OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Analysis of Genomic and Proteomic Sequence Using Fir Filter

P.Saranya[1], V.Harigopalkrishna[2], D.Murali[3], M.Ravikumar[4], M.Sujatha[5]
[1,2,3,4,5]*ECE, LIET, INDIA*

ABSTRACT: *Bioinformatics is a field of science that implies the use of techniques from mathematics, informatics, statistics, computer science, artificial intelligence, chemistry, and biochemistry to solve biological problems usually on the molecular level. Digital Signal Processing (DSP) applications in genomic sequence analysis have received great attention in recent years.DSP principles are used to analyse genomic and proteomic sequences. The DNA sequence is mapped into digital signals in the form of binary indicator sequences. Signal processing techniques such as digital filtering is applied to genomic sequences to identify protein coding region. Frequency response of genomic sequences is used to solve many optimization problems in science, medicine and many other applications. The aim of this paper is to describe a method of generating Finite Impulse Response (FIR) of the genomic sequence. The same DNA sequence is used to convert into proteomic sequence using transcription and translation, and also digital filtering technique such as FIR filter applied to know the frequency response. The frequency response is same for both gene and proteomic sequence.*

Keywords: *DSP, FIR, Frequency response, Genomic sequence, Proteomic sequences*

I. INTRODUCTION

This paper represents a method for generating the response of genomic sequence using fir filter. Genomics is a highly cross-disciplinary field that creates paradigm shifts in such diverse areas as medicine and agriculture. It is believed that many significant scientific and technological endeavours in the 21st century will be related to the processing and Interpretation of the vast information that is currently revealed from sequencing the genomes of many living organisms, including humans. Genomic information is digital in a very real sense; it is represented in the form of sequences ofwhich each element can be one out of a finite number of entities. Such sequences, likeDNA and proteins, have been mathematically represented by character strings, in which each character is a letter of an alphabet. In the case of DNA, the alphabet is size4 and consists of the letters A, T, C and G; in the case of proteins, the size of the corresponding alphabet is 20.Biomolecular sequence analysis has already been a major research topic among computer scientists, physicists, and mathematicians.

The main reason that the field of signal processing does not yet have significant impact in the field is because it deals with numerical sequences rather than character strings. The possibility of finding a wide application of DSP techniques to the analysis of genomic sequences arises when these are converted appropriately into numerical sequences, for which several rules have been developed. Notice that genomic signals do not have time or space as the independent variable, as occur with most physical signals. However, if we properly map character string into one or more numerical sequences, then digital signal processing (DSP) provides a set of novel and useful tools for solving highly relevant problems. Even the process of mapping DNA into proteins and the interdependence of the two kinds of Sequences can be analyzed using simulations based on digital filtering. The principles of DSP techniques are used to analyze both DNA and protein sequences.

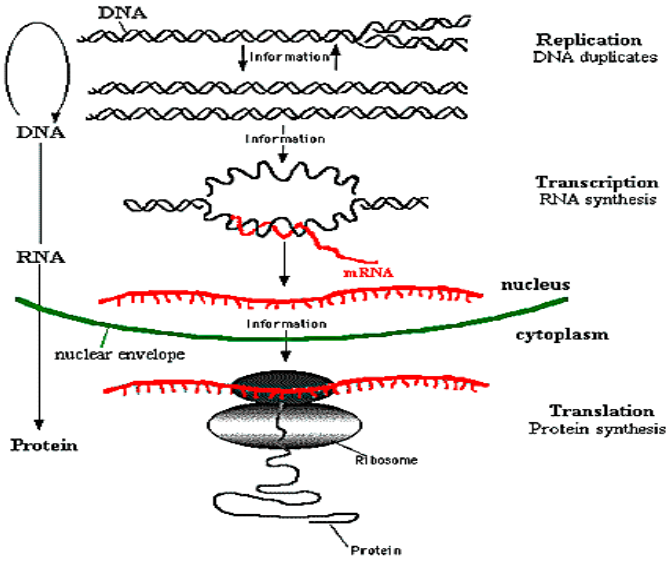
DSP techniques such as FIR filter are applied to find the frequency response. This response is used to identify protein coding region. When the DNA sequence is converted into protein sequence with the applications of DSP techniques, the structure of DNA and PROTEIN sequence will be approximately same. The frequency response of DNA and converted protein sequence can be obtained with the help of DSP algorithms. Among the different DSP algorithms digital filters are one of technique used for the analysis of DNA and PROTEIN sequences. With the use of FIR filter the response of these sequences can be obtained. Using the Kaiser window the linear phase application will be stable. These DSP-based approaches result in alternative mathematical formulations and may provide improved computational techniques for the solution of useful problems in genomic information science and technology. As DNA sequences are character strings, for DSP techniques to be applicable to these data, methods converting DNA sequences to numerical sequences are required. the generated numerical sequences are then processed using DSP techniques.

II. CONCEPTS OF MOLECULAR BIOLOGY

2.1 Central Dogma Of Molecular Biology

The Central dogma of molecular biology is that DNA codes for RNA and RNA codes for proteins. Thus the production of a protein is a two-stage process, with RNA playing a key role in both stages.

Transcription is the process by which the information contained in a section of DNA is transferred to a newly assembled piece of messenger RNA(mRNA). It is facilitated by RNA polymerase and transcription factors. In eukaryotic cells the primary transcript pre-mRNA must be processed further in order to ensure translation. This normally includes a 5' cap a poly-A tail and splicing. Alternative splicing can also occur, which contributes to the diversity of proteins any single mRNA can produce.



The Central Dogma of Molecular Biology

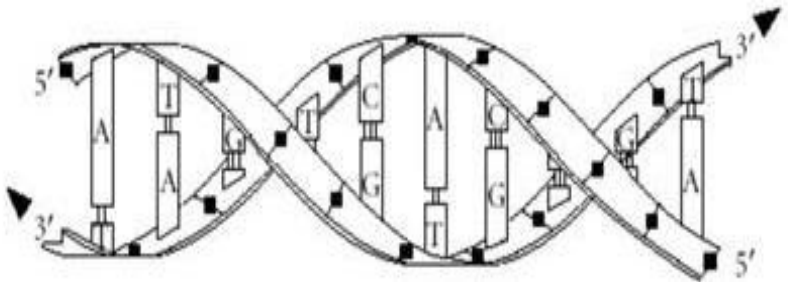
Eventually, this mature mRNA finds its way to a ribosome, where it is translated. In prokaryotic cells, which have no nuclear compartment, the process of transcription and translation may be linked together. In eukaryotic cells, the site of transcription the cell nucleus is usually separated from in the site of translation the cytoplasm), so the mRNA must be transported out of the nucleus into the cytoplasm, where it can be bound by ribosome. The mRNA is read by the ribosome as triplet codon, usually beginning with an AUG (adenine-Uracil−guanine), or initiator methionine codon downstream of the ribosome binding site. As the amino acids are linked into the growing peptide chain, they begin folding into the correct conformation. Translation ends with a UAA, UGA, or UAG stop codon. The nascent polypeptide chain is then released from the ribosome as a mature protein**.**

2.2 Dna

A single strand of DNA is a bio molecule consisting of many linked, smaller components called nucleotides. Each nucleotide is one of four possible types designated by the letters A, T, C, and G and has two distinct ends, the 5′ end and the 3′end, so that the 5′ end of a nucleotide is linked to the 3′ end of another nucleotide by a strong chemical bond (covalent bond), thus forming a long, one-dimensional chain (backbone) of a specific directionality. Therefore, each DNA single strand is mathematically represented by a character string, which, by convention specifies the 5′ to 3′ direction when read from left to right. Single DNA strands tend to form double helices with other single DNA strands. Thus, a DNA double strand contains two single strands called complementary to each other because each nucleotide of one strand is linked to a nucleotide of the other strand by a chemical bond (hydrogen bond), so that A is linked to T and vice versa, and C is linked to G and vice versa. Each such bond is weak (contrary to the bonds forming the backbone), but together all these bonds create a stable, double helical structure. The two strands run in opposite directions, as shown in Figure, in which we see the sugar-phosphate chemical structure of the DNA backbone created by strong (covalent) bonds, and that each nucleotide is characterized by a base that is attached to it. The two strands are linked by a set of weak (hydrogen) bonds. The bottom left diagram is a simplified, straightened out depiction of the two linked strands.
For example, the part of the DNA double strand shown in Figure is
5′ - C-A-T-T-G-C-C-A-G-T - 3′
3′ - G-T-A-A-C-G-G-T-C-A - 5′

Because each of the strands of a DNA double strand uniquely determines the other strand, a double-stranded DNA molecule is represented by either of the two character strings read in its 5′ to 3′ direction. Thus, in the example above, the character strings CATTGCCAGT and ACTGGCAATG can be alternatively used to describe the same DNA double strand, but they specify two different single strands which are complementary to each other. DNA strands that are complementary to themselves are called self-complementary, or palindromes. For example AATCTAGATT is a palindrome. DNA molecules store the digital information that constitutes the genetic blueprint of living organisms. This digital information has been created and reliably stored throughout billions of years of evolution during which some vital regions of DNA sequences have been remarkably preserved, despite striking differences in the body plans of various animals.

A DNA sequence can be separated into two types of regions: genes and intergenic spaces. Genes contain the information for generation of proteins. Each gene is responsible for the production of a different protein. Even though all the cells in an organism have identical genes, only a selected subset is active in any particular family of cells.



2.3Rna

Ribonucleic acid (RNA) is a chemical similar to a single strand of DNA. Together with DNA, RNA comprises the nucleic acids, which, along with proteins, constitute the three major macromolecules essential for all known forms of life. Like DNA, RNA is assembled as a chain of nucleotides, but is usually single-stranded. Cellular organisms use messenger RNA (mRNA) to convey genetic information often notated using the letters G, A, U, and C for the nucleotides guanine, adenine, Uracil and cytosine that directs synthesis of specific proteins, while many viruses encode their genetic information using an RNA genome.

Some RNA molecules play an active role within cells by catalyzing biological reactions, controlling gene expression, or sensing and communicating responses to cellular signals. One of these active processes is protein synthesis, a universal function whereby mRNA molecules direct the assembly of proteins on ribosome. This process uses transfer RNA (tRNA) molecules to deliver amino acids to the ribosome, where ribosomal RNA (rRNA) links amino acids together to form proteins

2.4 Protiens

Proteins are large biological molecules, or macromolecules, consisting of one or more chains of amino acid residues. Proteins perform a vast array of functions within living organisms, including catalyzing metabolic reactions, replicating DNA, responding to stimuli, and transporting molecules from one location to another. Proteins differ from one another primarily in their sequence of amino acids, which is dictated by the nucleotide sequence of their genes, and which usually results in folding of the protein into a specific three-dimensional structure that determines its activity. Like other biological macromolecules such as polysaccharides and nucleic acids, proteins are essential parts of organisms and participate in virtually every process within cells. Many proteins are enzymes that catalyze biochemical reactions and are vital to metabolism. Proteins also have structural or mechanical functions, such as actin and myosin in muscle and the proteins in the cytoskeleton, which form a system of scaffolding that maintains cell shape. Other proteins are important in cell signalling, immune responses, cell adhesion, and the cell cycle. Proteins are also necessary in animals' diets, since animals cannot synthesize all the amino acids they need and must obtain essential amino acids from food. Through the process of digestion, animals break down ingested protein into free amino acids that are then used in metabolism.

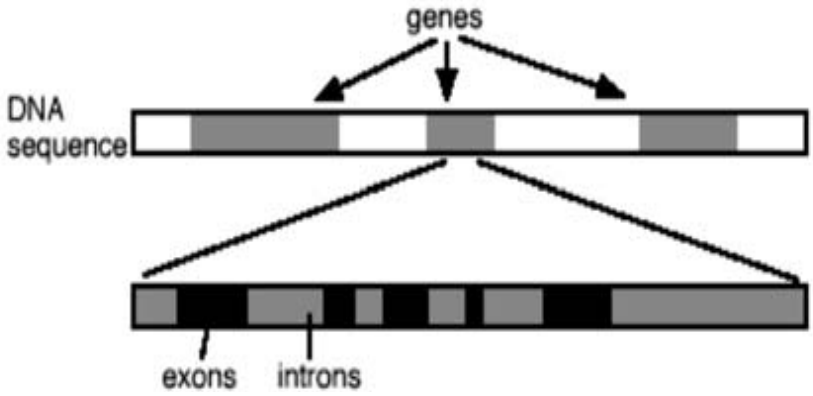
2.5 GENETIC CODE

Protein synthesis is governed by the genetic code which maps each of the 64 possible triplets (codons) of DNA characters into one of the 20 possible amino acids. Fig. shows the genetic code in which the 20 amino acids are designated by both their one-letter and three-letter symbols. A particular triplet, ATG, serves as the START codon and it also codes for the M amino acid (methionine); thus, methionine appears as the first amino acid of proteins, but it may also appear in other locations. We also see that there are three STOP codons indicating termination of amino acid chain synthesis, and the last amino acid is the one generated by the codon

preceding the STOP codon. Coding of nucleotide triplets into amino acids can happen in either the forward or the reverse direction based on the complementary DNA strand. Therefore, there are six possible reading frames for protein coding DNA regions.

FIRST POSITION	SECOND POSITION OF CODON: T	C	A	G	THIRD POSITION
T	TTT Phe (F)	TCT Ser (S)	TAT Tyr (Y)	TGT Cys (C)	T
	TTC Phe (F)	TCC Ser (S)	TAC Tyr (Y)	TGC Cys (C)	C
	TTA Leu (L)	TCA Ser (S)	**TAA (STOP)**	**TGA (STOP)**	A
	TTG Leu (L)	TCG Ser (S)	**TAG (STOP)**	TGG Trp (W)	G
C	CTT Leu (L)	CCT Pro (P)	CCT Pro (P)	CGT Arg (R)	T
	CTC Leu (L)	CCC Pro (P)	CCC Pro (P)	CGC Arg (R)	C
	CTA Leu (L)	CCA Pro (P)	CCA Pro (P)	CGA Arg (R)	A
	CTG Leu (L)	CCG Pro (P)	CCG Pro (P)	CGG Arg (R)	G
A	ATT Ile (I)	ACT Thr (T)	AAT Asn (N)	AGT Ser (S)	T
	ATC Ile (I)	ACC Thr (T)	AAC Asn (N)	AGC Ser (S)	C
	ATA Ile (I)	ACA Thr (T)	AAA Lys (K)	AGA Arg (R)	A
	ATG Met (M) **(START)**	ACG Thr (T)	AAG Lys (K)	AGG Arg (R)	G
G	GTT Val (V)	GCT Ala (A)	GAT Asp (D)	GGT Gly (G)	T
	GTC Val (V)	GCC Ala (A)	GAC Asp (D)	GGC Gly (G)	C
	GTA Val (V)	GCA Ala (A)	GAA Glu (E)	GGA Gly (G)	A
	GTG Val (V)	GCG Ala (A)	GAG Glu (E)	GGG Gly (G)	G

The total number of nucleotides in the protein coding area of a gene will be a multiple of three, that the area will be bounded by a START codon and a STOP codon, and that there will be no other STOP codon in the coding reading frame in between. However, given a long nucleotide sequence, it is very difficult to accurately designate where the genes are. Accurate gene prediction becomes further complicated by the fact that, in advanced organisms, protein coding regions in DNA are typically separated into several isolated sub regions called exons. The regions between two successive exons are called introns. When DNA is copied into mRNA during transcription, the introns are eliminated by a process called splicing. The same gene can code for different proteins. This happens by joining the exons of a gene in different ways. This is called alternative splicing. Alternative splicing seems to be one of the main purposes for which the genes in eukaryotes are split into exons. The mRNA obtained after splicing is uninterrupted and is used for making proteins.



III. SIGNAL PROCESSING FOR DNA SEQUENCES

3.1 Collection of Input Data

Most of the identified genomic data is publicly available over the Web at various places worldwide, one of which is the entrez search and retrieval system of the National Center for Biotechnology Information NCBI) at the National Institutes of Health (NIH). The NIH nucleotide sequence database is called GenBank and contains all publicly available DNA sequences. For example, one can go to http://www.ncbi.nlm.nih.gov/entrez and identify the DNA sequence with Accession Number AF324494 and AF320294; choose Nucleotide under Search and then fill out the other entry by typing: [Accession Number] and pressing "Go." Clicking on the resulting accession number will show the annotation for the genes as well as the whole nucleotide sequence in the form of raw data. Similarly, *Entrez* provides access to databases of protein sequences as well as 3-D macromolecular structures, among other options. As another example, a specialized repository for the processing

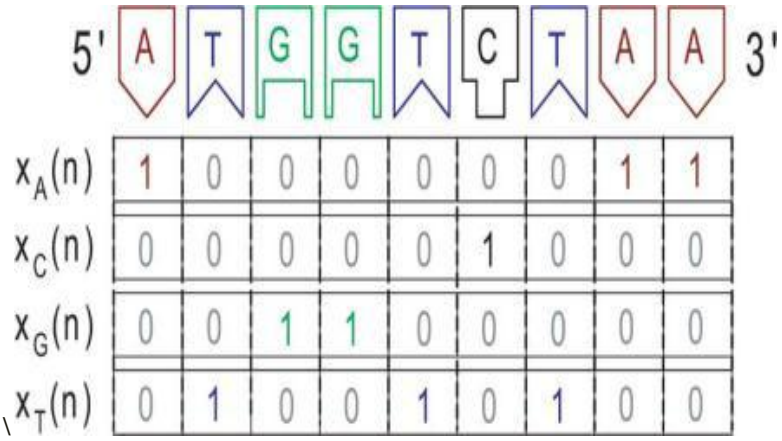
and distribution of 3-D, macromolecular structures can be found in Public databases. The collected nucleotide sequences from such databases are such as AF324494 and AF320294.

3.2 Mapping of DNA Strand into Digital Signals

As explained Genomic information is in the form of alphabets A, T, C and G. Signal processing deals with numerical sequences. Hence character strings have to be mapped into one or more numerical sequences. Then signal processing techniques can be applied for analysis of DNA sequences.

In a DNA sequence we have to assign numbers to the characters *A*, *T*, *C*, *G*, respectively. A proper choice of the numbers *can* provide potentially useful properties to the numerical sequence. The first approach to convert genomic information in numerical sequences was given by Voss with the definition

Of the indicator sequences, defined as binary sequences for each base, where 1 at position *k* indicates the presence of the base at that position, and 0 its absence.



For example, given the DNA sequence

ACTTAGCTACAGA…

The binary indicator sequences *X* for each base A, T, C and G are respectively:

$X_A[K] = 1000100010101\ldots$

$X_T[K] = 0011000100000\ldots$

$X_C[K] = 0100001001000\ldots$

$X_G[K] = 0000010000010\ldots$

The main advantages of the indicator sequences are theirsimplicity, and the fact that they can provide a four dimensional representation of the frequency spectrum of a character string, by means of computing the DFT of each one of the indicator sequences. The binary indicator sequences of both DNA and PROTEIN structure can be obtained. This can be applied to the digital filters in order to obtain the frequency response of sequences

3.3 Digital Filters

Digital filters that incorporate *digital-signal-processing* (DSP) techniques have received a great deal of attention in technical literature in recent years.

A digital filter is a discrete system capable of realizing some transformation to an input discrete numerical sequence. There are different classes of digital filters namely linear, nonlinear, time-invariant or adaptive digital filters. Digital filters are characterized by numerical algorithms that can be implemented in any class of digital processors The system transfer function relates the input and output sequences $x[n]$ and $y[n]$, through their respective *Z* transforms $X[z]$ and $Y[z]$. A variety of digital filter design techniques allow to obtain any desired magnitude response with frequency selectivity properties. Digital Filters can be very complicated devices, but they must be able to map to the difference equations of the filter design. This means that since difference equations only have a limited number of operations available (addition and multiplication), digital filters only have limited operations that they need to handle as well. There are only a handful of basic component**s** to a digital filter**,** although these few components can be arranged in complex ways to make complicated filters. A digital filter computes a quantized time-domain representation of the convolution of the sampled input time function and a representation of the weighting function of the digital filter. They are realized by an extended sequence of multiplications and additions carried out at a uniform spaced sample interval. . In particular, LTI digital filters can pertain to one of two categories, according to the duration of their response to the impulse, or Dirac delta function, when it is used as the input signal: infinite (IIR) or finite (FIR) impulse

response. A variety of digital filter design techniques allow to obtain any desired magnitude response with frequency selectivity properties, whereas it is desired that the phase response be a linear function of □, in order to have low distortion .

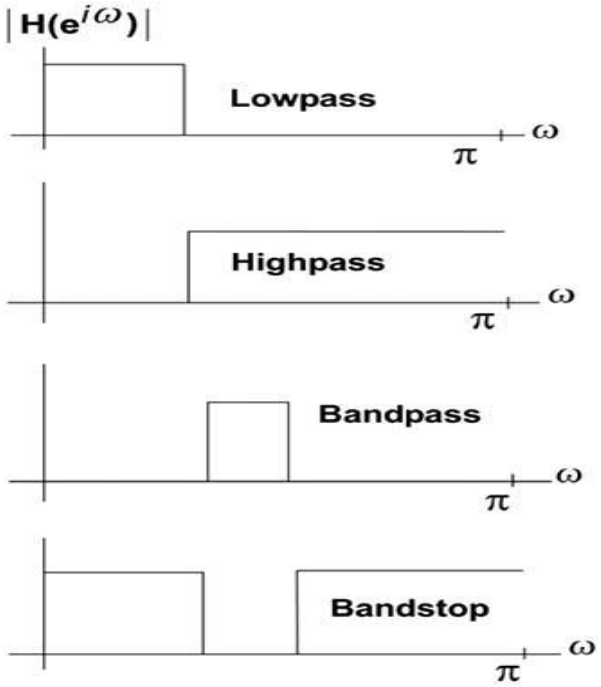
3.3.1 Fir Filter

In signal processing, a finite impulse response (FIR) filter is a filter whose impulse response (or response to any finite length input) is of *finite* duration, because it settles to zero in finite time. The impulse response of an Nth-order discrete-time FIR filter lasts for *N* + 1 samples, and then settles to zero. An FIR filter is based on a feed-forward difference equation

FIR digital filters are characterized by a discrete convolution operation of the form

$$y[n] = \sum_{m=0}^{N-1} h[m]x[n-m]$$

In this equation, $h[m]$ is the impulse response of the filter, which has a length of N samples. On the other hand, a property of FIR digital filters is that they can exhibit a perfect linear phase response under certain conditions of symmetry in their impulse response. This has been a motivation for the use of digital FIR filters in many applications. According to the frequency interval (band) transmitted, the magnitude of the basic ideal prototype filter frequency responses, can be *low pass*, *high pass*, *band pass* and *band stop.*



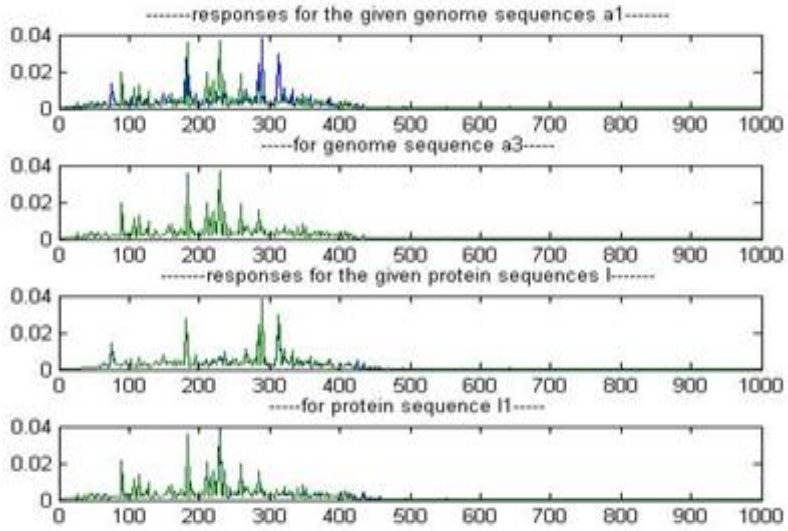
They can easily be designed to be linear phase by making the coefficient sequence symmetric. This property is sometimes desired for phase-sensitive applications the response of DNA sequence and protein sequence are generated using the FIR filter.FIR has different Types of implementations but this paper is implemented using the Kaiser Windowing technique.

3.3.2 Kaiser Window

The Kaiser window is an approximation to the prolate-spheroidal window, for which the ratio of main lobe energy to the side lobe energy is maximised. For a Kaiser window of particular length, the parameter β controls the side lobe height. For a given β, the side lobe height is fixed with respect to window length. As β increases the side lobe height decreases and the main lobe width increases. Kaiseord returns filter order n and beta parameter to specify a Kaiser window for use with the fir1 function. Given a set of specifications in the frequency domain, Kaiseord estimates the minimum FIR filter order that will exactly meet the specifications. Kaiseord converts the given filter specifications into pass band and stop band ripples and converts cut-off frequencies into the form needed for windowed FIR filter design.kaiser window method is constrained to produce filters with minimum deviation in all of the bands. The Kaiser window is a kind of adjustable window function which provides independent control of the main lobe width and ripple ratio. But the Kaiser window has the disadvantage of higher computational complexity due to the use of Bessel functions.

Results

The frequency response for the genomic and proteomic sequences are generated using fir filter are shown below. The peaks represent the exons which help in predicting the cancer cells. This can be useful for many genetic applications using different dsp techniques.



IV. CONCLUSION

The application of Digital Signal Processing in Genomic Sequence Analysis has received great attention in the last few years, providing a new insight in the solution of various problems. The responses for the DNA and genomic sequences when compared with the ideal characteristics of those sequences, provides the information regarding sequences. This help in developing new diagnostic tools, therapeutic procedures and pharmacological drugs for applications like cancer classification and prediction**.**

REFERENCES

[1]. D. Anastassiou, ―Genomic signal processing,‖ *IEEE Sign Proc Mag*, vol. 18, no. 4, pp. 8-20, 2001.
[2]. D. Anastassiou, ―DSP in genomics processing and frequency-domain analysis of character strings,‖ in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 1053-1056.
[3]. J. V. Lorenzo-Ginori, A. Rodriguez-Fuentes, R. G. Abalo, and R. S. Rodriguez, ―Digital signal processing in the analysis of genomic sequences,‖ *Current Bioinformatics*, vol. 4, pp. 28 – 40, 2009.
[4]. D. L. Brutlag, ―Understanding the human genome,‖ in *Scientific American: Introduction to Molecular Medicine*, P. Leder, D. A. Clayton, and E. Rubenstein, Eds., New York NY: Scientific American Inc. 1994, pp. 153-168,.
[5]. A. Khare, A. Nigam, and M. Saxena, ―Identification of DNA sequences by signal processing tools in protein-coding regions,‖ *Search & Research*, vol. 2, no. 2, pp. 44-49, 2011.
[6]. J. Tuqan and A. Rushdi, ―A DSP approach for finding the codon bias in DNA sequences,‖ *IEEE J Select Topics Sign Proc*, vol. 2, pp. 343- 356, 2008.
[7]. R. K. Deergha and M. N. S. Swamy, ―Analysis of genomics and proteomics using DSP techniques,‖ *IEEE Transactions on Circuits and systems*—I: Regular papers, vol. 55, no. 1, pp. 370-379, 2008.
[8]. E. N. Trifonov, ―3-, 10.5-, 200- and 400-base periodicities in genome sequences,‖ *Physica A* , vol. 249, pp. 511-516, 1998.
[9]. D. Kotlar and Y. Lavner, ―Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions,‖ *Genome Res* , vol. 13, pp. 1930-1937, 2003.
[10]. T. W. Fox and A. Carreira, ―A digital signal processing method for gene prediction with improved noise suppression,‖ *EURASIP J Appl Sign Proc*, vol. 1, pp. 108-111, 2004.
[11]. S. Datta and A. Asif, ―DFT based DNA splicing algorithms for prediction of protein coding regions,‖ in *Proc. IEEE Conference Record of 38th Asilomar Conference on Signals, Systems and Computer*, 2004, vol. 1, pp. 45-49.
[12]. P. P. Vaidyanathan and B. J. Yoon, ―The role of signal-processing concepts in genomics and proteomics,‖ *J Franklin Inst* , vol. 341, pp. 111-35, 2004.
[13]. J. Tuqan and A. Rushdi, ―A DSP perspective to the period-3 detection problem,‖ in *Proc. IEEE International Workshop on Genomic Signal Processing and Statistics*, 2006, pp. 53-54.
[14]. A. Rushdi and J. Tuqan, ―Trigonometric transforms for finding repeats in DNA sequences,‖ in *Proc. IEEE International Workshop on Genomic Signal Processing and Statistics*, 2008, pp. 1-4.
[15]. S. S. Sahu and G. Panda, ―A DSP approach for protein coding region identification in DNA sequence,‖ *International Journal of Signal and Image Processing*, vol. 1, no. 2, pp. 75-79, 2010.