2012 IACSIT Coimbatore Conferences
IPCSIT vol. 28 (2012) © (2012) IACSIT Press, Singapore

Filtering Approach to DNA Signal Processing

Inbamalar T M, Sivakumar R [+]

R M K Engineering College, Chennai, India

Abstract. The theory and methods of signal processing applied to molecular biology is popular in recent years. The task is to develop methods for genomic sequence analysis. Identifying protein coding regions called exons in a DNA sequence is the fundamental step in the computational recognition of genes. The Fourier spectrum of a DNA protein coding region exhibits an f=1/3 peak. This paper proves the aforementioned and several other empirical observations attributed to DNA sequences and apply filters to predict genes and proteins. An approach based on adaptive filtering is proposed, to locate exons based on the period-3 behavior of biological sequences. Simulation experiments were done and the best suited algorithm for the above said problem was identified. As a future development, same techniques can be applied to analysis protein sequences.

Keywords: Deoxyribo Nucleic Acid (DNA), Exon, Digital filter.

1. Introduction

The genome contains the history of human evolution and specifies the mechanism of human development: all of humanity's physical capabilities and deficiencies are encoded in the genome [1]. In humans there are 23 pairs of chromosomes. All of these occur in homologous pairs. Two homologous chromosomes contain the same genes, but a gene may have several alternate forms called alleles and the alleles of the gene on the two chromosomes may be different. The total content of the DNA molecules within the chromosomes is called the genome of an organism. Within an organism, each cell contains a copy of the genome. The human genome contains about 3 billion base pairs and about 35,000 genes.

In particular, the genome contains the blueprint of all protein coding genes and the control signals used to coordinate the expression of the genes. The well being of any cell relies on the successful recognition of these signals, and a large number of biological mechanisms have evolved towards this goal. Specific protein complexes are responsible for the copying of a gene segment from DNA to messenger RNA and for its eventual translation into protein following the genetic code to assign an amino acid to every tri-nucleotide codon. Specific classes of proteins called transcription factors help recruit the transcription machinery to a target gene by binding their specific DNA signals in response to environmental conditions.

Genomic detection and analysis is of crucial importance in understanding the biological information contained in the genome. A major challenge for genomic research for the next few years is to elucidate the relationship among sequence structure and function of genes. DNA sequence analysis reveals some hidden information to distinguish from coding and non coding regions and to explore some structural similarity among DNA sequence. The National Centre for Biotechnology Information (NCBI) website allows public access to the DNA sequences.

The intron and exon identification problem identifies the location of protein coding regions, which implies those exons and the non-coding regions by computational means. The base sequence in the coding region has strong period-3 component. This phenomenon is due to the non-uniform codon usage that means, even though there are several codons which could code a given amino acid, they are not used with uniform probability and this creates a codon bias. The period-3 property is a good indicator of gene location and this property is exploited for the identification of exons in DNA sequences.

[+] Corresponding author.
E-mail address: hod.ece@rmkec.ac.in.

Genomic signal processing is primarily the processing of DNA sequences, RNA sequences and Proteins. A DNA sequence is made from alphabets of four elements, namely A, T, C and G. For example

……ATCCCAAGTATAAGAAGTA……. The letters A, T, C and G represent nucleotides or nitrogenous bases: ADENINE, GUANINE, CYTOSINE and THYAMINE.

Proteins are biochemical compounds consisting of one or more polypeptides typically folded into a globular or fibrous form, facilitating a biological function. A polypeptide is a single linear polymer chain of amino acids bonded together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. The sequence of amino acids in a protein is defined by the sequence of a gene, which is encoded in the genetic code. In general, the genetic code specifies 20 standard amino acids; however, in certain organisms the genetic code can include selenocysteine—and in certain archaea—pyrrolysine. Shortly after or even during synthesis, the residues in a protein are often chemically modified by posttranslational modification, which alters the physical and chemical properties, folding, stability, activity, and ultimately, the function of the proteins. Sometimes proteins have non-peptide groups attached, which can be called prosthetic groups or cofactors. Proteins can also work together to achieve a particular function, and they often associate to form stable protein complexes.

If we assign numerical values to the four letters in the DNA sequence, we can perform a number of signal processing operations such as Fourier transformation[2,4],digital filtering[5,8]

2. DNA Signal Processing

The Discrete Fourier Transform (DFT) is a very useful tool, because it can reveal periodicities in the input data as well as the relative intensities of these periodic components. The DFT however cannot distinguish appropriately close spectral components for time signals of short duration. Digital filters (window techniques) help in overcoming these limitations in some extent.

A digital filter is a particular class of discrete system capable of realizing some transformation to an input discrete numeric sequence. There are different classes of digital filters according to the properties of their input – output relationships, for example linear, nonlinear, time invariant or adaptive.

3. Filtering Approach

3.1. DNA Numeric Representation

In recent years, a number of schemes have been introduced to map DNA nucleotides into numerical values. Some possible desirable properties of a DNA numerical representation include: (1) each nucleotide has equal "weight" (e.g., magnitude), since there is no biological evidence to suggest that one is more "important" than another; (2) distances between all pairs of nucleotides should be equal, since there is no biological evidence to suggest that any pair is "closer" than another; (3) representations should be compact, in particular, redundancy should be minimized; and (4) representations should allow access to a range of mathematical analysis tools.

The assignment of numerical value to each amino acid is based on some physical properties that are relevant to the its biological functioning. A variety of amino acid indices has been reported in literature. An effective way of assigning the numerical value is the electron-ion-interaction potential (EIIP). The EIIP is defined as the average energy of delocalized electrons of the amino acid which can be evaluated by the pseudo potential model reported in [11]. The EIIP values for the 4 amino acids are listed in Table 1 using which the primary sequence of DNA can be converted to the numerical sequence by replacing each amino acid by the corresponding EIIP values.

The DNA character sequence of interest is mapped onto a numerical sequence using EIIP values. The EIIP of a nucleotide is a physical quantity denoting the average energy of valence electrons in the nucleotide. The EIIP sequence is a weighted sum of four indicator sequences and can be represented by

$$\mathbf{X}_{\mathrm{EIIP}} = \mathbf{W}_A\mathbf{X}_A + \mathbf{W}_G\mathbf{X}_G + \mathbf{W}_T\mathbf{X}_T + \mathbf{W}_C\mathbf{X}_T$$

Table.1 EIIP values of Nucleotides

Nucleotide	EIIP
A	**0.1260**
G	**0.0806**
T	**0.1335**
C	**0.1340**

3.2. Filtering Approach DNA Signal Processing

In digital signal processing field, DFT is used to analyze the periodicity of a signal. Let x(n) represents a discrete periodic signal, then its DFT X(K) is given by:

$$\mathbf{X(k)} = \sum_{n=0}^{N-1} \mathbf{x(n)\, e^{-j2\pi kn/N}} \qquad \mathbf{k=0,1,2,\ldots.,N-1}$$

where N is the length of the series x(n). Consider T represents the periodicity of the signal and f represents its frequency. T=1/f shows the power spectrum of the sequence with unitary frequency. To satisfy the sampling theorem, the number of frequency samples is N/2. Let w be the unitary frequency. Therefore, T=2/w.

The total energy of the four indicator sequences of the DNA sequence is given by,

$$S(w)=|X_A(w)|^2+|X_T(w)|^2+|X_C(w)|^2+|X_G(w)|^2$$

where $X_F(w)$ is the Fourier transform of the indicator sequence $x_k(n)$ of a DNA sequence, K Є {A,C,G,T}.

The adaptive filtering method is used for predicting biological function segments. This is shown in fig.2. Let the input signal of the filter be denoted by x (n), the output be y(n), the desired response be d(n) and the error be e(n). The desired signal d (n) is chosen in terms of the period behavior of special biological segments such as period-3 behavior. The error e(n) between output and desired response signal d(n) is used to regulate the weights vector of the filter.

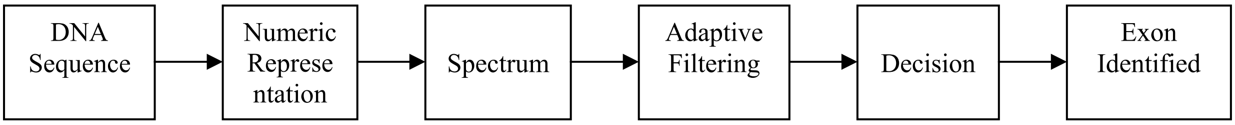

Fig. 1: Block diagram for adaptive filtering method

The symbolic DNA sequence is mapped into a set of digital signals, by using Voss representation. The four indicator sequences are the input to the adaptive filter. Period-3 property exists within the exons (coding regions inside the genes) for eukaryotes (cells with nucleus) and does not exist within the introns (non coding regions in the genes) because of coding biases in the translation of codons into amino acids. In this paper, period-3 property is applied to find the protein coding segments in a DNA sequence. The desired signal is generated by sinusoidal function sin(fk) with the frequency f = 2π / 3 , k=0,1,2,3…… which is a desired period-3 signal.

The LMS, RLS and NLMS filter are applied to the updating algorithms of the adaptive filter.

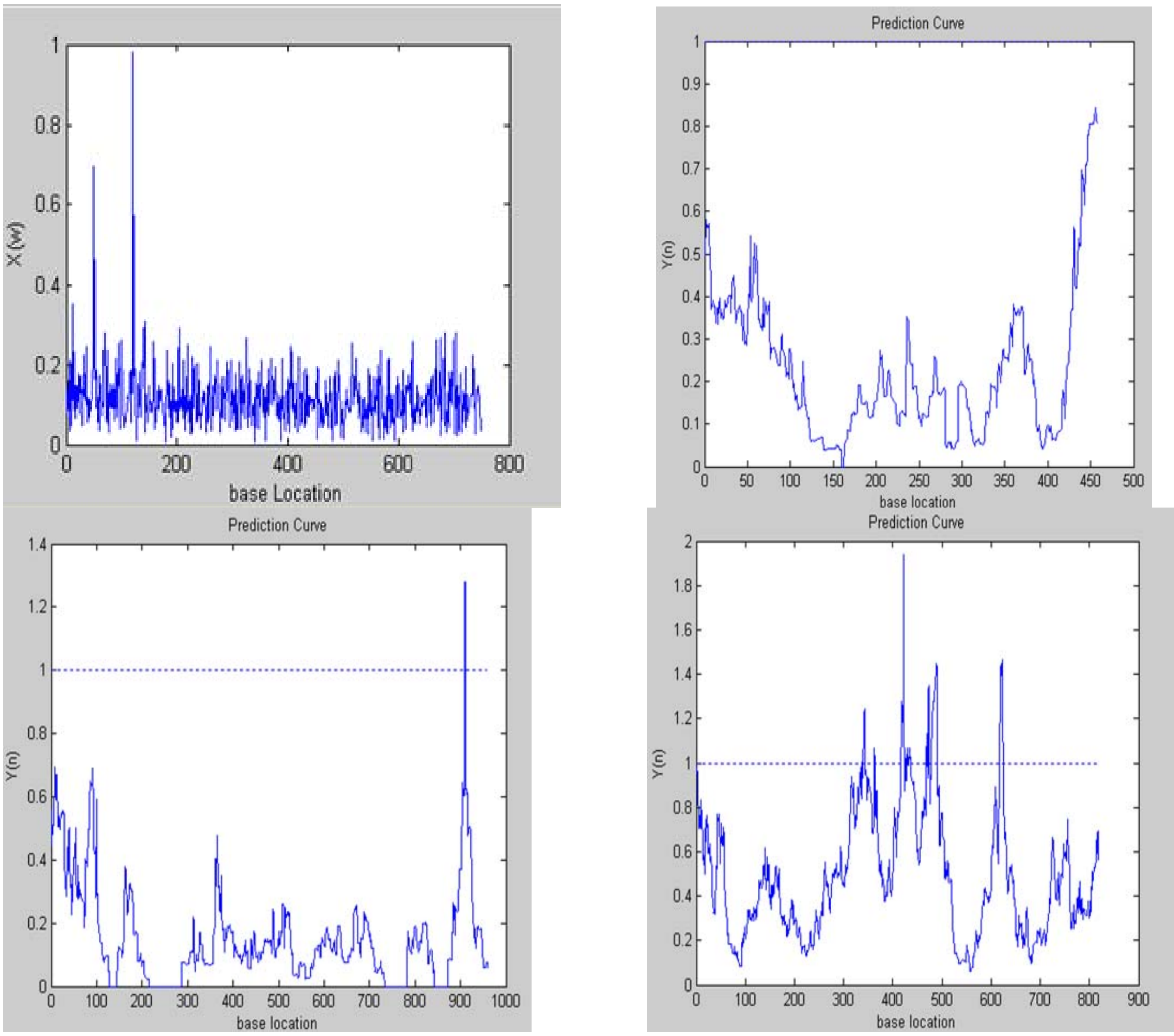

Fig. 2: Simulation results for adaptive filtering method (a)Spectrum (b)LMS Algorithm (c)RLS Algorithm (d)NLMS Algorithm

4. Conclusion

The EIIP representation is discussed. Secondly, a novel adaptive filtering approach is presented to predict the biological function segments. The predictive location curve of the exons is obtained by simulation experiments. It is shown that the presented adaptive filtering approach is valid. There are some obvious peaks at the locations of biological function segments with period property and no peaks at other regions. Comparing the different algorithms implemented NLMS Algorithm provide better solution.

5. References

[1] D.L.Brutlag,"Understanding the human genome", In Leder, P., Clayton, D. A. and Rubenstein, E. (Ed.), Scientific American*: Introduction to Molecular Medicine* (pp. 153-168). New York NY: Scientific American Inc. 1994.

[2] S.Tiwari, S.Ramachandran, A.Bhattacharya, S.Bhattacharya and R.Ramaswamy, *"Prediction of probable genes by Fourier analysis of genomic sequences"*,CABIOS,vol. 13, no. 3, pp. 263-270, 1997.

[3] E.NTRifonov,"*3-,10.5-,200-,and 400- base periodicities in genome sequences*", Physica A, vol.249,1998, pp.511-516.

[4] .D.Anastassiou, "*Genomic signal processing*", IEEE Signal Processing Magazine, pp. 8-20, July 2001

[5] P.P.Vaidyanathan and B-J. Yoon, "*Gene and exon prediction using all pass based filters*", presented at workshop

on Genomic signal processing and stat.,Raleigh, NC,Oct.2002

[6] P. P. Vaidyanathan, "*Genomics and proteomics: a signal processor's tour*", IEEE circuits and systems magazine, vol.4, no.4, 2004, pp.6-29

[7] Tuqan J,Rushdi A "*A DSP perspective for finding the codon bias in DNA sequences*", IEEE j select Topics Sign Proc. 2008; 2:343-356

[8] Juan V.Lorenzo-Ginori, Anibal Rodriguez-Fuentes, Ricardo Grau Abalo, and Robersy sanchez Rodriguez , "*Digital signal Processing in the Analysis of Genomic sequences*" Current Bioinformatics, 2009,4, 28 – 40.

[9] Shmulevich, E. R. Dougherty, "*Genomic Signal Processing*", Princeton University Press, 2007

[10] S.Haykin ,"*Adaptive Filter Theory*", fourth edition, Prentice Hall, 2002

[11] I. Cosic, *"Macromolecular bioactivity: Is it resonant interaction between macromolecules?Theory and applications,"* IEEE Trans.Biomed. Eng., vol. 41, no. 12, pp. 1101-1114, Dec. 1994.