculation of L/2-periodic sequence: In this stage, period length, p is set to half-length S, i.e., $L/2$. Then periodicity transform is used to calculate the L/2-periodic sequence, called $T_{L/2}$, which is closest to S.

Calculation of Periodogram Coefficient: Periodogram coefficient as defined in (equation 4.11) is calculated to identify exact inverted repeat. However, to detect inexact inverted repeat due to substitution, the equation 4.11 requires

Table 4.1: Average of the sum of nucleotides.

	A	C	G	T
A	$1+j$	1	j	0
C	1	$1-j$	0	$-j$
G	j	0	$-1+j$	-1
T	0	$-j$	-1	$-1-j$

modification. For instance, if we calculate the periodogram coefficient of the strings GCATTTGC, GCAGTTGC, GCGGTTGC, we get 0.75, 0.875, 0.75, respectively, using (equation 4.11). The mismatch subsequences in the above strings are denoted by underlining them. The string GCATTTGC has one mismatch and GCGGTTGC has two mismatches, yet the same value of 0.75 is obtained for both of these strings. Thus, the straightforward formulation of periodogram coefficient does not provide a direct correlation between the number of mismatches and the observed values.

Let us understand the reasons for this ambiguity. We had earlier used a mapping of each DNA nucleotide to a complex number, i.e., A= $1 + j$, C = $1 - j$, G= $-1 + j$ and T = $-1 - j$. We observe that the average sum of any two nucleotides and its norm (which is used in equation 4.11) are not the same for different sets of nucleotides, thereby leading to the above ambiguity. In Table 4.1, we show the average sum of any two nucleotides. For example, the average of nucleotide A with nucleotide A (represented as column 1 and row 1 of Table 4.1) would be $((1+j)+(1+j))/2 = 1+j$; and the average of nucleotides A and G would be (row 1 and column 3 of Table 4.1) $((1+j)+(-1+j))/2 = j$. Table 4.2 shows the squared norm of the entries in Table 4.1. That is, the first entry is $||1+j||^2$ which is the squared norm of $(1+j)$. The calculation

Table 4.2: Square of the norm of Table 4.1.

	A	C	G	T
A	2	1	1	0
C	1	2	0	1
G	1	0	2	1
T	0	1	1	2

of the periodogram coefficient must take into account the varying contributions from nucleotide pairs as depicted in Table 4.2. The squared norm of each pair of nucleotides can have a magnitude of 0, 1, or 2. In case of a match, the magnitude contribution is 2, whereas in case of a mismatch, the magnitude contribution can be either 0 or 1. Next, we analyze the result of these varying contributions on the periodogram coefficient. Let L be the string length, α be the number of matches, β be the number of mismatches that have magnitude 1, and σ be the number of mismatches that have magnitude 0. Then the periodogram coefficient can be computed as (using equation 4.11):

$$\begin{aligned} \lambda &= \frac{2*(2*\alpha)+1*(2*\beta)+0*(2*\sigma)}{2*L} \\ &= \frac{L-2*(\beta+\sigma)+\beta}{L}, \\ &= 1-\frac{\beta+2*\sigma}{L} \quad = 1-\frac{k+\sigma}{L} \end{aligned} \tag{4.19}$$

Note that (equation 4.19) is dependent on σ, and is not simply a function of the number of mismatches k (where $k=\beta+\sigma$). It can be seen from Table 4.3 that for k number of mismatches we obtain different values for periodogram coefficient because of the dependence of (equation 4.19) on σ. In order to make the periodogram coefficient independent of σ, we use the following compensated periodogram coefficient taking into account the nature of the nucleotide pair.

Table 4.3: Ambiguous variation of λ with the number of mismatches.

String length(L)	# Match(α)	# Mismatch(k)		Periodogram coeff. (λ)
		$k = \beta + \sigma$		
		β	σ	
2	0	1	0	0.5
		0	1	0.0
4	1	1	0	0.75
		0	1	0.5
	0	2	0	0.5
		0	2	0.0
		1	1	0.25

Let α be the number of matches and k be the total number of mismatches. The contribution of a match is two, and that of a mismatch is one, then the expression for compensated periodogram coefficient is obtained as follows:

$$\begin{aligned}\lambda_m &= \frac{2*(2*\alpha)+1*(2*\beta)+1*(2*\sigma)}{2*L} \\ &= \frac{L-2*(\beta+\sigma)+(\beta+\sigma)}{L} \\ &= 1-\frac{\beta+\sigma}{L} \quad = 1-\frac{k}{L}\end{aligned} \tag{4.20}$$

From (equation 4.20) it is observed that the compensated periodogram coefficient is dependent on total number of mismatches and the length of palindrome. Hence, we can obtain an unique value of compensated periodogram coefficient for k number of total mismatches.

4.4.2 Algorithm 2

The inverted repeat detection algorithm operates in two stages. The first stage identifies small contiguous inverted repeats position and its length in the DNA sequence. A contiguous inverted repeat (exact inverted repeat) is represented by tuple $< X, Y, l >$ where $(X,\ Y)$ represents a pair of coordinates revealing the position of the repeat in the genome sequence and l is the length of the repeat. The second phase of the algorithm merges the small contiguous inverted repeats to obtain inexact inverted repeats present in the DNA sequence. An inexact inverted repeats may consists of many small contiguous inverted repeats like the one shown in the Figure 4.3. The inexact repeat **CCAACC**TGACCT**GGTTGG** is formed by merging two contiguous repeats (**CCA**, **TGG**) and (**ACC**, **GGT**) represented by tuples $< 1, 19, 3 >$ and $< 5, 16, 3 >$ respectively.

Inputs: A DNA sequence (S), Minimum length of contiguous repeat (L_{min}), Window length (W).

Preprocessing Stage: In this step, four binary sequences (consisting of 0s and 1s) are constructed each for the input DNA sequence and its reverse complemented sequence as discussed in the section 4.3.1.

Identification of contiguous inverted repeat sequence: The major difficulties while detecting an inverted repeat in DNA sequence are the length of the repeat and the position of such repeat in the DNA sequence. As discussed in section 4.3.1, the delay parameter in the correlation operation gives the location of inverted repeat and the value of correlation provide the number of matches, which can be used in finding the length of the inverted repeat. An exact repeat consists of a single continuous repeat, however an inexact repeat may consists of many small contiguous repeats. In this stage, the algorithm identify the length and position of contiguous inverted repeat in the given DNA sequence. A pseudo code of this stage is provided in the PSEUDO CODE 1.

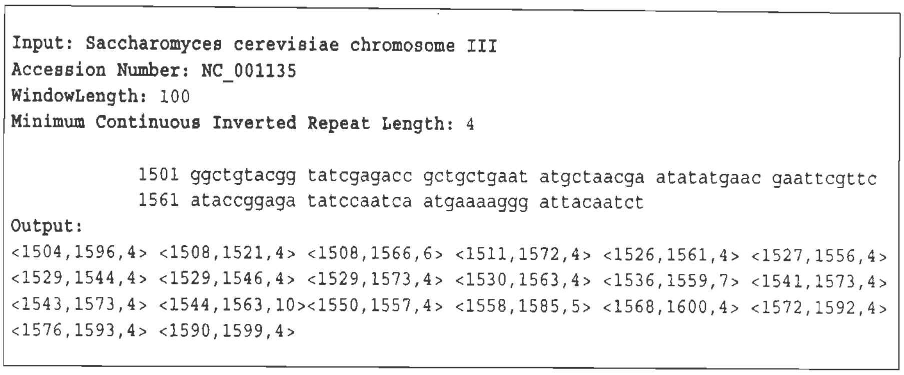
The identification process is based on dividing the given DNA sequence into small subsequence until some stopping criteria is met. One of the important stopping criteria while searching for inverted repeats is based on the count of nucleotides A, C, G, and T in the DNA sequence. The $MaxMatch$ variable represents the maximum length of continuous inverted repeat sequence that can be present in the DNA sequence and is sum of min(N_{A}, N_{T}) and min(N_{C} , N_{G}), where N_A, N_T, N_C, N_G are the counts of nucleotides A, C, G, T in the DNA sequence between $Start$ and End position. For example, for AGCGGCATGAT-CATGATCATGCCCG. N_A=6, N_T =5, N_C =7, N_G=7 and $MaxMatch$ = 12 which mean at maximum, there can be a continuous inverted repeat of length 12 in the DNA sequence. Thus, for any DNA sequence if it is found that $MaxMatch$ is less than L_{min} which is provided by the user can be straight away rejected and hence reducing our inverted repeat search cases. After all stopping criteria for the DNA sequence fails, a search is made for an exact contiguous inverted repeat. If length of the length satisfies the minimum matching length criteria (L_{min}) then its position and length is recorded otherwise a further search for inverted repeat is made in the DNA sequence.

The identification of the position and length of the inverted repeat is based on the value of the correlation coefficients that is obtained after performing a correlation between the DNA sequence and its reverse complemented DNA sequence. The delay parameter of the correlation tells the location of inverted repeat in the sequence and the value of correlation is directly related to the length of the inverted repeat. The list of all contiguous repeats that were identified in a portion of DNA sequence of *Saccharomyces cerevisiae* chromosome III is provided in Figure 4.5.

PSEUDO CODE 1: FIND CONT IREPEAT(S, $Start$, End, L_{min})

```
begin
if CHECK FLAG($Start$,$End$) = TRUE then
    return
if ($End$-$Start$+1) < 2*$L_{min}$ then
    return
$N_A \leftarrow$ No. of A's in $S[Start \ldots End]$
$N_C \leftarrow$ No. of C's in $S[Start \ldots End]$
$N_G \leftarrow$ No. of G's in $S[Start \ldots End]$
$N_T \leftarrow$ No. of T's in $S[Start \ldots End]$
$MaxMatch \leftarrow \min(N_A, N_T) + \min(N_C, N_G)$
SET FLAG($Start$,$End$) = TRUE
if ($MaxMatch < L_{min}$) then
    return
$i \leftarrow start$, $j \leftarrow end$, $TotalMatch \leftarrow 0$
while ($i < j$) and ($TotalMatch < MaxMatch$) and $S[i] = S[j]$ do
    $i \leftarrow i + 1$, $j \leftarrow j - 1$
    $TotalMatch \leftarrow TotalMatch + 1$
if ($TotalMatch \geq L_{min}$) then
    OUTPUT($Start$, $End$, $TotalMatch$)
    $Start \leftarrow Start + TotalMatch$,  $End \leftarrow End - TotalMatch$
    if ($TotalMatch < MaxMatch$) then
```

```
            FIND CONT IREPEAT($S$, $Start$, $End$, $L_{min}$)
        return
    else $Corr$ = FIND CORRELATION ($S$, $I$, $Start$, $End$)
        $i \leftarrow 0$
        while ($i < \lfloor WindowLength \rfloor$) do
            if ($Corr[i] \geq 2 * L_{min}$) then
                FIND CONT IREPEAT($S$, $Start$, $Start + i$, $L_{min}$)
                FIND CONT IREPEAT($S$, $Start + i + 1$, $End$, $L_{min}$)
                $i \leftarrow i + 1$
        return
end
```



```
Input: Saccharomyces cerevisiae chromosome III
Accession Number: NC_001135
WindowLength: 100
Minimum Continuous Inverted Repeat Length: 4

           1501 ggctgtacgg tatcgagacc gctgctgaat atgctaacga atatatgaac gaattcgttc
           1561 ataccggaga tatccaatca atgaaaaggg attacaatct
Output:
<1504,1596,4> <1508,1521,4> <1508,1566,6> <1511,1572,4> <1526,1561,4> <1527,1556,4>
<1529,1544,4> <1529,1546,4> <1529,1573,4> <1530,1563,4> <1536,1559,7> <1541,1573,4>
<1543,1573,4> <1544,1563,10><1550,1557,4> <1558,1585,5> <1568,1600,4> <1572,1592,4>
<1576,1593,4> <1590,1599,4>
```

Figure 4.5: A list of exact contiguous inverted repeat identified in the DNA sequence of *Saccharomyces cerevisiae* chromosome III between 1501 and 1600.

Merging of contiguous inverted repeats: In this stage, the contiguous inverted repeats that are present in the same window are merged together in order to form inexact inverted repeats. The output from previous stage consists of a list of tuples $< X, Y, l >$. Two tuples $< X1, Y1, l1 >$ and $< X2, Y2, l2 >$ can be merged only if the following condition holds true:

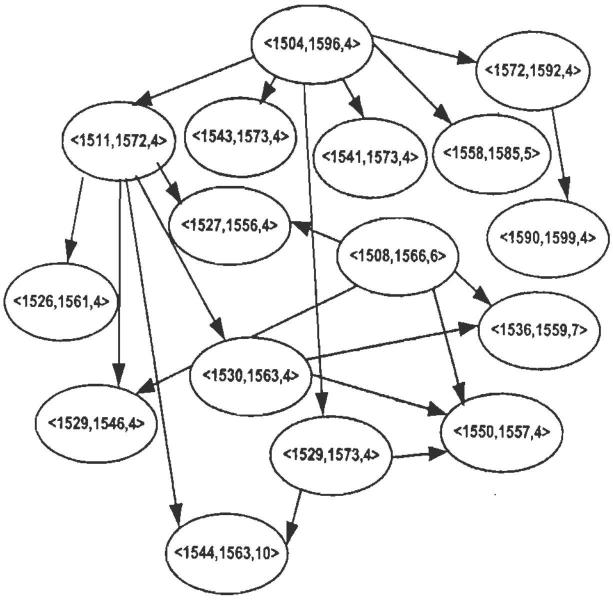
$$X2 \geq (X1 + l1) \quad and \quad Y2 \leq Y1 - l1, \quad where \quad X1 < X2 \tag{4.21}$$

For example, ACGGATATGT have contiguous inverted repeat as $< 1, 10, 2 >$ i.e., AC $- - - - - -$ GT and $< 5, 8, 2 >$ i.e., ATAT, so both can be combined according to the above rule to obtain an inverted repeat as AC$- -$ATAT$- -$ GT. An acyclic graph is constructed in order to obtain inexact inverted repeats from a list of contiguous inverted repeats generated in the previous stage. The nodes of the acyclic graph are labeled as $< X, Y, l >$. An edge is created from node $N1 \equiv < X1, Y1, l1 >$ to node $N2 \equiv < X2, Y2, l2 >$ if and only if the condition provided in (equation 4.21) holds true.

After the construction of graph is completed, a topological sorting of the acyclic graph is done. The sorting may result in various paths and each such path forms an inverted repeat of the DNA sequence. For selecting the starting node for topological sorting, the following conditions must be satisfied:

- starting node must be non-traversed node.
- if $P \equiv < X1, Y1, l >$ is selected as a starting node then $X1$ must be the smallest starting location from the set of non-traversed node and $Y1$ must be farthest among all nodes that are starting from $X1$.

After reaching an end node, all the nodes of the current path are displayed in the order they were visited. Each such path obtained forms an inverted repeat of the input DNA sequence. Figure 4.6 shows the acyclic graph constructed out of the contiguous inverted repeat provided in Figure 4.5. The inverted repeat which have the highest number of matching is $< 1504, 1596, 4 >< 1511, 1572, 4 >< 1530, 1563, 4 >< 1536, 1559, 7 >$ and the inverted repeat is **TGAC** $- - - - - - - - - - - - - - - - - - -$ GG $-$ **TATC**GAGACCGCTGCTGAA**TATG**CT**AACGAAT**



List of Inexact inverted repeat in the DNA sequence

<1504, 1596,4> <1511,1572,4> <1526,1561,4>
<1504, 1596,4> <1511,1572,4> <1529,1546,4>
<1504, 1596,4> <1511,1572,4> <1530,1563,4><1550,1557,4>
<1504, 1596,4> <1511,1572,4> <1530,1563,4><1536,1559,7>
<1504, 1596,4> <1511,1572,4> <1527,1556,4>
<1504, 1596,4> <1511,1572,4> <1544,1563,10>
<1504, 1596,4> <1543,1573,4>
<1504, 1596,4> <1541,1573,4>
<1504, 1596,4> <1558,1585,4>
<1504, 1596,4> <1572,1592,4> <1590,1599,4>
<1504, 1596,4> <1529,1573,4> <1544,1563,10>
<1504, 1596,4> <1529,1573,4> <1550,1557,4>
<1508, 1566,4> <1527,1556,4>
<1508, 1566,4> <1529,1546,4>
<1508, 1566,4> <1550,1557,4>
<1508, 1566,4> <1536,1559,7>

Figure 4.6: A list of inexact inverted repeats detected in the DNA sequence of *Saccharomyces cerevisiae* chromosome III between 1501 and 1600 base pair.

ATATGAACGA**ATTCGTT** - - **CATA**CCGGA - - - - - - - - - - - **GATA**TCCA ATCAATGAAAAGGGAT**TACA**.

4.5 Experimental Results

To demonstrate the capabilities of the inverted repeat detection algorithm, experiments were performed on several actual DNA sequences available on public databases. In order to demonstrate the working of the inverted repeat algorithm, the result of a test performed on a DNA sequence is as follows:

Escherichia_coli_O157:H7.trna74

Location: (3542018–3541942), Length: 77 bp

Sequence: GCATCCGTAGCTCAGCTGGATAGAGTACTCGGCTACGAACC-GAGCGGTCGGAGGTTCGAATCCTCCCGGATGCACCA

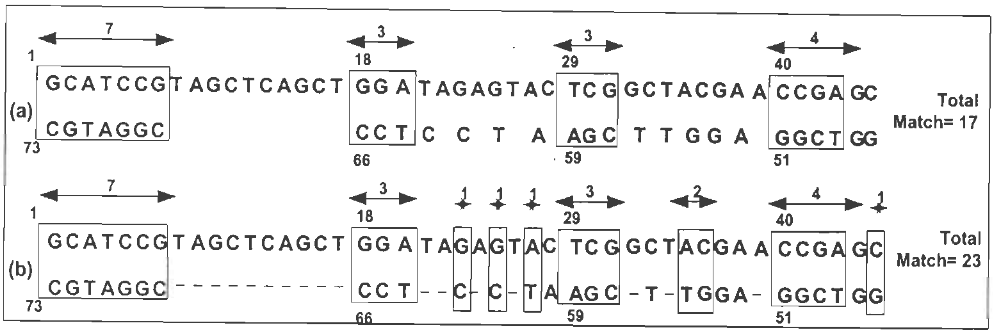

7
3
3
4
1
18
29
40
GCATCCGTAGCTCAGCTGGATAGAGTACTCGGCTACGAACCGAGC
(a)
CGTAGGC
CCTC C T A AGC T T G G A GGCTGG
73
66
59
51
Total Match= 17
7
3
1
1
1
3
2
4
1
1
18
29
40
GCATCCGTAGCTCAGCTGGATAGAGTACTCGGCTACGAACCGAGC
(b)
CGTAGGC-----------CCT--C-C-TAAGC-T-TGGA-GGCTGG
73
66
59
51
Total Match= 23

Figure 4.7: (a) One of the maximal inverted repeat given by the algorithm at the end of second stage of the algorithm. The contiguous inverted repeats are shown in rectangular boxes. The number above the box shows the position of the repeat in the sequence, and the number above the arrow denotes the length of inverted repeat. Total number of matches in the sequence is 17. Applying a global sequence alignment algorithm on the region where no repeat was reported, the number of matches has increased to 23. The dark boxes in (b) shows additional matches detected.

For the experiment minimum length of continuous inverted repeat is taken as 3 and the window length is taken as 100. At the end of second stage, 36 inexact inverted repeats were reported from the sequence. One of the inexact inverted repeat obtained from Escherichia_coli_O157:H7.trna74 at the end is the following:

$< 1, 73, 7 >< 18, 66, 3 >< 29, 59, 3 >< 40, 51, 4 >$

This inexact inverted repeat is shown in Figure 4.7. The contiguous inverted repeats identified by the algorithm are shown in the rectangular boxes. The position of the contiguous inverted regions are (1, 73), (18, 66), (29, 59), and (40, 51) and the corresponding lengths are 7, 3, 3 and 4 respectively. The length of the contiguous repeat is greater than or equal to 3. This is because the minimum number of contiguous repeat for the DNA sequence was taken as 3. The inverted repeat provided by the inverted repeat detection algorithm consists of matched (shown in rectangular box) and unmatched region. The region between two contiguous inverted repeat matches i.e., unmatched region is further processed in order to obtain a maximal repeat. The maximal repeat of the DNA sequence is provided in Figure 4.7. The unmatched regions are aligned using a global alignment technique.

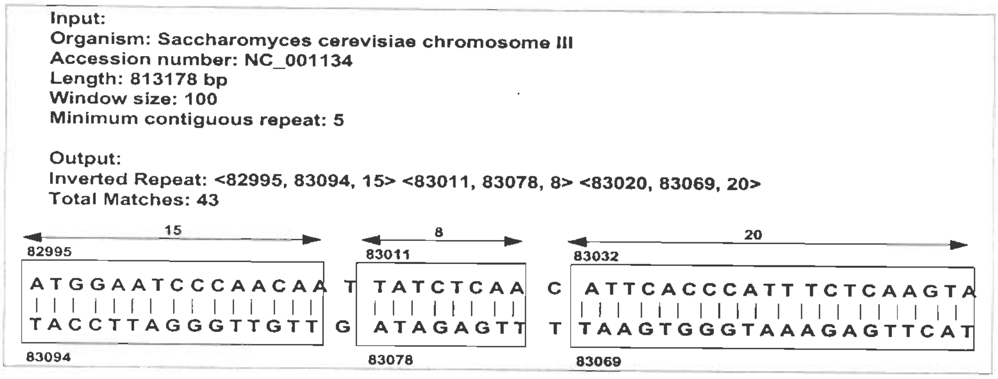
The inverted repeat detection algorithm was applied on various chromosomes of *Saccharomyces cerevisiae* (baker's yeast) genome data available at NCBI website *http://www.ncbi.nlm.nih.gov.*

***Saccharomyces cerevisiae* chromosome III**: A detailed test was performed for different window sizes and minimum contiguous repeat length. A typical result is shown in Figure 4.8. The inverted repeat shown in the figure was reported when applied to chromosome III of *Saccharomyces cerevisiae* with window size equal to 100 and minimum continuous repeat length as 5. Total number of matches in the reported inverted repeat was 43. As the window size was increased to 300 the length of the inverted repeat reported in the Figure 4.8 was increased. The inverted repeat is given by:

$$< 82899, 83191, 11 >< 82914, 83176, 24 >< 82939, 83151, 12 >< 82952, 83138, 5 >$$
$$< 82958, 83132, 23 >< 82981, 83108, 29 >< 83011, 83078, 8 >< 83020, 83069, 20 >$$

TATGTAGAAAT ATAG **ATTCCATTTTGAGGATTCCTATAT** C **CTC**
GAGGAGAAC T **TCTAG** T **ATATTCTGTATACCTAATATTAT** - **AG**
CCTTTATCAATGGAATCCCAACAA T **TATCTCAA** C **ATTCACCC**

ATTTCTCAAGTA CTATTCATCT **TACTTGAGAAATGGGTGAAT** T **TTGAGATA** G **TTGTTGGGATTCCATTGTTGATAAAGGCT** A **ATAATATTAGGTATACAGAATAT** G **CTAGA** G **GTTCTCCTCGAG** C **ATATAGGAATCCTAAAATGGAAT** TAGC **ATTTCTACATA**



```
Input:
Organism: Saccharomyces cerevisiae chromosome III
Accession number: NC_001134
Length: 813178 bp
Window size: 100
Minimum contiguous repeat: 5

Output:
Inverted Repeat: <82995, 83094, 15> <83011, 83078, 8> <83020, 83069, 20>
Total Matches: 43
```

Figure 4.8: An inverted repeat reported by the algorithm in chromosome III of *Saccharomyces cerevisiae* DNA sequence.

The starting location of the inverted repeat in the DNA sequence is 82899 and the length is equal to 293. The inverted repeat is obtained by merging eight contiguous inverted repeats of length 11, 24, 12, 5, 23, 29, 8 and 20. The contiguous repeats are shown in bold and are also underlined. The total number of matches in the inverted repeat sequence is 132.

***Saccharomyces cerevisiae* chromosome IV**: One of the maximal inverted repeat reported by the algorithm for window size=100 and minimum contiguous repeat length as 5 is the following:

$< 307956, 308027, 5 >< 307962, 308021, 5 >< 307982, 308010, 12 >$

The inverted repeat is formed by merging 3 contiguous repeats of length 5, 5 and 12. The starting location of the inverted repeat is 307956. The total number of matches in the inverted repeat sequence is 22.

***Saccharomyces cerevisiae* chromosome VIII**: One of the maximal in-

verted repeat reported by the algorithm for window size=100 and minimum contiguous repeat length as 5 is the following:
$< 4147, 4238, 5 >< 4164, 4232, 10 >< 4176, 4203, 6 >$

The inverted repeat is formed by merging 3 contiguous repeats of length 5, 10 and 6. The starting location of the inverted repeat is 4147 in the DNA sequence. The total number of matches in the inverted repeat sequence is 21.

4.6 Conclusion

Identification of inverted repeats and especially inexact inverted repeats in a DNA has remained one of the challenging problem in DNA sequence analysis. Most of the existing methods for inverted repeat identification are either very difficult to handle or inefficient in identifying inexact inverted repeat. Also, till now, there does not exist any signal processing framework for identifying IRs. The objectives of this chapter was to introduce an easier, sensitive and yet efficient signal processing approach for IRs identification. Based on fast correlation technique and periodicity measure, algorithms are presented for identifying both exact and inexact IRs. Additionally, the algorithm require the user to input only two parameters: maximum inverted repeat size or window length and minimum length of contiguous repeat. Experimental results of IRs detection algorithms show the effectiveness of the proposed techniques.

Chapter 5

Correlation Measure for RNA Secondary Structure Prediction

Predicting the secondary structure of a RNA molecule from the knowledge of its primary structure is a challenging task. The function of many RNA molecules depends crucially on their structure. This chapter presents a novel signal processing based framework for predicting the secondary structure of a RNA molecule from its primary sequence. Correlation function is applied for finding base pairing regions of RNA by quantifying the degree of matching between RNA sequence and its reverse complemented sequence.

The advantage of this framework is that it requires only two input parameters and does not need the user to be an expert in using the program. Additionally, the framework gives a list of probable RNA secondary structures which are further processed by free energy algorithm to estimate the structure with minimum free energy of the RNA molecule. Experiments conducted over tRNA and tmRNA sequences demonstrate the effectiveness of the framework.

5.1 Introduction

A ribonucleic acid (RNA) molecule consists of a chain of nucleotides. Each nucleotide is comprised of a base, a phosphate group and a sugar group. The nucleotides differ only because of the base involved. There are four choices for the base, namely Adenine (A), Cytosine (C), Guanine (G) and Uracil (U). The specific sequence of the bases along the chain is called primary structure of the molecule. A secondary structure for a RNA molecule is simply a set of pairing interaction between bases in the molecule. Each base can be paired with at most one other base.

RNA molecules have a large number of functions in the cell, which often depend on its special structural properties. RNA molecules plays important role in a variety of important biological processes that include catalysis, RNA splicing, regulation of transcription, translation, and RNA-DNA, RNA-RNA and RNA-protein interactions. The function of an RNA molecule is determined by its structure. The formation of RNA structure is hierarchical. The primary structure, which is the sequence of nucleotides form the first level of organization. At the next level is the secondary structure, the sum of canonical (AU, CG and UG) base pairs. And there is a tertiary structure, which is the three-dimensional arrangement of bases and the quaternary structure is the interaction of RNA with other molecules, which are proteins, RNA or DNA molecules. To a large extend the structure of RNA molecule is determined by its secondary structure. Thus it is very important to predict the secondary structure of a RNA molecule to know its tertiary and quaternary structure. The experimental approaches for the discovery of structure are very expensive and time consuming, and thus computational methods are followed to predict the structure of RNA from the primary structure of the molecule.

The computational algorithms for predicting RNA secondary structure is primarily divided into two categories: phylogenetic comparison/covariance methods [79, 80] and minimum free energy methods [81, 82, 83]. The former methods start from the assumption that structure is much more conserved than sequence during evolution. Base pairs are inferred by finding positions in aligned sequences that co-vary so as to conserve base pairing potential. Comparative analysis is quite robust when a number of homologous sequences are available. Over 97% of base-pairs predicted for ribosomal RNA were demonstrated in subsequent crystal structures [84]. In contrast, free energy minimization methods require only a single sequence and proceed automatically without the labor-intensive steps of iterative alignment and base pair detection that comparative sequence analysis requires. Free energy minimization is the most popular method for the prediction of RNA secondary structure. The energy minimization algorithms are based on the hypothesis that a RNA molecule forms a structure that has minimum free energy. The free energy of the RNA molecule is calculated based on the summation of the free energy of all the loops in it [85]. Also, many tools have been developed that use a combination of minimum free energy and a covariation score [86, 87] or probabilistic models compiled from large reference data-sets [88, 79].

The current available programs require a number of parameters from user to predict the structure of the RNA molecule. The contribution of this study is to provide a signal processing based technique to predict the secondary structure of RNA molecule. The proposed algorithm takes mainly two input parameters: RNA sequence and the minimum number of continuous base pairs. These parameters, unlike in other programs, do not require an expert user who understands the inner details of the system. A novel correlation measure is defined later in this paper for identifying the regions in RNA sequence with high base pairing. The algorithm tested over a number of RNA sequences and the predicted structure is compared with mfold server (*http://www.bioinfo.rpi.edu/applications/mfold/old/rna*).

5.2 Correlation measure for RNA secondary structure prediction

The secondary structure of a RNA molecule is the collection of base pairs that occur in its 3-dimensional structure. The base pairing in RNA sequence leads to formation of various types of loops (hairpin loop, internal loop, stack/stem loop, multi-branched loop) in a RNA secondary structure. A RNA sequence is represented as $R = r_1, r_2, r_3, \ldots, r_n$ where r_i is called the ith nucleotide. Each r_i belongs to the set A, C, G, U. The pairs A–U and C–G are called Watson-Crick pairs and G–U is called wobble pair. A secondary structure or folding, on R is a set S of ordered pairs r_i, r_j written as i–j, satisfying the following:

1. The distance between the bases in the pairs (i–j) should be greater than 3.

2. If i–j and i'–j' are 2 different base pairs, ($i \leq i'$), then either

   - $i \leq j \leq i' \leq j'$ (i–j precedes i'–j'), or
   - $i \leq i' \leq j' \leq j$ (i–j includes i'–j').

For example, the secondary structure of a random RNA molecule shown in Figure 5.1 is written as $S \equiv$\{(2,53,3), (5,49,2), (8,31,4), (14,26,4), (35,45,3)\}.

Each secondary structure of a RNA sequence has some Gibbs free energy $\Delta G(S)$ associated with it [81, 82]. For pseudoknot-free secondary structures,

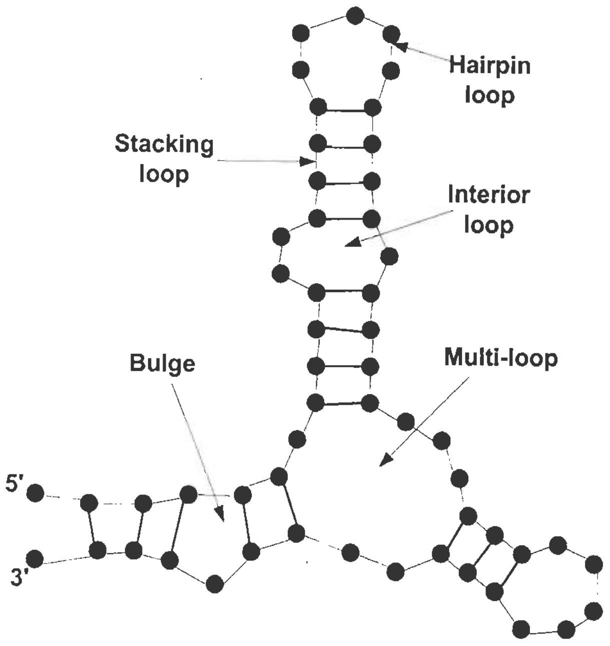


Figure 5.1: Secondary structure of a psuedo RNA molecule. The thick line stands for the backbone of the molecule and thick line stand for base pairings. The solid dots represent monomers $5'$ and $3'$ show the head and tail of this RNA of length 54. Many different loops formed when RNA folds are also shown in the figure.

this is typically calculated as the sum of the free energies of all loops. The Gibbs free energy is commonly used to describe the secondary structure since it contains entropic contributions from the formation of base pairs. The total free energy is the sum of the energy contributed from each elementary piece such as the stacking of base pairs and the connecting loops.

The RNA secondary structure prediction problem can be stated as follows: Given a RNA sequence R, let Ω be a set of secondary structure, and let $G(S)$ be the free energy of S, where $S \in \Omega$, then the objective of prediction algorithm is to find S' such that $G(S')$ is minimum. In the proposed algorithm, the secondary structure of a RNA is predicted by first maximizing base pairing and later on minimizing free energy of the RNA molecule. The base pairing regions are identified based on the correlation framework as discussed in section 4.3.1. The correlation coefficients for Watson-Crick and Wobble pairs in a RNA sequence is obtained using (equation 4.12 and equation 4.13) and is given by,

$$\begin{aligned}\rho_{R,I}[k] &= \rho_{R_A,I_A}[k] + \rho_{R_C,I_C}[k] + \rho_{R_G,I_G}[k] + \\ &\quad \rho_{R_U,I_U}[k] + \rho_{R_G,I_A}[k] + \rho_{R_U,I_C}[k], \quad 1 \leq k \leq L \end{aligned} \tag{5.1}$$

where R is a primary RNA sequence, I is the reverse complemented sequence of S and L is the length of the RNA sequence. R_A, R_C, R_G and R_U are the binary sequences of R obtained using (eqution 4.15). Similarly, I_A, I_C, I_G and I_U are the binary sequences of I.

The correlation measure used by the proposed algorithm for identifying continuous Watson-Crick and wobble base pairs is given by,

$$\mu_{R,I} = \frac{1}{N}\rho_{R,I}[k], \quad 1 \leq k \leq L \tag{5.2}$$

5.3 RNA Secondary Structure Prediction Algorithm

The algorithm is divided into three different stages. The complete framework of the algorithm is shown in Figure 5.2. The algorithm starts with nucleotide mapping which form the preprocessing step of the algorithm. Correlation coefficients for the input RNA sequence (R) are obtained by processing the binary mapped sequences. These coefficients are further processed and the regions having high base pairs in the RNA sequence are identified. Some of these regions are merged together to construct secondary structures for the input RNA sequence.

The algorithm provides a list of secondary structures and the structure that has minimum free energy [81, 82] is reported as the secondary structure of the input RNA sequence.

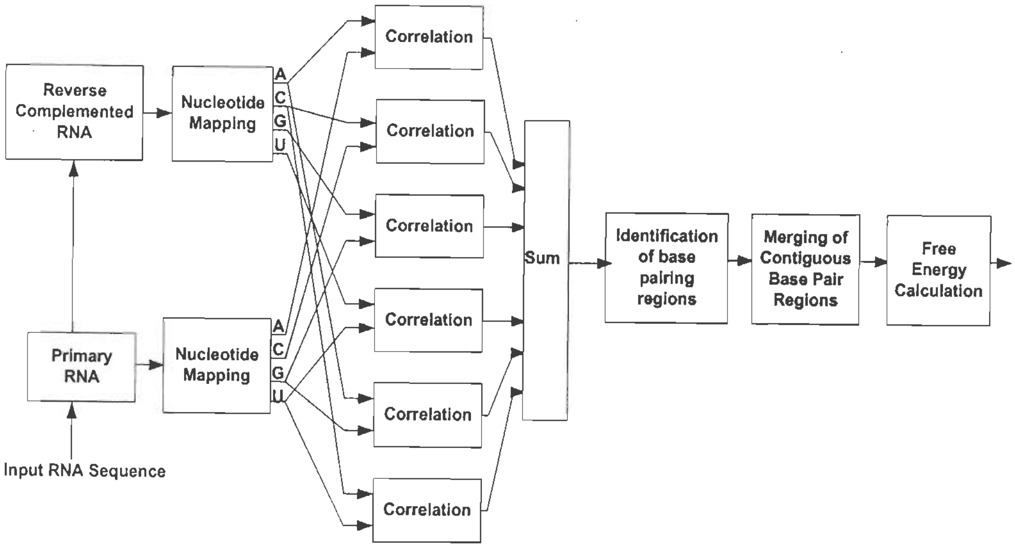


Figure 5.2: Block diagram of the various steps involved in the secondary structure prediction algorithm.

Inputs: RNA sequence (R), minimum number of contiguous base pairs (β).

Preprocessing Stage: The symbolic RNA sequence (R) is converted into four binary indicator sequences R_{A}, R_{C}, R_{G}, and R_{U} as discussed in the section 4.3.1. Let n_{A}, n_{C}, n_{G}, and n_{U} represent the number of nucleotides A, C, G, and U in R. And, let 'α' represent the number of base pairs for a RNA secondary structure (S). Then, the maximum value of α (if Watson-Crick pairs are given priority over wobble pairs) is given by:

$$\begin{aligned}\alpha_{max} &= min(n_{\mathrm{A}}, n_{\mathrm{U}}) + min(n_{\mathrm{C}}, n_{\mathrm{G}}) + \\ & \quad min(n_{\mathrm{U}} - min(n_{\mathrm{A}}, n_{\mathrm{U}}), n_{\mathrm{G}} - min(n_{\mathrm{C}}, n_{\mathrm{G}})) \end{aligned} \tag{5.3}$$

The value of α_{max} is useful in devising threshold parameters which are defined in next stage of the algorithm for identifying high base pair regions.

Identification of high base pair region: In this step the regions that satisfy the minimum number of contiguous base pairs (β) as provided by users are identified in the input RNA sequence. Two threshold parameters λ_{upper} and λ_{lower} are defined which are useful in locating the base pairs. The threshold parameters utilize the value of α_{max} which was defined in (equation 5.3).

$$\begin{aligned}\lambda_{lower} &= \delta_{lower} * \alpha_{max}, \quad \lambda_{upper} = \delta_{upper} * \alpha_{max}, \\ & \quad where \quad 0 < \delta_{lower}, \ \delta_{upper} > 1, \ \delta_{lower} < \delta_{upper}\end{aligned} \tag{5.4}$$

The values δ_{upper} and δ_{lower} are fixed for the proposed experiment to 0.75 and 0.5. These values were obtained after testing the algorithm to a number of test data sets. Using these two threshold values and the correlation coefficient sequence that was defined in the previous step, the base pair identification is divided into three different cases. For the first case in which the correlation coefficient values are below lower threshold value λ_{lower}, the RNA sequence is rejected (i.e., subsequence not explored further). In the second case where the value of threshold lies between upper and lower thresholds value, the RNA sequence is divided into two subsequences and each subsequence is explored further for high base pairing regions. For the last case, if the correlation coefficient lies above the upper threshold parameter the RNA sequence is not explored further and is accepted as a high base pairing region. The pseudo code of this stage is provided in PSEUDO CODE 2.

```
PSEUDO CODE 2: FIND HIGH BPAIR($R$, $i$, $j$, $\beta$)
begin
1.  if($j - i < 3$) and CHECK FLAG($i$,$j$)=0 then
    //checking minimum distance between base pairs
        return
2.  else
    //if base pairs satisfy the minimum distance criteria
        SET FLAG($i$, $j$)$\leftarrow$ 1, $match \leftarrow 0$
        while($j - i$)$<$ 3 and $r_i$ is base pair of $r_j$ do
            //looking for contiguous base pairs
            $i \leftarrow i + 1$, $j \leftarrow j - 1$
            $match \leftarrow match + 1$
        if($match \geq \beta$) then
            //check for minimum contiguous match criteria
            OUTPUT($i - match$, $j + match$, $match$)
            FIND HIGH BPAIR($R$, $i + 1$, $j - 1$, $\beta$)
        else
            $i \leftarrow i - match$, $j \leftarrow j + match$
            if ($\alpha_{max} < 2 * \beta$) then
                //check for presence of enough base pairs
                return
            else
                //calculate coefficient
                $\mu[i, \ldots, j]$ = FIND CORR($R[i, \ldots, j]$)
        $k \leftarrow i$
        while($k < j$) do
            if($k < j$)($\mu[k] > \lambda_{lower}$) and ($\mu[k] < \lambda_{upper}$) then
                FIND HIGH BPAIR ($R$, $i$, $k$, $\beta$)
                FIND HIGH BPAIR ($R$, $k+1$, $j$, $\beta$)
            else if($\mu[k] \geq \lambda_{upper}$) then
                FIND BASE PAIRS ($R[i,k]$)
                FIND BASE PAIRS ($R[k+1,j]$)
end
```

Merging of contiguous base pairing regions: The previous step of the algorithm results in construction of a tree structure. The nodes of the tree are represented by a tuple $< X, Y, L >$, where X, Y are the positions of a contiguous base pair region and L is the length of the contiguous base pair region in the RNA sequence. The children of a node $P \equiv < X_p, Y_p, L_p >$ satisfy the following properties:

- $< X_l, Y_l, L_l >$ is a left child if $X_l \geq (X_p + L_p)$ and $Y_l \leq (Y_p + L_p)$.
- $< X_r, Y_r, L_r >$ is a right child if $X_r > Y_p$.

The tree is traversed in depth first search manner and the nodes covered from the starting node to the end node form a secondary structure of RNA. For selecting the starting node, the following conditions must be satisfied:

- Starting node must be a non-traversed node.
- If $P \equiv < X1, Y1, L1 >$ is the selected starting node then $X1$ must be the smallest among the non-traversed node and $Y1$ must be farthest among all nodes that are starting from $X1$.

Later on the free energy for each structure is calculated and the structure that has minimum energy is selected as the secondary structure of the RNA molecule (R).

5.4 Experimental Results

In order to test the effectiveness of the proposed algorithm, experiments were performed on datasets of transfer RNA (tRNA) obtained from publicly available website *http://lowelab.ucsc.edu/GtRNAdb/* and datasets of tmRNA obtained from *http://www.indiana.edu/tmrna*. From the tRNA database a tRNA sequence for an organism was taken up for the experiment. The details of the tRNA sequence are as follows:
Organism: Drosophila melanogaster(Release 4), Sequence name: arm 2L.trna35,
Location: (3173084-3173003), Length: 82bp,
Sequence:gcagucguggccgagcgguuaaggcgucugacuagaaaucagauucccucugggagcgua
gguucgaauccuaccgacugcg

For the experiment, the minimum length of contiguous base pair region was taken as 3. The minimum folding energy reported by the algorithm was -27.3 Kcal/mol and structure of the molecule is given by: $< 1, 81, 7 >, < 10, 25, 3 >$ $, < 27, 43, 5 >, < 45, 55, 4 >, < 58, 74, 5 >$. The predicted secondary structure of RNA molecule is shown in Figure 5.3. The base pairs in the structure are shown with dots. This tRNA was later on tested using mfold program available at *http://www.bioinfo.rpi.edu/applications/mfold/cgi-bin/rna-form1.cgi*. For the test the setting of the mfold server was left to its default values. The folding energy of the tRNA predicted by the mfold server was -27.1 Kcal/mol and the structure is given by: $< 1, 81, 7 >, < 8, 21, 5 >, < 23, 47, 2 >, < 26, 44, 6 >, < 51, 73, 5 >$

, < 57, 66, 2 >. Thus the algorithm provides a secondary structure for the tRNA which has lower folding energy than that predicted by mfold program.

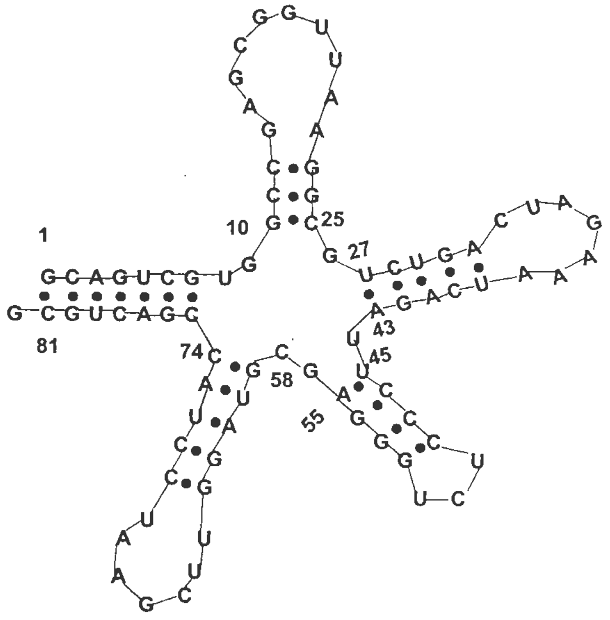


Figure 5.3: The secondary structure predicted by the algorithm. The base pairs in the structure are shown with a dot. The free energy of the RNA secondary structure is -27.3 Kcal/mol.

Experimental results of some Eukarya and Archaea tRNA sequences are provided in Table 5.1. The secondary structures that were predicted by the proposed algorithm and the mfold server are shown in Table 5.1. The free energy of the structures that were predicted by algorithms (i.e., proposed algorithm and mfold program) is quite close and is shown in the Table 5.1. As the minimum number of contiguous base pair is taken as 3, the secondary structure obtained by the proposed algorithm for some of the tRNA is different from that given by mfold program. However, the folding energy obtained by both methods is quite near. When the minimum length of matching region was reduced to 2, the proposed program gives out the same secondary structure as that predicted by the mfold program.

The result obtained on some RNA sequences using the proposed algorithm is presented in this section. The minimum contiguous match was set to 4.

Sequence name: Recli_amer2_0200 Reclinomonas americana mitochondrion pre-tmRNA homolog, version 2, RNA sequence: uauauuaacuauggacccgagggcaguucucggcaucuccauuuagauauuguuuuuaaggggauguuuuuaggauucgacauaguaauaua
Secondary structure predicited: $< 1, 92, 7 >, < 9, 85, 5 >, < 17, 37, 7 >, < 34, 66, 7 >, < 43, 59, 5 >$
Folding energy of the structure: -26.3 Kcal/mol.

The secondary structure predicted by the proposed algorithm for Recli_amer2 _0200 Reclinomonas americana mitochondrion pre-tmRNA homolog is shown in Figure 5.4.

Sequence name: Recli_amer3_0201 Reclinomonas americana mitochondrion pre-tmRNA homolog, version 3
RNA sequence: uuucguaguaacuauggacccgagggcaguucucggcaucuccaucuaaaaaaauuuuuuuuaaggggauguuuuuagaauucgacauaguacaauauuaugaga
Secondary structure predicted: $< 1, 105, 10 >, < 11, 91, 6 >, < 20, 36, 7 >, < 37, 71, 7 >, < 47, 63, 7 >$
Folding energy of the structure: -30.4 Kcal/mol.

The secondary structure predicted by the proposed algorithm for Recli_amer3 _0201 Reclinomonas americana mitochondrion pre-tmRNA homolog is shown in Figure 5.5. The tmRNAs were also tested with mfold program and it gives out the same secondary structure as that given by the proposed algorithm. From the

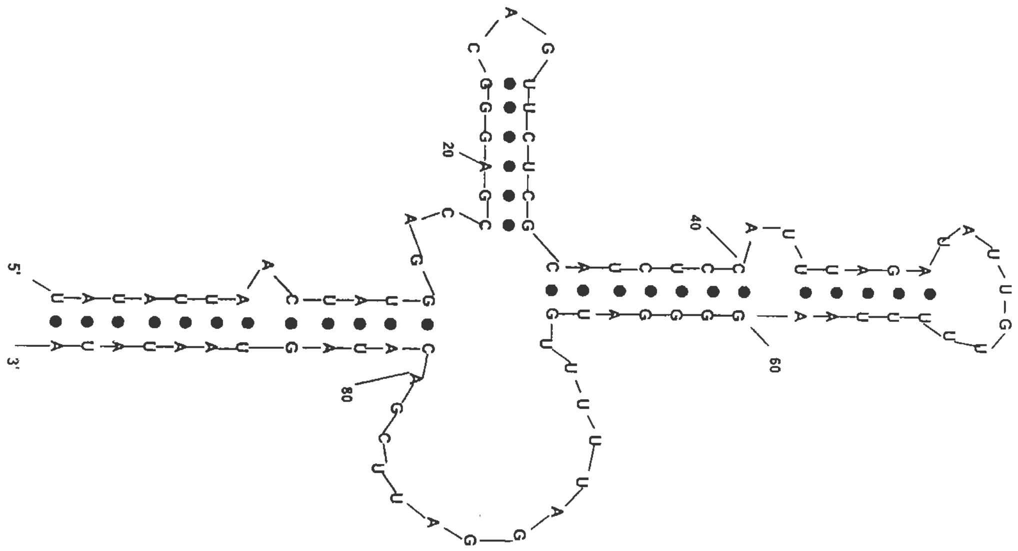


Figure 5.4: Secondary structure of Recli_amer2_0200 Reclinomonas americana mitochondrion pre-tmRNA homolog, version 2 with minimum contiguous match taken as 4.

experimental results, it is seen that the mfold program does not always provide a secondary structure of the RNA molecule that has the minimum free energy. For example, for *Drosophila melanogaster* (Release4) RNA sequence the mfold server provides a secondary structure having free energy as -27.1 Kcal/mol, however the proposed algorithm provides a structure having free energy equal to -27.3 Kcal/mol. Also, the mfold program requires a large number of input parameters, however the proposed algorithm require two input parameters. Thus, the proposed algorithm provides a very simple and novel signal processing based approach for any non-computer experts/biologists for predicting secondary structure of RNA molecule. Applying the correlation theorem of signal processing the calculation of coefficient takes $O(L \log_2 L)$ where L is the length of the RNA sequence. Thus, the algorithm can be applied to even very large RNA sequences without much increase in the computational time. Additionally, the algorithm

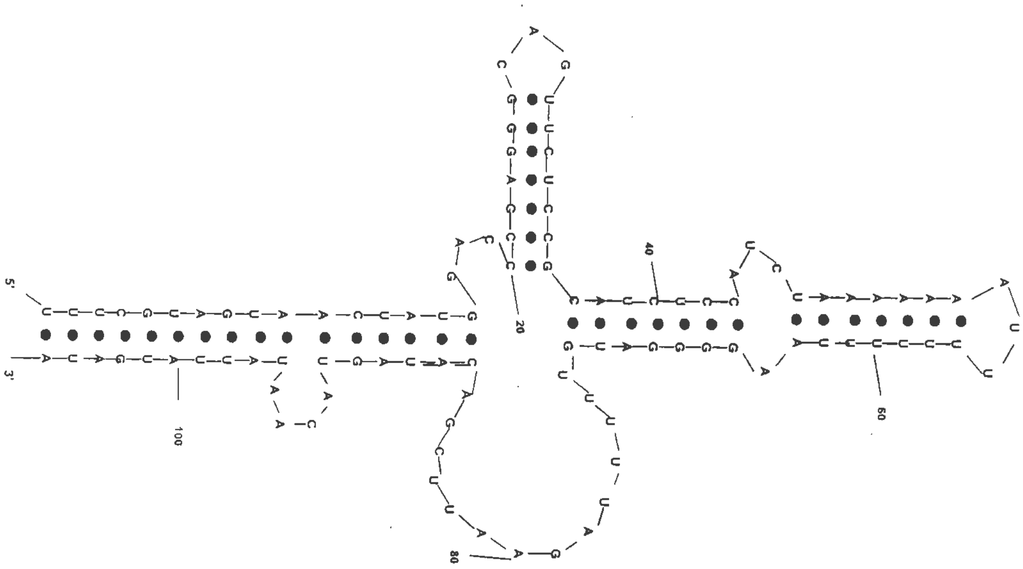


Figure 5.5: Secondary structure of Recli_amer3_0201 Reclinomonas americana mitochondrion pre-tmRNA homolog, version 3 with minimum contiguous match taken as 4.

can be used for designing RNA molecule satisfying the minimum contiguous match parameter.

Table 5.1: The secondary structure of RNA with minimum folding/free energy calculated by the algorithm.

Sequence name	Length	Our algorithm		Mold program	
		Structure	Free Energy	Structure	Free Energy
Athal-chr1.trna15	73	$< 1,72,7 >, < 9,32,3 >,$ $< 36,63,5 >, < 42,56,3 >$	-20.6	$< 1,72,7 >, < 8,64,3 >,$ $< 11,43,3 >, < 16,34,3 >$	-26.53
C.Elegans, Chr_IV.trna26	72	$< 1,71,8 >, < 10,24,4 >,$ $< 26,42,5 >, < 49,62,3 >$	-28.2	$< 1,71,8 >, < 10,24,4 >,$ $< 26,42,5 >, < 49,62,3 >$	-28.2
D.melanogaster Arm_2L.trna35	82	$< 1,81,7 >, < 10,25,3 >,$ $< 27,43,5 >, < 45,55,4 >,$ $< 58,74,5 >$	-27.3	$< 1,81,7 >, < 8,21,6 >,$ $< 23,47,2 >, < 26,45,6. >,$ $< 51,73,5 >, < 57,66,2 >$	-27.1
Gallus gallus Chr7.trna15	82	$< 1,81,7 >, < 10,25,3 >,$ $< 35,52,7 >, < 58,74,5 >$	-29.1	$< 14,32.5 >, < 19,66,4 >,$ $< 24,59,4 >, < 29,54,2 >$ $< 35,52,7 >$	-31.0
Homo sapiens Chr12.trna16	75	$< 1,74,12 >, < 15,58,3 >,$ $< 29,45,6 >$	-20.6	$< 1,74,12 >, < 15,58,3 >,$ $< 24,50,2 >, < 29,45,6 >$ $< 35,52,7 >$	-21.3

5.5 Conclusion

The main contribution of this chapter is the use of correlation based framework for predicting the secondary structure of a RNA sequence. The advantage of using the algorithm is that it gives a list of probable secondary structure for the RNA sequence. The secondary structures are then processed by free energy calculation algorithm and the structure having minimum free energy is selected as the most probable secondary structure for the RNA sequence.

The secondary structure prediction algorithm requires the specification of mainly two well understood parameters: primary RNA sequence and minimum number of continuous base pair required for the secondary structures. The other parameters required are the values for degrees of upper and lower matching parameters. These can be generally fixed to a reasonable value by the algorithm.

Chapter 6

Recognition of Coding and Non-Coding DNA Sequences

In this chapter, a pattern recognition framework for classification of coding and non-coding DNA sequences is presented. In addition, a novel feature vector consisting of wavelet variance coefficients (WVC) is also proposed. The various tasks performed for the classification system are discussed first. Later on, a detailed 10-fold performance evaluation of the system on *Saccharomyces cerevisiae* (*Yeast*) and *Escherichia coli* (*E. coli*) genome is performed. A comparison of the proposed approach with a standard classification technique is also done.

The classification algorithms operate on the basic assumption that every protein coding region should have some distinct sequence features or properties that can distinguish it from the surrounding regions, such as non-coding regions and intergenic regions. The standard existing classification techniques are based on linear/slope model of Z-curve components. However, the linear model provides a poor approximation for highly non-linear Z-curve components. The second problem is that the features which are derived from linear model of Z-curve do not consider the local information content of the DNA sequence. In the proposed technique a wavelet based time series analysis is performed for extracting features from Z-curve components. The wavelet variance feature vector is obtained based on scale by scale decomposition of Z-curve components variance and provides both local and global information contents of DNA sequences. Additionally, the presented wavelet based time series analysis technique for DNA sequences provides a generic approach for analysis of genomic data and can be extended to other problems related to DNA sequence analysis. Experimental results obtained from analysis of complete genome data of *Yeast* and *E. coli* demonstrate the effectiveness of the proposed approach.

6.1 Introduction

A (protein-coding) gene may be defined as any pattern in a DNA sequence which results (under proper conditions) in the generation of a protein product. A gene is further divided into exons and introns. Although some exons (or parts of them) may be non-coding, most gene finding softwares use the term 'exon' to denote the coding part of the exons only. The problem of gene recognition is to define an algorithm which takes as input DNA sequence and produces as output a feature table describing the location and structure of the pattern making up any genes present in the sequence. At the core of most gene identification algorithms are one or more coding measures – functions which produce, given any sample window of sequence, a number of vectors intended to measure the degree to which a sample sequence resembles a window of 'typical' exonic DNA. The prediction and classification of coding and non-coding DNA sequences are popular research areas and many review papers on gene prediction have been published [89, 90, 91, 92, 93, 94, 95].

In the past two decades, a number of useful coding measures/indices have been proposed, for example codon usage bias [96], base composition bias between codon positions [97], and periodicity in base occurrence [7]. Review on various coding region statistics is provided in section 6.2. Most of the previous classification measures were obtained either by studying the composition of the nucleotides in the DNA sequences or Z-curve components. The Z curve representation for DNA was given by Zhang and co-workers [98, 99]. A brief introduction to Z-curve is provided in section 6.3.1. The popular and accurate protein-coding DNA classification tools are based on coding measures derived from Z-curve components [100, 101, 102]. The Z-curve based coding measures are obtained by calculating the slope of all three components of Z-curve. The linear/slope model fitting approach concentrates on global trend in the DNA sequence. Hence, there may be cases where some local information which otherwise is important for identifying as a coding sequence is lost, especially for short DNA sequence.

Till date, a number of techniques exist for protein-coding identification, however very little work has focused [103, 102] on the classification of coding and non-coding DNA sequences. The aim of this chapter is to provide a novel coding measure that provides both local and global information of the DNA sequence for designing more accurate classification system. In addition, the goal is not meant to replace previous coding measures, rather, to act as a complement to these already widely used measures. The proposed novel coding measure is represented by a fixed length feature vector (in pattern recognition terminology). The feature vector is obtained by applying a wavelet based time series analysis approach to Z-curve components. The feature vector calculation task makes no assumptions about the model followed by the components of Z-curve and is purely based on analysis of variance (ANOVA) [104], a time series analysis technique. Each component of the Z-curve is viewed as a time series data generated by some stochastic stationary process. Then, a scale-by-scale decomposition of Z-curve components variance is performed using maximal overlap discrete wavelet transform (MODWT). The collection of wavelet variance coefficients obtained at various scale is defined as a feature vector. Thus, the feature vector obtained is a summary of various levels of information (finer to coarser or local to global) present in DNA sequences. Later on, support vector machine (SVM) [105, 106], an efficient machine learning tool, is applied on the novel features for classification.

6.2 Literature Review

During the past twenty years, several gene finding algorithms (GRAIL [107], GeneParser [108], GeneFinder [109], MZEF [110], VEIL [111], Genie [112], GENESCAN [113], HMMgene [114], GeneMark [115]), etc. have been developed. At the core of the gene finding algorithms are one or more coding measures [96, 90, 89, 116]–functions which calculate, for a given window of sequence, a number or vector that measures attributes correlated with protein coding function. Some of the coding measures that are widely used by the gene finding algorithms are as follows:

Start codon measures: The frequencies of ATG triplets in the genomes of various species were systematically analyzed by Saito and Tomita [117]. ATG triplet is involved in the initiation of translation and is also called the start codon. Let the total number of the ATG triplet contained in all three frames in a sequence be denoted by N_{ATG}. The number of frames containing the start codon in a sequence is denoted by k, i.e., $k = 0, 1, 2, 3$. The start codon measures are given by:

$$f_{StartCodon1} = (1 + k^2) * N_{ATG} \tag{6.1}$$

$$f_{StartCodon2} = N_{ATG} \tag{6.2}$$

Stop codon measures: The distributions of the three stop codons, i.e., TAA, TAG and TGA, in three phases along coding, noncoding, and intergenic sequences are studied in detail by Wang *et al.* [118]. Let N_{TAA}, N_{TAG}, and N_{TGA} be the number of triplets TAA, TAG, and TGA occurring in all the three frames of the sequence. Like the start codon measures, the stop codon measures are given by:

$$f_{StopCodon1} = (1 + k^2) * N_{STOP}, \tag{6.3}$$

$$f_{StopCodon2} = N_{STOP}, \quad where \quad N_{STOP} = N_{TAA} + N_{TAG} + N_{TGA} \tag{6.4}$$

Position asymmetry measure: Let $f(b, r)$ be the frequency of nucleotide b at triplet position r. Let $f(b) = \sum_{r=1}^{3}(f(b, r))/3$ be the average frequency of nucleotide b at the three triplet positions, and define the asymmetry in the distribution of nucleotide b as the variance of this frequency, i.e., $asym(b) = \sum_{i=1}^{3}(f(b, i) - f(b))^2$. The position asymmetry measure (PA) of the sequence is defined as:

$$PA = asym(A) + asym(C) + asym(G) + asym(T) \tag{6.5}$$

Purine measure: It is well known that the first codon position in coding region is predominantly taken by purines and this fact is independent of species , whereas bases in non-coding regions tend to be randomly distributed. The occurrence frequencies of purines in the three reading frames are denoted by $(a_i + g_i)$, $i = 1, 2, 3$, and the purine measure is defined as: $max(a_i + g_i)$, $i = 1, 2, 3$.

Pyrimidine measure: It is well known that the third codon position in coding region is predominantly taken by pyrimidines and this fact is independent of species, whereas bases in non-coding regions tend to be randomly distributed. The occurrence frequencies of pyrimidines in the three reading frames are denoted by $(c_i + t_i)$, $i = 1, 2, 3$, and the pyrimidine measure is defined as: $max(c_i + t_i)$, $i = 1, 2, 3$.

Z-curve measures: The Z-curve representation for a DNA sequence was given by Zhang and Zhang [98]. The Z-curve measures are based on the differences of single nucleotide frequencies at the three codon positions between the protein coding ORFs and the non-coding ORFs. The frequencies of bases A, C, G, and T occurring in an ORF with bases at positions $1, 4, 7, \ldots$; $2, 5, 8, \ldots$; $3, 6, 9, \ldots$, are denoted by a_1, c_1, g_1; a_2, c_2, g_2; a_3, c_3, g_3, respectively. The Z-curve measures, x_1, y_1, z_1, x_2, y_2, z_2, x_3, y_3, z_3 are given by:

$$\begin{aligned} x_i &= (a_i + g_i) - (c_i + t_i) \\ y_i &= (a_i + c_i) - (g_i + t_i) \\ z_i &= (a_i + t_i) - (c_i + g_i) \end{aligned} \tag{6.6}$$

where $i = 1, 2, 3$.

The most recent and accurate protein-coding identification techniques are based on Z-curve [100, 119, 120, 121].

Simple Z-curve measures: The simple Z-curve (SZ) measures [101] are given by:

$$\begin{aligned} SZ_1 &= max[(a_i + g_i) - (c_i + t_i)] \\ SZ_2 &= max[(a_i + c_i) - (g_i + t_i)] \\ SZ_3 &= max[(a_i + t_i) - (c_i + g_i)] \end{aligned} \qquad (6.7)$$

where $i = 1, 2, 3$.

Periodic asymmetry index: Given a DNA sequence, three distinct probabilities are considered, the probability P_{in} of finding pairs of the same nucleotide at distance nucleotide at distances $k = 2, 5, 8, \ldots$, the probability P^1_{out} of finding pairs of the same nucleotide at distance $k = 0, 3, 6, \ldots$, and the probability P^2_{out} of finding pairs of the same nucleotide at distances $k = 1, 4, 7, \ldots$. The value of P_{in} will be greater than the other two other probability for protein-coding regions, whereas for non-coding regions the three probabilities will be similar. The periodic asymmetry index (PAI) is given by:

$$PAI = \frac{max(P_{in}, P^1_{out}, P^2_{out})}{min(P_{in}, P^1_{out}, P^2_{out})} \qquad (6.8)$$

Average mutual information measure: Let p_i and p_j be the probabilities of nucleotides i and j in the DNA sequence, and $p_{ij}(k)$ is the probability in the DNA sequence of the pair of nucleotides i and j at a distance of k nucleotides. For each distance k, sixteen different individual correlations can be calculated. A measure that summarizes all sixteen correlations at a given distance k is the mutual information function and is given by:

$$I(k) = \sum_{i,j \in \{A,C,G,T\}} p_{ij}(k) \log_2(\frac{p_{ij}(k)}{p_i p_j}) \qquad (6.9)$$

The mutual information $I(k)$ quantifies the amount of information that can be obtained from one nucleotide about another nucleotide at a distance k. In protein-coding DNA, $I(k)$ oscillates between two values, while in non-coding DNA, $I(k)$ is rather flat. Herzel and Grosse [122] called the two values between which $I(k)$ oscillates in coding DNA the in-frame mutual information I_{in} at a distance $k = 3, 6, 9, \ldots$, and the out-of-frame mutual information I_{out} at $k = 4, 5, 8, \ldots$. In order to reduce the pair of numbers I_{in} and I_{out} to a single quantity, they compute the average mutual information (AMI) as

$$AMI = \frac{I_{in} + 2I_{out}}{3} \tag{6.10}$$

Fourier spectral measure: A major signal in protein-coding regions of genomic sequences is a three-base periodicity [7]. The Fourier spectral measure is defined as follows. Let $A_d(t)$, $C_d(t)$, $G_d(t)$ and $T_d(t)$ be the number of distinct pairs of nucleotide bases A, C, G and T respectively, in a DNA sequence separated by a distance t, where t ranges from 1 to N. Let $s(t) = A_d(t) + C_d(t) + G_d(t) + T_d(t)$. Let $S(k)$ be the discrete Fourier transform (DFT) of $s(t)$, i.e.,

$$S(k) = \sum_{t=0}^{N-1} s(t)^{-j2\pi kt/N} \tag{6.11}$$

For three-base periodicity, $S(k)$ exhibits a strong peak at the frequency index $k = N/3$. Let $P(k) = |S(k)|^2$ be the power spectrum of $S(k)$, then the Fourier spectrum measure is defined as follows:

$$f_{FT} = \frac{P(k)}{(1/(2w) + 1) \sum_{j=k-w}^{j=k+w} P(j)} \tag{6.12}$$

where $k = N/3$ and $2w + 1$ are the window used to obtain the average power spectrum within the window.

In past, a number of signal processing techniques (Fourier spectral measure) based on period-3 property of protein-coding regions have been proposed [7, 123, 5, 2, 3, 17, 4, 18, 119]. A major drawback with Fourier spectral and Z-curve coding measures is that they provide only global level of information present in DNA sequences.

6.3 Classification System

Pattern recognition is the scientific discipline dealing with methods for object description and classification. Applications of pattern recognition systems and techniques are numerous areas, such as agriculture, astronomy, biology, economy, engineering, geology, medicine, military and security. A fundamental notion in pattern recognition, independent of whatever approach is followed, is the notion of similarity. In a classification system objects are assigned to a particular class based on the measurement of features or patterns. Features are generally mathematical or numeric representations of the object. For instance, when apple is to be distinguished from orange, their color and shape is looked. The feature vector in this case is represented by 2-dimensional vector **F** and is given as:

$$\mathbf{F} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \equiv \begin{bmatrix} color \\ shape \end{bmatrix} \tag{6.13}$$

The color is represented by red-green-blue components and in order to obtain the numeric representation of the shape feature we may, for instance, measure the distance, away from the top, or the maximum width of the object and normalize the distance by the height. Similarly, for classifying DNA sequences, the biological property of the DNA sequence is to be considered. Most of the protein coding

DNA sequences have some distinct sequence features that can distinguish it from the surrounding regions, such as intergenic regions. In [100], it was demonstrated that the Z-curve which has the information of purine/pyrimidine, amino/keto and strong-H bond/weak-H bond bases distribution can help in identifying protein coding sequence.

In this study, the information provided by the Z-curve has been considered for constructing a fixed length feature vector for DNA sequences. Unlike the previous classification techniques [100, 101] which are based on calculating slope of the distributions curves (i.e., Z-curve components) the proposed approach applies the wavelet based technique for extracting features from the distribution curves. This helps in deriving different level of information from Z-curve components. The complete framework for our classification problem consists of four major components which are shown in Figure 6.1. First, the DNA sequences (or patterns) and other necessary information are acquired from different databases. Secondly, the Z-curve components for DNA sequences are generated. At the third step the Z-curve components are decomposed into wavelet variance coefficients (WVC) in order to calculate a fixed length feature vector. Finally, in the fourth step an optimized SVM model is constructed using a hierarchical grid search based technique [124, 125] for automatic optimization of machine learning parameters. The components of feature extraction unit and classification unit are shown in Figure 6.1 and are discussed in the next section.

6.3.1 Time series model for Z-curve components

The Z-curve [98] is a three-dimensional curve representation for a given DNA sequence. The Z-curve for a DNA sequence is a connection of points whose coordinates are x_n, y_n, z_n ($n = 0, 1, 2, \ldots, N$, where N is the length of the DNA

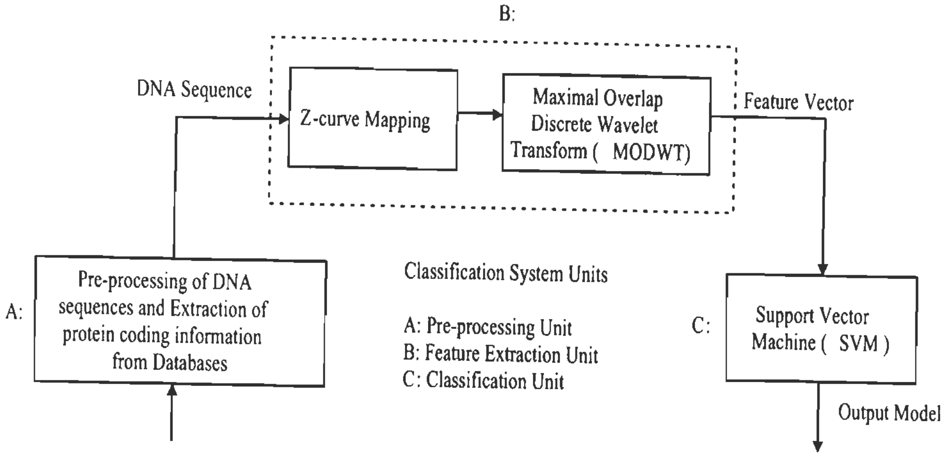

B:
DNA Sequence
Z-curve Mapping
Maximal Overlap Discrete Wavelet Transform (MODWT)
Feature Vector
A:
Pre-processing of DNA sequences and Extraction of protein coding information from Databases
Classification System Units
A: Pre-processing Unit
B: Feature Extraction Unit
C: Classification Unit
C:
Support Vector Machine (SVM)
Output Model

Figure 6.1: A block diagram showing the various units of pattern recognition based classification system.

sequence being studied).

$$\begin{aligned} x_n &= (A_n + G_n) - (C_n + T_n) \equiv R_n - Y_n, \\ y_n &= (A_n + C_n) - (G_n + T_n) \equiv M_n - K_n, \\ z_n &= (A_n + T_n) - (C_n + G_n) \equiv W_n - S_n, \\ x_n, y_n, z_n &\in [-N, N], n = 0, 1, 2, \ldots, N. \end{aligned} \tag{6.14}$$

where A_n, G_n, C_n and T_n are the cumulative occurrence numbers of nucleotides A, G, C and T, respectively in the DNA sequence. R, Y, M, K, W and S represent the bases of purine, pyrimidine, amino, keto, weak hydrogen bonds and strong hydrogen bonds, respectively, according to the Recommendation 1984 by the NC-IUB [126]. The starting coordinate of the Z-curve is taken as $(0, 0, 0)$, where $A_0 = C_0 = G_0 = T_0 = 0$. For example, the three components of Z-curve for a DNA sequence of *Yeast* are shown in Figure 6.2.

The three components of the Z-curve, x_n, y_n and z_n, represent three inde-

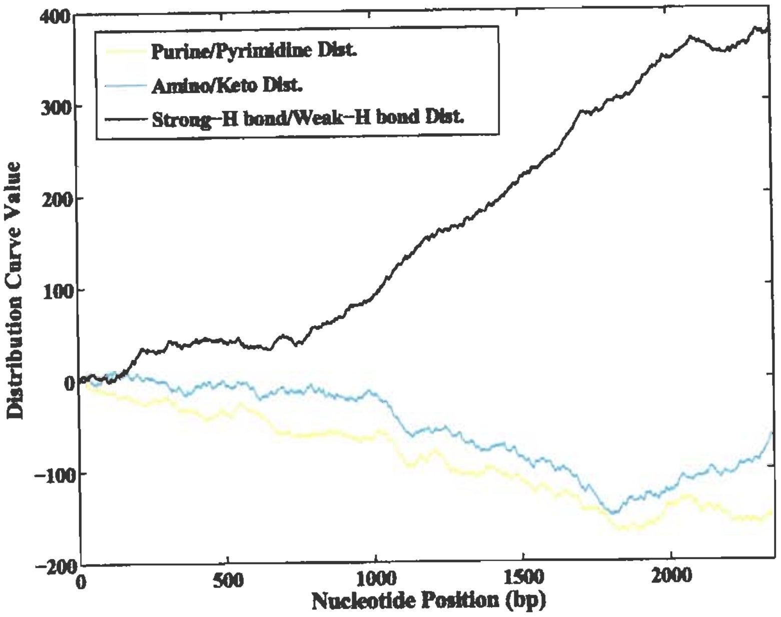


Figure 6.2: The three components of Z-curve for a DNA sequence.

pendent distributions that completely describe the DNA sequence being studied. Each component of the Z-curve, i.e., x_n,y_n and z_n has a clear biological meaning. The x_n component represents the distribution of purine/pyrimidine (A or G/C or T). Similarly, the component y_n represents the distribution of amino/keto (A or C/G or T), and the component z_n displays the distribution strong-H bond/weak-H bond bases (A or T/G or C) along the sequence respectively. For more details about the Z-curve refer to *http://tubic.tju.edu.cn/zcurve/*.

In [100], authors have shown that using the global variation information (slope) of Z-curve components, a DNA sequence can be identified as coding or non-coding sequence. In this study we have also extracted information from Z-components. However, our analysis is based on wavelet theory which helps in extracting information both at local and global level from Z-curve components. The feature vector $\mathbf{F_Z}$ can be represented by a 3-dimensional vector based on three different properties of DNA sequences and is given as:

$$\mathbf{F_Z} = \begin{bmatrix} f_x \\ f_y \\ f_z \end{bmatrix} \equiv \begin{bmatrix} purine/pyrimidine\ distribution \\ amino/keto\ distribution \\ strong - Hbond/weak - H\ bond\ distribution \end{bmatrix} \tag{6.15}$$

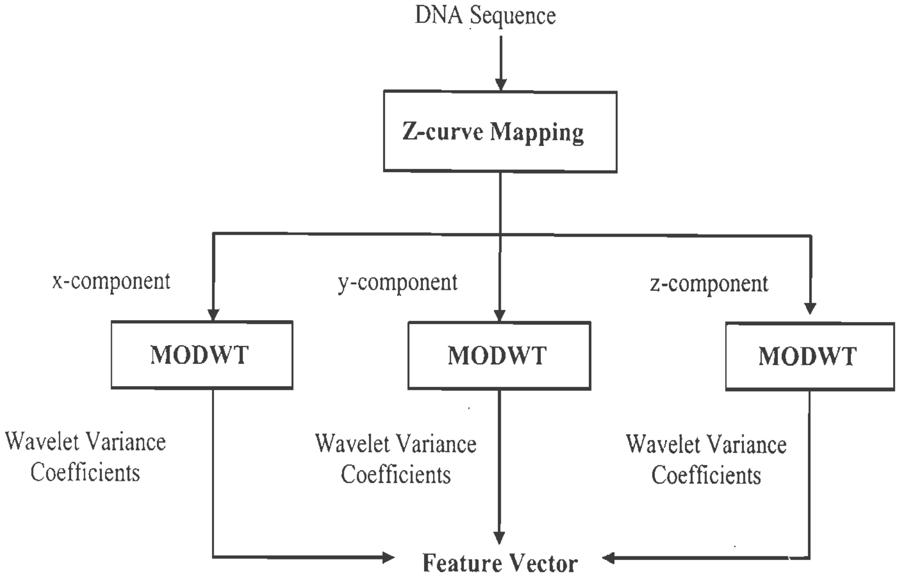


Figure 6.3: A block diagram showing the construction of the proposed feature vector.

Formally, each Z-curve component is viewed as a discrete time series data, where the time interval between two observations, i.e., $\triangle t$ is equal to 1 amino acid or 1 bp and generated by some stationary stochastic process. In order to extract the coding features from the DNA sequences we perform a time series analysis using MODWT, a modified version of discrete wavelet transform (DWT) well suited for time series data. In [32], authors have shown that the wavelet variance is a very useful parameter of wavelet analysis for identifying patterns in biological data (C+G islands and transmembrane proteins). In this study, a novel fixed length feature based on scale-by-scale decomposition of wavelet variance of Z-components is defined for classifying coding and non-coding DNA sequences. The steps involved in construction of feature vector from a DNA sequence are shown in Figure 6.3.

6.3.2 Maximal overlap Discrete Wavelet Transform

Wavelet is a very powerful mathematical tool which has been applied in a wide variety of disciplines, such as image coding and compression, signal processing, astronomy, medicine, geophysics, analysis of climate time series, chaos, fractal, turbulence, financial market, economics time series and in certain area of mathematics, as in solution to partial differential equation or numerical analysis. In bioinformatics the wavelet based techniques have been applied especially in DNA, protein and microarray data analysis [127, 32, 128, 129, 33]. The main features of wavelet analysis are multiresolution analysis (MRA) and wavelet packet analysis. MRA enables the researchers to separate out a variable or signal into its constituent multiresolution components and is a very popular for analysis of different types of time series data. Wavelets are, by definition, small waves. Wavelets possess many desirable properties, some of which are useful in genomic sequence analysis, but many of which are not. In this chapter, our focus is on extracting features by applying the wavelet based analysis technique for time series data and their ability to decompose statistical information scale by scale.

MODWT [130, 26, 131] is an important extension of Discrete Wavelet Transform (DWT) and is an important tool for analysis of different type of time series data [132, 133, 32]. In our current study we have applied the concept of analysis of variance (ANOVA) of time series data [104] using MODWT for an important problem of DNA sequence analysis. The MODWT is a linear filtering operation that transforms a series into coefficients related to variation over a set of scales. It is similar to the DWT in that both are linear filtering operations producing a set of time-dependent wavelet and scaling coefficients. Both have basis vector associated with a location t and a unitless scale $\tau_j \equiv 2^{j-1}$ for each decomposition level $j = 1, \ldots, J$. Both are suitable for analysis of variance (ANOVA) and multiresolution analysis (MRA). Detail about wavelet theory is provided in [134, 135, 136, 137].

The MODWT differs from the DWT in that it is a highly redundant and non-orthogonal transform. Although the MODWT gives up orthogonality (through not sub-sampling) it has several advantages over DWT:

1. The MODWT can handle any sample size N, while Jth order DWT restricts the sample size to multiple of 2^J. The property is very useful for analysis of the DNA sequence, as the length of the DNA sequence is not a multiple of 2^J.

2. The detail and smooth coefficients of a MODWT multiresolution analysis are associated with zero-phase filters.

3. The MODWT is invariant to circularly shifting the original time series.

4. The MODWT yields an estimator of the variance of the wavelet coefficients that is statistically more efficient than the corresponding estimator based on the DWT. This helps in constructing better feature set and thus helps in designing an efficient classifier for classification of DNA sequences into

coding and non-coding.

Advantages (1) and (4) are very useful for our study. They help in analyzing DNA sequences of arbitrary size and provide statistically more efficient wavelet coefficients that helps in designing more accurate classification model.

Let $\mathbf{X}$ be an input column vector containing a sequence $X_0, X_1, \ldots, X_{N-1}$ of N (not necessarily dyadic) data points of x-component of Z-curve time series. We assume that X_t was collected at time $t\Delta t$, where Δt is the time interval between consecutive observation (in our case Δt is equal to 1 bp and $t\Delta t$ represents base pair position). The MODWT of level J is an orthonormal transform of $\mathbf{X}$ defined by

$$\mathbf{W} = \boldsymbol{\mathsf{W}}\mathbf{X} \tag{6.16}$$

where $\mathbf{W}$ is a column vector of length $(J+1)N$, and is an $(J+1)N$ x N real-valued non-orthogonal matrix. The matrix $\boldsymbol{\mathsf{W}}$ can be decomposed into $J+1$ submatrices, each of them N x N and is given by

$$\mathbf{W} = [W_1, W_2, \ldots, W_J, V_J] \tag{6.17}$$

Instead of using the wavelet and scaling filters, the MODWT utilizes the rescaled filters

$$\tilde{h}_j = h_j/2, \;\; \tilde{g}_J = g_J/2, \;\; j = 1, 2, \ldots, J. \tag{6.18}$$

where h_j and g_J are wavelet (high-pass) and scaling (low-pass) filters.

To construct the N x N dimensional submatrix W_1, we circularly shift the rescaled wavelet filter vector by integer unit to the right so that

$$W_1 = \left[\tilde{h}_1^{(1)}, \tilde{h}_1^{(2)}, \ldots, \tilde{h}_1^{(N-1)}, \tilde{h}_1\right]^T \tag{6.19}$$

Submatrices W_1, W_2,...,W_J are formed similarly.

The vector of MODWT coefficients given in (equation 6.16) may be decomposed into $J+1$ vectors:

$$\mathsf{W} = [\mathsf{W}_1, \mathsf{W}_2, \ldots, \mathsf{W}_J, \mathsf{V}_J] \tag{6.20}$$

where W_j is a length of $N/2^j$ vector of wavelet coefficients associated with change on scale of length $\tau_j \equiv 2^{j-1}$ and V_J is a length of $N/2^J$ vector of scaling coefficients associated with averages on a scale of length $2^J = 2\tau_J$.

In [130, 26] it was proved that the MODWT is an energy-preserving transform in the sense that

$$\| \mathbf{X} \|^2 = \| \mathbf{W} \|^2 = \sum\nolimits_{j=1}^{J} \| \mathsf{W}_j \|^2 + \| \mathsf{V}_J \|^2 \tag{6.21}$$

so that $\| \mathbf{W} \|^2$ represents the contribution of the energy of $\mathbf{X}$ due to change at scale τ_j, while $\| \mathbf{V} \|^2$ represents the contribution due to variation at scales τ_{J+1} and higher. Because of the orthonormality of W and the special form of V_J, we can decompose (analyze) the sample variance (empirical power) of $\mathbf{X}$ into pieces that are associated with scales τ_1, τ_2,. . . ,τ_J:

$$\hat{\sigma}^2_{\mathbf{X}} \equiv \frac{1}{N} \| \mathbf{X} \|^2 - (\overline{X})^2 = \frac{1}{N} \| \mathbf{W} \|^2 - (\overline{X})^2 \tag{6.22}$$

$$= \frac{1}{N} \sum_{j=1}^{J} \| \mathsf{W}_j \|^2 + \frac{1}{N} \| \mathsf{V}_J \|^2 - (\overline{X})^2 \tag{6.23}$$

where $\hat{\sigma}^2_{\mathbf{X}}$ is the sample variance of $\mathbf{X}$ and $\overline{X}^2$ is its mean. Hence,$\| \mathsf{W}_j \|^2/N$ is the contribution to the sample variance of $\mathbf{X}$ due to change at scale τ_j, i.e., $\tilde{\sigma}^2_{\mathbf{X}}(j)$. Similarly, the sample variance for y and z components of Z-curve time series data can also be obtained for a DNA sequence.

The wavelet variance decomposes (analyzes) the variance of the time series data a scale-by-scale basis and is closely related to the concept of spectral density function (SDF) and offers a simple summary of the SDF. The individual wavelet coefficients are associated with a band of frequencies and specific timescale. In the next section we define the feature vector using wavelet variance obtained from Z-curve representation of the DNA sequence that helps in distinguishing whether the given DNA sequence is of coding or non-coding ORFs. For example, the series of wavelet coefficients obtained for three components of Z-curve as shown in Figure 6.2 are provided in Figure 6.4, Figure 6.5, and Figure 6.6. The decomposition was performed using Haar wavelet and the maximum level of decomposition was set to 10. Furthermore, the values of WVCs obtained for the corresponding components are shown in Figure 6.7.

The implementation of MODWT was done using free downloadable MATLAB WMTSA [26] toolkit version 0.2.5 available at *http://www.atmos.washington.edu/wmtsa.*

6.3.3 Feature vector

As said earlier, each Z-curve component, i.e., x_n, y_n, and z_n is treated as discrete time series data. The feature vector for a DNA sequence is obtained from calculation of wavelet variance using (equation 6.22) for three Z-curve components. The feature vector for classification is given by:

$$\begin{aligned} f_x \equiv \hat{\sigma}_x^2 &= \left[\hat{\sigma}_x^2(1), \hat{\sigma}_x^2(2), \ldots, \hat{\sigma}_x^2(L)\right], \\ f_y \equiv \hat{\sigma}_y^2 &= \left[\hat{\sigma}_y^2(1), \hat{\sigma}_y^2(2), \ldots, \hat{\sigma}_y^2(L)\right], \\ f_z \equiv \hat{\sigma}_z^2 &= \left[\hat{\sigma}_z^2(1), \hat{\sigma}_z^2(2), \ldots, \hat{\sigma}_z^2(L)\right], \end{aligned}$$

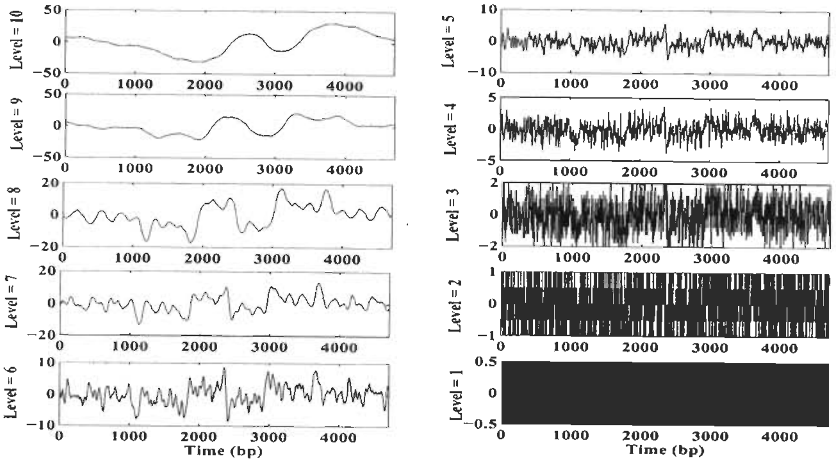


Figure 6.4: The wavelet coefficients obtained at various level (L = 1 to 10) for x-component (i.e., purine/pyrimidine distribution) of Z-curve (as shown in Figure 6.2).

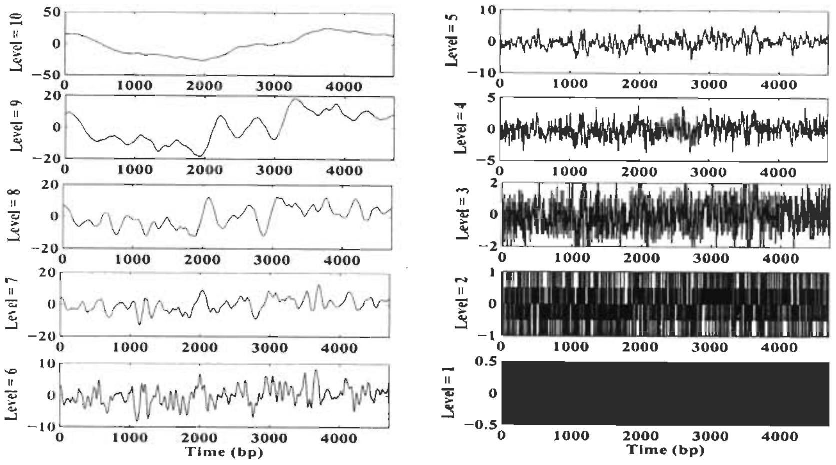


Figure 6.5: The wavelet coefficients obtained at various level (L = 1 to 10) for y-component (i.e., amino/keto distribution) of Z-curve (as shown in Figure 6.2).

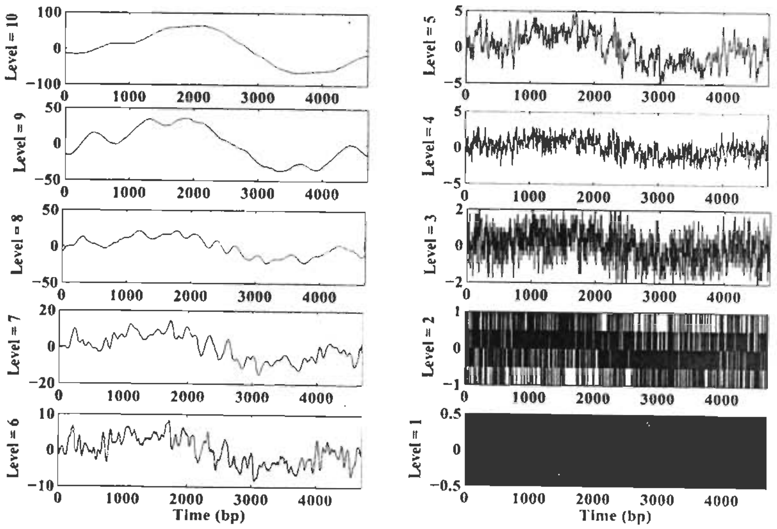


Figure 6.6: The wavelet coefficients obtained at various level ($L = 1$ to 10) for z-component (i.e., strong H-bond/weak H-bond distribution) of Z-curve (as shown in Figure 6.2).

$$\mathbf{Fz} = \begin{bmatrix} f_x \\ f_y \\ f_z \end{bmatrix} \equiv \left[\hat{\sigma}_x^2(1), \ldots, \hat{\sigma}_x^2(L), \hat{\sigma}_y^2(1), \ldots, \hat{\sigma}_y^2(L), \hat{\sigma}_z^2(1), \ldots, \hat{\sigma}_z^2(L) \right] \quad (6.24)$$

where L is the maximum level of decomposition of the time series data.

Thus, $\mathbf{F_Z}$ is a collection of wavelet variance coefficients (WVC). The dimension of $\mathbf{F_Z}$ is $3L$ and is dependent upon the number of levels (L) to which the time series data has to be decomposed. The level of decomposition is dependent on number of observation points in the time series data (i.e., length of DNA sequence) and $L \geq \log_2 N$, where N is the number of data points in the time series. Another parameter that has to be considered for feature extraction is the selection of wavelet. The type of wavelet used for a particular analysis using wavelet technique is dependent on nature of data. The Haar wavelet is used

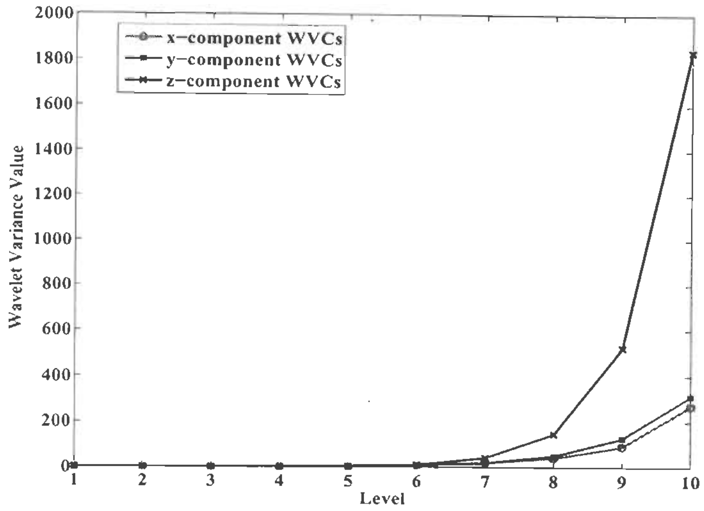


Figure 6.7: The wavelet variance value obtained at various level ($L = 1$ to 10) for x, y and z components of Z-curve (as shown in Figure 6.2).

for extracting features because the value of Z-components either increases or decreases by unity at each step. However, result obtained using other wavelets is also provided in section 6.4.

6.3.4 Support vector machine

The SVM was proposed by Vapnik and co-workers [105, 106] as a very effective technique for general purpose supervised pattern recognition. SVM is based on the idea of structural risk minimization, which bounds the generalization error to the sum of training set error and a term depending on the Vapnik-Chervonenkis dimension [105, 106] of the learning machine. The SVM induction principle minimizes an upper bound on the error rate of a learning machine on test data (i.e., generalization error), rather than minimizing the training error itself which is used in empirical risk minimization. This helps them to generalize well on the unseen data. The SVM method has been successfully applied to isolated handwritten digit recognition [105], microarray data analysis [138], protein structure prediction [139], CpGs island prediction [140] etc.

When used for classification, SVMs separate a given set of binary labeled data with a hyperplane that is maximally distant from them. SVM maps the input patterns into a higher dimensional feature space through some nonlinear mapping function (kernel) chosen *a priori*. A linear decision is then constructed in this high dimensional feature space. A classification procedure usually involves training and testing datasets which consist of a list of records. Each record in the training dataset contains a class label and several attributes (features). The objective of SVM is to produce a model using the training dataset records which can predict the class labels for the records in the testing datasets.

The SVM learns linear decision rules $h(x) = \text{sign}(\mathbf{w}.\mathrm{x} + b)$ described by a weight vector $\mathbf{w}$ and a threshold b. Let the training dataset contain 'm' training records with each record having 'n' features and a class label. The ith training record is given by (x^i, y^i), $i = 1, 2, \dots, m$, where $x^i = (x^i_1, x^i_2, \dots, x^i_n)$ and label $y^i \in \{1, -1\}$. For a linearly separable input, the SVM finds the hyperplane with maximum Euclidean distance to the closest training examples. This distance is called the margin 'δ' as depicted in Figure 6.8. For non separable training sets, the amount of training error is measured using slack variable ξ_i as shown in Figure 6.8 for two class problems. Computation of hyperplanes requires the solution to the following optimization problem.

$$Minimize : P(\mathbf{w}, b, \xi) = \frac{1}{2}\mathbf{w} \cdot \mathbf{w} + C\sum_{i=1}^{m} \xi_i \tag{6.25}$$

$$Subject\ to : \forall_i^m : y^i \left[\mathbf{w} \cdot \mathbf{x}^i + b\right] \geq 1 - \xi_i,\ \forall_i^m : \xi_i > 0 \tag{6.26}$$

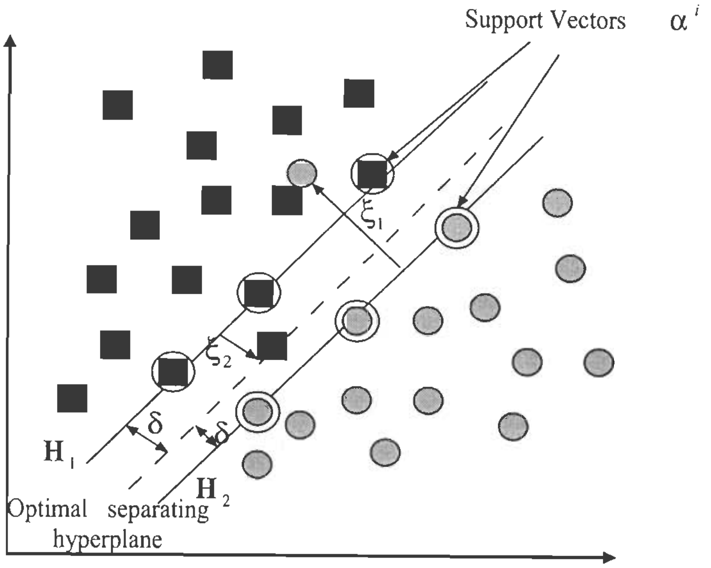


Figure 6.8: The optimal separating hyperplane (OSH), support vectors α^i and the slack variable ξ_i.

The point α_i is called support vector (SV) and are the points that lie closest to the separating hyperplane as shown in Figure 6.8. The SV associated with $\mathbf{x}^i$ expresses the strength with which that point is embedded in the final decision function and often only a small subset of points will be associated with non-zero α_i. For solving the general case of linearly non-separable inputs, SVM maps the input vector $\mathbf{x}$ into a higher dimensional feature space by some kernel function and constructs an optimal separating hyperplane (OSH).

The performance of SVM classification is strongly related to the choice of machine learning parameters. There are a large number of kernel functions available in literature. In general, radial basis function (RBF) is a reasonable first choice. The RBF kernel is given by

$$K(\mathbf{x}, \mathbf{z}) = \exp\left(-\gamma \cdot \parallel \mathbf{x} - \mathbf{z} \parallel^2\right) \tag{6.27}$$

Thus, for the RBF kernel function two parameters i.e., C and γ need to be calculated. The knowledge of C and γ for the training dataset is not known in prior. Usually, these parameters are obtained on a trial and error basis i.e., the user performs SVM classification using different combinations of (C, γ) pair and selects the one that gives maximum performance. The goal is to select good values of (C, γ) so that the classifier provides high performance output on the testing and unseen datasets.

For finding the optimum values of parameters (C, γ) automatically, a grid search technique as provided in [124, 125] is applied using 10-fold cross validation. Different combinations of (C, γ) are tried and the one with the best cross validation accuracy is picked. The grid search is performed in a hierarchical manner. Keeping one of the parameter fixed, the other parameter is grown exponentially and classification performance is evaluated using cross validation. The combination of (C, γ) that provides the best performance is selected and further a finer grid search on that region can be conducted to improve the classification performance. SVM based classification implementation was achieved using free downloadable LIBSVM library available at *http://www.csie.nyu.edu.tw/ cjlin/libsvm* for academic use [141].

6.4 Performance Evaluation

To evaluate the classification performance of SVM using wavelet variance coefficient (WVC) feature vector a 10-fold cross-validation was performed on the dataset of *Yeast* and *E. coli* genomes. The generated optimized model was tested on unseen data to demonstrate the generalization capability of the system. The experiments were conducted using RBF kernel because of its superior performance over other kernels.

6.4.1 Dataset description

The genome of *Yeast* and *E. coli* was downloaded from GenBank database *http://www.ncbi.nlm.nih.gov/* and the coding ORFs (open reading frames) details for the Yeast genome were extracted from MIPS databases [142], available at *http://www.mips.gsf.de/genre/proj/Yeast.* For the present study all the 16 chromosomes (leaving the ORFs of mitochondria) is considered. The ORF information for the *E. coli* genome was extracted from the header information which is present in the downloaded file (GenBank Accession Number = NC_000913) from GenBank database.

6.4.2 Performance measures

In [89], many prediction measures for protein coding regions are provided. For the present study, the performance measures of the predictions is calculated at ORF level. The ORF based approach in evaluating the accuracy of protein coding region for *Yeast* genome was also taken up in [100].

The prediction performance was determined by measuring the sensitivity (SE), specificity (SP), and accuracy (ACC) obtained from the experiments. The SE, SP and ACC parameters were calculated using the following equations:

$$SE = \frac{TP}{(TP + FN)} \tag{6.28}$$

$$SP = \frac{TN}{(TN + FP)} \tag{6.29}$$

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{6.30}$$

where, TP (true positive), TN (true negative), FN (false negative) and FP

(false positive) correspond # coding ORFs that have been correctly predicted as coding ORFs, # non-coding ORFs predicted as non-coding ORFs, # coding ORFs predicted as non-coding ORFs, # non-coding ORFs predicted as coding ORFs respectively.

6.4.3 Results on *Yeast* genome

For the *Yeast* genome the total number of entries of coding ORFs for 16 chromosomes (excluding mitochondria) in MIPS database is equal to 6559. As the training and test datasets should be accompanied by the counterparts of negative samples, the intergenic sequences (non-coding ORFs) with length longer than 11 base pair (bp) is randomly selected from the 16 chromosomes and each of them starts with ATG and ends with one of the stop codons (i.e., TAA, TAG, or TGA). The procedure to select intergenic sequences is slightly different from that used in [100]. If the procedure provided in [100] is followed, the average length of non-coding ORF sequences obtained is $\approx$ 350 bp. However, if the average length of positive samples (coding ORFs) is calculated, it is found to be $\approx$ 1341 bp. So, the negative samples created from the procedure given in [100] lead to creation of a biased dataset, where most of the shorter ORFs are non-coding DNA sequence. To create a better negative dataset (i.e., non-coding ORFs) than Zhang *et al.* [100] a slightly different procedure is followed and is described as follows:

- Find the length of the intergenic sequences between any two adjoining coding ORFs. Ignore the sequence if its length is $<$ 12 bp.

- For all sequences $\geq$ 12 bp, starting from the first base, search for the first start codon (i.e., ATG) along the sequence. In the downstream direction, starting from the 4th codon (1 codon = 3 bps) beginning from ATG, search for the last stop codon encountered. Then the DNA sequence starting from ATG and ending with one of the stop codon is regarded as one candidate for the intergenic sequences. Note that this is not an ORF because there often may be several stop codons within it.

The above procedure provides 5635 non-coding ORFs sequences with average length $\approx$ 607 bp which is better than that provided by [100]. The MIPS database [142] categorizes the complete ORFs of *Yeast* genome into six different classes, i.e., known proteins, strong similarity to known proteins, similarity or weak similarity to known proteins, similarity to unknown proteins, no similarity, and questionable ORFs. For our experiment the coding ORFs dataset that belong to known protein class which constitutes a large portion of coding ORFs is selected. This is done so that the system performance can first be verified with the known ORFs and later on can be applied for predicting novel ORFs. The *Yeast* genome consists of 5277 known protein coding ORFs.

Our next step is to calculate the feature vector value for each ORF. The feature extraction procedure requires specification of a wavelet filter and the maximum scale of decomposition (L). The maximum level of decomposition (L) for wavelet analysis is chosen depending on the minimum length of the DNA sequences (N) that is taken up for the study. Haar wavelet was selected as the wavelet filter for our experiments. However, experiments were also performed using other wavelets, i.e., Daubechies, least symmetric and Coiflet and it was found that the Haar wavelet provides the best performance parameters. The distribution of features values generated by x, y, and z components of Z-curve for coding/non-coding DNA sequences of *yeast* genome is provided in Figure 6.9, Figure 6.10, Figure 6.11 respectively.

In Table 6.1, the performance measures i.e., sensitivity (SE), specificity (SP), and accuracy (ACC) obtained by our technique for the *Yeast* genome is provided. Each training dataset of Table 6.1 consists of equal number of positive and negative samples. For example, when the training dataset has 250 samples then this means that the training dataset has 125 positive samples (i.e., coding ORF) and 125 negative samples (i.e., non-coding ORF). The maximum accuracy provided by the optimized classifier is equal to 92.91%. From Table 6.1, it is also observed that the classifier performance with different value of (L) is almost same. In fact, for smallest level of decomposition (i.e., for smaller DNA sequences) highest accuracy level is achieved.

As said earlier that the wavelet variance feature vector need selection of a wavelet filter. So, experiments were also performed with various other wavelets (other than Haar) on the same datasets that were taken for the Haar wavelet (Table 6.1). The performance parameters obtained are given in Table 6.2. From the table it is observed that the accuracy is almost similar for all the wavelets that were taken up for the study. However, the Haar wavelet performs (Table 6.1) slightly better than other wavelet filters (Table 6.2).

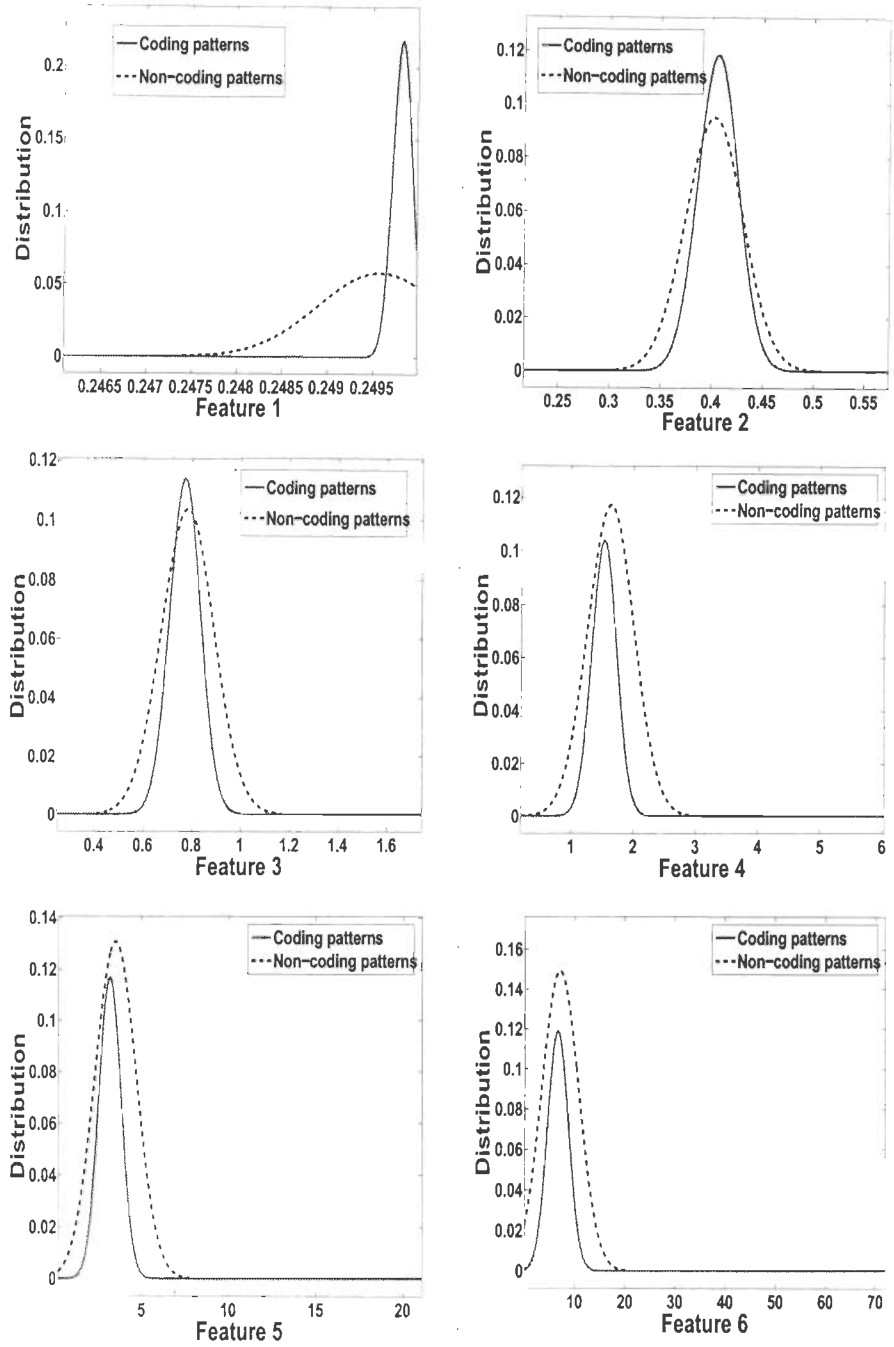


Figure 6.9: Distribution of features (no. 1 to no. 6, i.e., x-component or purine/pyrimidine) for coding/non-coding DNA sequences of *yeast* genome.

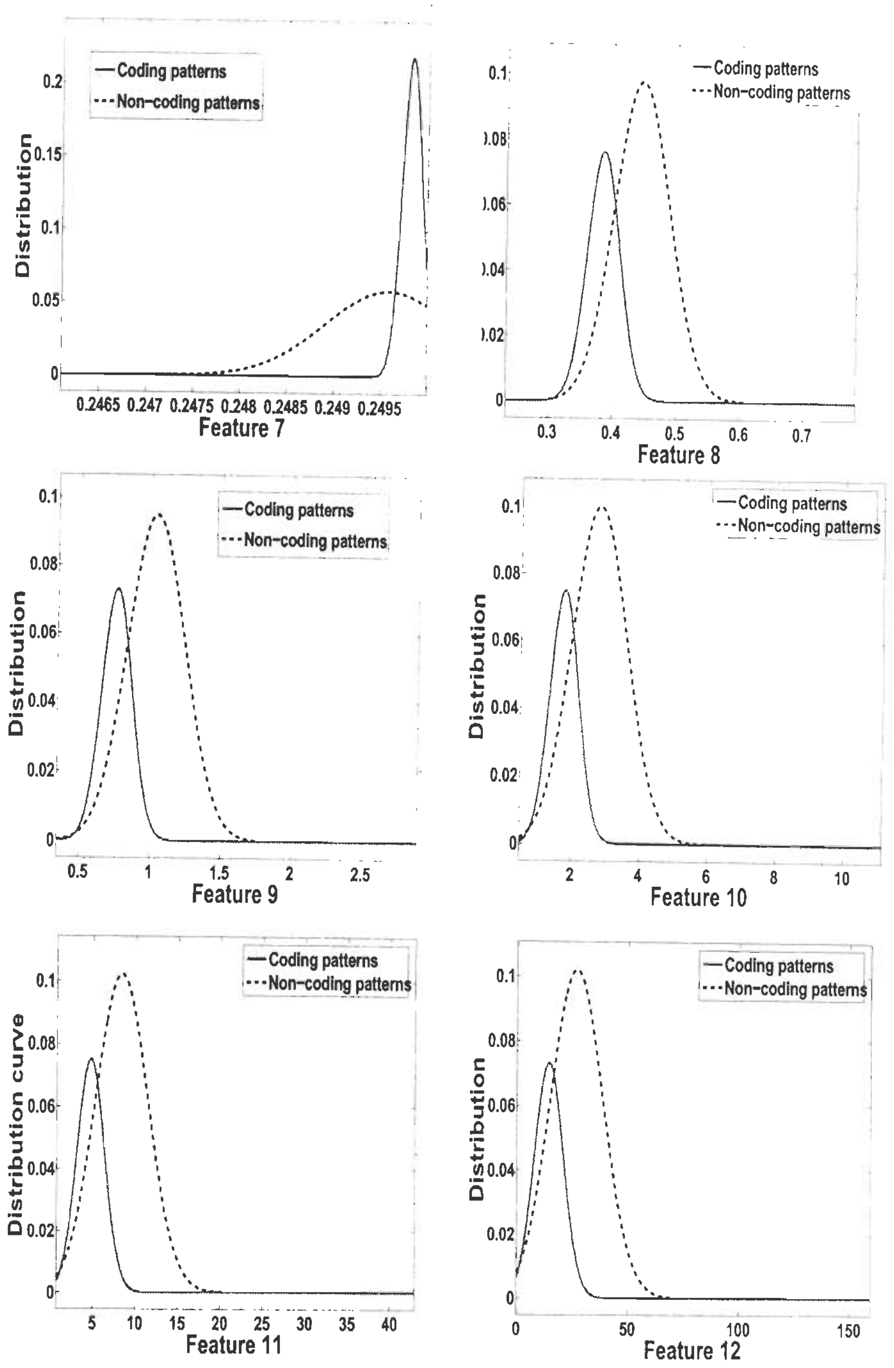


Figure 6.10: Distribution of features (no. 7 to no. 12, i.e., y-component or amino/keto) for coding/non-coding DNA sequences of *yeast*.

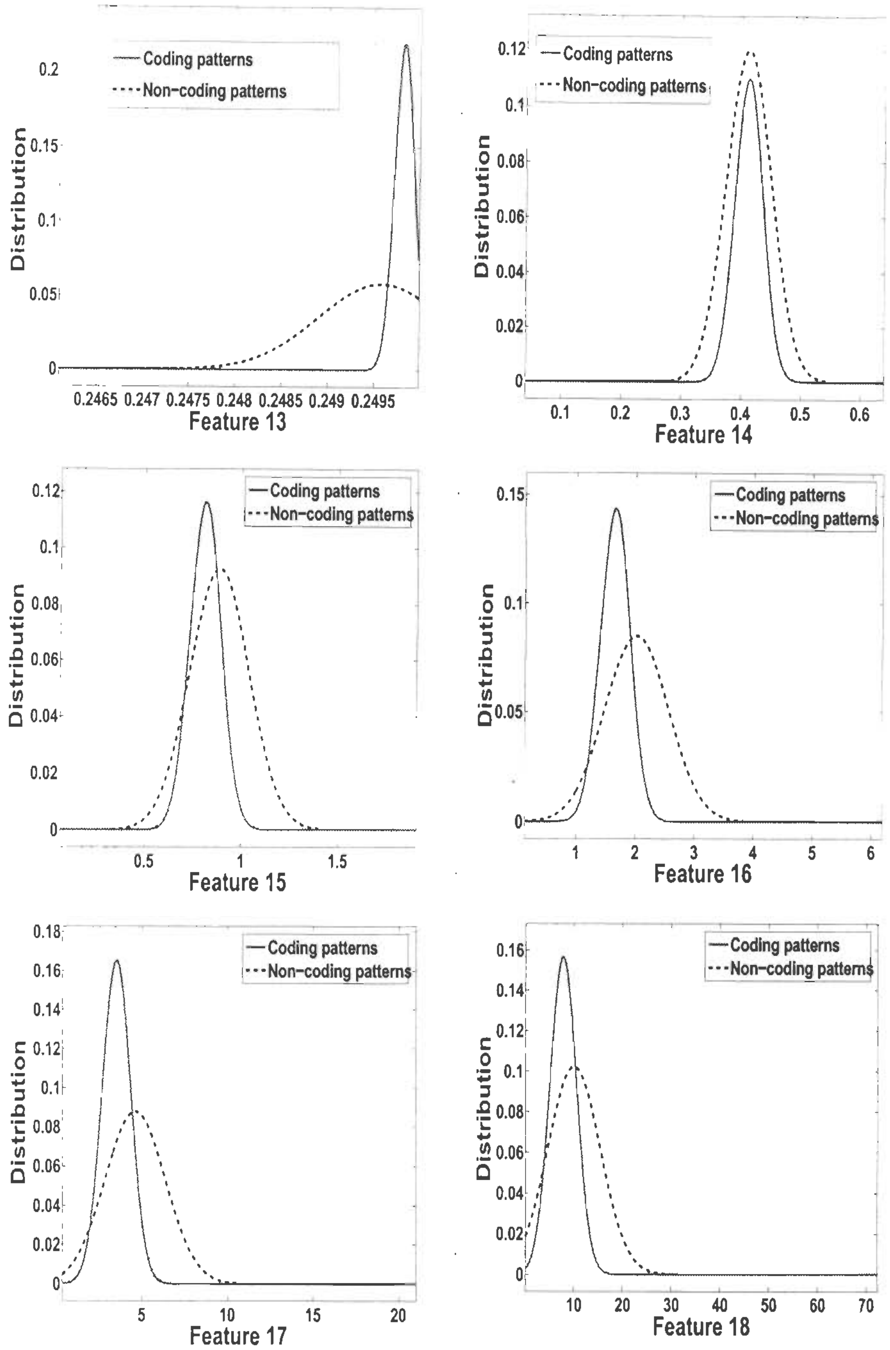


Figure 6.11: Distribution of features (no. 13 to no. 18, i.e., z-component or strong-H/weak-H bond) for coding/non-coding DNA sequences of *yeast* genome.

Table 6.1: The performance of our approach for *Yeast* genome datasets.

# Training	# Test	$L=6$,$N\geq 64$bp			$L=7$,$N\geq 128$bp			$L=8$,$N\geq 256$bp		
data	data	$SE(\%)$	$SP(\%)$	$ACC(\%)$	$SE(\%)$	$SP(\%)$	$ACC(\%)$	$SE(\%)$	$SP(\%)$	$ACC(\%)$
250[a]	1750[a]	91.4	91.3	91.4	88.5	90.7	89.6	77.8	92.3	85.1
500[a]	1500[a]	92.4	89.1	91.3	92.0	88.0	90.0	89.9	86.8	88.3
750[a]	1250[a]	90.7	90.6	90.6	95.2	92.2	93.7	89.9	89.1	89.5
1000[a]	1000[a]	90.8	90.6	90.7	94.6	89.8	92.2	90.6	89.6	90.1
2000[a]	[b,c,d]	94.6	91.3	92.9	92.7	91.6	92.1	93.3	90.3	92.2

L is the level of decomposition for the wavelet variance.

N is the length of the Z-curve time series data (DNA sequence).

[a] datasets consist of equal number of positive (coding ORFs) and negative (non-coding ORFs) samples.

[b] Test dataset(L=6): (4277 positive samples, 4357 negative samples).

[c] Test dataset(L=7): (4268 positive samples. 3862 negative samples).

[d] Test dataset(L=8): (4206 positive samples, 2520 negative samples).

Table 6.2: The performance of our approach for *Yeast* genome datasets using different wavelet types.

# Training sample = 2000[a], $L = 6$,$N \geq 64$bp, Feature vector dimension = 18								
Daubechies[b]			Least symmetric[b]			Coiflet[b]		
$SE(\%)$	$SP(\%)$	$ACC(\%)$	$SE(\%)$	$SP(\%)$	$ACC(\%)$	$SE(\%)$	$SP(\%)$	$ACC(\%)$
94.0	91.5	92.7	93.5	91.3	92.4	94.0	91.4	92.7

[a] Training dataset:(1000 positive samples, 1000 negative samples).

[b] Test dataset:(4277 positive samples, 4357 negative samples).

Next, a performance comparison of wavelet variance coefficients (WVC) features with slope based features proposed by Zhang *et al.* in [100] is performed. The performance parameters obtained by both approaches are provided in Table 6.3. This time, the classifier was trained with large number of samples in order to construct more accurate model. From the result it was observed that the accuracy achieved by WVC features is slightly better than the slope features. Even though the wavelet and slope feature provide different information about the DNA sequence, the accuracy achieved by both features is similar.

The aim of this study was not meant to replace previous coding measures, rather, to act as a complement to these already widely used measures. Next, a new feature vector which is a combination of wavelet variance and slope features is defined. The results that were obtained after combining both of the feature sets have been provided in Table 6.4. From the table it is observed that the accuracy of optimized SVM classifier has increased to 96%. Thus, a classifier designed using both wavelet and slope features provides better results than the classifier based on single features set.

6.4.4 Results on *E.Coli* genome

The genome of *E. coli* was downloaded from GenBank database. The information about the protein coding regions in the genome was extracted from the header of the *E. coli* genome file of NCBI (GenBank accession number = NC_000913). The non-coding ORFs were constructed similar to that were constructed for the *Yeast* genome, discussed in the previous section 6.4.3. The complete positive (ORF) and negative (non-ORF) samples were divided into training and test datasets. The samples for training and test datasets were selected randomly. Later on, 10-fold cross-validation was performed for constructing an optimized SVM model.

Table 6.3: Comparison of SVM classifier performance measures based on wavelet variance and slope features of Z-curve components for *Yeast* genome.

Level[a]	Min. ORF	Wavelet variance features			Slope-based features[e]		
	length	$SE(\%)$	$SP(\%)$	$ACC(\%)$	$SE(\%)$	$SP(\%)$	$ACC(\%)$
6[b]	64	95.2	90.6	92.9	90.9	91.7	91.3
7[c]	128	95.5	91.7	93.6	91.6	92.6	92.1
8[d]	256	95.6	88.3	91.9	94.0	89.8	91.9

[a] Test dataset:(1000 positive samples, 1000 negative samples).

[b] Training dataset:(4277 positive samples, 4357 negative samples), Feature vector length=18.

[c] Training dataset:(4268 positive samples, 3862 negative samples), total features = 21.

[d] Training dataset:(4206 positive samples, 2520 negative samples), total features = 24.

[e] Feature vector length = 10.

Table 6.4: Classification performance achieved using both wavelet variance and slope based features for *Yeast* genome.

Level[a]	Minimum	Wavelet variance and Slope-based features		
	ORF length (bp)	$SE(\%)$	$SP(\%)$	$ACC(\%)$
6[b]	64	96.5	94.9	95.7
7[c]	128	94.9	96.9	95.9
8[d]	256	96.1	93.2	94.6

[a] Test dataset:(1000 positive samples, 1000 negative samples).

[b] Training dataset:(4277 positive samples, 4357 negative samples), Feature vector length = 28.

[c] Training dataset:(4268 positive samples, 3862 negative samples), Feature vector length = 31.

[d] Training dataset:(4206 positive samples, 2520 negative samples), Feature vector length = 34.

Table 6.5: Classification performance achieved using combined feature vector (wavelet variance and slope features) for *E.Coli* genome.

Level	Minimum	Combined feature vector		
	ORF length (bp)	$SE(\%)$	$SP(\%)$	$ACC(\%)$
6^a	64	96.9	95.6	96.2
7^b	128	95.4	95.6	95.5

[a] Training: (3234 positives and 1110 negative samples), Test: (1000 positive and 1000 negative samples).

[b] Training: (3213 positives and 951 negative samples), Test: (1000 positive and 500 negative samples).

Experiments were performed for $L = 6$ and 7 with Haar wavelet. The performance parameters obtained using combined feature vector (i.e., wavelet variance and slope) are provided in Table 6.5. An accuracy of more than 96% is achieved for *E. coli* genome.

6.5 Conclusion

Based on the novel wavelet variance coefficient feature vector, a pattern classification framework is provided for classifying the DNA sequences into coding and non-coding sequence class. The study of wavelet variance of the time series data is of interest because it decomposes (analyzes) the variance of stochastic processes on a scale-by-scale basis and hence provides information from local to global variation present in data. The wavelet variance is also of interest because it is related to the concept of spectral density function (SDF). Further, MODWT has been used for calculating variance because it offers several advantages over DWT and is especially very useful for analysis of time series data. This approach also provides a generic methodology for analyzing DNA sequences.

The optimized machine learning parameters of the SVM system was obtained using hierarchical grid based parameter selection techniques. The wavelet variance coefficient features obtained by our method for the Z-curve components yield more than 93% accuracy for *Yeast* genome dataset. The major challenge in classification of DNA sequences is that there is no exact information about the features or measures that makes a DNA sequence protein coding or not. It is totally based on the statistical study of the DNA sequences. Also, it is always better to use more than one type of information of an object for improving the classification accuracy. Thus, later on SVM classifiers using WVC features and slope based features is designed. For the combined feature vector it was observed that the accuracy of 10-fold cross-validated classifier reaches upto 96% for the *Yeast* genome and more than 96% for the *E. coli* genome datasets.

Chapter 7

Identification and Classification of GPCRs

G-protein-coupled receptors (GPCRs) are one of the largest group of proteins in vertebrate species. Their classification and functional annotation is very important in today's medical and pharmaceutical research because GPCRs play important roles in cellular signalling networks involving such processes as neurotransmission, cellular metabolism, secretion, cellular differentiation and growth, inflammatory and immune responses, smell, taste and vision. GPCRs have been proved to be one of the most attractive targets for pharmaceutical intervention. In this chapter, a novel feature vector for efficient identification and classification of GPCRs, GPCR family, GPCR subfamily and GPCR sub-subfamily is presented. The feature vector is based on wavelet variance of protein profiles. First, the various tasks involved in the classification system are presented followed by performance evaluation of the system on datasets downloaded from GPCR database (GPCRDB). A comparison of the proposed technique with SVMpred on standard is also presented.

The existing SVM based approaches for GPCRs classification are either based on amino acid or dipeptide composition in a protein sequences. The dipeptide based SVM approach is the most accurate technique to identify and classify the GPCRs. However, this approach has two major disadvantages. Firstly, the dimension of dipeptide based feature vector is equal to 400. The large dimension makes the classification task computationally and memory wise inefficient. Secondly, it does not consider the biological properties of protein sequence for identification and classification of GPCRs. This chapter introduces a novel feature vector based on variation of seven key biological properties of amino acids in a protein sequence. The feature vector is calculated after performing a wavelet based time series analysis technique of protein sequences. In addition, the dimension of the feature vector is also reduced to 35.

7.1 Introduction

G protein-coupled receptors (GPCRs), also known as seven transmembrane receptors, 7TM receptors, heptahelical receptors, and G protein linked receptors (GPLR), are a protein family of transmembrane receptors that transduce an extracellular signal (ligand binding) into an intracellular signal (G protein activation). A diagram of GPCR protein is shown in Figure 7.1. They are among the largest and most diverse protein families in mammalian genomes. GPCRs have been aggressively pursued as drug targets due to their central role in physiological processes affecting almost all aspects of the life cycle of an organism [143]. It is estimated that about 50% of all current drug targets are GPCRs and are the most successful of any target class in terms of therapeutic benefit [144, 145]. The 3-dimensional structure of GPCRs are largely unsolved, except for that of one GPCR (Bovine rhodopsin) [146]. In contrast, the amino acid sequences of thousands of GPCR-related proteins are known [147, 148, 149]. In the absence of experimental data, computational methods are frequently used to facilitate identification and characterization of novel receptors. Due to vital role of GPCRs and enormous data, developing an accurate and faster technique for automatic classification of GPCRs and its families, subfamilies and sub-subfamilies is of prime importance.

GPCRs have been divided into six principal classes or families generated based on sequence similarities: class A (rhodopsin-like), class B (secretin-like), class C (metabotropic glutamine/pheromone), class D (fungal pheromone), class E (cAMP receptors), and the Frizzled/Smoothened class [147]. Each class is further divided into families based on their ligand specificity, with some families combined into larger groups based on closely related ligands. For example, the class A GPCRs include groups such as amine binders, peptide binders, prostanoid receptors, and olfactory receptors. The amine binding group, for instance is formed by seven families (acetylcholine receptor, adrenoceptor, dopamine receptor, histamine receptor, serotonin receptor, octopamine receptor and trace amine receptor). In humans, there are three major families of GPCRs (the rhodopsin, secretin and metabotropic glutamine receptor-like superfamilies) comprising more than 50 subfamilies and 350 sub-subfamilies [150, 151].

In the past, many strategies have been proposed for identifying novel GPCRs. Sequence alignment techniques such as BLAST and FASTA coupled with pattern database (PRINTS) are the simplest and frequent programs that have been used for identifying novel GPCRs through their similarity to known receptors. In [152], an automatic classification technique based on the signature search against pattern/motif databases with highly improved diagnostic performance is presented. However, the above mentioned techniques fail when query protein

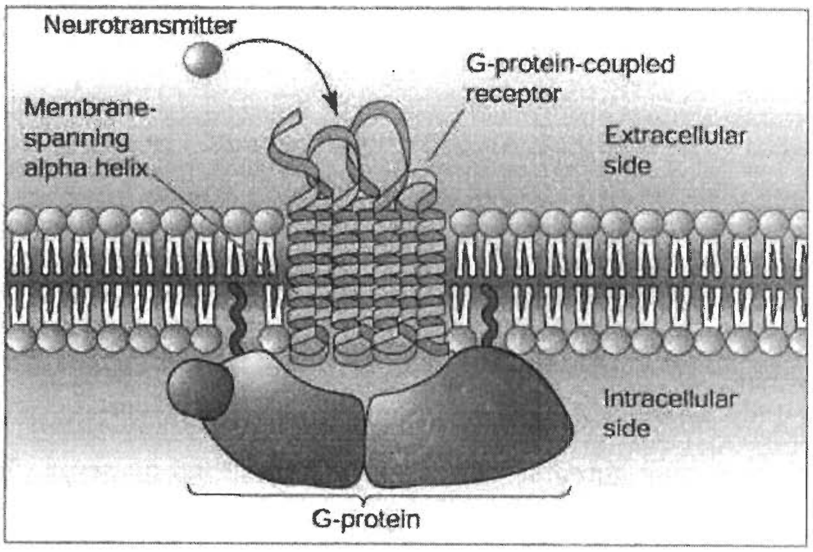


Figure 7.1: G protein-coupled receptor protein.

lacks significant similarity to the database sequences. To overcome the limitation of sequence comparison techniques, a support vector machine (SVM) based technique was presented in [153, 154]. This technique is based on the feature vectors which are derived from amino acid composition of the protein sequences. A more accurate SVM based technique (GPCRpred) was proposed in [155]. GPCRpred is based on feature vectors derived from dipeptide composition of the protein sequence. Dipeptide composition encapsulates information about the fraction of amino acids as well as the local order.

The GPCRpred program has two major disadvantages. Firstly, the dimension of the feature vector is very high (i.e., 400). The large dimension of feature vector makes the classification task quite expensive in terms of computational and memory used. Secondly, the key physicochemical properties of an amino acid sequence that can probably reveal more properties of a protein family have been ignored by the exiting SVM based approaches for identification and classification of GPCRs. The objectives of this chapter are:

1. To provide a novel feature vector based on variation of biological (physicochemical) properties of amino acids in GPCRs sequences.

2. To reduce the dimension of fixed length feature vector for designing faster classifiers.

3. To improve the accuracy of identification and classification system for GPCRs.

The novel feature vector presented in this chapter is inspired from the work of Vannucci and Lio [32, 128, 129] for transmembrane proteins. In [32], the authors had shown how non-decimated wavelet transforms and the wavelet variance [104] scale-by-scale decomposition can be applied to extract features from hydrophobicity profiles of transmembrane proteins. In this chapter, a feature vector which is a collection of wavelet variances value is proposed. The wavelet variance values are obtained after performing a scale-by-scale decompositions of seven key physicochemical properties of amino acids. The dimension of the proposed feature vector is also reduced to 35. Later on, an optimized SVM based classification system was designed to identify a given protein sequence as GPCR or not. Once identified as GPCR protein sequence, it was further classified into its family, subfamily and sub-subfamily. In section 7.3.3, the strategy followed for classification of GPCRs protein sequences has been provided.

7.2 Literature Review

In this section, a brief review of available techniques for identification and classification of GPCRs is presented.

7.2.1 Sequence similarity method

The simplest and most frequently used techniques for identifying proteins related to a query sequence is to search a sequence database using pairwise alignment tools, such as the Basic Local Alignment Search Tool (BLAST) and FASTA families of programs [36, 38, 39, 156]. The strength of the match is judged by a score based on the similarity of two biological sequences after alignment.

Sequence identity, which is the fraction of the pairwise alignment identical between the query and the reference sequence, is an alternative metric for judging the likelihood of two sequences being homologous. It is thought that when sequence identity between two aligned sequences falls below 20-25%, homology, and therefore shared function, can no longer be reliably inferred. Moreover, the accuracy of machine-generated alignments decays as the intrinsic pairwise sequence similarity decreases [157], further complicating matters. Because receptors that share a natural ligand can have pairwise sequence identities below 25% (e.g., the histamine receptors [158]), there are no hard cutoffs to positively associate an orphan to a known GPCR, much less a receptor that shares the same ligand. For example, using BLAST to compare the protein sequence of human histamine H4 receptor to that of histamine H1 receptor yields 26% identity over the length of the match, whereas the sequence identity over the length of the match, whereas the sequence identify between the human somatostatin receptor type 1 with a receptor with different ligand (nociceptor receptor) is actually higher at 43%. Nevertheless, pairwise sequence comparison is a convenient approach and one that has certainly worked in the identification of putative GPCRs of unknown function and in some cases can provide strong clues as to the natural ligand as well.

7.2.2 Motif-based classification

To overcome some of the problems of pairwise methods, motif or signature based method were developed for classification of GPCRs. Typically, motifs are derived by parsing multiple alignments into consensus sequences, the conserved columns of which reflect important structural and/or functional residues. The first approach, illustrated by PROSITE [159, 160], is to use only a single conserved region encoded as a consensus sequence or regular expression. When searching PROSITE, such regular expressions have to be matched exactly, sometimes leading to high error rates. Many true relationships are missed because sequences deviate slightly from the expression, and many false matches are made because the patterns are short and non-selective. The EMOTIF database also uses regular expressions to encode regions of conservation [161]. Here, however, different variants of each expression are derived, offering greater flexibility to capture distant family members.

The PRINTS database uses a different approach: knowing that protein families are likely to contain more than one conserved region, using all such motifs to create a characteristic fingerprint improves diagnostic performance [162, 163]. Although individual motifs are short, greater selectivity is achieved by exploiting the mutual context of motif neighbours. False-positive matches can then be more easily distinguished from family members, as they usually fail to match several motifs within a given fingerprint. The Blocks database also uses multiple motifs (known as blocks) to describe families [164]. In contrast with PRINTS manual approach, blocks are automatically derived. The scoring methods also differ from those of PRINTS, as blocks are encoded as weight position-specific scoring matrices, rather than the frequency matrices typical of fingerprinting.

7.2.3 Profile-based method

Profiles are statistical descriptions of the primary sequence consensus of a gene family derived from a multiple sequence analysis. Whereas pairwise alignment methods such as BLAST use position-independent scoring (i.e., in the case of proteins, the incremental value of aligning two given amino acids is the same irrespective of its location in the overall sequence), profile methods use position-specific scores for the placement of various amino acids. Profiles are commonly represented in a statistical model called a hidden Markov model (HMM) [35]. The HMM is able to estimate the probability that a query sequence was generated by the model itself. Like BLAST, one metric for evaluating a match between a query sequence and the HMM is also the E-value, corresponding to the number of hits that would be expected to have a score equal or better by chance alone.

In practice, one would gather member of each protein family (or subfamily) and train an HMM to represent the group. After building HMMs for all families of interest, one would then match the sequence with unknown function (the query) against them all, and assign membership to the family corresponding to the best E-value (or alternatively, the best score). The predictive power of the model is a function of several variables, such as exact algorithm used to train the models, as well as the accuracy of any multiple sequence alignment used to guide the model training.

Papasaikas *et al.* presented PRED-GPCR program for GPCR recognition and and classification at the family level [165, 152] using signature profile HMMs corresponding to some well characterized GPCR families. In [166], the authors proposed a phylogenic tree-based profile hidden Markov model(T-HMM) based on chemical and physical property of amino acids for classification and identification of GPCRs. The chemical and physical properties of amino acid sequence reveal more properties of protein family where pure sequence-based methods fail to find some detailed information beyond the discriminative power of 20 characters. The technique is divided into two main branches: (1) ligand-based classification and (2) G-protein coupling-based classification. It achieved overall predictive accuracy of 99.9% for ligand group-based classification (i.e., amine vs. peptide binding), and over 99% for ligand family-based classification. In addition, the G-protein coupling specificity-based classification provides 83% accuracy over large dataset.

A major disadvantage of signature based approach is absence of large signatures database. Furthermore, these signatures are collected by experts. Also, it is not known that whether this method is useful for G-protein coupling specificity but the SVM method can be used to identify the G-protein coupling specificity [167].

7.2.4 SVM-based method

Amino acid composition–Protein information can be encapsulated in a vector of 20 dimensions, using amino acid composition of the protein. This amino acid composition is the fraction of each amino acid type within a protein. The fractions of all 20 amino acids are calculated using the following equation,

$$\text{Fraction of AA}(i) = \frac{\text{Total number of amino acids of type } i}{\text{Total number of amino acids in protein sequence}} \tag{7.1}$$

where i is an amino acid i out of 20 amino acids (AA).

Dipeptide composition–The dipeptide composition provides global information of a protein sequence in the form of a fixed-length vector. Dipeptide encapsulates information about the fraction of amino acids as well as their local order. The dipeptide composition of each protein is calculated using the following equation,

$$\text{Fraction of Dipep}(i) = \frac{\text{Total number of Dipep}(i)}{\text{Total number of all possible dipeptides}} \tag{7.2}$$

where Dipep(i) is a dipeptide i out of 400 dipeptides.

Support vector machines (SVMs) are statistical machine learning algorithm that has been successfully used to classify biological sequence [168]. Like HMMs, SVMs are trained from labeled (classified) data. Unlike HMMS, the SVM approach enables training on both positive and negative examples of family membership and hence is discriminate examples from different classes.

Karchin *et al.* [153, 154] applied SVM based approach for classification and identification of GPCRs. The fixed length feature vector for SVM was constructed using amino acid composition as given in (equation 7.1). This method provided an accuracy of 99.3% for recognition of GPCRs, and its accuracy for prediction of family and sub-subfamily is 88.4% and 86.3% respectively.

In [155], a more accurate technique (GPCRpred) than amino acid composition based SVM is presented. GPCRpred uses dipeptide composition (equation 7.2) based fixed length feature vector for classification purpose. This technique is composed of three-steps for annotating GPCRs: (i) it predicts whether the query sequence belongs to the GPCR superfamily or not; (ii) it predicts class or family of GPCRs; and (iii) it predicts the GPCR sub-family if it belongs to class A of GPCRs. GPCRpred provides an accuracy of 99.5% for recognition of GPCRs. Similarly, it achieved an overall accuracy of 91.3% and 96.4% at family and sub-family level respectively.

In [42], a fast Fourier transform-based support vector machine technique (Pred-GPCR) has been proposed for predicting GPCR subfamilies according to protein's hydrophobicity. In this method, a 512 point FFT has been applied to hydrophobicity profile of protein sequences for constructing a fixed length feature vector for classification purpose. In classifying class B, C, D and F subfamilies, the method achieved an accuracy of 93.3%.

The dipeptide composition based method (GPCRpred) is better for identifying and classifying GPCRs superfamily than the amino acid composition based methods: phylogenic tree-based profile hidden markov model (T-HMM), bagging classification tree and SVM (Karchin *et al.* [153, 154]) achieves lower accuracy than dipeptide method SVM [155]. However GPCRpred classification technique has two major drawbacks. First, this technique suffers from high dimensionality problem because the length of dipeptide based feature vector is equal to 400. Secondly, the feature vector is based only on dipeptide composition.

7.2.5 Clustering-based method

In [169], a bagging classification tree algorithm is proposed to predict the type of receptor based on its amino acid composition. For construction of classification tree the C4.5 was implemented. The C4.5 algorithm uses a divide-and-conquer approach for growing decision trees. This algorithm selects a feature (i.e., an amino acid composition) to split the training data into subsets (i.e., nodes of the decision tree). The default criterion used by C4.5 for feature selection is 'information gain ratio', a measure based on information theory. This measure can quantify how well a given feature separates the training data. At each node the training dataset will be further divided until some stopping criteria are satisfied. Then each terminal subset (leaf node) is assigned to a class label (receptor type). After generating the maximal classification tree, a pruning technique is used to simplify the classification tree and avoid over-fitting. Pruning a tree consists of replacing a whole sub-tree by a leaf node. The replacement takes place if the expected error rate in the subtree is greater than that in the single leaf. Further, the bootstrap aggregating (bagging) procedure is applied to improve the prediction accuracy reported by a single classification tree. In the bagging procedure, bootstrap samples are formed by drawing at random with replacement from the original learning set. Classifiers (different classification trees) are built for each bootstrap sample, then the multiple classifier are aggregated by majority votes, i.e., the final class is the one predicted by the majority of the predictors. The bagging classification tree algorithm provides an overall accuracy of 86.9% for GPCRs sub-family classification and 81.5% for sub-subfamily classification.

7.3 Classification System

The pattern recognition framework developed for the current classification problem is shown in Figure 7.2. The complete framework is divided into three different stages: amino acid mapping, feature construction and classification. The first step involves the transformation of the amino acid sequence into a numerical sequence. Numerical series obtained are then analyzed using wavelet based time series analysis technique for extracting information in terms of wavelet variance [104]. The original numerical sequence is thus transformed into variances using non-decimated wavelet transform or maximal overlap discrete wavelet transform (MODWT) [130, 26, 131] technique. The wavelet variance values obtained from seven numerical sequences are later on combined to form a fixed length feature vector. Finally, an SVM based classification is performed using the wavelet variance feature vector to identify protein classes.

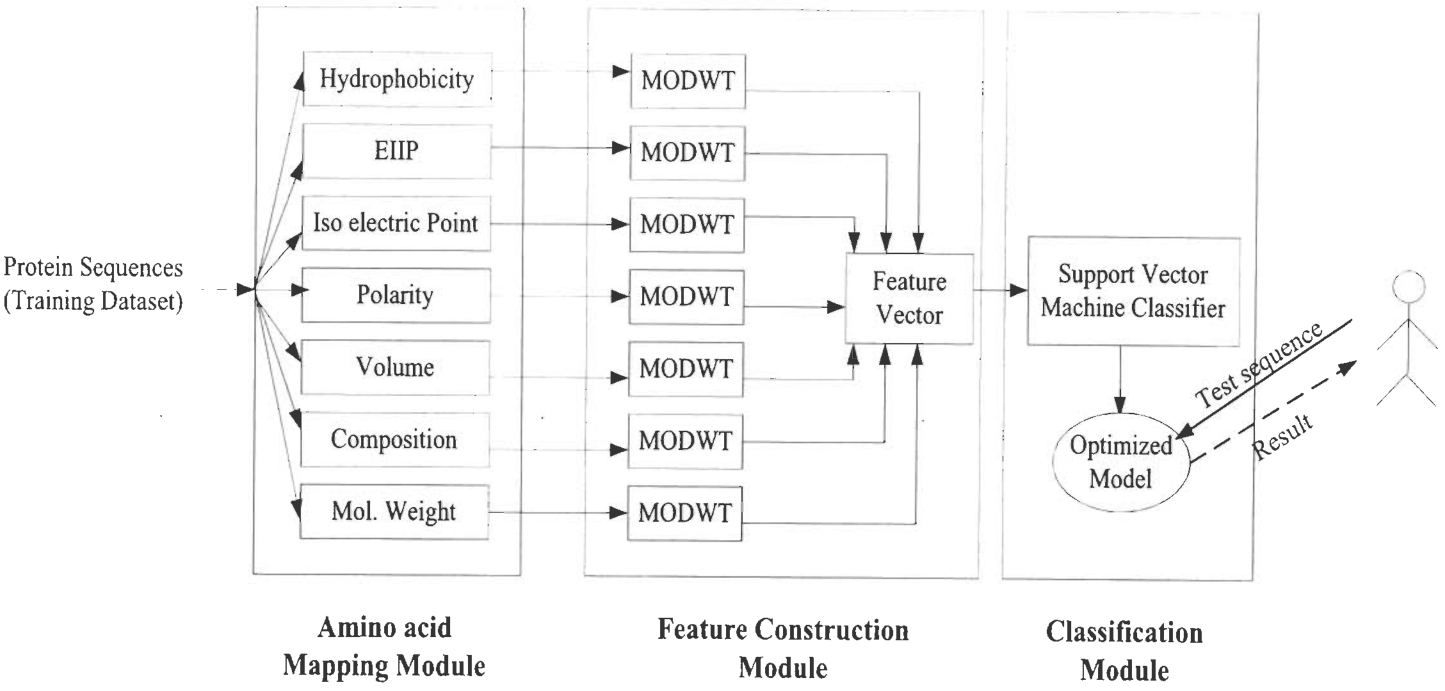
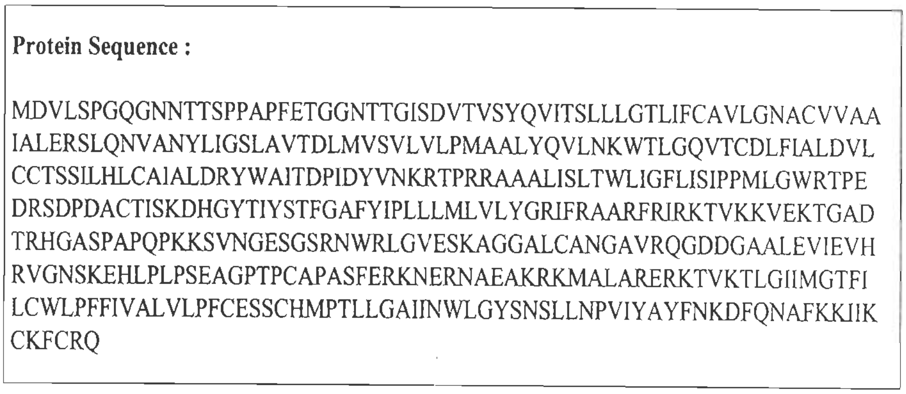


Figure 7.2: A block diagram of the proposed classification system.



Protein Sequence :

```
MDVLSPGQGNNTTSPPAPFETGGNTTGISDVTVSYQVITSLLLGTLIFCAVLGNACVVAA
IALERSLQNVANYLIGSLAVTDLMVSVLVLPMAALYQVLNKWTLGQVTCDLFIALDVL
CCTSSILHLCAIALDRYWAITDPIDYVNKRTPRRAAALISLTWLIGFLISIPPMLGWRTPE
DRSDPDACTISKDHGYTIYSTFGAFYIPLLLMLVLYGRIFRAARFRIRKTVKKVEKTGAD
TRHGASPAPQPKKSVNGESGSRNWRLGVESKAGGALCANGAVRQGDDGAALEVIEVH
RVGNSKEHLPLPSEAGPTPCAPASFERKNERNAEAKRKMALARERKTVKTLGIIMGTFI
LCWLPFFIVALVLPFCESSCHMPTLLGAIINWLGYSNSLLNPVIYAYFNKDFQNAFKKIIK
CKFCRQ
```

Figure 7.3: Amino acid sequence of 5-hydroxytryptamine 1A receptor in Human.

7.3.1 Amino acid mapping

Seven principal properties of amino acids, i.e., hydrophobicity, electronic, isoelectric point, polarity, volume, composition and molecular weight are used for amino acid mapping. The amino acid mapping scales are discussed in this section.

Hydrophobicity

The hydrophobicity determines the structure and function of protein, especially for the transmembrane proteins. The hydropathy scale provided by Kyte and Doolittle in [170] is used for experiment and is provided in Table 7.1. The scale has been composed wherein the hydrophilic and hydrophobic property of the 20 amino acids side-chains is taken into consideration. The scale is based on an amalgam of experimental observations derived from the literature. Figure 7.4 shows the hydrophobicity profiles generated from the amino acid sequences of 5-hydroxytryptamine 1A receptor in Human as shown in Figure 7.3.

Table 7.1: Hydropathy scale for amino acids.

Amino acid	Hydropathy index
Isoleucine (I)	4.5
Valine (V)	4.2
Leucine (L)	3.8
Phenylalanine (F)	2.8
Cysteine/cystine (C)	2.5
Methionine (M)	1.9
Alanine (A)	1.8
Glycine (G)	-0.4
Threonine (T)	-0.7
Tryptophane (W)	-0.9
Serine (S)	-0.8
Tyrosine (Y)	-1.3
Proline (P)	-1.6
Histidine (H)	-3.2
Glutamic acid (E)	-3.5
Glutamine (Q)	-3.5
Aspartic acid (D)	-3.5
Asparagine (N)	-3.5
Lysine (K)	-3.9
Arginine (R)	-4.5

Table 7.2: The electron-ion interaction potential (EIIP) values for amino acids.

Amino acid	**EIIP**
Isoleucine (I)	0.0000
Valine (V)	0.0057
Leucine (L)	0.0000
Phenylalanine (F)	0.0946
Cysteine/cystine (C)	0.0829
Methionine (M)	0.0823
Alanine (A)	0.0373
Glycine (G)	0.0050
Threonine (T)	0.0941
Tryptophane (W)	0.0548
Serine (S)	0.0829
Tyrosine (Y)	0.0516
Proline (P)	0.0198
Histidine (H)	0.0242
Glutamic acid (E)	0.0058
Glutamine (Q)	0.0761
Aspartic acid (D)	0.1263
Asparagine (N)	0.0036
Lysine (K)	0.0371
Arginine (R)	0.0959

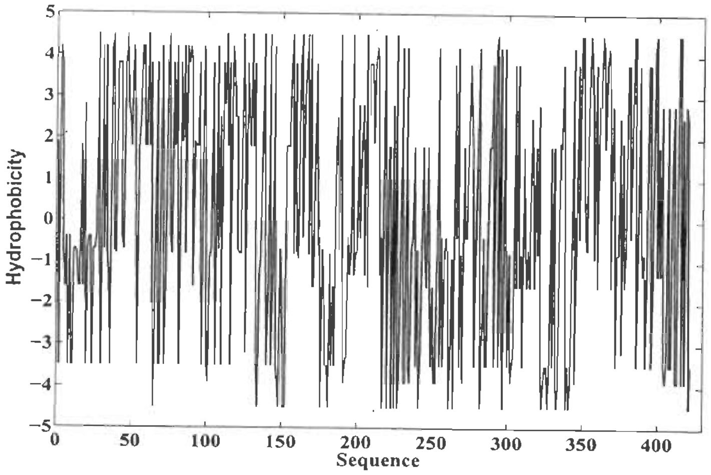


Figure 7.4: Hydrophobicity profile of 5-hydroxytryptamine 1A receptor in Human.

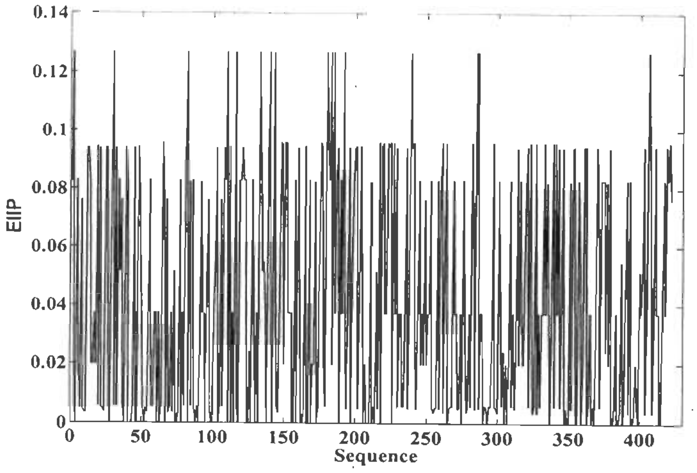


Figure 7.5: EIIP profile of 5-hydroxytryptamine 1A receptor in Human.

Electron-ion interaction potential (EIIP)

The EIIP describes the average energy of all valence electrons of amino acid sequences. The EIIP values for each amino acid are calculated using the following general model pseudopotential [171]:

$$\langle \overrightarrow{k+q}|w|\vec{k}\rangle = 0.25Z\sin(\pi 1.04Z)/(2\pi) \tag{7.3}$$

where q is a change of momentum of valence electron in the interaction with potential w,while

$$Z = \Sigma(Z_i)/N \tag{7.4}$$

where Z_i is the number of valence electrons of the i-th component of each amino acid and N is the total number of atoms in the amino acids. The EIIP values for 20 amino acids are shown in Table 7.2. Each amino acid irrespective of its position in a sequence, can thus be represented by a unique number. Figure 7.5 shows the EIIP profiles generated from the amino acid sequences of 5-hydroxytryptamine 1A receptor in Human which is shown in Figure 7.3.

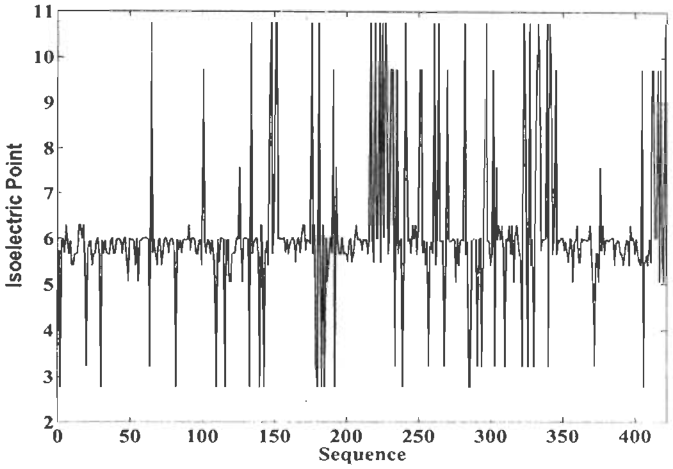


Figure 7.6: Isoelectric point profile of 5-hydroxytryptamine 1A receptor in Human.

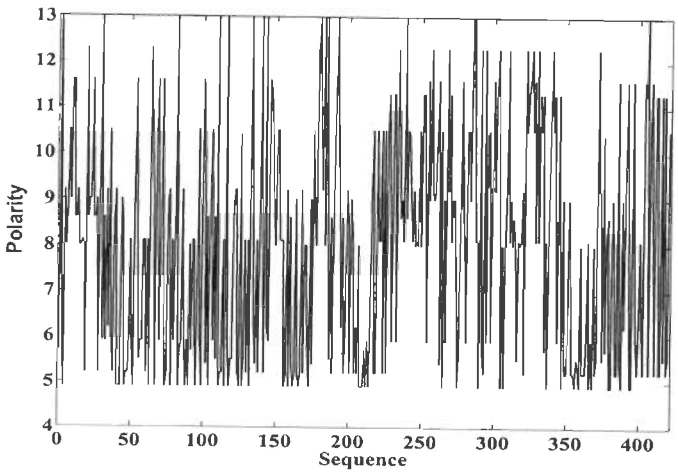


Figure 7.7: Polarity profile of 5-hydroxytryptamine 1A receptor in Human.

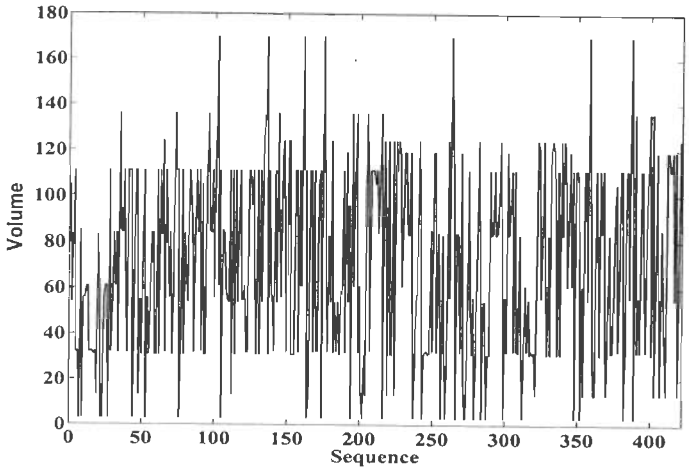


Figure 7.8: Volume profile of 5-hydroxytryptamine 1A receptor in Human.

Remaining physicochemical properties

The values of isoelectric point, polarity, volume, composition and molecular weights of the amino acids are obtained from *www.http://pages.pamona.edu/*

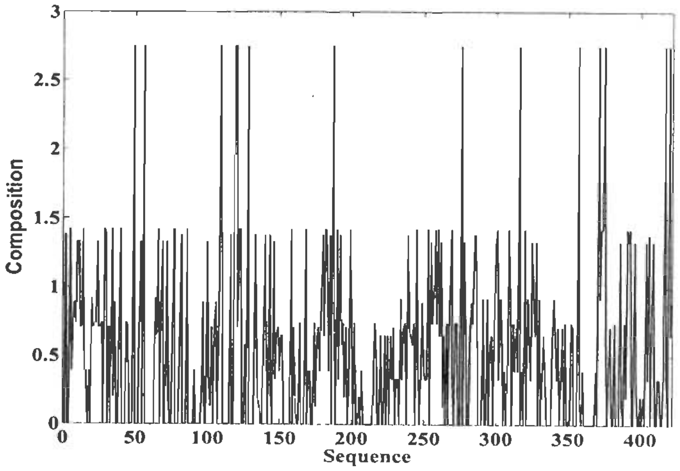


Figure 7.9: Composition profile of 5-hydroxytryptamine 1A receptor in Human.

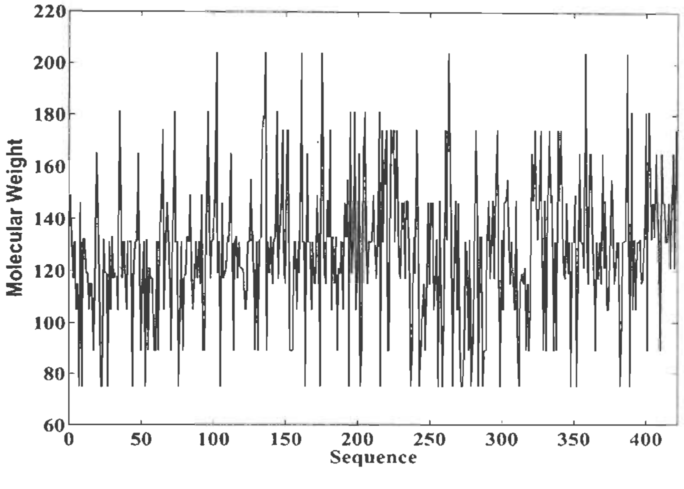


Figure 7.10: Molecular weight profile of 5-hydroxytryptamine 1A receptor in Human.

ac044747/aroc/Genetic_Code.swf.

Table 7.3: The isoelectric point, polarity, volume, composition and molecular weights for amino acids.

Amino acid	Isoelec. Point	Pol.	Vol.	Comp.	Mol. Weight
Isoleucine (I)	6.02	5.2	111	0.0	131.17
Valine (V)	5.96	5.9	84	0.0	117.15
Leucine (L)	5.98	4.9	111	0.0	131.17
Phenylalanine (F)	5.48	5.2	132	0.0	165.19
Cysteine/cystine (C)	5.05	5.5	55	2.75	121.16
Methionine (M)	5.74	5.7	105	0.0	131.17
Alanine (A)	6.0	8.1	31	0.0	89.09
Glycine (G)	5.94	9	3	0.74	75.07
Threonine (T)	5.66	8.6	61	0.71	119.12
Tryptophane (W)	5.89	5.4	170	0.13	204.23
Serine (S)	5.68	9.2	32	1.42	105.09
Tyrosine (Y)	5.66	6.2	136	0.2	181.19
Proline (P)	6.3	8.0	32.5	0.39	115.13
Histidine (H)	7.59	10.4	96	0.58	155.16
Glutamic acid (E)	3.22	12.3	83	0.92	147.13
Glutamine (Q)	5.65	10.5	85	0.89	146.15
Aspartic acid (D)	2.77	13.0	54	1.38	133.10
Asparagine (N)	5.41	11.6	56	1.33	132.12
Lysine (K)	9.74	11.3	119	0.33	146.19
Arginine (R)	10.76	10.5	124	0.65	174.20

7.3.2 Feature extraction

A wavelet transform decomposes a signal into several groups (vectors) of coefficients. Different coefficient vectors contain information about characteristics of the sequence at different scales. Coefficients at coarse scales capture gross and global features of the signal while coefficients at fine scales contains local details. In this section, a novel feature vector based on wavelet variance [104] is proposed. Unlike previous feature vectors [154, 155], the proposed feature vector contains information about the variability of seven key physiochemical properties of protein sequences over different scales. This helps in understanding the functionality of proteins from its primary sequence in a much better way than simple composition based studies [154, 155] and hence provide more accurate identification of protein classes.

Wavelets seem to be well suitable for the analysis of biological signals that present a multi-scale nature. A scale-by-scale wavelet decomposition of variance and correlation help in highlighting hidden structures of single sequences and similarities among different sequences. For example, wavelet coefficient profiles for 5-hydroxytryptamine 1A receptor (Human) and 5-hydroxytryptamine 1A receptor (Mouse) are shown in Figure 7.11 and Figure 7.12 respectively. From the figures it can be easily observed that the hydrophobic and EIIP profiles of two proteins is similar. In [32], authors have shown that wavelet variance decomposition of the hydrophobic profiles of proteins can be used for detecting the transmembrane regions of a protein. This means that the wavelet variance for a transmembrane protein will differ from non-transmembrane protein. The wavelet variance values obtained at different scale for 5-hydroxytryptamine 1A receptor (Human) and 5-hydroxytryptamine 1A receptor (Mouse) sequences using hydrophobic and EIIP profiles are shown in Figure 7.13 and Figure 7.14

respectively. From the figures it is clear that the wavelet variance values obtained from various level are similar for both GPCR proteins sequences. The feature vector that is proposed in this section is based on the relationship of transmembrane proteins and wavelet variances obtained after a scale-by-scale decomposition of hydrophobic and EIIP profiles of proteins.

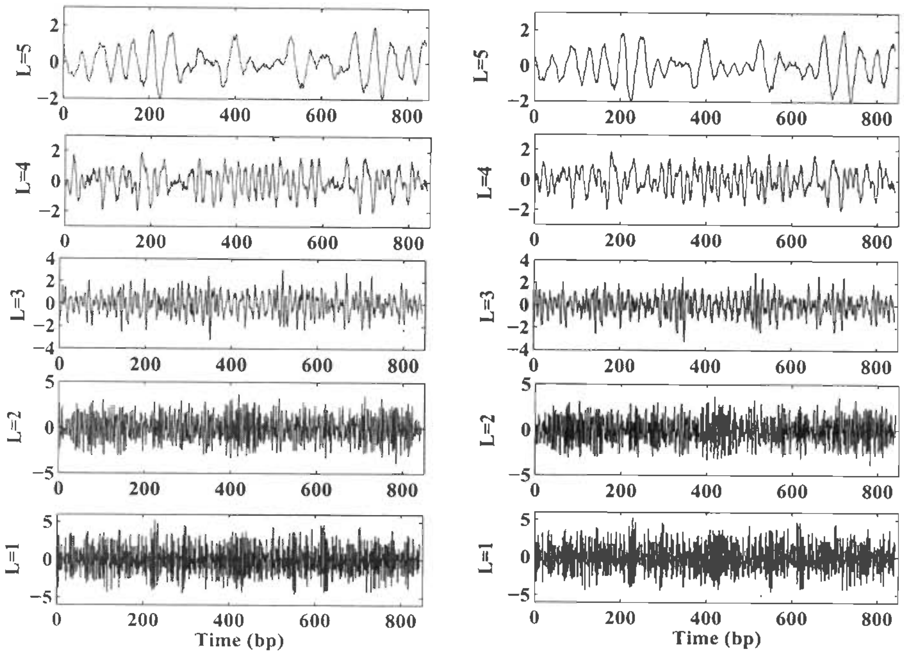


Figure 7.11: Wavelet coefficients for Hydrophobicity profile of 5-hydroxytryptamine 1A receptor(Human) and 5-hydroxytryptamine 1A receptor (Mouse).

A non-decimated version of the DWT, a modified transform also known as maximal overlap DWT (MODWT) [130, 26, 131] is applied for calculating the wavelet variances. Discussion about MODWT and scale-by-scale decomposition of wavelet variance for a time series data is provided in chapter 6.

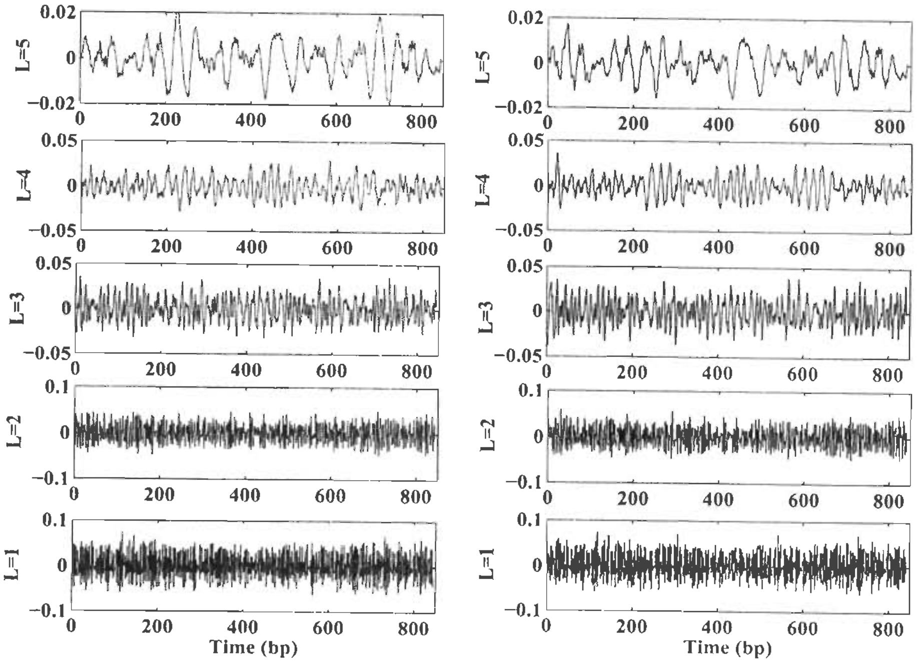


Figure 7.12: Wavelet coefficients for EIIP profile of 5-hydroxytryptamine 1A receptor(Human) and 5-hydroxytryptamine 1A receptor (Mouse).

Feature vector

The feature vector $\mathbf{F_{Prot}}$ based on hydrophobicity, EIIP, isoelectric point, polarity, volume, composition and molecular weight for GPCRs classification is given as follows:

$$\mathbf{F_{Hyp}} = \left[\hat{\sigma}^2_{Hyp}(1), \hat{\sigma}^2_{Hyp}(2), \ldots, \hat{\sigma}^2_{Hyp}(L)\right]^T \tag{7.5}$$

$$\mathbf{F_{EIIP}} = \left[\hat{\sigma}^2_{EIIP}(1), \hat{\sigma}^2_{EIIP}(2), \ldots, \hat{\sigma}^2_{EIIP}(L)\right]^T \tag{7.6}$$

$$\mathbf{F_{Iso}} = \left[\hat{\sigma}^2_{Iso}(1), \hat{\sigma}^2_{Iso}(2), \ldots, \hat{\sigma}^2_{Iso}(L)\right]^T \tag{7.7}$$

$$\mathbf{F_{Pol}} = \left[\hat{\sigma}^2_{Pol}(1), \hat{\sigma}^2_{Pol}(2), \ldots, \hat{\sigma}^2_{Pol}(L)\right]^T \tag{7.8}$$

$$\mathbf{F_{Vol}} = \left[\hat{\sigma}^2_{Vol}(1), \hat{\sigma}^2_{Vol}(2), \ldots, \hat{\sigma}^2_{Vol}(L)\right]^T \tag{7.9}$$

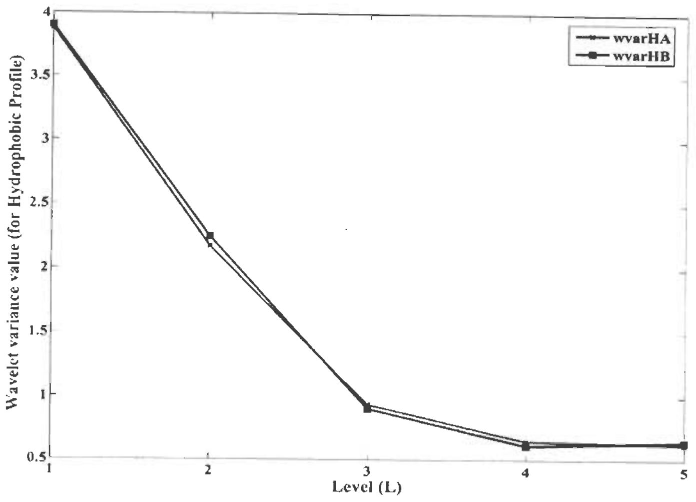


Figure 7.13: Wavelet variance for Hydrophobicity profile of 5-hydroxytryptamine 1A receptor(Human) and 5-hydroxytryptamine 1A receptor (Mouse).

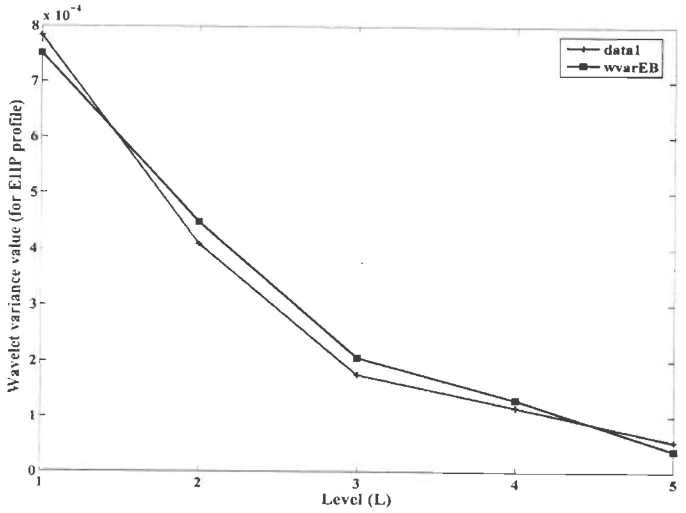


Figure 7.14: Wavelet variance for EIIP profile of 5-hydroxytryptamine 1A receptor(Human) and 5-hydroxytryptamine 1A receptor (Mouse).

$$\mathbf{F_{Comp}} = \left[\hat{\sigma}^2_{Comp}(1), \hat{\sigma}^2_{Comp}(2), \ldots, \hat{\sigma}^2_{Comp}(L)\right]^T \tag{7.10}$$

$$\mathbf{F_{Mol}} = \left[\hat{\sigma}^2_{Mol}(1), \hat{\sigma}^2_{Mol}(2), \ldots, \hat{\sigma}^2_{Mol}(L)\right]^T \tag{7.11}$$

$$\mathbf{F_{Prot}} = \mathbf{F_{Hyp}} \oplus \mathbf{F_{EIIP}} \oplus \mathbf{F_{Iso}} \oplus \mathbf{F_{Pol}} \oplus \mathbf{F_{Vol}} \oplus \mathbf{F_{Comp}} \oplus \mathbf{F_{Mol}} \tag{7.12}$$

$$\begin{aligned}\mathbf{F_{Prot}} =&[\hat{\sigma}^2_{Hyp}(1), \ldots, \hat{\sigma}^2_{Hyp}(L), \hat{\sigma}^2_{EIIP}(1), \ldots, \hat{\sigma}^2_{EIIP}(L), \hat{\sigma}^2_{Iso}(1), \ldots, \\ &\hat{\sigma}^2_{Iso}(L), \hat{\sigma}^2_{Pol}(1), \ldots, \hat{\sigma}^2_{Pol}(L), \hat{\sigma}^2_{Vol}(1), \ldots, \hat{\sigma}^2_{Vol}(L), \hat{\sigma}^2_{Comp}(1), \\ &\ldots, \hat{\sigma}^2_{Comp}(L), \hat{\sigma}^2_{Mol}(1), \ldots, \hat{\sigma}^2_{Mol}(L)]^T\end{aligned} \tag{7.13}$$

where L is the maximum level of decomposition of the time series data (protein sequence length) and $\oplus$ represents concatenation operation. Thus, $\mathbf{F_{Prot}}$ is a collection of wavelet variance coefficients (WVC). The dimension of $\mathbf{F_{Prot}}$ is equal to $2L$ and is dependent on the number of levels (L) to which the time series data has to be decomposed. The value of L is further dependent on the length of time series data (i.e, protein sequence) and $L \geq \log_2(N)$, where N is the number of observation points in the time series. As most of the GPCR sequence have length greater than 32 the value of L is taken as 5. In this study Daubechies [135] wavelet has been used for analysis.

7.3.3 Classification strategy

This section describes a support vector machine-based implementation developed for annotating GPCRs on the basis of fixed length feature vector (equation 7.13). A four step strategy similar to that proposed in [155] is followed for annotating GPCRs are as follows:

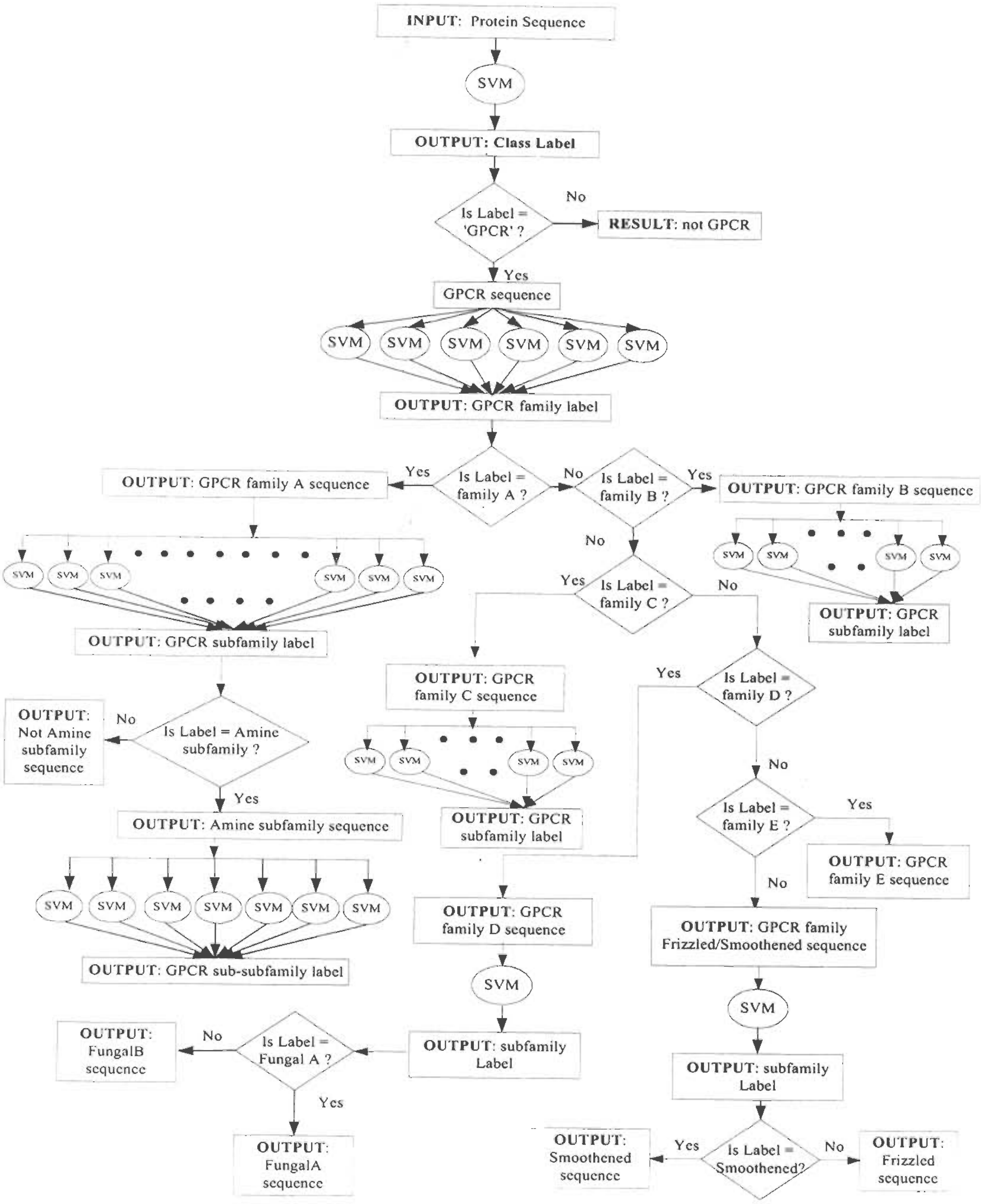

INPUT: Protein Sequence
SVM
OUTPUT: Class Label
Is Label = 'GPCR' ?
No
RESULT: not GPCR
Yes
GPCR sequence
SVM
SVM
SVM
SVM
SVM
SVM
OUTPUT: GPCR family label
OUTPUT: GPCR family A sequence
Yes
Is Label = family A ?
No
Is Label = family B ?
Yes
OUTPUT: GPCR family B sequence
SVM
SVM
SVM
SVM
OUTPUT: GPCR subfamily label
No
Yes
Is Label = family C ?
No
SVM
SVM
SVM
SVM
SVM
SVM
OUTPUT: GPCR subfamily label
OUTPUT: GPCR family C sequence
SVM
SVM
SVM
SVM
OUTPUT: GPCR subfamily label
Yes
Is Label = family D ?
No
OUTPUT: Not Amine subfamily sequence
No
Is Label = Amine subfamily ?
Yes
OUTPUT: Amine subfamily sequence
SVM
SVM
SVM
SVM
SVM
SVM
SVM
OUTPUT: GPCR sub-subfamily label
Is Label = family E ?
Yes
OUTPUT: GPCR family E sequence
No
OUTPUT: GPCR family D sequence
SVM
OUTPUT: subfamily Label
OUTPUT: GPCR family Frizzled/Smoothened sequence
SVM
OUTPUT: subfamily Label
OUTPUT: FungalB sequence
No
Is Label = Fungal A ?
Yes
OUTPUT: FungalA sequence
OUTPUT: Smoothened sequence
Yes
Is Label = Smoothened?
No
OUTPUT: Frizzled sequence

Figure 7.15: A block diagram of 4-step strategy followed for identification of GPCR, its family, subfamily and sub-subfamilies.

1. Predicts whether the protein sequence belongs to the GPCR superfamily or not.
2. Predicts the family of GPCRs.
3. Predicts the subfamily of GPCRs.
4. Predicts the sub-subfamily of GPCRs.

The four-step strategy for identifying the superfamily, family and subfamily of GPCRs is shown in Figure 7.15.

Support vector machine

The SVM was implemented using the downloadable tool *pattern classification program* (PCP) written by Buturović [172]. PCP is an open-source machine learning program for supervised classification of patterns and is freely available at *http://pcp.sourceforge.net.* It incorporates a widely used open-source SVM implementation called LIBSVM [141] and performs a cross-validation estimation of classifier performance. In addition, the program includes the provision of automated and efficient model selection (optimization of parameters) for support vector machine (SVM) classifier. The program enables the user to change the individual class costs. This may be useful to achieve a desired balance among class-conditional error rates. PCP (through LIBSVM [141]) provides four kernel types: radial basis function (RBF), linear, polynomial and sigmoid. Experiments were conducted using different kernels, however the RBF was selected because of its superior performance parameters. A brief introduction on SVM theory and RBF is given in chapter 5.

7.4 Performance Measures

7.4.1 GPCR superfamily

The performance of SVM in distinguishing GPCRs from non-GPCRs was evaluated using 5-fold cross-validation. In the 5-fold cross-validation, the dataset was partitioned randomly into five equal-sized sets. The training and testing of each classifier was carried out five times using one distinct set for testing and other four sets for training. The performance of the SVM classifier was measured in terms of four performance measures–sensitivity (SN), specificity (SP), accuracy (ACC) and Matthew's correlation coefficient (MCC). The performance measures are defined as follows:

$$SN = \frac{TP}{(TP + FN)} \tag{7.14}$$

$$SP = \frac{TN}{(TN + FP)} \tag{7.15}$$

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{7.16}$$

$$MCC = \frac{((TP * TN) - (FP * FN))}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{7.17}$$

where, TP (true positive), TN (true negative), FN (false negative) and FP (false positive) represents # GPCR sequences that have been correctly predicted as GPCR sequence, # non-GPCR sequences predicted as non-GPCR sequence, # GPCR sequences predicted as non-GPCR sequence, # non-GPCR sequences predicted as GPCR sequence respectively.

7.4.2 GPCR family, subfamily and sub-subfamily

The four performance measures–accuracy (ACC), Matthew's correlation coefficient (MCC), total ACC and total MCC as provided by Hua and Sun [173] are given by

$$ACC(i) = p(i)/obs(i) \tag{7.18}$$

$$MCC(i) = \frac{p(i)n(i) - u(i)o(i)}{\sqrt{(p(i)+u(i))(p(i)+o(i))(n(i)+u(i))(n(i)+o(i))}} \tag{7.19}$$

$$ACC_{total} = \frac{\sum_{i=1}^{k} p(i)}{N} \tag{7.20}$$

$$MCC_{total} = \frac{\sum_{i=1}^{k} obs(i)MCC(i)}{N} \tag{7.21}$$

where, N is the total number of sequence, i is the any family/subfamily/sub-subfamily, k is the family/subfamily/sub-subfamily number, $obs(i)$ is the number of sequences observed in family/subfamily/sub-subfamily i, $p(i)$ is the number of correctly predicted sequences of family/subfamily/sub-subfamily i, $n(i)$ is the number of correctly predicted sequences not of family/subfamily/sub-subfamily i, $u(i)$ is the number of under-predicted sequences, and $o(i)$ is the number of over-predicted sequences.

7.5 Experimental Results

7.5.1 Dataset description

To evaluate the performance of the proposed approach the GPCRs protein sequences were downloaded from GPCRDB *http://www.gpcr.org/7tm.* GPCRs have been divided into six principal classes or families generated based on sequence: class A (rhodopsin-like), class B (secretin-like), class C (metabotrophic glutamine/pheromone), class D (fungal pheromone), class E (cAMP receptors), and the Frizzled/Smoothened class [147, 148]. Each class/family is further divided into subfamilies and sub-subfamilies based on their ligand specificity, with some families combined into larger groups based on closely related ligands. The negative dataset i.e., non-GPCRs protein sequences were downloaded from *http://www.soe.ucsc.edu/research/compbio/gpcr/superfamily_seqs/* [154], and includes 99 decoy negative examples (i.e., non-GPCRs): 18 archaea rhosphins and 80 G-protein alpha domains and 2466 additional negative examples obtained from SCOP version 1.37 PDB90 domain data data [154, 168]. For comparison with GPCRpred the GPCRs protein sequences were downloaded from above mentioned website (negative dataset) which were used by Karchin *et al.* [154] and Bhasin and Raghava [155].

The downloaded datasets were pre-processed before using for the experiment. The protein sequences having length less than 32 amino acids (as $L = 5$) and the protein sequences having invalid amino acid symbol (i.e., not among 20 alphabet) were removed from the dataset. The processed dataset obtained for Karchin *et al.* [154] dataset is given in Table 7.4.

7.5.2 GPCR identification

The performance of the system developed for identifying GPCRs from other protein sequences (i.e., non-GPCRs) using $\mathbf{F_{Prot}}$ (equation 7.13) feature vector is summarized in Table 7.5. The obtained results show that the technique can identify GPCRs using wavelet variance feature vector from other protein sequence with an accuracy of 99.9% and a *MCC* of 0.998, when evaluated through 5-fold cross-validation. The best results were obtained using RBF kernel with $\gamma =$

Table 7.4: Processed dataset for GPCRs recognition and GPCR family recognition.

Superfamily	GPCR Family	# of sequence
GPCR	Class A	664
	Class B	55
	Class C	16
	Class D	11
	Class E	3
	Total	749
Non-GPCR	decoy	99
	SCOP 1.37 PDB90 domain data	2254
	Total	2353

0.045316 and the optimized value of C was equal to 3.1623x10^6.

Table 7.5: The performance of the proposed technique for identification of GPCRs.

Dataset Information: # GPCR protein sequences = 2573 # Non-GPCR protein sequences = 2353
Experiment Setting: 5-fold cross-validation Normalized Feature Vector Kernel = *Radial Basis Function*(RBF)
Experimental Result: *Optimal parameters:* Cost (C) = 3.1623x10^6 Gamma (γ) = 0.045316 *Performance Measures:* $TP = 2571$, $TN = 2350$, $FN = 2$, $FP = 3$ Sensitivity (SE) = 99.92% Specificity (SP) = 99.87% Accuracy (ACC) = 99.9% $MCC = 0.998$

Table 7.6: The performance of the proposed technique on unseen GPCR sequences.

Dataset Information			Performance parameters					
Class	# Train	# Test	# Predicted		Sensitivity(%)	Specificity(%)	Accuracy(%)	MCC
			GPCR	Non-GPCR	(SE)	(SP)	(Recall Rate)	
GPCR	1853	730	730(TP)	0(FN)	100.0	99.62	99.8	0.996
Non-GPCR	1569	784	781(TN)	3(FP)				
# Total	3422	1514						

It is very important to evaluate the performance of a classification system on an unseen or recall patterns/dataset to demonstrate its generalization capability. None of the protein sequences in recall dataset were used for training the system. A dataset of 1514 protein sequences (#GPCRs = 730, #non-GPCRs = 784) was selected from the processed downloaded dataset for testing and the remaining protein sequences were used for training the classification system. The performance obtained is summarized in Table 7.6. The system provided an accuracy of 99.8% and *MCC* of 0.996, when evaluated through a 5-fold cross-validation based model. The results demonstrated the generalization capability of the proposed technique and robustness of wavelet variance feature vector. The results proved that GPCRs and non-GPCRs can be classified with high accuracy using wavelet variance of physiochemical profiles of protein sequence as a feature vector. The key point to note is that with only 35-dimensional feature vector a better performance than GPCRpred [155] is achieved.

7.5.3 GPCR family identification

The classification of GPCRs into its families is a multi-class classification problem. Unlike, SVMpred program where a series of binary SVMs were constructed to identify families of GPCR, a multi-class classification approach was followed. From the experimental using the proposed feature vector it was found the multi-class classification approach provides better performance than a series of binary SVMs approach.

Table 7.7: The performance of the proposed technique for identification of GPCR families.

GPCR Family	# Training	# Predicted	*ACC*(%)	*MCC*
Rhodopsin and andrenergic-like receptors (Class A)	1861	1850	99.41	0.962
Calcitonin and PTH-like receptors (Class B)	308	291	94.48	0.963
Metabotropic-like receptors (Class C)	204	188	92.16	0.948
Pheromone-like receptors (Class D)	62	52	83.87	0.914
cAMP-like receptors (Class E)	10	9	90.0	0.948
Frizzled/Smoothened	128	122	95.31	0.955
Total	2573	2512	97.63	0.959

Table 7.8: The performance of the proposed technique on unseen dataset of GPCR families.

GPCR Family	Dataset		# Correctly	$ACC(\%)$	MCC
	# Training	# Test	predicted		
Rhodopsin and andrenergic-like receptors (Class A)	1361	500	497	99.40	0.965
Calcitonin and PTH-like receptors (Class B)	208	100	94	94.0	0.959
Metabotropic-like receptors (Class C)	144	60	57	95.0	0.972
Pheromone-like receptors (Class D)	42	20	16	80.0	0.892
cAMP-like receptors (Class E)	10	10	10	100.0	1.0
Frizzled/Smoothened	88	40	33	82.5	0.889
Total	1853	730	707	96.85	0.959

The overall 5-fold cross-validation accuracy achieved for GPCR family classification was equal to 97.63% and is provided in Table 7.7. A test was also performed on unseen dataset of GPCR family. In this test leaving class E data the training and testing dataset of class A, class B, class C, class D and Frizzled/Smoothened consists of different protein sequences. Due to lower number of protein sequences for class E, the training and test dataset were kept same because a learning technique requires a good number of examples for reliable prediction. The performance of the SVMs in recognizing the unseen or recall patterns of the classes or families of GPCRs is summarized in Table 7.8. The overall accuracy and *MCC* achieved for unseen pattern/dataset for identifying the six GPCR classes were 96.85% and 0.959 respectively.

7.5.4 GPCR subfamily identification

The prediction of GPCR subfamilies is crucial in assigning a function to GPCRs. Therefore, experiments have also been performed for identifying the subfamilies of GPCRs. The system was evaluated using 5-fold cross-validation (except for Gastric inhibitory peptide, Very large GPCR subfamilies of class B for which a 4-fold cross-validation was performed) and the performance of the system in predicting the subfamilies in terms of accuracy (*ACC*) is shown in Table 7.9. The overall prediction accuracy and *MCC* of the proposed technique for identifying subfamilies of GPCR is equal to 96.64% and 0.97 respectively and is better than GPCRpred program. Further, test on unseen dataset of class A (rhodopsin-like) subfamilies which constitutes more than 80% of GPCRs was also performed. Accuracy of 96.78% and *MCC* of 0.97 was achieved and the result is summarized in Table 7.10.

Table 7.9: The performance of the proposed technique for identification of GPCR subfamily.

Family	Subfamily	#Actual	#Predicted	*ACC*(%)	*MCC*
Class A	Amine	574	528	91.99	0.93
	Cannabinoid	24	22	91.67	0.96
	Gonadotropin-releasing hormone	78	70	89.74	0.93
	Hormone proteins	65	65	100.0	1.0
	Leukotriene B4 receptor	12	12	100.0	1.0
	Lysosphingolipid & LPA	60	57	95.0	0.97
	Melatonin	22	22	100.0	1.0
	Nucleotide-like	146	137	93.84	0.96
	Olfactory	2489	2470	99.24	0.99
	Peptide	1314	1257	95.66	0.96
	Platelet activating factor	9	7	77.78	0.88
	Prostanoid	95	93	97.89	0.99
	Rhodopsin	625	615	98.4	0.99
	TRHS	36	27	75.0	0.85
	Viral	88	74	84.09	0.91
	Overall for Class A	5637	5456	96.79	0.97
Class B	Calcitonin	29	28	96.55	0.96

Continued on next page

Table 7.9 – continued from previous page

Family	Subfamily	#Actual	#Predicted	$ACC(\%)$	MCC
	Corticotropin releasing factor	34	34	100.0	1.0
	Gastric inhibitory peptide	4	4	100.0	1.0
	Glucagon	23	21	91.3	0.95
	Growth hormone-releasing hormone	15	13	86.67	0.93
	Parathyroid hormone	23	21	91.3	0.95
	PACAP	21	21	100.0	0.97
	Secretin	6	6	100.0	1.0
	Vasoactive intestinal polypeptide	22	18	81.82	0.90
	Diuretic hormone	7	6	85.71	0.92
	EMR1	27	27	100.0	1.0
	Latrophilin	41	39	95.12	0.96
	Brain-specific angiogenesis inhibitor	18	17	94.44	0.97
	Methuselah-like proteins (MTH)	23	20	86.96	0.93
	Cadherin EGF LAG (CELSR)	11	11	100.0	1.0
	Very large GPCR	4	4	100.0	1.0
	Overall for Class B	308	290	94.16	0.94

Continued on next page

Table 7.9 – continued from previous page

Family	Subfamily	#Actual	#Predicted	$ACC(\%)$	MCC
Class C	Metabotropic glutamate	53	51	96.23	0.97
	Calcium-sensing like	41	37	90.24	0.91
	Putative pheromone receptors	20	19	95.0	0.97
	GABA-B	35	32	91.43	0.93
	Orphan GPRC5	17	17	100.0	0.97
	Orphan GPRC6	5	5	100.0	1.0
	Bride of sevenless proteins	4	4	100.0	1.0
	Taste receptors	29	26	89.65	0.92
	Overall for Class C	204	191	93.63	0.95
Class D	Fungal pheromone A-Factor like	21	21	100.0	1.0
	Fungal pheromone B like	39	39	100.0	1.0
	Overall for Class D	60	60	100.0	1.0
Frizzled/ Smoothened	Frizzled	115	115	100.0	0.96
	Smoothened	13	12	92.31	0.96
	Overall for Frizzled/Smoothened	128	127	99.22	0.96
	Overall for subfamily	6337	6124	96.64	0.97

7.5.5 GPCR sub-subfamily identification

The receptors of amine subfamily are specifically major drug targets for therapy of nervous disorders and psychiatric diseases. The recognition of novel amine type of receptors and their cognate ligands is of paramount interest for pharmaceutical companies. The complete dataset that is available at GPCRDB was taken for the current experiment. The 5-fold cross-validation accuracy that was obtained for amine subfamily after the experiment is given in Table 7.11. The overall accuracy and *MCC* obtained was equal to 93.38% and 0.95 respectively.

7.5.6 Human GPCR identification

To evaluate the performance of the proposed approach for human GPCRs a 5-fold cross-validation test was performed on the dataset downloaded from GPCRDB [147]. The protein sequences were pre-processed and sequences of length less than 32 amino acids were removed (as $L = 5$) from the dataset. A total number of 2583 human GPCRs belonging to the six families or classes of GPCR were obtained after pre-processing. The dataset that was taken for GPCRs classification experiment in section 7.5.2 is taken as non-GPCR dataset. The performance of the SVMs is summarized in Table 7.12. The system provided an accuracy of 99.88% and a *MCC* of 0.998, when evaluated through a 5-fold cross-validation based model. In addition, a test on recall patterns was also performed to evaluate the generalization capability of the proposed technique. The performance obtained is summarized in Table 7.13. The system provided an accuracy of 99.83% and an *MCC* of 0.996, when evaluated through a 5-fold cross-validation based model.

Table 7.10: The performance of the proposed technique on unseen subfamilies of class A.

Subfamily	#Train	#Test	#Predicted	*ACC*(%)	*MCC*
Amine	474	100	96	96.0	0.96
Cannabinoid	20	4	4	100.0	1.0
Gonadotropin-releasing hormone	61	17	13	76.47	0.87
Hormone proteins	50	15	15	100.0	1.0
Lysosphingolipid & LPA	45	15	14	93.33	0.97
Melatonin	17	5	5	100.0	1.0
Nucleotide-like	102	44	38	86.36	0.93
Olfactory	2021	468	466	99.57	0.99
Peptide	1027	287	277	96.52	0.95
Prostanoid	70	25	24	96.0	0.98
Rhodopsin	516	109	109	100.0	1.0
TRHS	32	4	2	50.0	0.71
Viral	64	24	18	75.0	0.86
Overall	4499	1117	1081	96.78	0.97

Table 7.11: The performance of the proposed technique for identification of GPCR sub-subfamily.

Sub-subfamily	# Actual	# Predicted	*ACC*(%)	*MCC*
Muscarinic acetylcholine	63	61	96.82	0.96
Adrenoceptors	118	114	96.61	0.97
Dopamine	94	84	89.36	0.94
Histamine	48	45	93.75	0.96
Serotonin	157	145	92.36	0.93
Octopamine	25	18	72.0	0.80
Trace amine	69	69	100.0	1.0
Overall	574	536	93.38	0.95

Table 7.12: The performance of 5-fold cross-validation of the proposed technique for identification of Human GPCRs.

# GPCRs	# non-GPCRs	True positive (TP)	True negative (TN)	False positive (FP)	False negative (FN)	SN (%)	SP (%)	ACC (%)	MCC
2583	2353	2581	2349	4	2	99.92	99.83	99.88	0.998

Table 7.13: The performance of the proposed technique on unseen Human GPCRs.

Dataset Information			Performance parameters					
Class	# Train	# Test	# Predicted		Sensitivity(%)	Specificity(%)	Accuracy(%)	MCC
			GPCR	Non-GPCR	(SE)	(SP)		
GPCR	1612	971	971(TP)	0(FN)	100.0	99.62	99.83	0.996
Non-GPCR	1569	784	781(TN)	3(FP)				
# Total	3181	1755						

Table 7.14: Comparison of the proposed technique with GPCRpred for identification GPCRs.

Type	Our Technique				GPCRpred	
	# Actual	# Predicted	*ACC*(%)	*MCC*	*ACC*(%)	*MCC*
GPCR	749	746	99.84	0.996	99.5	0.99
Non-GPCR	2353	2351				
Overall	3102	3097				

7.5.7 Performance comparison with SVMpred

To compare the performance of the proposed method with GPCRpred program an experiment was performed on the dataset of GPCR and non-GPCR protein sequences that was used by Bhasin and Raghava [155]. Accuracy of 99.84% was achieved (refer to Table 7.15) and is better than that achieved by 400-dimension feature vector based program (GPCRpred). The key point to note here is that using only 35-dimensional feature vector better accuracy is achieved as compared to GPCRpred program.

For comparing the performance of the proposed method with GPCRpred for family identification a 5-fold cross-validation was performed on the dataset that was used by Raghava and Bhasin. The result obtained is summarized in Table 7.15 and the overall accuracy achieved by the proposed method is higher than SVMpred. Poor result was obtained for class D and class E due to low number of training samples.

Table 7.15: Comparison of the proposed technique with GPCRpred for identification GPCR family.

GPCR Family Type	Our Technique				GPCRpred	
	# Actual	# Predicted	*ACC*(%)	*MCC*	*ACC*(%)	*MCC*
Rhodopsin and andrenergic-like receptors (Class A)	664	662	99.7	0.953	98.1	0.8
Calcitonin and PTH-like receptors (Class B)	55	53	96.36	0.98	85.7	0.84
Metabotropic-like receptors (Class C)	16	15	93.75	0.968	81.3	0.81
Pheromone-like receptors (Class D)	11	6	54.54	0.68	36.4	0.49
cAMP-like receptors (Class E)	3	2	66.67	0.816	100.0	1.0
Total	749	738	98.53	0.951	97.3	0.81

7.6 Conclusion

In this chapter a novel wavelet variance based feature vector for identification and classification of GPCRs is presented. Based on pattern recognition framework the proposed approach is divided into three different tasks: amino acid mapping, feature construction and classification. The feature vector unlike previous SVM based approaches summarizes the variation of seven important biological properties (hydrophobicity, electronic, isoelectric point, polarity, volume, composition and molecular weight) of amino acids in a protein sequence. The feature extraction technique is based on wavelet based time series analysis. Further, the dimension of the proposed feature vector is equal to 35 and is much smaller than GPCRpred program whose feature vector dimension is equal to 400. This huge reduction in feature vector dimension facilitates in developing computational and memory efficient classifiers for drug discovery applications.

Performance evaluation performed on the complete dataset of GPCR superfamily, family, subfamily, sub-subfamilies and human GPCRs that is available at GPCRDB. The proposed technique provides a 5-fold cross-validation accuracy of 99.9%, 97.63%, 96.64% and 93.38% for GPCR superfamily, families, subfamilies and sub-subfamilies (amine group) respectively. Test conducted for identifying human GPCRs provided an accuracy of 99.88%. In addition, tests were also conducted on unseen or recall datasets and accuracy of 99.8%, 96.85% and 96.78% was obtained for GPCR superfamily, families and subfamilies respectively. This proves the generalization capability of the proposed classification technique and sensitivity of the novel features. Comparison of experimental results with GPCRpred shows that the proposed approach is more accurate. The proposed pattern recognition framework coupled with wavelet based feature extraction provides a generic approach for sequence analysis, especially for sequence classification.

Chapter 8

Conclusions and Future Work

8.1 Contributions of the Thesis

The contributions of the thesis are as follows:

- A novel signal processing measure for identifying tandem repeat patterns in DNA sequences is presented in this thesis. The motivation for developing a signal processing based technique for tandem repeat algorithm is due to similarity of tandem repeats and short periodic signals. The proposed approach is based on orthogonal decomposition of input signal using exactly subspace decomposition algorithm. The algorithm resolves the common problem of the existing signal processing algorithms where by they could not identify whether an inexact repeat pattern is of due to period P or its multiple (i.e., $2P$, $3P$ and so on). In addition, it resolves other issues that are present with existing signal processing solution for tandem repeat identification problem.

  The proposed algorithm is computationally efficient and computes tandem repeats in $O(NL_w \log L_w)$, where N is the length of DNA sequence and

L_w is the window length, for identifying repeats. The algorithm operates in two stages. In the first stage, each nucleotide is analyzed separately for periodicity, and in the second stage, the periodic information of each nucleotide is combined together to identify the tandem repeats. Datasets having exact and inexact repeats were taken up for the experimental purpose. Comparison of sensitiveness of the algorithm with existing methods both signal processing and non-signal processing shows the effectiveness of the approach.

- Identifying exact and inexact inverted repeats (IRs) or DNA palindrome in DNA sequences is an important problem in bioinformatics. The existing programs require a number of input parameters for finding inverted repeats and are difficult to use for non-computer experts especially biologists. Also, till date there does not exist any signal processing algorithm for IR identification. As one of the aims the thesis is to explore novel application of signal processing technique for bioinformatics problems, a signal processing approach is proposed for IR identification.

  In this thesis, two algorithms have been presented for identifying inverted repeats. The first algorithm is based on periodicity transform and is able to identify exact IR and inexact IR (due to substitution). The second algorithm applies a fast correlation based signal processing approach for identifying all cases of inexact IRs. The motivation for developing a correlation based approach is due to presence of high correlated patterns between a DNA sequence and its reverse complemented sequence whenever inverted repeats are present in the input DNA sequence. The algorithm requires only two input parameters and is a easier method for identifying IRs. The correlation based approach with slight modification is extended to predict

RNA secondary structure from its primary sequence.

- Gene prediction and the classification of protein coding and non-coding DNA sequences are unsolved and popular research problems in bioinformatics. Several powerful computational methods have been developed and their performance is highly dependent on coding measures that are used for characterizing sequences. Period-3 property of the coding sequences and slope based model for Z-curve components are two important and popular coding measures for classifying a DNA sequence into protein coding and non-coding classes. However, both of these measures provide only a global level of information that is present in the DNA sequences. For many DNA sequences, especially, for small coding DNA sequences the local information that is present in the sequence is also important for correct classification.

  In this thesis, a pattern recognition framework for classification of DNA sequences into coding and non-coding classes is presented. A novel coding measure is calculated using wavelet based time series analysis technique on Z-curve components. The coding measure contains both local and global level variance information of Z-curve components. An optimized support vector machine (SVM) based on the novel coding measure is constructed for accurate classification. Later on, a fixed length vector based on wavelet variance and slope based coding measure is also used for classification purpose. Performance evaluation of the proposed approach shows that wavelet variance classification provides better accuracy than slope based method (existing most accurate method). Further experiments show that a combined coding measure (wavelet variance and slope) provides accuracy up to 96% for *Yeast* and more than 96% for *E.Coli* genome.

- An automatic identification and classification of G-protein coupled receptors (GPCRs) is proposed in this thesis. GPCRs play important roles in cellular signalling networks and are a large superfamily of receptors. They are major therapeutic targets of numerous prescribed drugs. However, the ligand specificity of of many receptors is unknown and there is little structural information available. Thus, computational techniques for automatic recognition and characterization of GPCRs are very important. Till date, there exists several methods for identification and classification of GPCRs. SVMpred program is a SVM based approach and is the most accurate method for classification of GPCRs. However, this technique has two major drawbacks. First, the fixed length vector dimension is huge and is equal to 400. Secondly, the feature vector is based on dipeptide composition and ignores the physicochemical properties of amino acids which are important for knowing the functionality of proteins.

  In this thesis, a novel feature vector based on variation of seven physicochemical properties (hydrophobicity, electronic, isoelectric point, polarity, volume, composition and molecular weight) of amino acids is proposed. The feature vector is obtained based on wavelet method for analyzing time series sequences. Further, the dimension of the proposed feature vector is only 35. This low dimension of the feature vector facilitates developing computational and memory efficient classifiers for drug discovery applications. A four step strategy is followed for automatic classification of protein sequences into GPCRs, GPCR family, GPCR subfamily and GPCR subsubfamily using SVMs. Experiments were performed on complete dataset that is available at GPCR database (GPCRDB). Performance evaluation of the proposed system on standard dataset shows that our approach is more accurate than GPCRpred. Further test conducted on unseen datasets shows the effectiveness of our approach.

8.2 Future Work

There are a number of extensions and scope of the present research work. The three step methodology proposed in this thesis can be followed for other sequence analysis problems and for finding novel applications of signal processing algorithms. Some of the future works of the current research work are as follows:

- The proposed exactly period subspace decomposition algorithm can be extended for identifying protein coding regions from a genome based on period-3 component.

- More coding measures can be included in the proposed coding feature vector for further improvement in classification accuracy. The proposed feature vector can be utilized by the existing gene finding algorithms for accurate identification of genes.

- The proposed pattern recognition approach can be extended for identifying methylated CpGs patterns from DNA sequences.

- The proposed feature based on physicochemical properties on amino acids can be applied for prediction of protein structural classes, identification of membrane protein type, enzyme family classification and many others. In addition, other physicochemical properties of amino acids can also be included for further improvement in accuracy of the proposed system.

- A clustering based approach can be developed based on the proposed feature for finding classes or families of proteins.

- The proposed pattern recognition approach system for time series data (DNA and protein sequences in this thesis) can be applied to economic, atmospheric and other time series data.
- The proposed wavelet based variance analysis can be extended to wavelet based correlation analysis and sequence similarity algorithm can be developed.

Bibliography

[1] D. A. Benson, I. K. Mizrachi, D. J. Lipman J. Ostell, and D. L. Wheeler. GenBank. *Nucleic Acids Research*, 35:D21–D25, 2007.

[2] P. P. Vaidyanathan and B.-J. Yoon. Gene and exon prediction using allpass-based filters. In *IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS '02)*, Raleigh, NC, October 2002.

[3] P. P. Vaidyanathan and B.-J. Yoon. Digital filters for gene prediction. In *36th Asilomar Conference on Signals, Systems and Computers*, Monterey, CA, November 2002.

[4] T. W. Fox and A. Carreira. A digital signal processing method for gene prediction with improved noise suppression. *EURASIP Journal on Applied Signal Processing*, 1:108–114, 2004.

[5] D. Anastassiou. Genomic signal processing. *IEEE Signal Processing Magazine*, pages 8–20, July 2001.

[6] D. Anastassiou. DSP in genomics: processing and frequency-domain analysis of character strings. In *Proc. IEEE Int. Conference on Acoustics, Speech, Signal Processing (ICASSP '01)*, volume 2, pages 1053–1056, Salt Lake City, Utah, USA, 7-11 May 2001.

[7] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy. Prediction of probable genes by Fourier analysis of genomic sequences. *CABIOS*, 13(3):263–270, 1997.

[8] D. Sharma, B. Issac, G. P. S. Raghava, and R. Ramaswamy. Spectral repeat finder (SRF): identification of repetitive sequences using Fourier transformation. *Bioinformatics*, 20(9):1405–1412, 2004.

[9] T. T. Tran, V. A. Emanuele II, and G. T. Zhou. Techniques for detecting approximate tandem repeats. In *Proc. IEEE Int. Conference on Acoustics, Speech, Signal Processing (ICASSP '04)*, volume 5, pages 449–452, Montreal, Canada, 17-21 May 2004.

[10] X. Zhang, A. Kassim, and V. B. Bajic. Digital signal processing for potential promoter prediction. In *IEEE International Workshop on Biomedical Circuits and Systems (BioCAS '04)*, pages S2.7.INV–16–S2.7.INV–18, Singapore, 1-3 December 2004.

[11] C. H. Trad, Q. Fang, and I. Cosic. Protein sequence comparison based on the wavelet transform approach. *Protein Engineering*, 15(3):193–203, 2002.

[12] P. Lio and M. Vannucci. Finding pathogenicity islands and gene transfer events in genome data. *Bioinformatics*, 16(10):932–940, 2000.

[13] A. A. Tsonis, P. Kumar, J. B. Elsner, and P. A. Tsonis. Wavelet analysis of DNA sequences. *Physical Review E, Stat. Phys. Plasmas Fluids Relat. Interdsip. Top*, 53:1828–1834, February 1996.

[14] E. Pirogova, Q. Fang, M. Akay, and I. Cosic. Investigation of the structural and functional relationships of oncogene proteins. *Proc. of the IEEE*, 90(12):1859–1867, December 2002.

[15] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy. Charaterizing long-range correlations in DNA sequences from wavelet analysis. *Physical Review Letter*, 74:3293–3296, April 1995.

[16] E. N. Trifonov and J. L. Sussman. The pitch of chromatin DNA is reflected in its nucleotide sequence. In *Proc. Natl. Acad. Sci. USA*, volume 77, pages 3816–3820, 1980.

[17] D. Kotlar and Y. Lavner. Gene prediction by spectral rotational measure: a new method for identifying protein-coding regions. *Genome Research*, 13:1930–1937, 2003.

[18] S. Datta and A. Asif. A fast DFT based gene prediction algorithm for identification of protein coding regions. In *Proc. IEEE Int. Conference on Acoustics, Speech, Signal Processing (ICASSP '05)*, volume 5, pages V–653 – V–656, 18-23 March 2005.

[19] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simon, and H. E. Stanley. Long-range correlations in nucleotide sequences. *Nature*, 356:168–170, March 1992.

[20] A. Arneodo, Y. D' A. Carafa, A. Audit, E. Bacry, J. F. Muzy, and C. Thermes. What can we learn with wavelets about DNA sequences ? *Physica A*, 249:439–448, 1998.

[21] H. Herzel and I. Grobe. Measuring correlations in symbol sequences. *Physica A*, 216:518–542, 1995.

[22] H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, and M. Simons. Scaling features of noncoding DNA. *Physica A*, 273:1–18, 1999.

[23] P. B.-Galvan, P. Carpena, R. R.-Roldan, and J. L. Oliver. Study of statistical correlation in DNA sequences. *Gene*, 300:105–115, 2002.

[24] L. Zhang and Z. Jiang. Long-range correlations in DNA sequences using 2D DNA walk based on pairs of sequential nucleotides. *Chaos solitons & Fractals*, 22:947–955, 2004.

[25] J. Chen, L.-X. Zhang, and D. Zhao. Long-range correlations in DNA sequences using two-dimensional DNA walks. *Chinese Journal of Polymer Science*, 23(1):11–16, 2005.

[26] D. B. Percival and A. T. Walden. *Wavelet methods for time series analysis.* Cambridge Press, Cambridge, 2000.

[27] R. F. Voss. Evolution of long-range fractal correlation and 1/f noise in DNA base sequences. *Physical Review Letters*, 68(25):3805–3808, 1992.

[28] M. de Sousa Vieira. Statistics of DNA sequence: a low-frequency analysis. *Physical Review E*, 60(5):5932–5937, November 1999.

[29] W. Li. The study of correlation structures of DNA sequences: a critical review. *Computer Chemistry;* 21(4):257–271, 1997.

[30] W. Li. Expansion-modification systems: a model for spatial 1/f spectra. *Physical Review A*, 43(10):5240–5260, May 1991.

[31] P. Lio and M. Vannucci. Wavelet change-point prediction of transmembrane proteins. *Bioinformatics*, 16(4):376–382, 2000.

[32] M. Vannucci and P. Lio. Non-decimated wavelet analysis of biological sequences: applications to protein structure and genomics. *Sankhya: The Indian Journal of Statistics*, 63, Series B:218–233, 2001.

[33] S.-Y. Wen and C.-T. Zhang. Identification of isochore boundaries in the human genome using the technique of wavelet multiresolution analysis. *Biochemical and Biophysical Research Communication*, 311:215–222, 2003.

[34] B.-J. Yoon and P. P. Vaidyanathan. Identification of CpG islands using a bank of IIR lowpass filters. In *Proc. 11th Digital Signal Processing Workshop*, Taos Ski Valley, New Mexico, August 2004.

[35] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis.* Cambridge Univ. Press, Cambridge, UK, 1998.

[36] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[37] S. McGinnis and T. L. Madden. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32:W20–W25, 2004.

[38] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[39] W. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. In *Proc. Natl. Acad. Sci. USA*, volume 85, pages 2444–2448, 1988.

[40] I. Cosic. Macromolecular bioactivity: is resonant interaction between macromolecules ? – theory and applications. *IEEE Trans. Biomedical Engg.*, 41(12):1101–1114, 1994.

[41] K. Gupta, D. Thomas, S. V. Vaidya, K. V. Venkatesh, and S. Ramakumar. Detailed protein sequence alignment based on spectral similarity score (SSS). *BMC Bioinformatics*, 6(105), 2005.

[42] Y-Z. Guo, M-L. Li, K-L. Wang, Z-N. Wen, M-C. Lu, L-X. Liu, and L. Jiang. Fast Fourier transform–based support vector machine for prediction of G-protein coupled receptor subfamilies. *Acta Biochimica at Biophysica Sincia*, 37(11):759–766, 2005.

[43] J. Zhao, X. W. Yang, J. P. Li, and Y. Y. Tang. DNA sequence classification based on wavelet packet analysis. In *Proc. 2nd International Conference on Wavelet Analysis and its Application*, pages 424–429, 2001.

[44] M. Buchner and S. Janjarasjitt. Detection and visualization of tandem repeats in DNA sequences. *IEEE Transactions on Signal Processing*, 51(9):2280–2287, 2003.

[45] J. A. Berger, S. K. Mitra, M. Carli, and A. Neri. New approches to genome sequences analysis based on digital signal processing. In *Proc. Workshop on Genomic Signal Processing and Statistics (GENSIP '02)*, Raleigh, NC, USA, October, 2002.

[46] 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*, 341:37–53, 2004.

[47] G. Dodin, P. Vandergheynst, P. Lenoir, C. Cordier, and L. Marcourt. Fourier and wavelet waveform analysis: a tool for visualizing regular patterns in DNA sequences. *Journal of Theoretical Biology*, 206(3):323–326, 2000.

[48] W. Hahn. Telomerase and cancer: where and when ? *Clinical Cancer Research*, 7(10):2953–2954, 2001.

[49] M. Mitas. Trinucleotide repeats associated with human disease. *Nucleic Acids Research*, 25(12):2245–2253, 1997.

[50] R. R. Sinden. Human genetics 99: trinucleotide repeats biological implications of the DNA structures associated with disease-causing triplet repeats. *Am. J. Hum. Genet.*, 64:346–353, 1999.

[51] E. Y. Siyanova and S. M. Mirkin. Expansion of trinucleotide repeats. *Molecular Biology*, 35(2):208–223, 2001.

[52] R. R. Sinden, V. N. Potaman, E. A. Oussatcheva, C. E. Pearson, Y. L. Lyubchenko, and L. S. Shlyakhtenko. Triplet repeat DNA structures and human genetic disease: dynamic mutations from dynamic DNA. *J. Biosci.*, 27(1):53–65, February 2002.

[53] K. Tamaki and A. J. Jeffreys. Human tandem repeat sequences in forensic DNA typing. *Legal Medicine*, 7(4):244–250, 2005.

[54] G. Benson. Tandem repeat finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 27(2):573–580, 1999.

[55] S. Kurtz, J. V. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich. REPuter: the manifold applications of repeat analysis on a genome scale. *Nucleic Acids Research*, 29(22):4633–4642, 2001.

[56] R. Kolpakov, G. Bana, and G. Kucherov. Mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Research*, 31(13):3672–3678, 2003.

[57] G. M. Landau, J. P. Schmidt, and D. Sokol. An algorithm for approximate tandem repeats. *Journal of Computational Biology*, 8(1):1–18, 2001.

[58] E. F. Adebiyi, T. Jiang, and M. Kaufmann. An efficient algorithm for finding short approximate non-tandem repeats. *Bioinformatics*, 17:S5–S12, 2001.

[59] A. M. Hauth and D. A. Joseph. Beyond tandem repeats: complex pattern structures and distant regions of similarity. *Bioinformatics*, 18:S31–S37, 2002.

[60] D. D. Muresan and T. W. Parks. Orthogonal, exactly periodic subspace decomposition. *IEEE Trans. Signal Processing*, 51(9):2270–2279, 2003.

[61] G. Benson. Tandem Repeat Finder [online], software available at *http://tandem.bu.edu/trf/trf.html* [last accessed May, 2006].

[62] S. Kurtz and C. Schleiermacher. REPuter: fast computation of maximal repeats in complete genome. *Bioinformatics*, 15(5):426–427, 1999.

[63] S. Kurtz, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich. Computation and visualization of degenerate repeats in complete genomes. In *Proc. of the International Conference on Intelligent Systems for Molecular Biology*, pages 228–238, Menlo Park, CA, 2000.

[64] A. M. Hauth. *Identification of tandem repeats simple and complex pattern structures in DNA*. PhD thesis, Univ. of Madison, WI, 2002.

[65] P. Rice, I. Longden, and A. Bleasby. EMBOSS: the european molecular biology open source software. *Trends Genetics*, 16:276–277, 2000.

[66] N. Volfovsky, B. J. Haas, and S. L. Salzberg. A clustering method for repeat analysis in DNA sequences. *Genome Biology*, 2(8):27.1–27.11, 2001.

[67] A. T. Castelo, W. Martins, and G. R. Gao. TROLL– tandem repeat occurrence locator. *Bioinformatics*, 18(4):634–636, 2002.

[68] W. A. Sethares and T. W. Staley. Periodicity transform. *IEEE Trans. Signal Processing*, 47(11):2953–2964, 1999.

[69] A. D. Otten and S. J. Tapscott. Triplet repeat expansion in myotonic dystrophy alters the adjacent chromatin structure. In *Proc. Natl. Acad. Sci. USA*, volume 92, pages 5465–5469, 1995.

[70] H. Bussey, D. B. Kaback, W. Zhong, D. T. Vo, M. W. Clark, N. Fortin, J. Hall, B. F. F. Ouellette, T. Keng, A. B. Barton, Y. Su, C. J. Davies, and R. K. Storms. The nucleotide sequence of chromosome I from saccharomyces cerevisiae. In *Proc. Natl. Acad. Sci. USA*, volume 9, pages 3809–3813, 1995.

[71] H. Ogata, S. Audic, V. Barbe, F. Artiguenave, P.-E. Fournier, D. Raoult, and J.-M. Claverie. Selfish DNA in protein-coding genes of rickettsia.

[72] H. Ogata, S. Audic, C. Abergel, P.-E. Fournier, and J.-M. Claverie. Protein coding palindrome are a unique but recurrent feature in rickettsia. *Letter: Genomic Research*, 12:808–816, 2002.

[73] M. D. LeBlanc, G. Aspeslagh, N. P. Buggia, and B. D. Dyer. An annotated catalog of inverted repeats of caenorhabditis elegans chromosomes III and

X, with observations concerning odd/even biases and conserved motifs. *Letter: Genomic Research*, 10:1381–1392, 2000.

[74] C. E. Pearson, H. Zorbas, G. B. Price, and M. Z.-Hadjopoulos. Inverted repeats, stem loops, and cruciforms: significance for initiation of DNA replication. *J. Cell. Biochem.*, 63:1–22, 1996.

[75] J. J. Bissler. DNA inverted repeats and human disease. *Frontiers of Bioscience*, pages 408–418, March 1998.

[76] D. Gusfield. *Algorithms on strings, trees, and sequences: computer science and computational biology.* Cambridge Univ. Press, Cambridge, U.K., 1997.

[77] P. E. Warburton, J. Giordano, F. Cheung, Y. Gelfand, and G .Benson. Inverted repeat structure of the human genome: The X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Letter: Genome Research*, 14:1861–1869, 2004.

[78] C. Berberidis, I. Vlahavas, W. G. Aref, M. Atallah, and A. K. Elmagarmid. On the discovery, of weak periodicities in large time series. In *Proc. 6th European Conf. on Principles and Practice of Knowledge Discovery in Databases*, pages 51–61, 2002.

[79] B. Knudsen and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research*, 31(13):3423–3428, 2003.

[80] S. R. Eddy and R. Dublin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088, 1999.

[81] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1999.

[82] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406–3415, 2003.

[83] I. L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13):3429–3431, 2003.

[84] R. R. Gutell, J. C. Lee, and J. J. Cannone. The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.*, 12:301–310, 2002.

[85] B. Rastegari, Y. S. Zhao, and M. Safari. Linear time algorithm for calculating the free energy of RNA secondary structure including pseudoknots (vs other algorithms). pages 43–62, March 15, 2004.

[86] I. L. Hofacker, M. Fekete, and P. Stadler. Secondary structure prediction for aligned RNA sequences. *Journal of Molecular Biology*, 319(5):1059–1066, 2002.

[87] J. Ruan, G. Stormo, and W. Zhang. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, 20:56–66, 2004.

[88] B. Knudsen and J. Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 5(6):446–454, 1999.

[89] M. S. Burset and R. Guigo. Evaluation of gene structure prediction programs. *Genomics*, 34:353–367, 1996.

[90] J. W. Fickett and C-S. Tung. Assessment of protein coding measures. *Nucleic Acids Research*, 20(24):6441–6450, 1992.

[91] J. W. Fickett. The gene identification problem: an overview for developers. *Computer Chemistry*, 20(1):103–118, 1996.

[92] J.-M. Claverie. Computational methods for the identification of genes in vertebrate genomic sequences. *Human Molecular Genetics*, 6(10):1735–1744, 1997.

[93] G. D. Stormo. Gene-finding approaches for eukaryotes. *Genome Research*, 10:394–397, 2000.

[94] M. Q. Zhang. Computational prediction of eukaryotic protein-coding genes. *Nature Genetics*, 3:698–709, September 2002.

[95] C. Mathe, M.-F. Sagot, T. Schiex, and P. Rouze. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research*, 30(19):4103–4117, 2002.

[96] R. Staden and A. D. McLachlan. Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Research*, 10(1):141–156, 1982.

[97] J. W. Fickett. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research*, 10(17):5303–5318, 1982.

[98] R. Zhang and C.-T. Zhang. Z curves, an intuitive tool for visualizing and analyzing DNA sequences. *Journal Biomolecular Structure Dynamics*, 11:767–782, 1994.

[99] C.-T. Zhang. A symmetrical theory of DNA sequences and its applications. *Journal of Theoretical Biology*, 187:297–306, 1997.

[100] 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. *Nucleic Acids Research*, 28(14):2804–2814, 2000.

[101] Y. Wu, A. W. C. Liew, H. Yan, and M. Yang. Classification of short human exons and introns based on statistical features. *Physical Review E*, 67:1–7, 2003.

[102] A. W.-C. Liew, Y. Wu, and H. Yan. Selection of statistical features based on mutual information for classification of human coding and non-coding DNA sequences. In *17th International Conference on Pattern Recognition*, volume 3, pages 766–769, 2004.

[103] K. Crosby and P. Gabbert. BioSPRINT: classification of intron and exon sequences using the SPRINT. In *Proc. of the IEEE Computational Systems Bioinformatics Conference (CSB '04)*, pages 668–669, Standford, CA, 16-19 August 2004.

[104] D. B. Percival. On estimation of wavelet variance. *Biometrika*, 82(3):619–631, 1995.

[105] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[106] V. Vapnik. *Statistical learning theory.* Wiley, New York, 1998.

[107] Y. Xu, R. J. Mural, J. R. Einstein, M. B. Shah, and E. C. Uberbacher. GRAIL: a multi-agent neural network system for gene identification. *Proceedings of the IEEE*, 84(10):1544–1552, October 1996.

[108] E. E. Snyder and G. D. Stormo. Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Research*, 21(3):607–613, 1993.

[109] V. V. Solovyev, A. A. Salamov, and C. B. Lawrence. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Research*, 22:5156–5163, 1994.

[110] M. Q. Zhang. Identification of protein coding regions in the human genome by quadratic discriminant analysis. In *Proc. Natl. Acad. Sci. USA*, volume 94, pages 565–568, January 1997.

[111] J. Henderson, S. Salzberg, and K. H. Fasman. Finding genes in DNA with a hidden markov model. *Journal of Computational Biology*, 4:127–141, 1997.

[112] M. G. Reese, D. Kulp, H. Tammana, and D. Haussler. Genie-gene finding in drosophila melanogaster. *Genome Research*, 10:529–538, 2000.

[113] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268:78–94, 1997.

[114] A. Krogh. Two methods for improving performance of an HMM and their application for gene finding. In *Proc. of the Fifth International Conference on Intelligent Systems for Molecular Biology*, pages 179–186, Menlo Park, CA, 1997. AAAI Press.

[115] J. Besemer and M. Borodovsky. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Research*, 33:W451–W454, 2005.

[116] M. dos Reis, R. Savva, and L. Wernisch. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Research*, 32(17):5036–5044, 2004.

[117] R. Saito and M. Tomita. On negative selection against ATG triplets near start codons in eukaryotic and prokaryotic genomes. *Journal of Molecular Evolution*, 48:213–217, 1999.

[118] Y. Wang, C.-T. Zhang, and P. Dong. Recognizing shorter coding regions of human genes based on the statistics of stop codons. *Biopolymers*, 63:207–216, 2002.

[119] F.-B. Guo, H.-Y. Ou, and C.-T. Zhang. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Research*, 31(6):1780–1789, 2003.

[120] L.-L. Chen, H.-Y. Ou, R. Zhang, and C.-T. Zhang. ZCURVE_CoV: a new system for recognize protein coding genes in coronavirus genomes, and its applications in anlyzing SARS-CoV genomes. *Biochemical and Biophysical Research Communications*, 307:382–388, 2003.

[121] A. Rusdhi and J. Tuqan. Gene identification using the Z-curve representation. In *Proc. IEEE Int. Conference on Acoustics, Speech, Signal Processing (ICASSP '06)*, pages II–1024 – II–1027, Toulouse, France, 14-19 May 2006.

[122] H. Herzel and I. Grosse. Measuring correlations in symbol sequences. *Physica A*, 216(6):518–542, 1995.

[123] M. Yan, Z.-S. Lin, and C.-T. Zhang. A new Fourier transform approach for protein coding measure based on the format of the Z curve. *Bioinformatics*, 14(8):685–690, 1998.

[124] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. Technical report, Department of computer science and information technology, National Taiwan University, 2003.

[125] A. Mittal and S. Gupta. Automatic content-based retrieval and semantic classification of video content. *International Journal on Digital Libraries*, 6:30–38, 2006.

[126] A. Cornish-Bowden. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research*, 13(9):3021–3030, 1985.

[127] A. Arneodo, Y. D' A. Carafa, E. Bacry, P. V. Graves, J. F. Muzy, and C. Therme. Wavelet based fractal analysis of DNA sequences. *Physica D*, 96:291–320, 1996.

[128] P. Lio and M. Vannucci. Wavelets change-point prediction of transmembrane proteins. *Bioinformatics*, 16(4):376–382, 2000.

[129] P. Lio. Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics*, 19(1):2–9, 2003.

[130] D. B. Percival and H. O. Mofjeld. Analysis of subtidal coastal sea level fluctuation using wavelets. *Journal of American Statistical Association*, 92:868–880, 1997.

[131] B. Whitcher, P. Guttorp, and D. B. Percival. Wavelet analysis of covariance with application to atmospheric time series. *Journal of Geophysical Research*, 105:941–962, 2000.

[132] W. Zhang, X.-T. Zhang, X. Xiong, and C.-Y. Li. Long-memory of shanghai stock market: A wavelet-based approach. In *Fourth International Conference on Machine Learning and Cybernatics*, pages 3496–3500, Guangzhou, 18-21 August 2005.

[133] X. Xiong, X.-T. Zhang, W. Zhang, and C.-Y. Li. Wavelet-based beta estimation of china stock market. In *Fourth International Conference on Machine Learning and Cybernatics*, pages 3501–3505, Guangzhou, 18-21 August 2005.

[134] S. G. Mallat. A theory of multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Machine Intell.*, 11:674–693, 1989.

[135] I. Daubechies. *Ten lectures on wavelets.* SIAM, Philadelphia, 1992.

[136] M. Vetterli. Wavelets and filter banks: theory and design. *IEEE Trans. Signal Proc.*, 40:2207–2232, 1992.

[137] C. Torrence and G. P. Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78, 1998.

[138] M. P. Brown and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. In *Proc. Natl. Acad. Sci. USA*, pages 262–267, 2000.

[139] S. Hua and Z. Sun. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine. *Journal of Molecular Biology*, 308:397–407, 2001.

[140] M. Bhasin, H. Zhang, E. L. Reinherz, and P. A. Reche. Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Letters*, 579:4302–4308, 2005.

[141] C. C. Chang and C. J. Lin. LIBSVM: a library for support vector machine. Technical report, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[142] U. Guldener, M. Munsterkotter, G. Kastenmuller, N. Strack, J. van Helden, C. Lemer, J. Richelles, S. J. Wodak, J. G.-Martinez, J. E. Perez-Ortin, H. Michael, A. Kaps, E. Talla, B. Dujon, B. Andre, J. L. Souciet, J. De Montigny, E. Bon, C. Gaillardin, and H.-W. Mewes. CYGD: the comprehensive yeast genome database. *Nucleic Acids Research*, 33:D364–D368, 2005.

[143] K. K. Wesley, D. J. Sheffler, and B. L. Roth. G-protein-coupled receptors at a glance. *J. Cell Sci.*, 116(24):4867–4869, 2003.

[144] G. Muller. Towards 3D structure of G protein-coupled receptors: a multidisciplinary approach. *Current Medicinal Chemistry*, 7:861–888, 2000.

[145] A. Wise, K. Gearing, and S. Rees. Target validation of G-protein coupled receptors. *Drug Discovery Today*, 7(4):235–246, 2002.

[146] K. Palczewski, T. Kumasaka, T. Hori, C. A. Behnke, H. Motoshima, B. A. Fox, I. Le Trong, D. C. Teller, T. Okada, R. E. Stenkamp, M. Yamamoto, and M. Miyano. Crystal structure of rhodopsin: a G protein-coupled receptor. *Science*, 289:739–745, 2000.

[147] F. Horn, J. Weare, M. W. Beukers, S. Horsch, A. Bairoch, W. Chen, O. Edvardsen, F. Campagne, and G. Vriend. GPCRDB: an information system for G-protein coupled receptors. *Nucleic Acids Research*, 26:277–281, 1998.

[148] F. Horn, G. Vriend, and F. E. Cohen. Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. *Nucleic Acids Research*, 29(1):346–349, 2001.

[149] F. Horn, E. Bettler, L. Oliveira, F. Campange, F. E. Cohen, and G. Vriend. GPCRDB information system for G-protein coupled receptors. *Nucleic Acids Research*, 31:294–297, 2003.

[150] S. Takeda, S. Kadowaki, S. Haga, H. Takaesu, and S. Mitaku. Identification of G-protein-coupled receptor genes from the human genome sequence. *FEBS Letter*, 520:97–101, 2002.

[151] S. M . Foord. Receptor classification: post genome. *Current opinion Pharmacology*, 3:114–120, 2003.

[152] P. K. Papasaikas, P. G. Bagos, Z. I. Litou, and S. J. Hamodrakas. PRED-GPCR: GPCR recognition and family classification server. *Nucleic Acids Research*, 32:W380–W382, 2004.

[153] R. Karchin. Classifying G-protein coupled receptors with support vector machines. Master's thesis, University of California, Computer Science, UC Santa Cruz, CA 95064, 2000.

[154] R. Karchin, K. Karplus, and D. Haussler. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, 18(1):147–159, 2002.

[155] M. Bhasin and G. P. S. Raghava. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Research*, 32:W383–W389, 2004.

[156] W. Pearson. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, 85:2444–2448, 2000.

[157] A. M. Lesk, M. Levitt, and C. Chothia. Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. *Protein Engineering*, 1:77–78, 1986.

[158] T. Oda, N. Morikawa, Y. Saito, Y. Mauho, and S. Matsumoto. Molecular cloning and characterization of a novel type of histamine receptor preferentially expressed in leukocytes. *Journal of Biological Chemistry*, 275(47):36781–36786, 2000.

[159] K. Hofmann, P. Bucher, L. Falquet, and A. Bairoch. The PROSITE database, its status in 1999. *Nucleic Acids Research*, 27:215–219, 1999.

[160] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. J. A. Sigrist, K. Hofmann, and A. Bairoch. The PROSITE database, its status in 2002. *Nucleic Acids Research*, 30(1):235–238, 2002.

[161] J. Y. Huang and D. L. Brutlag. The EMOTIF database. *Nucleic Acids Research*, 29(1):202–204, 2001.

[162] T. K. Attwood, M. D. Croning, D. R. Flower, A. P. Lewis, J. E. Mabey, P. Scordis, J. N. Selley, and W. Wright. PRINT-S: the database formerly known as PRINTS. *Nucleic Acids Research*, 28(1):225–227, 2000.

[163] T. K. Attwood, M. J. Blythe, D. R. Flower, A. Gaulton, J. E. Mabey, N. Maudling, L. McGregor, A. L. Mitchell, G. Moulton, K. Paine, and

P. Scordis. PRINTS and PRINTS-S shed light on protein ancestry. *Protein Engineering*, 30(1):239–241, 2002.

[164] J. G. Henikoff, A. G. Elizabeth, S. Pietrokovski, and S. Henikoff. Increased coverage of protein families with the blocks database servers. *Nucleic Acids Research*, 28(1):228–230, 2000.

[165] P. K. Papasaikas, P. G. Bagos, Z. I. Litou, and S. J. Hamodrakas. A novel method for GPCR recognition and family classification from sequence alone using signatures from profile hidden markov models. *SAR and QSAR in Environmental Research*, 14:413–420, 2003.

[166] B. Qian, O. S. Soyer, R. R. Neubig, and R. A. Goldstein. Depicting a protein's two faces: GPCR classification by phylogenetic tree-based HMMs. *FEBS Letter*, 554:95–99, 2003.

[167] S. Moller, J. Vilo, and M. D. R. Croning. Prediction of the coupling specificity of G protein coupled receptors to their G protein. *Bioinformatics*, 17:S174–S181, 2001.

[168] T. Jaakkota, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7:95–114, 2000.

[169] Y. Huang, J. Cai, L. Ji, and Y. Li. Classifying G-protein coupled receptors with bagging classification tree. *Computational Biology and Chemistry*, 28:275–280, 2004.

[170] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157:105–132, 1982.

[171] V. Veljkovic and I. Slavic. General model for pseudopotentials. *Physical Review Letter*, 29:105–108, 1972.

[172] L. J. Buturovic. PCP: a program for supervised classification of gene expression profiles. *Bioinformatics*, 22(2):245–247, 2006.

[173] S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721–728, 2001.

Author's Publications

International Journals (published)

1. R. Gupta, D. Sarthi, A. Mittal and K. Singh. A novel signal processing measure to identify exact and inexact tandem repeat patterns in DNA sequences. *EURASIP Journal on Bioinformatics and Systems Biology*, Article ID 43596, pages 17, 2007, doi:10.1155/2007/43596.

2. R. Gupta, A. Mittal and S. Gupta. An efficient algorithm to detect palindromes in DNA sequences using periodicity transform. *Signal Processing Journal*, Elsevier Publication, 86:2067-2073, 2006.

International Conferences

1. R. Gupta, A. Mittal and K. Singh. Discrete wavelet transform features for classification of protein coding regions in DNA sequences using SVM. In *Proc. International Conference on Bioinformatics (InCoB '06)*, pages P-046.1- P-046.6, New Delhi, India, 18-20 December 2006.

2. R. Gupta, A. Mittal and K. Singh. Identifying inverted repeat structure in DNA sequences using correlation framework. In *Proc. of European Conference on Signal Processing (EUSIPCO '06)*, pages 1865.1-1865.5, Italy, September 2006.

3. R. Gupta, A. Mittal and K. Singh. Exactly periodic subspace decomposition based approach for identifying tandem repeats in DNA sequences. In *Proc. of European Conference on Signal Processing (EUSIPCO '06)*, pages 1857.1-1857.5, Italy, September 2006.

4. R. Gupta, A. Mittal and S. Gupta. Predicting secondary structure of RNA using correlation framework. In *Proc. of Eight International Conference on Information Technology (CIT '05)*, pages 257-262, Bhubaneswar, India, 20-23 December 2005.

5. R. Gupta, A. Mittal, V. Narang and W-K. Sung. Detection of palindromes in DNA sequences using periodicity transform. In *IEEE International Workshop on Biomedical Circuits and Systems (BioCAS '04)*, pages S2.7.INV-20 - S2.7.INV-23, Singapore, 1-3 December 2004.

International Journals (under communication)

1. R. Gupta, A. Mittal and K. Singh. A novel and efficient technique for identification and classification of GPCRs. *IEEE Transaction on IT in Biomedicine*, 2007.

2. R. Gupta, A. Mittal and K. Singh. Wavelet variance feature for classifying protein coding regions in a DNA sequence using SVM. *International Journal on Bioinformatics and Computational Biology*, 2007.

3. R. Gupta, A. Mittal and K. Singh. Correlation framework for identification of inverted repeats in DNA sequence. *Journal on Computational Intelligence in Bioinformatics*, 2007.

4. R. Gupta, A. Mittal and K. Singh. Prediction of RNA secondary struc-

ture based on correlation measure. *International Journal of Bioinformatics Research and Applications*, 2007.