CHAPTER 1

INTRODUCTION

With the recent development in the field of bioinformatics, signal processing methods have been applied for the analysis of the genomic datasets. The usage of signal processing tools for extracting the hidden information from the genomic datasets is called Genomics Signal Processing (GSP). Genomic data comprises of DNA sequences in the nucleus of the cell. DNA sequences contains four nucleotides bases adenine (‘A’), thymine (‘T’), guanine(‘G’) and cytosine. The DNA sequences in the eukaryotic cell (cell with nuclei) can be divided into intergenic and genic regions. The genic regions also have been divided in to exonic and intronic regions. The intronic regions do not have period-3 component. The exonic regions exhibit the period-3 behavior due to codon bias and these are associated with the protein coding regions. The protein coding regions are responsible for the formation of different proteins in the living organisms [Shakya et al. (2013a)]. Along with the period-3 component splice sites play an important role in identifying the protein coding regions accurately. The splice sites can be classified as acceptor splice sites (boundary of intron and exon) and donor splice sites (boundary of exon and intron).

Apart from the period-3 repeats, repeat patterns of other periods are also present in DNA sequences. The more than two contiguous repeated patterns of a particular period in the DNA sequences is called tandem repeats (TRs). The size, pattern structure, number of copies, location of the repeated patterns play an important role in the analysis of the tandem repeats. On the basis of the repeated pattern size the TRs can be classified as microsatellites, minisatellites, and satellites repeats. At present more than 10,000 STRs sequences have been published in human genome [Zhou et al. (2009)]. Tandem repeats are responsible for neuro-degenerative diseases such as Huntington’s disease, fragile X syndrome, and mytonic dystrophy, and cancer. STRs are used as simple tool for genetic profiling of individuals [Gupta et al. (2008)]. For the welfare of human beings or other living organisms the analysis of the genomic data is highly required as it would be helpful in medical field, forensic analysis and phylogenic studies [Hans et al. (2004)].
1.1 Genomics

Genomics is the study of an organism by extracting the hidden features in their genome. The complete set of DNA sequence of an organism is called genome. In general every human body has approximately 3 billion of DNA base pairs in a single cell. The DNA sequence contains the four nucleotides Adenine, Thymine, Guanine, and Cytosine. Information to build a human body is present in DNA sequences. The DNA sequences are bined together in chromosomes. Which are located in the nucleous of the cell. These cells are responsible for the formation of the specific proteins with the help of enzymes and messenger molecules. The enzymes copy the information about the genes from DNA in to the molecules called messenger ribonucleic acid (mRNA). The mRNA moves from nucleus to cell’s cytoplasm from where this mRNA is read by ribosomes. The ribosome helps to provide the link between mRNA and order of the amino acid for the formation of the specific protein. The protein helps in the formation of the body structures (organs, tissues), controls the chemical reaction and carries the signal between cells. If the DNA sequence has been mutated than an abnormal protein can be produced. These abnormal proteins can disturb the normal functioning of the human body and can result in a disease such as cancer. [NHGRI, (2011)]. Therefore, the study of the genomics is very much essential.

![Image](Image courtesy of U.S. Department of Energy Genome Programs)

**Fig.1.1:** The genome (inside the cell) contains all of an organism's genetic instructions.
1.2 Deoxyribose Nucleic Acid (DNA)

DNA molecules contain the digital information that constitutes the genetic blueprint used in the development and functioning of all known living organisms. It is found in the nucleus of the eukaryotic cell as double strand helix form and made up of the nucleotides. The basic unit of the nucleotides or bases comprises of one of the chemicals from (adenine (A), thymine (T), guanine (G) and cytosine (C)), deoxyribose sugar and phosphate. Its detailed Illustration is shown in Fig.1.2 (a) [Vaidyanathan et al. (2004) and Anastassiou et al. (2001)]. The single human cell contains three billion of the nucleotides approximately. The double helix two strand DNA can be separated into two parts like zipper by mechanical force or high temperature. Separated strands are complementary to each other in the sense that knowledge of one of them automatically reveals the other one. The nucleotides are covalently linked forming a long backbone of alternating sugar and phosphate as shown in Fig.1.2 (b). Each nucleotides of the one strand DNA have been associated with other strand DNA via a chemical bond (hydrogen bond). So that A is connected to T and vice versa, and C is connected to G and vice versa. The double strand helix DNA structure has been stabilized with hydrogen bond between nucleotides and base stacking relation among the aromatic nucleotide basis. Single hydrogen bond is weak, but together all these bonds create a stable and double helical structure just like a rope. Since only the bases differ between nucleotides, they are referred to by the names of their bases. The way in which the bases are linked together gives a chemical polarity to the DNA strand. This is indicated by referring to the phosphate end as the 5’ end and the sugar end as the 3’ end (left to right). The DNA sequences have been formed in a long one-dimensional string which is usually listed from the 5’end to the 3’ end because they are scanned in that direction when cell machinery uses bases as a signaling for the production of amino acids. This convention is based on the details of the chemical linkages (covalent bond) between the nucleotide subunits as shown in Fig.1.2 (c). The discovery of the DNA double helix is the most important achievement in the history of molecular biology. No other molecule has reached such an iconic status. An interesting fact of nature is that, while most natural signals such as heat, sound, and electromagnetic waves occur as continuous signals, the fundamental signal of life, the genetic information, occurs as a discrete signal. A good description of the discrete nature of the genetic information is given in [Vaidyanathan et al. (2004)]. A DNA sequence can contain various repetitive structures of the DNA. These repetitive structures (patterns) are associated with the particular periodicities. These
periodicities are separated into two types: periodicity due to codon bias (three base periodicity) and periodicity of the repeat motif of the particular size. The periodicity due to codon bias is only present in the genic region of the DNA. The genic region of the DNA contains the information for the production of proteins. In an organism all the cells have identical genes but only some of the genes are active in the particular family of the cells. Therefore, each gene is responsible for the formation of a different protein. For example active genes in nerve cells are different from genes which are active in the blood cells.

(a) Building blocks of DNA

(b) DNA single-strand

(c) DNA double-strand

Fig.1.2: Basic building blocks of the DNA structure.

1.3 DNA Sequencing

Oswald Theodore in 1994 demonstrated DNA (Deoxyribonucleic acid) as the genetic material. In 1953, James D.Watson and Francis Crick determined the double helix structure of the DNA along with four nucleotides bases that lead to the central dogma of molecular biology. The DNA
is the most essential unit of the species and individuals because it is responsible for structure, functionality of the cell and mysteries of life [Church et al.(1995)]. Therefore, DNA sequencing technologies have been introduced to help the biologist and medical peoples in the broad range of applications such as breeding, finding pathogenic genes, molecular cloning, and evolution studies. These technologies should be accurate, cheap, easy-to-operate, and fast. Since last thirty years DNA sequencing technologies find tremendous development and they act as an engine of the genome era which is characterized by large amount of genome data. In 1977, DNA sequencing technology was developed by Frederick Sanger which was based on chain termination technique and is also known as Sanger sequencing technology. Another sequencing technology developed by Walter Gilbert was based on chemical modification of DNA and subsequent cleavage at specific bases. Sanger sequencing technology has been utilized as the primary technology in the “first generation” of commercial and laboratory sequencing applications. After years of improvement first automatic sequencing machine, AB370 was introduced by Applied Biosystems in 1987. AB370 adopted capillary electrophoresis method which makes the sequencing more accurate and faster and was able to detect 500 bases in a day with read length of 600 bases and 96 bases at a time. The current model AB3730xl gives the output 2.88 M bases per day with read length touching 900 bases [Liu et al.(2012)]. The automatic sequencing instruments and associate software emerged in 1998 using the capillary sequencing machine and Sanger sequencing technology played an important role in the completion of the human genome project in 2001 [Collins et al. (2003)]. This project then motivated for the development of the fast, accurate, and low cost Next Generation Sequencing (NGS) instruments. The NGS technologies have high throughput, low cost and massive parallel analysis rather than Sanger method. Genomic data analysis is now a big data problem and the hidden information in the genomic sequences requires lot of data analysis with biological explanation.

1.4 Genome Annotation

The genome annotation is a process by which essential information to the raw DNA data, obtained after sequencing, is added to the genome data base. The genome annotation is of two types: structural and functional. In this work we have focused on structural genome annotation [Mark et al. (2012)]. The information about the features such as promoter regions, translational
start sites (TSS), untranslated regions (UTRs), exons, introns, acceptor splice sites and donar splice sites are important in gene structure and shown in Fig1.3.

![Genome annotation](http://carolguze.com/text/442-1-humangenome.shtml)

Following sections describe the various components of which the genome is comprised of.

(a) **Promoter regions**

A promoter is usually defined as a non-coding region of DNA that covers the transcription start site (TSS) or the 5’ end of the genes they regulate. Bulk of promoter region typically lies upstream of the TSS [Arniker et al. (2010)].

(b) **CpG Islands**

CpG islands (CGI) play an important role in the genomic analysis. The CGI contains high frequency of dinucleotides of CpG, where ‘p’ is referred as phosphodiester bond between two contagious nucleotides and is different from the hydrogen bond which is present between C and G within the two strands in a double helix DNA structure. The CGI length varies from 100 to 1000 base pairs and it rarely crosses the limit of 5000 base pairs. Most of the CpG islands (CGIs) are present around the promoter regions. The functionality of the genes is regulated by the promoter regions. Therefore, the prediction of the promoters in the genes is highly needed. Also, CGIs play significant role in discovering the epigenetic causes of cancer as well as it is located in
the promoter regions of certain tumor suppressor genes which are normally unmethylated in healthy cells [Kakumani et al. (2012)] as shown in Fig.1.4

![Fig.1.4: CpG islands in DNA](image)

(c) **Untranslated regions (UTR)**

The untranslated regions (UTRs) are the part of the start and stop exonic region which are not responsible for the production of proteins and are associated with the regulation of the particular genes. The UTRs and introns are closely related to the regulation of the particular gene. The UTRs and non-coding RNAs play an important role in the control of complex gene expressions [Barret et al. (2012)].

(d) **Exons**

An exon is a specific region of nucleic acid sequence that is referred to as coding regions of gene and it contains the information for protein synthesis. The organization of genes is fundamentally different in prokaryotes and eukaryotes. A prokaryotic gene occurs as an uninterrupted stretch of DNA that is transcribed into RNA without any further processing and can directly serve as an messenger RNA (mRNA). In contrast, a eukaryotic gene is separated into many fragments called exons. These fragments put together form the actual uninterrupted gene to form the final functional mRNA. The intermediate portions between the exons are called introns. Introns do not code for proteins and hence are referred to as non-coding regions. Usually exons are much shorter than introns and thus the coding portion of a gene is often only a small fraction of its total length. For example, in human genome the exonic fraction is as low as 2% [Lander et al. (2001)]. To form an uninterrupted gene all the introns are removed by the cellular mechanism and all the exons are joined together. By lacing genes with introns, a eukaryotic cell is able to
produce different proteins from a single gene by joining the exons in different combinations. This procedure is known as alternative splicing [Nilsen et al. (2010)] and represents a type of data compression developed through evolution for the purpose of manufacturing a wide variety of proteins from a small number of genes as illustrated in Fig.1.5 [Shakya et al. (2013a)]. It has been also observed that there is a prominent short-range correlation in the nucleotide arrangement in exonic regions which has been called $f = 1/3$ periodicity [Trifonov et al. (1980)]. This generic discriminative property that has caught the attention of many researchers is called the period-3 property. The periodicity-3 has been utilized by the DSP-methods to locate exonic regions in a DNA sequence. The period-3 property in a DNA sequence is closely related to the unbalanced distribution of the nucleotides in the three coding positions [Ficket et al. (1982), Tiwari et al. (1997), and Yin et al. (2005)]. In an exonic region, the distribution of the nucleotides at the three codon positions is unbalanced, while the distribution of nucleotides is uniform at the three codon positions in an intronic region. The unbalanced distribution of the nucleotides is present in exonic regions (protein coding regions) because the nucleotide usage in coding region is highly biased towards special amino acid compositions [Tiwari et al. (1997), Yin et al. (2005)] and Shakya et al. (2013a)].

Fig.1.5: Alternative splicing of exons and different proteins.
(e) **Introns**

The intron is a region inside the gene separated by the exonic regions. The introns do not code for the proteins and are called as non-coding regions of the gene. Generally, intron is a nucleotide sequence within a gene that is removed at the time of RNA splicing to produce the final functional mRNA. The sequences corresponding to the introns and exons must be identified very precisely during the removal of the introns because the incorrect detection of the boundaries between the exon and intron by even a single letter will mislead the results. The introns provide several important short sequences for efficient splicing such as donar sites and acceptor sites at either start or end of the intron as well as a branch point site, which are necessary for proper splicing [Shakya *et al.* (2013a)]. These details are shown in Fig.1.6.

![Fig.1.6: Details of eukaryotic DNA sequence.](image)

The frequency of introns varies widely across the spectrum of biological organism. The density and length of intronic genes varies across the species. For example, the human genome contains an average of 8.4 introns per gene and the mitochondrial genome of vertebrates does not contain any intronic genes. The intronic genes are present in eukaryotic genomes and absent in
prokaryotic genes [Shakya et al. (2013a)]. The intronic regions may contain other periodicities for the various repeating patterns except the periodicity due to non-uniform codon bias which is responsible for the formation of the proteins.

(f) Splice sites

Splice sites are the boundary of the exon and intron or intron and exon in the genic region of the DNA sequences. Splice sites can be classified as acceptor splice sites and donor splice sites. The acceptor splice site is the boundary between intron-exon. These sites contain the consensus pattern of the nucleotides “AG” just before starting of the exon. The donar splice site is presents at the boundary of exon-intron and is a consensus pattern of nucleotide “GT” present just before start of the intron. Splice sites are shown in Fig. 1.7. Splice sites prediction helps in the identification of the protein coding regions [Akhtar et al. (2006) and Akhtar et al. (2008)].

Fig. 1.7: Splice sites in DNA

1.5 Tandem Repeats in DNA

Computational analysis of the genomes is an important subject in contemporary research for understanding the biological functionality of organisms. The detection of DNA repeats by *ab initio* methods is of paramount importance, particularly evident in eukaryotes, which are involved in gene variations and regulatory functions on gene expressions [Treangen et al. (2012)]. The repeats in DNA sequences can be classified as tandem and dispersed [Sharma et al. (2004), Gupta et al. (2007), Brodzik et al. (2007), and Zhou et al. (2009)]. In tandem repeats (TRs), two or more contiguous copies of an arbitrary sequence of DNA symbols are present, whereas dispersed repeats consist of two or more nonadjacent copies of an arbitrary sequence. Repeats are shown in Fig. 1.8.
On the basis of the repeat motif (pattern) size, TRs can be further classified as satellites, minisatellites, and microsatellites (MSs) \cite{Zhou et al. (2009)} as shown in Fig. 1.9. The satellites range in length from 100 K base pairs (bps) to 1 Mbps with pattern size greater than 100 bps. The length of minisatellites varies between 1 to 20 Kbps and their pattern size range is 9–80 bps. Microsatellite repeats (MSs) also known as short tandem repeats (STRs) have a length less than 150 bps with a pattern size of 2–6 bps.
The tandem repeats can also be classified as perfect tandem repeats (PTRs) and approximate tandem repeats (ATRs). In PTRs, exact copies of the repeating patterns are present and inexact copies of repeating patterns are present in imperfect or approximate tandem repeats (ATRs) due to mutations. The PTRs and ATRs are shown in Fig. 1.10.

Among three types of TRs, MSs are important because of their documented functions and association with cancer and other diseases. MSs are responsible for the diseases such as Fragile-X syndrome, Huntington’s disease, Frederick’s ataxia, Spinocerebellar ataxia type 31 and 40 other neurological, neurodegenerative, and neuromuscular diseases [Mirkin et al. (2007), Polak et al. (2013), Usdin et al. (2008) and Yu et al. (2011)]. The gain or loss of repeating pattern of the short tandem repeats (microsatellites) in coding regions is often responsible to the change in the protein products. The gain or loss of repeating pattern of the short tandem repeats (microsatellites) in non-coding regions is identified to affect gene regulation. Mutations in microsatellite regions near to certain genes are found to be causing several genetic diseases and cancers. Microsatellites are more susceptible to mutations when compared to other regions of the genome and are often studied to understand evolution. Microsatellites are also useful in various fields such as forensics, DNA fingerprinting, population studies, linkage analysis, evolutionary studies, behavior of living organism, etc. [Hans et al. (2004)]. Example: Random changes that alter the length of microsatellite DNA near the gene for the vasopressin receptor.
affect social behavior in male voles. A longer microsatellite region resulted in more bonding and care giving. The behavior of the living organisms has been shown in Table 1.1 and Fig.1.11.

**Table 1.1:** Behavior of living organism due mutation in microsatellites

<table>
<thead>
<tr>
<th>Species</th>
<th>Microsatellite DNA</th>
<th>Vasopressin Receptor Gene</th>
<th>Social Behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prairie Voles</td>
<td>![Diagram]</td>
<td></td>
<td>![Diagram]</td>
</tr>
<tr>
<td>Montane Voles</td>
<td>![Diagram]</td>
<td></td>
<td>![Diagram]</td>
</tr>
<tr>
<td>Chimpanzees</td>
<td>![Diagram]</td>
<td></td>
<td>![Diagram]</td>
</tr>
<tr>
<td>Bonobos</td>
<td>![Diagram]</td>
<td></td>
<td>![Diagram]</td>
</tr>
<tr>
<td>Humans</td>
<td>![Diagram]</td>
<td></td>
<td>![Diagram]</td>
</tr>
</tbody>
</table>

**Fig.1.11:** Behavior of living organism due to mutation in microsatellites


1.6 Wet Lab Experiments and Computational Methods

After sequencing of the genome of a species, gene finding is one of the first and most important steps in understanding the molecular nature of genome. In the early days, gene finding
involved painstaking experiments on living cells and organism. The statistical analysis can be used to determine the order of genes on a certain chromosome by using their rate of homologous recombination. Information from several experiments is combined to create a specific map to identify the rough location of the known genes associated with each other. At present, with the availability of genome sequences of various species and access to powerful computational resources, the gene finding has largely become a computational problem [Akhtar et al. (2008a)].

Many computational gene finding programs such as AUGUSTUS [Stanke et al. (2004)], FGENES [Solovyev et al. (1995)], geneid [Parra et al. (2000)], GeneMark [Lukashin et al. (1998)], HMM [Haussler et al. (1996)], Genie [Reese et al. (2000)], GENSCAN [Burge et al. (1997)], HMMgene [Krogh et al. (1997)], Morgan [Salzberg et al. (1998)], and MZEF [Zhang et al. (1997)], have been introduced. Three types of approaches are used for gene finding, namely extrinsic, ab initio, and comparative [Johnson et al. (2008)]. In the extrinsic approach, reverse translation of genetic code is used to derive the family of possible coding DNA sequences. These candidate DNA sequences are then used to search a target for matches that are exact or inexact, and partial or complete. The basic local alignment search tool is widely used for this purpose [Johnson et al. (2008)]. Abinitio methods utilize certain features of protein coding genes to search them in DNA sequences. These features can be either statistical property, the biological signals related to the protein coding sequence or contents. Most widely used abinitio gene finding methods are GENSCAN and GENEid. In comparative gene findings, genome of related species are compared for length of coding regions, number of coding regions, position of the gene, sequence similarity, the amount of non-coding DNA in each genome, and other highly conserved regions.

1.7 Digital Signal Processing Methods for Genomics

Even with the presence of various data-dependent gene finding algorithms, the accuracy of gene prediction is still limited. One way to increase this accuracy is to combine different types of gene finding approaches (extrinsic, ab initio, and comparative) in one program, similar to AUGUSTUS+ [Stanke et al. (2004)]. Alternatively, higher accuracy can also be achieved by combining different gene finding programs [Murakami et al. (1998), and Pavlović et al. (2002)]. In both situations, the data dependence of the resultant system would increase [Akhtar et al. (2008a)]. Recently, numerical mapping of DNA sequences [Sharma et al. (2011), Kwan et al. (2009), and Cristea et al. (2002)] has opened up the possibility of applying signal processing
techniques to the analysis of genomic data. The extraction of the hidden information from genomic data using SP is called genomic signal processing [Dougherty et al. (2005) (a), and Dougherty et al. (2005) (b)]. The basic building block of GSP is given in Fig.1.12. The various reported signal processing methods have been reviewed in the next chapter.

To apply the digital signal processing method for genomic analysis, the nucleotides of the DNA sequences are to be converted into numerical sequences. The conversion from nucleotides to numeric sequence is called numerical mapping. After mapping we have to preprocess this numeric data to suppress the noise. Signal processing tools like spectral analysis, filtering and time frequency analysis are then used to identify the characteristics periodicities along with their temporal location [Tiwari et al. (1997), Marhann et al. (2011)]. A threshold is then used to extract the biological feature associated with the respective periodicity. In this work the flow graph shown in Fig.1.12 has been used as the basic model for various algorithms developed in this piece of research work.
1.8 Organization of the Thesis
The thesis consist of seven chapters. The brief description of each chapter is presented here.

Chapter 1: Introduction
This chapter provides an introduction to the work carried out. Emphasis is particularly on the basis of molecular biology. It is an application problem for which solutions have been devised using signal processing tools.

Chapter 2: Literature Review
To develop an understanding of molecular biology problems and the computational solutions to address them the existing literature was thoroughly studied. Existing gaps were identified and are discussed in this chapter.

Chapter 3: Detection of Short Exons in DNA by PCA of Multiple Optimized Spectrums
In this chapter a method for the short exons detection by combining optimized spectrums of multiple mappings using principal component analysis (PCA) is presented and its performance has been evaluated on the standard reported datasets along with a comparative study with other reported solutions.

Chapter 4: Digital Signal Processing Methods for Acceptor Splice Site Prediction
The algorithm developed in chapter 3 has been extended in this chapter for identifying the splices sites. The proposed method improves the detection performance of splice sites by combining it with statistical methods.

Chapter 5: Identification of Microsatellites in DNA
Identification of microsatellites in the DNA sequences using adaptive S-transforms has been done. This chapter describes the proposed algorithm and compares its performance with the existing algorithms.

Chapter 6: Identification of Minisatellites in DNA
Identification of minisatellites in the DNA sequences using Kaiser window based adaptive S-transform described in this chapter. The proposed algorithm gives better performance than other methods.

Chapter 7: Conclusion and Future work
Chapter 7 concludes the work and also provides the directions in which this work can be further continued. Some open research problems are described in this chapter.