PREDICTION OF PROTEIN CODING REGIONS IN DNA SEQUENCES

USING THE STOCKWELL TRANSFORM

A Thesis

Presented to the

Faculty of

San Diego State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Electrical Engineering

by

Shivashekar Murali Shankar

Fall 2011
SAN DIEGO STATE UNIVERSITY

The Undersigned Faculty Committee Approves the

Thesis of Shivashekar Murali Shankar:

Prediction of Protein Coding Regions in DNA Sequences Using the

Stockwell Transform

Ashkan Ashrafi, Chair
Department of Electrical Engineering

Satish Kumar Sharma
Department of Electrical Engineering

Morteza Monte Mehrabadi
Department of Mechanical Engineering

9/1/2011
Approval Date
DEDICATION

To the almighty Sri Sri Sri 1008 Raghavendra Swamiji
and
My Family
The power of God is with you at all times; through the activities of mind, senses, breathing, and emotions; and is constantly doing all the work using you as a mere instrument.

-Srimad Bhagavadgeetha
ABSTRACT OF THE THESIS

Prediction of Protein Coding Regions in DNA Sequences Using the Stockwell Transform

by

Shivashekar Murali Shankar

Master of Science in Electrical Engineering
San Diego State University, 2011

Bioinformatics is the application of information technology to the field of molecular biology. Bioinformatics has evolved drastically over time and involves the analysis and interpretation of biological data. De-oxy ribonucleic acid (DNA) stores the instructions which are responsible for functioning and development of living organisms. These instructions are referred to as genetic instructions and the segments of DNA which carry these instructions are referred to as genes. Genomics is a popular field in bioinformatics which involves genes and their analysis. An outstanding fact about genomic information is that, it can be represented digitally. Genomic information in DNA can be expressed as set of character strings in which each character correspond to a Nucleotide. DNA sequence consists of four nucleotides namely A, G, C, and T. All genetic instructions are stored in this DNA sequence. Genomic signal processing is an engineering discipline which deals with analysis and processing of genomic signals.

The aim of this research is to demonstrate the application of digital signal processing techniques in the field of Bioinformatics. Analysis of the sequence pattern in DNA to predict the protein coding regions, also called as exons, forms the crux of this thesis.

Initially, the analytical applications of digital signal processing in the field of genomics were thoroughly studied. Further, the DNA sequence was examined in detail to predict the exons from the given DNA sequence using Stockwell Transform. The C.Elegans F56F11.4 chromosome was taken as an input from the public database and the protein coding regions from this chromosome was predicted using Stockwell Transform. Our prediction results showed the correct magnitude and location of exons. The implementation of the research was done using the simulation capabilities of MATLAB.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>PAGE</th>
<th>ABSTRACT ............................................................................................................................. vi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LIST OF TABLES ................................................................................................................... ix</td>
</tr>
<tr>
<td></td>
<td>LIST OF FIGURES .............................................................................................................. x</td>
</tr>
<tr>
<td></td>
<td>LIST OF ABBREVIATIONS ................................................................................................... xii</td>
</tr>
<tr>
<td></td>
<td>ACKNOWLEDGEMENTS .......................................................................................................... xiii</td>
</tr>
</tbody>
</table>

## CHAPTER

1. **INTRODUCTION** ...........................................................................................................1

2. **BASICS OF MOLECULAR BIOLOGY** ............................................................................6

3. **PROTEIN FORMATION IN LIVING ORGANISMS** ......................................................14
   3.1 Steps Involved in Protein Formation ...............................................................14
      3.1.1 Transcription .......................................................................................... 14
      3.1.2 Splicing .................................................................................................. 15
      3.1.3 Translation ............................................................................................. 16
   3.2 Protein Formation in Prokaryotes ....................................................................16
   3.3 Protein Formation in Eukaryotes .....................................................................17

4. **EXISTING METHODS FOR PREDICTION OF EXONS** ............................................20
   4.1 Introduction to Windowed DFT Method .........................................................20
      4.1.1 Methodology .......................................................................................... 21
      4.1.2 Procedure for Prediction of Exons (Windowed DFT Method) .............. 23
      4.1.3 Flowchart of the Procedure (Windowed DFT Method) ......................... 24
      4.1.4 Results .................................................................................................... 24
   4.2 Introduction to EIIP Method............................................................................24
      4.2.1 Methodology .......................................................................................... 24
      4.2.2 Procedure for Prediction of Exons (EIIP Method) ................................ 30
      4.2.3 Flowchart of the Procedure (EIIP Method) ........................................... 31
      4.2.4 Results .................................................................................................... 31
5 PREDICTION USING STOCKWELL TRANSFORM ...........................................33
   5.1 Introduction to S-Transform .................................................................33
      5.1.1 The Gaussian Window .................................................................35
      5.1.2 Derivation of S-Transform from STFT ...........................................36
      5.1.3 Linearity and the Effect of Noise ....................................................36
      5.1.4 Inverse of the S-Transform ............................................................37
      5.2 Why S-Transform? ..........................................................37
   5.3 Flowchart of the Procedure (S-Transform Method) .........................40
   5.4 Results ..................................................................................41
   5.5 Prediction of Exons Using Time Frequency Filtering Approach ..........43
   5.6 Comparison of Different Methods for Prediction of Exons .................46

6 CONCLUSION .................................................................................48

BIBLIOGRAPHY .................................................................................50
LIST OF TABLES

PAGE

Table 2.1. Conversion for Amino Acids ................................................................. 12
Table 4.1. EIIP Value of Nucleotides ................................................................. 28
Table 5.1. Position Comparison of Protein Coding Regions from F56F11.4 ................. 46
LIST OF FIGURES

Figure 1.1. Flowchart of DNA analysis using different methods.................................................4
Figure 2.1. Prokaryotic cell structure ........................................................................................7
Figure 2.2. Eukaryotic cell structure .........................................................................................7
Figure 2.3. Format of a DNA sequence ....................................................................................7
Figure 2.4. Structure of double stranded DNA .........................................................................8
Figure 2.5. Schematic representation of amino acids. ..............................................................9
Figure 2.6. Flowchart of DNA ................................................................................................10
Figure 2.7. Format of a gene ...................................................................................................10
Figure 3.1. Transcription process .............................................................................................15
Figure 3.2. Translation process ...............................................................................................17
Figure 3.3. Protein formation in prokaryotes ..........................................................................18
Figure 3.4. Protein formation in eukaryotes ...........................................................................19
Figure 4.1. Flowchart of the procedure (Windowed DFT method) ...........................................25
Figure 4.2. Plot of windowed $X_A[k]$ sequence. .................................................................26
Figure 4.3. Plot of windowed $X_g[k]$ sequence ....................................................................26
Figure 4.4. Plot of windowed $X_c[k]$ sequence .................................................................27
Figure 4.5. Plot of windowed $X_t[k]$ sequence .................................................................27
Figure 4.6. Plot of $S(k)$ with five exons (Windowed DFT) ...................................................28
Figure 4.7. Plot of $S_{AN}(k)$ with five exons (EIIP method) .................................................29
Figure 4.8. Flowchart of the procedure (EIIP method) ..........................................................32
Figure 5.1. Time and frequency domain representation of Gaussian window .......................35
Figure 5.2. Illustration of S-Transform -1. ..........................................................................38
Figure 5.3. Illustration of S-Transform -2. ..........................................................................39
Figure 5.4. Flowchart of the procedure using S-Transform ...................................................41
Figure 5.5. Prediction of first exon ..........................................................................................42
Figure 5.6. Prediction of second exon. .................................................................43
Figure 5.7. Prediction of third exon. .....................................................................44
Figure 5.8. Prediction of fourth exon. ....................................................................44
Figure 5.9. Prediction of fifth exon. .......................................................................45
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>De-oxy Ribonucleic Acid</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier Transform</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>BIS</td>
<td>Binary Indicator Sequences</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger Ribonucleic Acid</td>
</tr>
<tr>
<td>ORF</td>
<td>Open Reading Frames</td>
</tr>
<tr>
<td>GSP</td>
<td>Genomic Signal Processing</td>
</tr>
<tr>
<td>DSP</td>
<td>Digital Signal Processing</td>
</tr>
<tr>
<td>A</td>
<td>Adenine</td>
</tr>
<tr>
<td>G</td>
<td>Guanine</td>
</tr>
<tr>
<td>C</td>
<td>Cytosine</td>
</tr>
<tr>
<td>T</td>
<td>Thiamine</td>
</tr>
<tr>
<td>EIIP</td>
<td>Electro Ion Pseudo Potential</td>
</tr>
<tr>
<td>S-Transform</td>
<td>Stockwell Transform</td>
</tr>
<tr>
<td>STFT</td>
<td>Short Time Fourier Transform</td>
</tr>
<tr>
<td>CWT</td>
<td>Continuous Wavelet Transform</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

The satisfaction that accompanies a successful completion of any task would be incomplete without the mention of people, whose constant guidance and encouragement crowned the efforts with success.

I express my heartfelt thanks to my thesis advisor Dr. Ashkan Ashrafi for providing me an opportunity in an excellent environment to carry out my research work under him. I have learnt a lot from him not only in Digital Signal Processing, but also in embracing personal values and relationships which shall last for a lifetime. I felt motivated and encouraged every time I attended his meetings. ‘Thanks’ will merely be only a formality because his guidance and inspiring persona was a major factor in the research which will always be cherished. I would also like to mention that extending a financial support towards my research assured his confidence in it and my abilities.

I would like to thank my family members without whose moral support my work would have become impossible.

I would like to thank the members of the Electrical Department who willingly extended their support and encouraged me towards my research.

I would like to thank my roommate Balaji Sundararaman without whose constant help and guidance on the concepts of Molecular Biology, this research would have lagged its pace.

Finally I would like to thank all my friends and especially my roommates (Amit, Varun, Vinoth, Balaji, Rakesh Prabhakara) for creating a healthy atmosphere, which encouraged me time and again to excel.
CHAPTER 1

INTRODUCTION

Molecular biology can be termed as the study of infinitesimal biological components at a molecular level. Molecular biology involves in-depth study of cells and their behavior. It is closely related to cell biology. Genetics and biochemistry are two areas which are interlinked with molecular biology. Molecular biology has its roots in biochemistry and genetics.

Biochemistry deals with the function and organization of chemical process inside the living organisms. Also, it deals with the structure and functions of basic cellular components such as proteins, carbohydrates, nucleic acids and other bio-molecules [1].

Cells present in living organisms contain many complex substances such as de-oxy ribonucleic acid (DNA), chromosomes and genes that shape the identity (heredity) which can be passed on to other cells. This identity helps to determine certain characteristics, such as hair color or the likelihood of having certain traits. These characteristics are commonly referred to as genes and study of these genes is known as genetics [2].

The prime focus of molecular biology has been the basics of genetics with the emphasis on DNA organization. It focuses heavily on structure and functionality of DNA. Regulation of genes and modifications of the gene expression have also been given high priority. One of the major factors has been the development and applications of genetic engineering. The core of the molecular biology involves the study about the genetic material, which is transcribed into ribonucleic acid (RNA) and then translated into protein. It also involves the study of structural properties of DNA and their application [2, 3].

Modern technology has allowed many selective methods to find and modify specific gene expressions, thereby laying a new foundation in molecular biology. The concepts and techniques in molecular biology have been rapidly and effectively employed to resolve numerous biological problems.

Most of the work in molecular biology is mathematical in nature and lot of work has been done recently at the interface of molecular biology and information technology [4].
Biological data is very huge and complex in nature. The biological data has vastly contributed in feeding the scientific research community with massive amount of information that can be stored and analyzed. The most pressing task of scientists and researchers involved in the field of molecular biology is to face the challenge of analyzing and interpreting massive numbers of adenines (A), thymines (T), cytosines (C) and guanines (G) sequences that encode functional biological processes [5, 6].

Advancements in technology have played a major role in integrating sciences with most comprehensive analysis tools. Storage and maintenance of huge biological data is one of the most important parts, which the technology has taken care of. Internet based search tools and desktop based search tools for analyzing the genome data are creating wonders in the field of bioinformatics. Many statistical tools and algorithms have also been developed to address this issue.

Bioinformatics is the application of information technology to analyze and process biological data. It is an interface between molecular biology and Information technology which uses computational algorithms and high end statistical software tools to solve many biological problems [5, 7].

Bioinformatics involves the use of computers in manipulating and organizing biological information so that we can analyze and understand tremendous amount of information present with us. By sorting and handling of biological data, bioinformatics helps us in discovering and detecting valuable information contained in huge volumes of biological data [4].

The field of bioinformatics is expected to provide important predictions from biological data. The challenge is the analysis of huge volumes of data that can no longer be performed by humans and requires computers, efficient algorithms to do the same. This is an emerging interdisciplinary area of molecular biology and information technology, which plays a key role in the pharmaceutical and agricultural sectors. This interdisciplinary field also helps in understanding of new targets for drug discovery and clinical research.

There are quite a few aims of bioinformatics, first, making storage access and retrieval of biological database at ease. Second and most importantly, develop tools and techniques for analysis of data that not only performs a statistical or semantic search but also apply biological knowledge and priorities while carrying out the analysis. For example, while
comparing proteins, the comparison should also involve a biological match, not just a text based match. Analysis like this involves advances in computational theory along with application of biological knowledge or parameters in the search function initiated by the system. Thirdly, the obtained results should be interpreted with a biological reference and not as string bits. This should be done to accommodate a global analysis of data with the aim of discovering general principles across many systems [8, 9, 10].

Cell is the basic building block of life. Human body contains more than 30 trillion cells. Millions of genetic instructions are contained in these cells which account for the protein production inside the human body. DNA contains the complete set of genetic instructions to determine the traits of the organism. This forms the base for drug discovery and diagnosis.

Genetic information is employed in the field of bioinformatics to analyze the complex DNA structure. Whole DNA data can be represented as character strings in which each alphabet corresponds to a nucleotide. Millions of such nucleotides make up a genetic instruction. To process this huge genetic instruction, we need robust computational algorithms and statistical tools. As these genetic instructions and DNA sequences can be expressed as discrete sequences, a new domain called genomic signal processing (GSP) came into existence [4].

Some regulatory decisions are taken by cells, which utilize many critical inputs to do it. As there are many critical inputs to be considered in order to model a genetic network, designing a decision making genetic network is a herculean task. Advanced and innovative analytical tools are very much necessary to build such kinds of complex modeling networks. This introduces us to a very efficient domain known as GSP. GSP deals with analysis and processing of genomic signals.

Conversion of nucleotides into character strings (discrete sequences) helps in applying digital signal processing (DSP) methods to analyze and manipulate genome data.

GSP integrates the theory and practical applications of signal processing with the global understanding of molecular biology. GSP is a domain focusing mainly on classification of genomic networks. Prediction and detection of genome from a high profile microarray serves as a solution to many pattern recognition systems. GSP also deals with the
statistical and dynamic modeling of gene networks which in turn helps in diagnosis of critical parameters related to diseases [4].

Many gene prediction methods have already been implemented and the exact location and position of the exons have been found [4, 5]. Some of the spectral analysis algorithms have also been developed to serve this purpose. Protein coding regions and non-coding regions from a DNA sequence have been predicted using these algorithms. Due to the very high volume of genome data, data representation is also a concern. Variations in this huge data and several numerical plots can easily be analyzed in GSP, making it an interesting tool.

Analysis of DNA data is a very complex task. It requires high end software tools and statistical algorithms to get the accurate result. Currently, this is done using efficient data structures and robust computer science algorithms. Computational complexity is high here, as the biological data is very huge. So, GSP is used as the preferred tool for DNA analysis in many complex cases. Use of DSP methods assures a very accurate analysis and better computational efficiency.

The analysis flowchart is shown in Figure 1.1.

![Flowchart of DNA analysis using different methods.](image)

GSP is widely used in sub areas of bioinformatics, such as:

- Proteomics
- Gene Prediction
- Analysis of DNA Microarrays
- Clinical discoveries and many more.

Here is a very brief outlook of the thesis. Chapter 2 provides a motivation by discussing the basics of molecular biology. It talks briefly about cells and their behavior,
DNA, genes, and their functionality in living organisms. It also talks about amino acids and protein structure. Core molecular biological techniques such as translation and transcription are thoroughly discussed. It also gives an insight into molecular biology from a technical point of view.

In Chapter 3, we discuss the formation of proteins in prokaryotes and eukaryotes. We take a look at the conversion of amino acids to proteins. The basic difference between the two cell biological processes has been explained. Gene prediction in prokaryotes and eukaryotes has been looked upon in detail.

In Chapter 4, we discuss various protein coding methods and their complexity. We consider C.Elegans chromosome and predict the exons from the given DNA sequence. The Fourier transform method and the electro ion pseudo potential (EIIP) prediction method has been looked into in detail and the comparison between the two add a new flavor to this chapter. Performance evaluation of the two methods along with the sufficient results has been covered here.

In Chapter 5, we mainly concentrate on Stockwell Transform (S-Transform) and its advantage over the other conventional transforms. We predict the protein coding regions by passing the same DNA sequence through a S-Transform. S-Transform and its application in GSP has been explained here. Comparison of several methods has been discussed and advantages of S-Transform have been proved over other methods.

Chapter 6 concludes this thesis.
CHAPTER 2

BASICS OF MOLECULAR BIOLOGY

Cell is the basic unit of all living organisms. Cells play an important role in maintaining the balance of living organisms. It mainly contains genetic or genomic information to code protein and thereby function independently. It is composed of carbohydrates, proteins, lipids and many biomolecules. Cells may be single-celled or multi-cellular. Based on this, organisms are majorly classified into two types, namely prokaryotes and eukaryotes.

Prokaryotes are relatively simple, unicellular organisms. They are regarded as the most primitive form of living organisms. Prokaryotes lack a membrane-bound nucleus and are very small when compared to eukaryotes [2]. Their genetic information is spread along the cell and surrounds the outer layer of the cell. The shape of the prokaryotic cell may be rod-shaped, spiral or spherical. Most of the bacteria and viruses fall under this category. Bacterial cells are very prominent to the digestive system of the living organisms and they act as decomposers too.

Eukaryotes are multi-celled organisms. They are a complex form of prokaryotes. Many homologous pairs of chromosomes are present in eukaryotic cells, which make them more special than prokaryotes. Eukaryotes are bound by a nucleus, where majority of the genetic information is stored. Almost all species of large organisms are eukaryotes, including animals, plants and fungi. Figure 2.1 and 2.2 show the basic structure of prokaryotes and eukaryotes [1].

Genetic instructions in the human body correspond to the functioning and other critical developmental activities inside it. DNA is the nucleic acid, which contains these kinds of genetic instructions [2]. These may include traits and other genetic factors.

DNA consists of a pair of complementary nucleotide strand, which is made up of de-oxy ribose sugar residue and a phosphoryl group, with a combination of one of the four bases namely A, T, C, and G. The sequence of four bases present in the nucleotide strands encodes the genetic information. These four character strings runs into millions in a specific
manner referred to as genes. DNA is a combination of many such genes. Figure 2.3 shows the format of a DNA sequence.

DNA

| GENE 1 | GENE 2 | GENE 3 | GENE 4 | GENE 5 | GENE 6 | ...... | ...... |

Figure 2.3. Format of a DNA sequence.

DNA strand, which is represented by the four letters A, T, C, and G, has two distinct “ends,” namely, the “5’ end” and the “3’ end.” The 5’ end of a nucleotide is linked to the 3’ end of another nucleotide by a very strong chemical bond, thus forming a long one-dimensional strand of a specific direction and functionality [1]. Therefore, each DNA strand can be mathematically represented by character strings, where each character represents one of the four nucleotides. There are two types of conventions that are involved in reading the information from the DNA strand. Convention 5’ to 3’ specifies the direction when read from left to right whereas 3’ to 5’ specifies the direction when read from right to left [4].

Base A can only bond with base T and base G can only bond with base C. These polymers are anti-parallel in nature. Hence these strands are normally referred to as double stranded complementary DNA strands [1, 2]. The schematic representation of a double stranded DNA is shown in Figure 2.4.
Consider an example of part of a DNA double strand:

5’- A-C-A-A-G -3’ (read from left to right; 5’ - 3’)
3’- T-G-T-T-C- 5 ‘(read from right to left; 3’ - 5’)

The above-mentioned DNA strand can be read in two directions. These character strings ‘ACAAG’ and ‘TGTTC’ can be alternatively used to describe the same DNA double strand, but they specify two different single strands, which are said to be complementary to each other. DNA strands that are complementary to themselves are called self-complementary, or palindromes. In the above-mentioned example, ‘ACAAG’ is a palindrome [4, 11].

The most important functionality of a gene is to code the protein inside the living organism. Amino acids play an important role in protein formation. Amino acids are alpha-amino carboxylic acids made up of amino group and carboxylic acid group attached to a carbon atom, to which another side chain is also connected. There are many types of amino acids depending on the type of side chains are attached; which determines the physic-chemical properties of the amino acid [1]. Two amino acids can form covalent amide
bonds between the carboxylic acid groups of one amino acid to the amino group of another. Proteins, which are the building blocks of a cell, are polymers of amino acids of a specific order. Figure 2.5 shows the schematic diagram of an amino acid [1].

DNA contains instructions, which are very helpful to form a vital component inside the human body called “Proteins.” Proteins are organic molecules that contain carbon, hydrogen, oxygen and nitrogen. Every function in the living cell depends on proteins. They are made up of unbranched chains of polymers. Originating from a very large segment of DNA, many smaller segments present inside it contribute to the “protein” formation. These smaller segments of DNA, which carry the vital genetic information, are referred to as Genes.

Gene is a sequence of DNA, which contains information in order to compose a full or a part of a functional protein or a RNA molecule. Gene is also a basic unit of an organism’s genome, which determines the characters of that particular organism. Genes are passed from ancestors to offspring, causing a particular offspring to inherit traits or characteristics of their ancestors [1, 2].

DNA is totally divided into two subgroups namely ‘Genes’ and ‘intergenic regions’. The functionality of a gene is to code the protein inside the living organism. They contain two major regions namely exons and introns. The intergenic regions do not contribute to the protein formation, but they help in synthesizing a gene, also known as transcription. Figure 2.6 summarizes the whole process.
DNA resides in the nucleus of each of the body’s trillions of cells. Each cell has many molecules of double-stranded DNA. Each molecule is made up of millions of small packets bound together called as chromosomes. The DNA present in chromosome has millions of genes inside it [4].

Every gene consists of two major regions namely Exons and Introns. Exons code the protein inside the human body. They are also called as the protein coding regions. The particular sequences of base pairs solely constitute for the protein coding inside the human body. Exons are transcribed into messenger RNA and eventually, they code for amino acids in the proteins. The length of exons may vary from organism to organism [1].

Introns are often referred to as non-coding regions, as they do not code the protein inside the human body. These regions separate the two exons, which are present in a given gene. Till date, the research and study conducted on the functional relevance has been very scarce and limited but it has been assumed that they regulate the expression of a gene in time and space [2]. Finally, for the production of protein inside a human body, the Introns must be spliced in a process called “splicing” which occurs before translation. Figure 2.7 shows the format of a gene sequence.
Protein is coded inside living organisms by a process called transcription. The basic input for this process is a codon. Codon is a triplet nucleotide, which codes specifically for an amino acid. There are four nucleotides (A, T, G, and C) and hence there are 64 possibilities to arrange them as three per group. However in nature, there are only 20 essential amino acids, which code the proteins. Hence, there is a possibility that one amino acid can be coded by more than one triplet code. Each organism prefers to use unique codons for a particular amino acid and hence evolutionarily distinct family of organisms use different codons for same amino acids and is known as codon bias [2].

The protein coding regions within genes are denoted by start and stop codons. Each codon represents a specific amino acid. Typically, there is one start codon and three stop codons. A start codon signifies the beginning while a stop codon signifies the end of the protein-coding region. The rest of the codons correspond to one of the 20 possible amino acids of a protein. Table 2.1 lists of amino acids, which code the protein inside the living organism [4].

Gene sequences, which are considered in triplets, are known as open reading frames (ORF). The region of the DNA sequences from the start codon (ATG) to the stop codon (TAG/TGA/TAA) is called the Open Reading frame.

In prokaryotes, an ORF is a sequence of DNA that starts with start codon “ATG” and ends with any of the three stop codons (TAA/ TAG/ TGA). Coding of gene sequences can happen in either the “forward” or the “reverse” coding direction. Depending on the starting point, there are six possible ways (three in forward direction and three in reverse direction) of translating any nucleotide sequence into amino acid sequence according to the genetic code [4, 12].

Consider an example DNA sequence:

**DNA Sequence:** ATT GCA AGA

**Coding in the forward direction:**

- **First ORF:** ATT GCA AGA
- **Second ORF:** AT TGC AAG A
- **Third ORF:** A TTG CAA GA

In the case mentioned above, the first ORF is obtained by considering the very first nucleotide in the DNA sequence and grouping it into three per group from thereon.
Table 2.1. Conversion for Amino Acids

<table>
<thead>
<tr>
<th>CODONS</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCG, GCA, GCC, GCT</td>
<td>Alanine</td>
</tr>
<tr>
<td>AGG, AGA, CGG, CGA, CGC, CGT</td>
<td>Arginine</td>
</tr>
<tr>
<td>AAC, AAT</td>
<td>Asparagine</td>
</tr>
<tr>
<td>GAC, GAT</td>
<td>Aspartic Acid</td>
</tr>
<tr>
<td>TGC, TGT</td>
<td>Cysteine</td>
</tr>
<tr>
<td>CAG, CAA</td>
<td>Glutamine</td>
</tr>
<tr>
<td>GAG, GAA</td>
<td>Glutamic acid</td>
</tr>
<tr>
<td>GGG, GGA, GGC, GGT</td>
<td>Glycine</td>
</tr>
<tr>
<td>CAC, CAT</td>
<td>Histidine</td>
</tr>
<tr>
<td>ATA, ATC, ATT</td>
<td>Isoleucine</td>
</tr>
<tr>
<td>CTG, CTA, CTC, CTT, TTG, TTA</td>
<td>Leucine</td>
</tr>
<tr>
<td>AAG, AAA</td>
<td>Lysine</td>
</tr>
<tr>
<td>ATG</td>
<td>Methionine</td>
</tr>
<tr>
<td>TTC, TTT</td>
<td>Phenylalanine</td>
</tr>
<tr>
<td>CCG, CCA, CCC, CCT</td>
<td>Proline</td>
</tr>
<tr>
<td>AGC, AGT, TCG, TCA, TCC, TCT</td>
<td>Serine</td>
</tr>
<tr>
<td>ACG, ACA, ACC, ACT</td>
<td>Threonine</td>
</tr>
<tr>
<td>TGG</td>
<td>Tryptophan</td>
</tr>
<tr>
<td>TAC, TAT</td>
<td>Tyrosine</td>
</tr>
<tr>
<td>GTG, GTA, GTC, GTT</td>
<td>Valine</td>
</tr>
<tr>
<td>TAG, TGA, TAA</td>
<td>Ochre</td>
</tr>
</tbody>
</table>


The second open reading frame is obtained by considering the second nucleotide of the DNA sequence and then grouping it till the end. A similar method is followed to get the third open reading frame.

Coding in the reverse direction:

CAG TCT TGC
.CA GTC TTG C
..C AGT CTT GC

In the above-mentioned example, the procedure to get the ORF is similar to that of forward direction. The only difference is that the strand is read from reverse direction [4].

Another important aspect of these DNA sequences is that, they can be intensified and stored in microarrays. A DNA microarray is an array, which consists of millions of
microscopic spots of DNA nucleotides. Every single microscopic spot corresponds to a specific DNA sequence also known as probes. These probes are attached to a solid surface to form a chemical matrix. Attaching millions of such probes on a chip forms a microarray chip. DNA microarrays are widely used to measure changes in genetic levels, to detect and predict the mutant genomes, or to compare the healthy genomes with the affected ones. These arrays have played a dramatic role in drug discovery and have formed a strong base for research and development.

The National Center for Biotechnology Information (NCBI) gives an excellent opportunity for researchers by providing access to biomedical and genomic information. It has the information about most of the genes, gene mapping and assays. The resources in the website include the complete detail about the DNA and RNA of most of the chromosomes. It has the domains and structures of many cells too [13].

File formats are very much essential in bioinformatics to retrieve and write data. As the biological data is very huge, making use of a suitable file format makes it easy to store and read data. Nucleotide data consists of only letters and it can be stored in a simple text pad or a notepad. “txt” format is very much efficient and it can store up to 50k nucleotides without much hassle. NCBI provides direct access to these files and it is stored in a file format called “FASTA.” FASTA format is widely used in the areas such as gene sequence alignment and protein analysis [13]. FASTA files of most of the chromosomes are available in NCBI, and it is a free download to the user. Information from the FASTA file can be retrieved easily by using the proper software [13].

A typical FASTA file has the following properties:

- First line in the FASTA file starts with “>” symbol. This symbol is followed by a very brief description of the gene and its unique accession number.
- Second line is a textual sequence comprising of one of the four nucleotides. This data is very huge and runs into many lines.
- Sequences of multiple chromosomes can be present here.
CHAPTER 3

PROTEIN FORMATION IN LIVING ORGANISMS

3.1 STEPS INVOLVED IN PROTEIN FORMATION

Proteins are important organic compounds present in living organisms. They are essential in regulating almost all cell functions. Protein structure is classified into four types: primary, secondary, tertiary and quaternary. Lengthy chains of amino acids bonded together forms a protein. Of the 64 amino acid group, only 20 amino acids code the protein. Amino acids form a firm bond between them and they are highly flexible in shape. The core to formation of proteins is amino acids. They determine the functionality of a protein [1, 2].

Proteins are classified on the basis of their functions. They act as enzymes, regulate process in hormones and also act as antibodies. In other words, all the functionality which takes place inside a cell is because of proteins.

Formation of protein inside a living organism takes place in three steps:
1. Transcription
2. Splicing
3. Translation

3.1.1 Transcription

The process of creating a copy of RNA from DNA sequence is called transcription. RNA is chemically similar to DNA but has an extra hydroxyl group in the ribose sugar backbone. It is believed to be evolutionarily primitive to DNA. RNA molecules act in many different roles such as enzymes, information storage molecules and regulatory molecules. There are three major types of RNA namely messenger ribonucleic acid (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA), each with specialized functions. mRNA codes information for protein synthesis, tRNA carries amino acids for protein synthesis and rRNA forms the catalytic core for ribosome which makes proteins. There are also small RNA molecules and micro RNA which regulate gene expression [1, 3].
A chemical component called RNA polymerase reads the DNA sequence and produces a complementary RNA strand. Thus, a copy of DNA is maintained and can be used for further processes. On the whole, during the transcription process, DNA sequence is transcribed into an RNA molecule. This RNA molecule can be used to encode a gene. Once the RNA molecule is well equipped to regulate the process and contribute for the encoding of the gene, a biomolecule called mRNA is created [2].

Messenger RNA is one of three major types of RNA (mRNA, tRNA and rRNA). This is biologically important molecule because it contains the information for synthesizing new proteins. mRNA is copied from DNA in a process called transcription where a complementary strand is synthesized for the template DNA strand. The final output of transcription process in the production of mRNA which can be further used in a process called translation to code a protein [2].

In prokaryotes, mRNA is almost similar to the DNA sequence containing the gene starting from ATG and ending with one of three STOP codons with some nucleotide over hangs.

In the eukaryotes, mature form of mRNA is made in two step process: Transcription generates pre-mRNA with introns and exons and splicing of introns produces mature mRNA [1, 2]. Figure 3.1 shows the transcription process.

![Figure 3.1. Transcription process.](image)

**3.1.2 Splicing**

The output of the transcription process (mRNA) contains both wanted and unwanted parts to code the protein. Splicing is basically done to remove the unwanted parts during the protein formation.
Removal of non-coding intron regions from the mRNA transcript is called as splicing. This is carried out by an enzyme complex called as spliciosome. Enzymes in the spliciosome recognize special nucleotide sequences during the beginning, middle and end stages of an intron. One can predict the exon-intron junctions by looking for these specific nucleotides [1, 2]. Due to their chemical properties, spliciosome chop this intron specifically and joins the two ends of the flanking Exons thereby making a continuous coding messenger RNA.

Prokaryotes are unicellular and hence do not contain introns. Gene structure of Prokaryotic cell comprises of only exons. Hence, it is a very rare possibility that they get spliced. Splicing is not that effective and it is very rarely used in prokaryotic protein coding. Eukaryotes are multicellular and they do contain both, introns and exons. Exons are got by splicing the intron regions from the mRNA. The exons which are obtained by splicing contribute to the protein coding. Splicing is very effective and it is very frequently used in eukaryotic protein coding [1].

### 3.1.3 Translation

The input to the translation process is the mRNA obtained by transcription process. Translation is the process of decoding the genetic message which is coded in the messenger RNA into a string of amino acids, which are joined sequentially to form the protein. Translation is carried out inside the cell by the macromolecular machinery called as ribosome [2].

The nucleotide triplets (codons) present in the mRNA are read by the transfer RNA molecules which carries the specific amino acids. Two adjacent amino acids are chemically linked by covalent peptide bonds. The order and the specificity of amino acid are determined by the mRNA. This process continues until the ribosome sees a STOP codon, at this particular point the process is terminated and mature protein (string of newly joined amino acids) is released [1, 2]. Finally, all exons are joined together in order to code the protein inside the living organism [1]. Figure 3.2 shows the translation process.

### 3.2 Protein Formation in Prokaryotes

In a single celled organism (prokaryotes), DNA sequence consists of genes, which in turn consists of only exons. These exons are first extracted from a process called
transcription and they are converted to protein by a process called translation, which have been discussed. These proteins which are obtained from the translation process are the building blocks for prokaryotes [1]. See Figure 3.3.

3.3 PROTEIN FORMATION IN EUKARYOTES

In a multi cellular organism (eukaryotes), DNA sequence consists of genes, which in turn consists of both introns and exons. These exons are first extracted to form a pre-RNA which also contains exons and introns. Further, introns are spliced and removed from pre RNA in a process called splicing. Finally, exons are clubbed together to form protein by a process called translation. The proteins which are obtained from this translation process form the building blocks for numerous cells present in the eukaryotes [2].

Conversion of a gene to a protein in eukaryotes is shown in Figure 3.4.
Figure 3.3. Protein formation in prokaryotes.
Figure 3.4. Protein formation in eukaryotes.
CHAPTER 4
EXISTING METHODS FOR PREDICTION OF EXONS

4.1 INTRODUCTION TO WINDOWED DFT METHOD

Prediction of protein coding regions in eukaryotes is a very difficult task. Protein coding regions (exons) in eukaryotes are not continuous in nature. They are accompanied by non-coding regions called introns. Distinguishing protein coding regions from non-coding regions is a challenging problem in the field of molecular biology. Many algorithms in time and frequency domain have been designed and implemented to predict the protein coding regions. Spectral content in DNA sequence are exploited by using so many DSP methods [4, 5, 14, 15].

DNA sequence is the combination of four nucleotides (or bases). A character string of nucleotides (A, T, C, and G) can be mapped to four signals known as binary indicator sequences (BIS) [5]. The binary indicator sequence takes the value of either binary “1” or binary “0” depending upon the presence and absence of base A at position "."

These BIS are divided into many subsequences known as words. Each word is a subsequence of many nucleotides. These words are termed as "w." Period-3 behavior in genes is defined by an observation from [16].

Observation: “When the length w of the subsequence is a multiple of three, i.e., \( w = 3m \), a 3-sample periodicity is observed in the nucleotide count functions in the coding regions” [16].

According to [4, 16, 17, 18], a protein coding region exhibits a peak at \( k = \frac{N}{3} \) (exhibits a relatively large value) whereas no such behavior is observed in the non-coding regions. This property is referred to as “periodicity property” or “period-3” behavior. Here, we implement one of those discrete Fourier transform (DFT) based splicing algorithm mentioned in [5].
Period-3 behavior is exploited in the DFT based splicing algorithm to find the protein coding regions in a DNA sequence. Here, we compute the magnitude of the frequency component at \( k = N/3 \). A digital antinotch filter is used to extract the protein coding regions and to suppress the noise [5], around the coding regions. Finally, a suitable windowing technique is applied to remove the noise components which accompanies the desired output and also to smoothen the curve.

### 4.1.1 Methodology

Considering an example DNA sequence

\[
x_a[n] = \text{“ATGGA”}
\] (4.1)

The binary indicator sequence for \( x_a[n] \) can be given by:

\[
x_a[n] = [10001]
\] (4.2)

The BIS \( x_G[n] \), \( x_C[n] \), and \( x_T[n] \) can be obtained in a similar fashion by substituting values of either binary “1” or binary “0” depending upon the presence and absence of base G or C or T at position “\( n \)”, respectively.

So, binary indicator sequence for \( x_G[n] \) is:

\[
x_G[n] = [00110]
\] (4.3)

So, binary indicator sequence for \( x_C[n] \) is:

\[
x_C[n] = [00000]
\] (4.4)

So, binary indicator sequence for \( x_T[n] \) is:

\[
x_T[n] = [01000]
\] (4.5)

It is evident that the sequence \([11111]\) is obtained by adding the above four BIS.

Next, each of these BIS is passed through an IIR antinotch filter to extract the protein coding regions from the DNA sequence.

In DSP, an ideal antinotch filter passes only one frequency and eliminates all other frequencies. It is the opposite of a notch filter [5].

We consider a digital filter \( H(z) \) which has a magnitude response of \( H(e^{jw}) \). According to the “period-3” property, the response of the digital filter will have a sharp peak.
at $\omega = 2\pi / 3$. IIR antinotch filters require shorter orders for their implementation compared to FIR filters, which makes them more effective. We consider second order IIR antinotch filter for our implementation [5].

The transfer function of the second order IIR antinotch filter is given by [5]:

$$A(z) = \frac{R^2 - 2R \cos \theta z^{-1} + z^{-2}}{1 - 2R \cos \theta z^{-1} + R^2 z^{-2}}$$

(4.6)

Note that $R^2 \leq 1$ for stability.

Each of the four BIS is passed through the anti notch filter to extract the protein coding regions respectively. Noise components (signals without period-3 characteristics) in those sequences can be suppressed if the order of the filter is increased. Outputs from the antinotch filters are termed as $Y_s[n]$, $Y_c[n]$ $Y_e[n]$ and $Y_t[n]$, respectively.

The protein coding regions in genes have period-3 components [5]. A protein coding region exhibits a peak at $k = N / 3$ (exhibits a relatively large value) whereas no such behavior is observed in the non coding regions. This is exploited to find exons by applying a Windowed DFT method.

Windowed DFT is one of the traditional methods used for prediction of exons [5]. Due to period-3 property, DFT coefficient magnitude at $k = N / 3$ is often significantly larger than the surrounding DFT coefficient magnitude. So, DFT at only “one single point per window” can be considered for computation.

These outputs are then passed through a DFT block of length $N$, given by [19, 20]:

$$X_A[k] = \sum_{n=0}^{N-1} x(n) e^{-2\pi nk/N}$$

(4.7)

Similarly, DFTs of $x_c[n]$ $x_e[n]$ and $x_t[n]$ are calculated as $X_c[k]$ $X_e[k]$ and $X_t[k]$, respectively. As period-3 property is used to find the protein coding regions, we consider the period to be a multiple of 3.

Fourier transforms do not clearly indicate how the frequency content of a signal changes over time. Amplitude of the spectrum does not specify this as the frequency related information is hidden in the phase of the signal. The work around for such types of problem
is to truncate the given signal into overlapping parts and then computing the Fourier transform of each block [g].

Though this gives us a better result than the conventional Fourier transform, noise components tend to add at the endpoints of each block. So, a proper windowing technique can be applied in order to nullify the noise at the endpoints. This whole procedure is referred to as Windowed Fast Fourier Transform (FFT) or Short-Time Fourier Transform [g].

There are many types of windowing techniques available for different applications of signal processing [21, 22]. A suitable window is chosen for better resolution and sharpness of the peak. Here, we are considering Kaiser window for suppressing the noise components [21]. The DFT at the single point $k = N/3$ is calculated and the Kaiser window is slided by one or more bases to calculate $S(N/3)$. The window length is chosen as 351, as it is a multiple of 3. This process is followed for all the four BIS to get respective Windowed FFTs.

Finally, we calculate the total strength of the peak (exons). The total strength of the peak in the protein coding region can be calculated as [5]:

$$S[k] = X_A[k]^2 + X_G[k]^2 + X_C[k]^2 + X_T[k]^2$$  

(4.8)

### 4.1.2 Procedure for Prediction of Exons  
(Windowed DFT Method)

The following procedure explains the prediction of exons using Windowed DFT.

1. A given FASTA file format is directly read into the MATLAB using Bioinformatics toolbox.
2. Sequence inside the FASTA file is converted to binary format and four duplicate copies of the same are created. These copies are called BIS.
3. In the first copy, presence of nucleotide ‘A’ is marked as binary ‘1’ while the absence of it is marked as ‘0’ (zero).
4. In the second copy, presence of nucleotide ‘T’ is marked as binary ‘1’ while the absence of it is marked as ‘0’ (zero).
5. In the third copy, presence of nucleotide ‘C’ is marked as binary ‘1’ while the absence of it is marked as ‘0’ (zero).
6. In the fourth copy, presence of nucleotide ‘G’ is marked as binary ‘1’ while the absence of it is marked as ‘0’ (zero).
7. All the four BIS are then passed through an IIR antinotch filter for the extraction of coding regions and noise suppression.
8. Outputs from the antinotch filter are then passed through DFT block of length \( N \) to predict the protein coding regions of the DNA sequences.

9. Calculation of DFT at the single point \( k = \frac{N}{3} \) is calculated and the Kaiser window is slid by one or more bases to calculate \( S(\frac{N}{3}) \).

10. Finally, squared magnitudes of the four distinct sequences are added to give the desired output using (4.8).

### 4.1.3 Flowchart of the Procedure (Windowed DFT Method)

See Figure 4.1 for flowchart of the procedure (Windowed DFT method).

### 4.1.4 Results

Using the FFT method, the protein coding regions of 8000 nucleotides were predicted. Protein coding regions have the higher magnitude and are represented by peaks. The corresponding values in the x-axis will give the location of the exons. The final plot \( S[k] \) indicates the presence and location of the exons. Results of the prediction using Windowed FFT are shown in Figures 4.2 through 4.6.

### 4.2 INTRODUCTION TO EIIP METHOD

In the Windowed DFT method which was seen earlier, four BIS were used and all the further computations were done on those four sequences to predict the exons. According to [23], EIIP Indicator sequence method considers only one sequence known as “EIIP sequence,” which is formed by substituting the EIIP of the nucleotides A, G, C, and T in the DNA sequence respectively. Instead of considering four BIS, only one EIIP sequence is considered here. Computational complexity is low here as the computational overhead is reduced by 75% [23].

#### 4.2.1 Methodology

Calculation of energy of “delocalized electrons” in amino acids and nucleotides are termed as the EIIP. There are several methods which employ this EIIP to predict the hot spots in protein and peptide design [23, 24]. In this method, we are making use of the EIIP of the nucleotides to predict the protein coding regions in a DNA sequence. The EIIP values for the nucleotides are given by Table 4.1 [23].
Figure 4.1. Flowchart of the procedure (Windowed DFT method).
Figure 4.2. Plot of windowed $X_a[k]$ sequence.

Figure 4.3. Plot of windowed $X_g[k]$ sequence.
Figure 4.4. Plot of windowed $X_c[k]$ sequence.

Figure 4.5. Plot of windowed $X_r[k]$ sequence.
Figure 4.6. Plot of $S^{(k)}$ with five exons (Windowed DFT).

Table 4.1. EIIP Value of Nucleotides

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>EIIP Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1260</td>
</tr>
<tr>
<td>G</td>
<td>0.0806</td>
</tr>
<tr>
<td>C</td>
<td>0.1340</td>
</tr>
<tr>
<td>T</td>
<td>0.1335</td>
</tr>
</tbody>
</table>


Considering a random sequence $x[n]$ with a sequence of “AGCTA,” and by using the values of EIIP and substituting the same, we get:

$$X[k] = [0.1260 \ 0.0806 \ 0.1340 \ 0.1335 \ 0.1260]$$

The above sequence can be termed as the EIIP indicator sequence. This sequence is taken as a reference sequence and all further computations are performed on this sequence. This numerical sequence represents the distribution of the free electron’s energies along the sequence [23].
Here also, we consider a suitable digital filter $H[z]$ which has a magnitude response of $H(e^{jw})$ as taken in the previous Windowed DFT method. According to the “period-3” property, the response of the digital filter will have a sharp peak at $w = \frac{2\pi}{3}$. IIR antinotch filters require shorter orders for their implementation compared to FIR filters, which makes them more effective. We consider second order IIR antinotch filter (4.6) for our implementation [5].

EIIP indicator sequence is then passed through the IIR antinotch filter to extract the protein coding regions respectively. Noise components in those sequences can be suppressed if the order of the filter is increased. Output obtained from the antinotch filter is termed as $Y_{AN}[n]$.

Note that, here we are considering just EIIP indicator sequence to perform all computations instead of four BIS sequences. See Figure 4.7.

![Graph](image)

**Figure 4.7.** Plot of $S_{AN}(k)$ with five exons (EIIP method).

Referring to Figure 4.7, it turns out that the power spectra of DNA segments that correspond to exons tend to exhibit a relatively strong component at the period-3 frequency, whereas segments that correspond to introns exhibit a relatively weak component at the
period-3 frequency. The plot of $S_{AN}(k)$ is shown above and it reveals a peak at $k = N/3$ in the protein coding regions whereas no peak is observed in the non-coding regions. This can be used as an excellent coding measure to predict the exons as the computational overhead is reduced by 75%.

Later, by selecting a proper windowing method, noise accompanying the desired output can be reduced to a great extent, thereby making the method more effective.

The corresponding DFT of 4.10 is represented as $Y_{AN}[k]$:

$$Y_{AN}[k] = \sum_{k=0}^{N-1} y(n)e^{-2\pi ik\frac{n}{N}}$$

\[(4.10)\]

where $k = 0,...,N-1$.

The output is then split and applied to a Kaiser window which suppresses the noise more effectively as it has small side lobes compared to rectangular window, which are used in traditional Windowed DFT method.

Finally we calculate the total strength of the peak (exons). The total strength of the peak in the protein coding region can be calculated as [23]:

$$S_{AN}(k) = Y_{AN}[k]^2$$

\[(4.11)\]

### 4.2.2 Procedure for Prediction of Exons (EIIP Method)

The following procedure explains the prediction of exons using EIIP method.

1. A given FASTA file format is directly read into the MATLAB using bioinformatics toolbox.
2. Sequence inside the FASTA file is converted to character strings and a new array is created, which contains the series of Nucleotides (A/G/C/T). This is copied into a new EIIP array.
3. In the new EIIP array, nucleotides A, G, C, and T are assigned their respective EIIP values.
4. The new EIIP array is then passed through an IIR antinotch filter for the extraction of coding regions and noise suppression.
5. Outputs from the antinotch filter are then passed through DFT block of length $N'$ to predict the protein coding regions of the DNA sequences.
6. Calculation of DFT at the single point $k = N/3$ is calculated and the Kaiser window is slided by one or more bases to calculate $S_{AN}(N/3)$.

7. Finally, the magnitudes $S_{AN}(k)$ are squared to give the desired output using (4.10).

4.2.3 Flowchart of the Procedure (EIIP Method)
See Figure 4.8. for flowchart of the procedure (EIIP method).

4.2.4 Results
Using the EIIP method, the protein coding regions of 8000 nucleotides were predicted. Protein coding regions have the higher magnitude and are represented by peaks. The clarity of the peaks is much better here compared to the traditional Windowed DFT method. This is because of the substitution of EIIP values in the DNA sequence. The locations of exons are more accurate when compared to the previous method.
Figure 4.8. Flowchart of the procedure (EIIP method).
CHAPTER 5

PREDICTION USING STOCKWELL TRANSFORM

5.1 INTRODUCTION TO S-TRANSFORM

Signals in practical applications are typically finite in duration and their frequency characteristics vary over time rapidly. Hence, processing and analysis of these signals have become an increasingly tough task to handle. Information technology has evolved innovatively over the time so that higher end mathematical tools can be used to analyze the signal and capture various important parameters of the signal.

Fourier analysis is widely used in the field of GSP and genomic prediction is achieved in an efficient manner as compared to previous traditional methods. Period-3 behavior of DNA sequences can be easily exploited using Fourier transform and the genetic sequences can be predicted efficiently. As DNA sequence consists of billions of nucleotides, finding the exact location of the particular genomic sequence is of prime importance. Though Fourier analysis has been widely used in GSP, it has its own limitations [4].

Fourier transform only produces the time-averaged spectrum of the given input signal.

If the characteristics of the time series vary, finding the spectral content of the signal will be a complicated task. Also, if there are redundant identical inputs to the Fourier analysis, amplitude spectra might look identical. This causes a problem as multiple frequencies cannot be captured using Fourier analysis [25].

Fourier analysis provides the frequency domain representation of the entire continuous signal. It does not specify the frequency content of the desired signal at different moments of time. This limitation led to the development of the Short Time Fourier Transform or STFT [26, 27].

STFT or Windowed FFT is a method which breaks down the non stationary signal into many smaller blocks and applies the conventional Fourier transform on each block. In
In this case, tracking of signal dynamics of non-stationary signals in time frequency analysis is restricted to the window width, i.e., a large window length provides good resolution in frequency domain and poor resolution in time domain. On the other hand, small window length provides good resolution in time domain and poor resolution in frequency domain. This is a major limitation of STFT and this makes the analysis of signals in real time more difficult [25, 26, 27].

Other useful tool to analyze time frequency characteristics of a signal is the Continuous Wavelet Transform (CWT) [25, 28]. CWT gives a very detailed description of signals in both time and frequency. CWT has more flexibility in time and frequency resolution compared to STFT. In STFT, the width of the window will remain constant throughout the analysis of the signal. In CWT, we can have window functions of variable lengths using translation/retrenchment of a window function called the mother wavelet function [26].

The mother wavelet is defined as [26]:

\[
\varphi(t) = \frac{1}{\sqrt{s}} \varphi \left( \frac{t-\tau}{s} \right)
\]  

(5.1)

where \( \varphi(t) \) is a mother wavelet function, \( s \) is the scale of wavelet transform and, \( \tau \) is the time shift.

The wavelet transform is defined as [26]:

\[
W_x(s, \tau) = \int_{-\infty}^{+\infty} x(t) \overline{\varphi(t)} dt
\]

(5.2)

where \( W_x(s, \tau) \) are called wavelet coefficients.

So, by having a mother wavelet function, the time resolutions are dependent on the frequency. Any small change in the signal can be effectively revealed and this makes it an important tool in signal analysis techniques [25, 26].

STFT uses fixed window analysis for computation of a one dimensional input signal. STFT is a multidimensional function, which often tries to map the one-dimensional signal into two-dimensional plane (time and frequency). This results in more redundancy and ends up in more resource consumption from the computational point of view [28, 29]. Though CWT is good tool to perform signal analysis, it only localizes the power spectrum as a
function of time. It does not retain the absolutely referenced phase information and therefore it cannot be directly inverted to FT spectrum. CWT is good for extraction of information in both time and frequency domain, but at the cost of introducing noise since it is sensitive to noise [25, 26].

We now look at the S-Transform and how different it is from the other transforms. Explained below are some related terminologies and properties of S-Transforms.

### 5.1.1 The Gaussian Window

Let us look at some important properties of “Gaussian window”:

1. It has a symmetric property in both time and frequency domain, i.e., the FT of a Gaussian is a Gaussian.
2. There is only main lobe in Gaussian and no side lobes are present. This helps in getting better frequency resolution.
3. On a decibel scale, Gaussians are quadratic. The “parabolic interpolation” of the Gaussian is perfect. This will come in handy when estimating the sinusoidal peak frequencies in the spectrum [30].

Figure 5.1 shows the time and frequency domain representations of the Gaussian window. The length is chosen to be 64.

![Figure 5.1. Time and frequency domain representation of Gaussian window.](image)
5.1.2 Derivation of S-Transform from STFT

Considering $\tau$ as time spectral localization, $f$ as the Fourier frequency, $g(t)$ as the Gaussian window function, $\sigma$ as window width, STFT of signal $h(t)$ can be defined as [25]:

$$STFT(\tau, f) = \int_{-\infty}^{\infty} h(t)g(\tau - t)e^{-2\pi\tau \sigma^2} dt$$  \hspace{1cm} (5.3)

The scalable Gaussian function can be described as [25]:

$$g(t) = \frac{|f|}{\sigma \sqrt{2\pi}} e^{-\frac{t^2 f^2}{2}}$$  \hspace{1cm} (5.4)

S-Transform is obtained by replacing the window function $g(t)$ with the scalable Gaussian function as given replacing window function $g(t)$ in (5.3):

$$S(\tau, f) = STFT(\tau, f) = \int_{-\infty}^{\infty} h(t)\frac{|f|}{\sigma \sqrt{2\pi}} e^{-\frac{(\tau - t)^2 f^2}{2}} e^{-2\pi\tau \sigma^2} dt$$  \hspace{1cm} (5.5)

It can be seen from (5.5) that the S-Transform is a special case of STFT with a scalable Gaussian window function. The Gaussian window function described in 5.4 is a function of time and frequency [30]. As the width and height of the window is dictated by the frequency, the frequency resolution at lower frequencies and time resolution at higher frequencies can be obtained by widening or narrowing the Gaussian window [25, 31, 32].

5.1.3 Linearity and the Effect of Noise

The S-Transform is a linear operation on the time series $h(t)$. One of the advantages here is that the data can be modeled as sum of signal and noise [25, 31]:

$$Data(t) = Signal(t) + noise(t)$$  \hspace{1cm} (5.6)

Applying S-Transform on both sides to (5.6) [25, 31]:

$$S\{Data\} = S\{Signal\} + S\{noise\}$$  \hspace{1cm} (5.7)

This is possible due to the fact that S-Transform is a linear operation on the time series $h(t)$.

So, application of S-Transform on a signal will lead to the transformed data and noise. So, noise components are separated instantaneously which will result in much better data [31, 32, 33].
5.1.4 Inverse of the S-Transform

S-Transform is a representation of the local spectrum, i.e., the simple operation of averaging the local spectra over time would give the Fourier transform spectrum.

\[ \int_{-\infty}^{+\infty} S(\tau, f) d\tau = H(f) \]  

(5.8)

where \( H(f) \) is the Fourier transform of \( h(t) \).

Further, we can also prove that:

\[ h(t) = \int_{-\infty}^{+\infty} S(\tau, f) e^{2\pi i f \tau} d\tau = H(f) \]  

(5.9)

which clearly shows that \( h(t) \) could be retrieved from \( S(\tau, f) \) and hence the Fourier transform of non-stationary time series could be attributed to S-Transform [25, 31, 33].

5.2 Why S-Transform?

The task of this thesis was to implement the most suitable mathematical operation that shall enable us to conveniently and effectively determine the time and frequency components of the extracted information. Hence, we looked upon some of the popular transforms in order to sift in the most suitable one. Now, we have the S-Transform, which gives emphasis to the resolution, being a function of frequency in this case, of the time frequency domain. Since our main intention is to accentuate the time resolution and frequency resolution at initial and later parts of the information respectively, we find S-Transform to be the most suitable of transforms since we can apply different windows, such as Gaussian, bi-Gaussian and the hyperbolic, based on our requirements [25, 26].

Because of the advantages in time and frequency resolution, it is used to predict the protein coding regions from a given DNA sequence. The S-Transform has gained popularity in the signal processing community because of its applications in geophysics, oceanology, engineering and biomedicine [34].

Let us look at an example which explains S-Transform before we apply it on a DNA sequence. Figures 5.2 and 5.3 which explain the S-Transform graphically.

Figure 5.2 which is shown above has 3 sub plots. The first subplot shows a sinusoid with three distinct time periods of 20, 31, and 60 sec. The second subplot and the third
Figure 5.2. Illustration of S-Transform -1.

subplots are the S-Transform of the signal. As we can see, in these subplots, X-axis gives the location of the time series. Y-axis gives the distinct frequency components of the signal. As we have three distinct time periods in the input signal, respective frequency components are obtained at .05 Hz, 0.032 Hz, and 0.016 Hz. In Figure 5.2, in these subplots, if we trace down horizontally from the Y-axis, frequency components are identified by “contour circles.” Peak amplitude is depicted as the brightest color on the contour circle. The width of
Figure 5.3. Illustration of S-Transform -2.
the contour circle depends on the “dilation factor” mentioned in (5.5). The narrower the width of the contour circle, the better is the frequency resolution.

One of the challenging factors in S-Transform is to have a contour circle which is of smaller width. This helps us in tracking down the distinct frequency components of the signal.

In Figure 5.2, the continuous sinusoid has a high frequency component of 0.05 Hz. In the second subplot, the contour circle corresponding to the high frequency does not have a narrow window width which is a concern. This is because, at lower frequencies, numbers of cycles (full cycle oscillations) of the input signal are more which will fit into the Gaussian window correctly. At high frequencies, numbers of cycles of the input signal are less which will not fit into the Gaussian window. In that case, at higher frequencies, width of the window is wider. The frequency resolution in a S-Transform can be improved by adjusting the parameter $\lambda$ in the scalable Gaussian window in (5.5). More the value of $\lambda$, narrower is the window. Though the narrowness of the window is achieved, distinct “frequency components” of a signal tend to overlap each other. So, a tradeoff between these two should be considered for better frequency resolution.

Third subplot is plotted with a higher value. We can see the difference evidently between second and third subplot. In Figure 5.2, the second subplot is plotted with $\lambda=0.8$. That is the reason for width of the window to be wider at high frequency. Also, in Figure 5.2, the third subplot is plotted with $\lambda=4$. That is the reason for width of the window to be narrower at high frequency.

For better analysis of the same procedure, we are considering the S-Transform of the same signal and plotting it through the “mesh” function in MATLAB. This gives the 3D view of the plotted signal. The peaks correspond to the amplitude of the given signal. We can observe the distinct frequency components of the signal through the “side-view.”

Figure 5.3 shows the same function plotted using “mesh” function.

### 5.3 Flowchart of the Procedure (S-Transform Method)

Figure 5.4 explains the whole procedure graphically.
5.4 RESULTS

Prediction of exons from the C.Elegans chromosome is done by considering first 8000 nucleotides at once. Due to computational complexity of S-Transform in a simulation environment, we consider breaking up the whole 8000 nucleotides into five parts and then predict the exons from those.

Protein coding regions have the highest intensity and are represented by contour circles. The corresponding values in the x-axis will give the location of the exon. While
predicting exons using the FFT method, whole 8000 nucleotides were considered at once and protein coding regions were predicted.

The breakup of the C.Elegans chromosome is as follows:

Part 1: 0-1400
Part 2: 1401-3000
Part 3: 3001-4500
Part 4: 4501-6000
Part 5: 6000-8000

Protein coding regions from the C.Elegans chromosome was predicted and the results are as shown in the following Figures.

The location of first exon is depicted in Figure 5.5. This location corresponds to the 1/3 frequency which is observed on the Y-axis. The location of the first exon is: 949-1040.

Figure 5.5. Prediction of first exon.
The location of second exon is depicted in Figure 5.6. This location corresponds to the 1/3 frequency which is observed on the Y-axis. The location of the second exon is: 2531-2921.

![Figure 5.6. Prediction of second exon.](image)

The location of third exon is depicted in Figure 5.7. This location corresponds to the 1/3 frequency which is observed on the Y-axis. The location of the third exon is: 4089-4418.

The location of fourth exon is depicted in Figure 5.8. This location corresponds to the 1/3 frequency which is observed on the Y-axis. The location of the fourth exon is: 5444-5650.

The location of fifth exon is depicted in Figure 5.9. This location corresponds to the 1/3 frequency which is observed on the Y-axis. The location of the fifth exon is: 7301-7599.

### 5.5 Prediction of Exons Using Time Frequency Filtering Approach

This research was started in January 2009 and was carried out as a part of academic thesis. Recently a research paper was published by Sitanshu Sekhar Sahu and Ganapati Panda
Figure 5.7. Prediction of third exon.

Figure 5.8. Prediction of fourth exon.
Our research aims at predicting the location of exons by application of S-Transform. The research carried out by the above mentioned authors enhances the above method by application of “time frequency filtering approach” and inverse S-Transform, making it a more accurate method.

We do not have any competitive interests with the paper which was published in June 2011. The research was carried concurrently by them and they were the first to publish the paper.

The procedure followed by the authors is explained below [35]:

1. C.Elegans chromosome is taken as input and the nucleotides are converted to their respective EIIP values.
2. S-Transform of the EIIP is calculated as per the formula.
3. In order to separate the period-3 components from the other introns, a band limited filter is applied to the output from the previous stage. This selects the period-3 components only, rejecting others.

4. As the locations are pre known, the time frequency filter mask is applied to separate out the exons.

5. Energy of the sequence is found out by taking the magnitude square of the output sequence.

5.6 COMPARISON OF DIFFERENT METHODS FOR PREDICTION OF EXONS

In this thesis, we have predicted the protein coding regions using three different methods.

Table 5.1 shows the location of exons obtained by each of those methods compared against each other. The locations mentioned by NCBI for the same chromosome (F56F11.4) is also shown [13].

<table>
<thead>
<tr>
<th>Exon Number</th>
<th>Location From NCBI</th>
<th>Windowed DFT Method</th>
<th>EIIP Method</th>
<th>S-Transform</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>929-1039</td>
<td>939-1182</td>
<td>949-1158</td>
<td>949-1040</td>
</tr>
<tr>
<td>2</td>
<td>2528-2857</td>
<td>2543-3008</td>
<td>2521-2980</td>
<td>2531-2921</td>
</tr>
<tr>
<td>3</td>
<td>4114-4377</td>
<td>4050-4439</td>
<td>4080-4451</td>
<td>4089-4418</td>
</tr>
<tr>
<td>4</td>
<td>5465-5644</td>
<td>5458-5669</td>
<td>5480-5682</td>
<td>5444-5650</td>
</tr>
<tr>
<td>5</td>
<td>7255-7605</td>
<td>7380-7800</td>
<td>7402-7730</td>
<td>7301-7599</td>
</tr>
</tbody>
</table>

By looking at column 3 from Table 5.1, we initially predicted the protein coding regions in a given DNA sequence using Windowed DFT method. Though the exons were predicted from a given DNA sequence, noise components accompanied the protein coding regions. So, the spectral content method followed by anti notch filter failed to detect the locations accurately.

There was some variation in the location of exons compared to the ones mentioned by NCBI [13]. The locations of the exons obtained from Windowed DFT method are shown in column 3.

The second method which was considered in this thesis was the EIIP method to predict the protein coding regions. This method used the EIIP of nucleotides to predict the
exons from the given DNA sequence. This method gave a good spectral component compared to Windowed DFT method and reduced the computational complexity by 75% as only one BIS was considered for computation [23]. As seen from column 4 from Table 5.1, locations of the exons are revealed more accurately than Windowed DFT method. This method has better accuracy compared to the Windowed DFT and the non-coding region is significantly suppressed. The locations of the exons obtained from the EIIP method are shown in column 4.

Finally, we looked at the S-Transform to predict the exons from the DNA sequence. Column 5 from Table 5.1 gives us the locations of the exons obtained through S-Transform. These locations match the locations from NCBI much closer than the other two methods. S-Transform method was more accurate than the above mentioned methods because of its advantages in time and frequency resolution. Because of the linearity property of the S-Transform, the signal and the noise does not get mixed [25]. Therefore, the S-Transform has a better noise performance compared to the other methods. It is very evident from column 5 of Table 5.1 that Stock well transform predicts the exons more accurately than any other method discussed in this thesis.
CHAPTER 6

CONCLUSION

Initially, prediction of protein coding regions (exons) was done through the established Windowed DFT method. An anti notch filter was used in this method to select only the coding regions and to suppress the non coding regions. Though, we were able to predict the protein coding regions, we had to compromise for the noise contents accompanying the output signal.

Secondly, we looked at the EIIP method, which reduced the computational complexity by 75% by just considering one BIS sequence. This method also had an antinotch filter to achieve the same purpose as in Windowed DFT method. As a single BIS was used, noise components accompanying the output were lesser than the previous method. This gave more accurate results than the previous method in terms of predicting the location of the exons.

Because of the advantages such as linearity and effectiveness in time and frequency resolution, S-Transform is widely used in Geophysical signal analysis and data mining. Also, S-Transform finds its application in Magnetic resonance imaging and identification of soil and structures. S-Transform helps in interpreting and analyzing complex data.

As the biological data is very huge and complex in nature, S-Transform was used for prediction of protein coding regions. S-Transform was applied on four BIS and location of the exons was predicted with less noise effect and thus qualifying the results as the optimal desired output. The predictions obtained using S-Transform conforms to the predictions using the Windowed DFT method and EIIP method. The magnitude and locations of the Exons were noted. Finally, the location of these exons showed that the prediction using S-Transform gives better result.

Traditional Windowed DFT method gave us an insight to prediction of protein coding regions. EIIP method enhanced the same process more effectively. But S-Transform helped us in getting accurate results by overcoming few obstacles that were present in previous methods. This thesis serves as the base for continuation of S-Transform to genomic analysis.
Combining advanced signal processing techniques with S-Transform may lead to the prediction of precise locations of protein coding regions of any given DNA sample.
BIBLIOGRAPHY


