

GENE PREDICTION USING DIGITAL FILTER

Gaurav Sapra & Preetika S. Sapra

U.I.E.T, P.U, Chandigarh

gaurav_sapra2001@yahoo.com

Abstract: Gene prediction is a problem of interpreting nucleotide sequences by computer, in order to provide information on protein coding genes. The digital filters have been considered here for this purpose. The DNA sequence is mapped into digital signals in the form of binary indicator sequences. Digital filtering operations are performed for each of the four indicator sequences.

1. INTRODUCTION

Deoxyribonucleic acid (DNA) is a molecule that contains the genetic code of all living things. DNA molecule consists of two strands that wind around one another to form a shape known as double helix. Each strand has a backbone made of alternating sugar and phosphate group. These two strands run in opposite direction to each other and are therefore antiparallel. Attached to each sugar is one of four types of molecule is called bases. The information in DNA is stored as a code made up of four chemical bases: adenine(A), cytosine(C), guanine(G), and thymine(T). In double stranded DNA, base A always pair with T because they make two hydrogen bonds and cytosine(C) and guanine(G) pair upto make three hydrogen bonds. Although the bases are always in fixed pairs, the pairs can come in any order.[1]
DNA sequence can be divided into genes and intergenic spaces as shown in Figure 1.1.

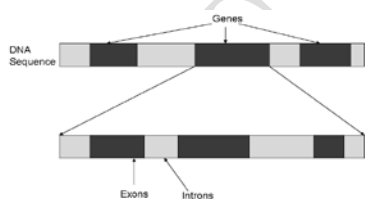


Figure 1.1. DNA sequence

Gene contains the information for generation of proteins. A gene is sequence made up from the four bases and can be divided into two subregions called exons and introns and only exons are involved in protein coding. The bases in the exon region can be imagined to be divided into group of three adjacent bases. Each triplet is called codon. There are 64 possible codons. Scanning the gene from left to right; a codon sequence can be defined by concatenation of the codons in all the exons. The codon sequence therefore uniquely identifies an amino acid sequence, which defines sequence,

which defines a protein. This mapping is called genetic code. Table 1.1 shows the genetic code in which 20 amino acids are designated by both one letter and three letter symbols.[2]

Table 1.1: List of amino acids and codons

1	A	Ala	Alanine	GCA, GCC, GCG, GCT
2	C	Cys	Cysteine (has S)	TGC, TGT
3	D	Asp	Aspartic acid	GAC, GAT
4	E	Glu	Glutamic acid	GAA, GAG
5	F	Phe	Phenylalanine ¹	TTC, TTT
6	G	Gly	Glycine	GGA, GGC, GGG, GGT
7	H	His	Histidine ²	CAC, CAT
8	I	Ile	Isoleucine ³	ATA, ATC, ATT
9	K	Lys	Lysine ⁴	AAA, AAG
10	L	Leu	Leucine ⁵	TTA, TTG, CTA, CTC, CTG, CTT
11	M	Met	Methionine ⁶ (has S)	ATG
12	N	Asn	Asparagine	AAC, AAT
13	P	Pro	Proline	CCA, CCC, CCG, CCT
14	Q	Gln	Glutamine	CAA, CAG
15	R	Arg	Arginine ⁷	AGA, AGG, CGA, CGC, CGG, CGT
16	S	Ser	Serine	AGC, AGT, TCA, TCC, TCG, TCT
17	T	Thr	Threonine ⁸	ACA, ACC, ACG, ACT
18	V	Val	Valine ⁹	GTA, GTC, GTG, GTT
19	W	Trp	Tryptophan ¹⁰	TGG
20	Y	Tyr	Tyrosine ¹¹	TAC, TAT

The protein coding regions of DNA sequences exhibit period 3 behaviour due to codon structure. This period 3 property can be exploited for locating exons. This paper deals with the prediction of genes using digital filters, as they can be very effective in extracting this period 3 information.[3]

2. METHODOLOGY

The methodology of gene prediction using digital filter has been divided into following steps.

2.1 Collection of Input Data

Most of the identified genomic data is publicly available over the web at various places worldwide. The National Institute Of Health (NIH) nucleotide sequence database is called

GenBank and contains all publicly available DNA sequences. The input data used in this paper is taken from gene F56F11.4 in C-elegans chromosome III having Accession Number: AF099922 from base number 7020-15080.[4]

2.2 Mapping of DNA Strand into Digital Signals

A single DNA strand is represented as a sequence of four bases, namely A, C, T and G. The method of mapping DNA sequence to a set of digital signals used in this dissertation consists of forming four binary indicator sequences $x_A(n)$, $x_C(n)$, $x_T(n)$ and $x_G(n)$ where a 1 would indicate the presence of a base and 0 indicates its absence. For example, the signal $x_A(n)$ in DNA segment GCCAATGCTGAA is 000110000011. The signals $x_G(n)$, $x_C(n)$ and $x_T(n)$ can also be obtained in a similar manner. This has been achieved with the help of C-programs written separately for each indicator sequence.

2.3 Designing of Filters

There are different ways of designing filters using the Filter Design and Analysis Tool in MATLAB. FDATool helps in quickly design digital FIR or IIR filters by setting filter performance specifications, by importing filters from your MATLAB workspace. Several response types are available as options namely Lowpass, Highpass, Bandpass, Bandstop, Differentiator, Hilbert transformer are chosen from the response type region in Digital filter panel as shown in Figure 2.1. The filter design specifications vary according to response type and design method. The filter designed is then computed using default filter design method, type, order, along with certain frequency or time domain specifications such as pass band frequencies and stop band frequencies.

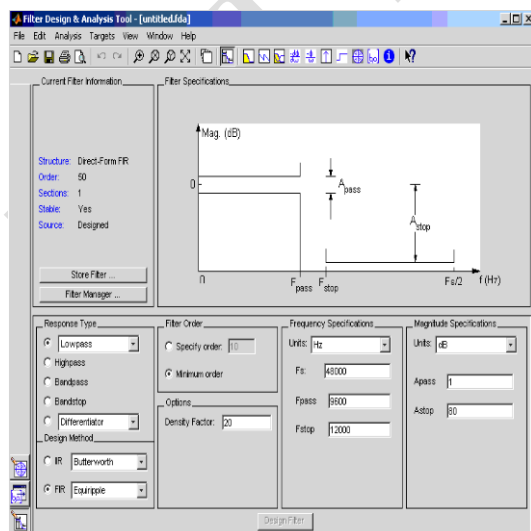


Figure 2.1 Design Filter Panel

Once the filter has been designed, indicator

sequences are passed through it. Four digital filtering operations are performed for each of the four indicator sequences. The indicator sequences are individually processed using the same digital filter to produce the signals Y_a , Y_c , Y_g and Y_t . A measure of the spectral content of a DNA sequence is given by (1).

$$Y = Y_a^2 + Y_c^2 + Y_g^2 + Y_t^2 \quad (1)$$

The signal Y produces large values in coding regions that exhibit strong period-three behaviour and is therefore an indicator of coding region.[5]

3. RESULTS

Results are in the form of plots giving measure of the spectral content of a DNA sequence. Different cases are considered one by one.

3.1 FIR Filters

Firstly, considering the case of FIR filters. For the FIR filter design, the following approaches are considered for passband edges chosen as $wc1 = 0.6665$ and $wc2 = 0.6667$.

a) Windowed Fourier series approach.

Figure 3.1 shows the exon prediction plots by Kaiser window.

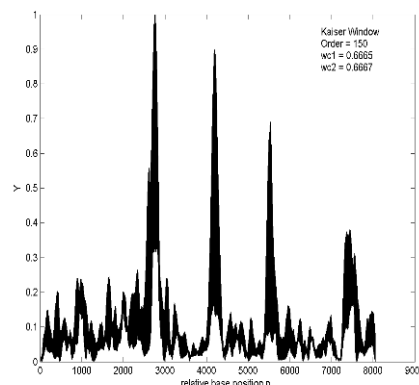


Figure 3.1 Kaiser Window

The peaks corresponding to the exons can be clearly seen but the presence of background “noise” due to $1/f$ characteristics in DNA sequence can also be noticed. Another window function, hamming, is also considered here as shown in Figure 3.2. The effect of noise is more visible here as compared to Kaiser window function even though order of filter is higher in this case.

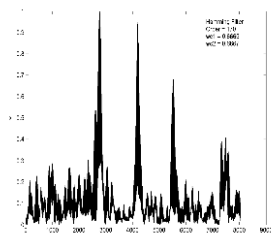


Figure 3.2 Hamming Window

b) Least Square and Constraint Least Square Algorithm.

The plots of Figure 3.3 and 3.4 shows the performance of digital filters based on Least-squares algorithm and Constrained Least-square algorithm respectively.

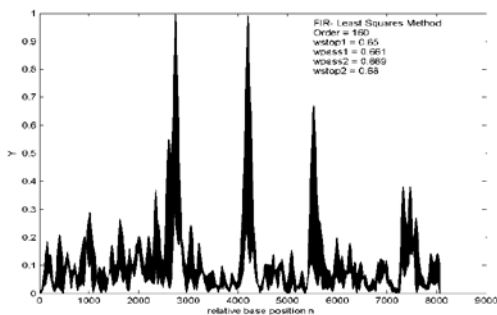


Figure 3.3 Least Square Algorithm

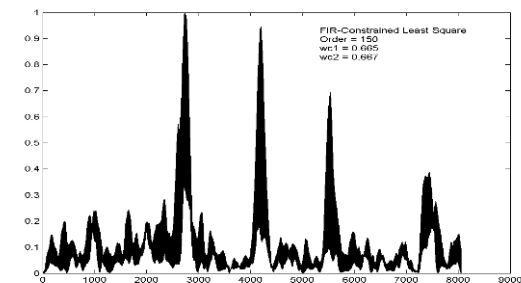


Figure 3.4 Constrained Least-Square

c) Equiripple Filters

The Figure 3.5 shows the exon prediction plot for Equiripple filters .It has been found that Equiripple filter is giving the comparable performance for a very high value of filter order.

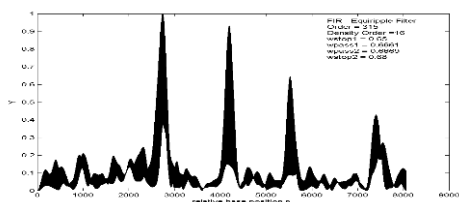


Figure 3.5 Equiripple Filter

3.2 IIR Filters

For the IIR filter design, the following approaches are considered for passband edges chosen as $wc1 = 0.6665$ and $wc2 = 0.6667$.

a) Butterworth filter

As it is clear from Figure 3.6, 3.7, and 3.8, butterworth filters upto order 14 with above mentioned specifications are clearly locating the exons.

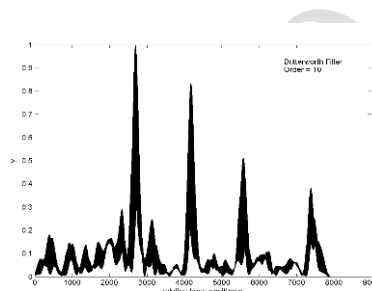


Figure 3.6 Order 10

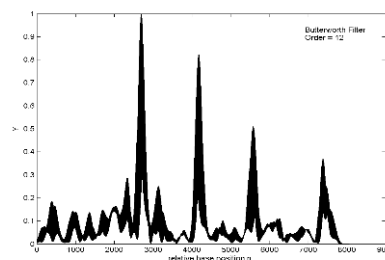


Figure 3.7 Order 12

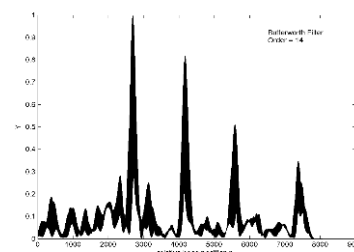


Figure 3.8 Order 14

Butterworth filters with these passband edges and order greater than 14 are found to be unstable.

b) Chebyshev Type I Filter

As shown in Figure 3.9 and 3.10 Chebyshev filters are giving satisfactory performance upto filter order 12.

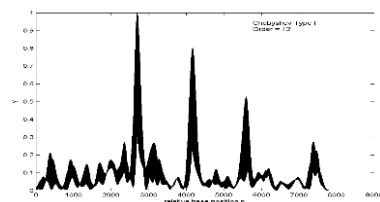


Figure 3.9 Order 10

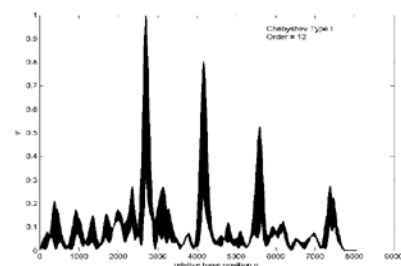


Figure 3.10 Order 12

c) Chebyshev Type II Filter

This category of filter does not give the desired result as it is quite clear from Figure 3.11.

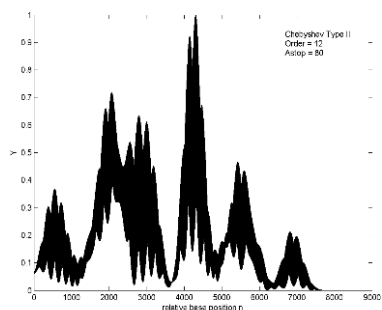


Figure 3.11 Chebyshev Type II Filter

d) Elliptic Filter

As it is clear from Figure 3.12 and 3.13 that the difference in the magnitudes of two most dominant peaks is comparatively less in lower order filters and it increases as the order increases. As the filter order is increased beyond 12, the filter becomes unstable.

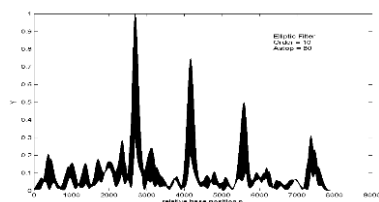


Figure 3.12 Order 10

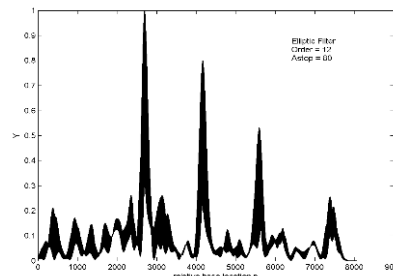


Figure 3.13 Order 12

Table 3.1 Summary of FIR filters order for desired performance

Type Of Filter	Order
Windowed Fourier Series Approach	150
Least-Square Algorithm	160
Constrained Least Square Algorithm	150
Parks-McClellan Algorithm	315

Table 3.2 Summary of IIR filters order for desired performance

Type Of Filter	Order
Butterworth	14
Chebyshev Type I	12
Chebyshev Type II	Undesired Performance
Elliptic	12

4. CONCLUSION AND FUTURE SCOPE

IIR filters are giving the desired performance at a much lower order in comparison to FIR filters. And as the order of these filters increases, the background noise is more suppressed. In future the use of multirate digital filter in identifying the protein coding regions can also be explored.

REFERENCES

[1].Anastassiou, D., "Genomic signal processing" IEEE Signal Processing Magazine, pp. 8–20, July 2001.
 [2].Burge, C. and Karlin, S. "Prediction of complete gene structure in human genome DNA" Journal of Mol. Biol. vol. 268, pp.78-94, 1997.
 [3].Gao, J., Qi, Y., Cao, Y. and Tung, W. "Protein Coding Sequence Identification by

- Simultaneously Characterizing the Periodic and Random Features of DNA Sequences”
Journal of Biomedicine and technology,
vol.2,pp. 139-146, 2005.
- [4].www.ncbi.nlm.nih.gov/
- [5]. Lutovac, M.D., Tomic, D.V. and Evans, B.L.,
Filter Design for Signal Processing: Using
MATLAB and Mathematica, Prentice Hall,
Inc., NJ, 2001.
- [6]. Yan,M., Lin,Z. and Zhang,C. “A new Fourier
transform approach for protein coding
measure based on the format of the Z curve”
Bioinformatics,vol.14, pp.685-690 ,1998.
- [7]. Zhuo,Wang., Yazhu ,Chen. and Yixue Li , “A
Brief Review of Computational Gene
Prediction Methods” Geno. Prot. Bioinfo.
vol.2, No. 4, Nov.2004.

NOT PRESENTED