Prediction of Genes Using Digital Signal Processing Technique with FFT (Fast Fourier Transformation)

Paper Submission: 03/02/2021, Date of Acceptance: 24/02/2021, Date of Publication: 25/02/2021

Abstract

We have proposed method DSP technique with FFT for Gene prediction. Gene prediction is the process of identifying protein coding regions (exons) in a given DNA sequence. The proteins coding regions of DNA sequence exhibit a period 3-behavier due to codon structure. The segment of DNA molecules called genes to carry useful information in their protein coding reasons (exons) and their responsible for proteins synthesis. In eukaryotes, (exons) reasons are separated why non-coding regions (introns) where in prokaryotes these regions are continuous. DSP based technique can be used for journey predictions automatically distinctiveness exam from entrance in DNA sequence DSP based technique result in alternative mathematical formulation and may provide improved computational techniques for the solution use full problem in genomic information, science and technology. DSP with FFT is the best gene preduction technique and it can be further used in the identification of hotspots inproteins.

Reeta Kumari Dixit
Assistant Professor,
Dept. of Mathematics,
S.S.P.G. College,
Shahjahanpur, U.P., India

Keywords: Gen, Exons, Introns, DNA, DSP, FFT

Introduction

Genomics is highly cross-disciplinary field that paradigm shift in such diversely areas as medicine, engineering, computer science 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 vest information that is currently raveled from sequencing the genomics of many living organism includes human [2]. Genomic signal processing is engineering discipline that studies the processing the genomic signals.

Aim of the Study

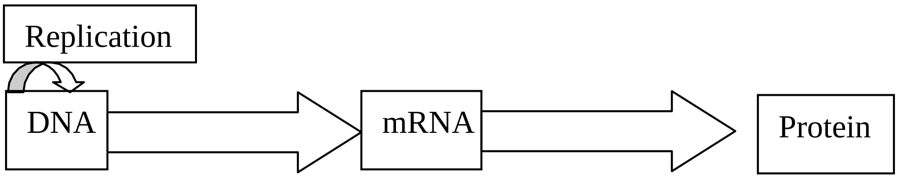
The aim of DSP is to integrate with the theory and methods of signal processing with the global understanding of fundamental genomics [2].

Gene protection is the field of computational biology that the concerned with algorithmetically identifying stretches of the sequence. Usually, genomic DNA; this specially includes proteins coding genes but may also includes other fundamental elements such as RNA genes and regulatory regions. Gene prediction is one of the most important steps in the understanding genome of species once it has been sequenced [2]. Bimolecular sequenced analysis has already been a major research topic among computer scientist, physists and mathematisions. 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 then digital signal processing techniques provides a set of novel and useful tools for solving more complex problems.

For example colour spectrograms provide significant information about bimolecular sequences which facilities understanding local nature, structure and function. Furthermore, the magnitude and the phase of properly define Fourier transforms can be used to predict important features of proteins coding regions in DNA. The proteins of mapping DNA into proteins and the interdependence of two kinds of sequences can be analyzed using simulation based on FFT. DSP based approaches results in mathematical formulation and may provide improved computational techniques for the solution of useful problems in genomics information and technology [3]

Period 3 Property

Which states that genetic information flow from DNA? A DNA strands can be divided into genes and intergenic spaces. Genes are responsible for protein synthesis. A DNA strands can be further subdivided into exons and introns for cells with a nucleus (eukaryotes).



Cells without a necleus are called prokaryotes and do not contain introns. The axons, coding regions within genes, are denoted by start and stop codons. Codons are a subsequence of three letters with in the DNA sequence. Because codons are comprised of three letters from the four-letter alphabet that makes up a DNA sequence, there are 64 possible codons. of the 64 possible codons, there are one start codon and three stop codons, and the remainder of the codons, corresponds to one of the twenty possible amino acids of protein. The relationship between DNA sequences, genes, intergenic spaces, exons, introns and codons is illustrated in

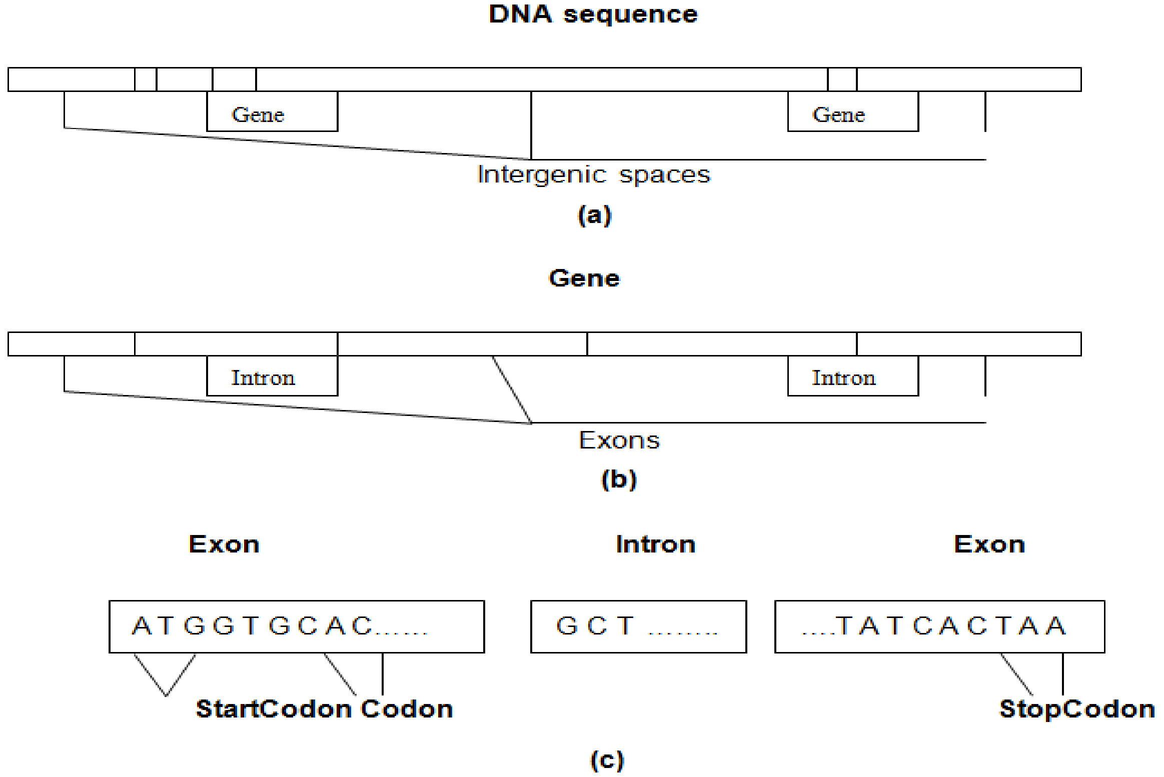


Figure 2: (a) An abstruction to illustrate the genes and and entergenic expressions which comprise a DNA sequence (b) An instructions of a gene to illustrate the subdivision of a gene into exons and introns. (c) Various subsequences' that comprise exons and introns in gene (each 3 letter grouping is a codon). The start condom is always ATG. However, one of the three possible stop condoms is illustrated As(TAA).

Methods of Gene Prediction

This measure computes the spectrum of a DNA sequence as a sum of the spectra of the four binary indicator sequences uA(n),uC(n),uG(n) and uT(n) at a specific frequency (1/3 for 1/3 predictcity),where n denotes the best pair position. The SCM is given as follows:

S(k)=1/N2(|UA(k)|2+UC(k)|2+UG(k)|2+UT(k)|2),k =0,1,2,……….N-1 (1)

Where, UA (k), UC (k), UG (k), and UT (k) are respectively, the fourier transforms of uA(n),uC(n),uG(n) and uT(n), and N is the length of the sequence.

P=S(N/3)/S (2)

$$S=2/N\sum_{k=1}^{N/2} S(\frac{K}{N}) = 1/N(1+\frac{1}{N} - \Sigma^{\alpha}P\alpha^{2}), \text{where } \alpha€\{A,T,C,G \quad (3)$$

Where RF is the relative frequency of occurrence of α

Investigated the value of P for the large number of coding and non-coding regions

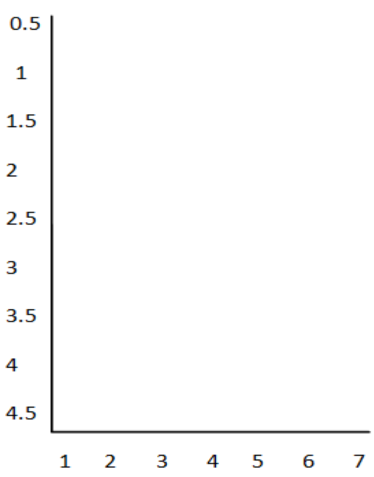
They computed the communicative distributions of SNR for these regions. They proposed the value of P=4 as a threshold to differentiate between coding and non-coding regions. if P is greater than or equal to 4, the reason is identified as cording; otherwise, the region is identified as non-coding.

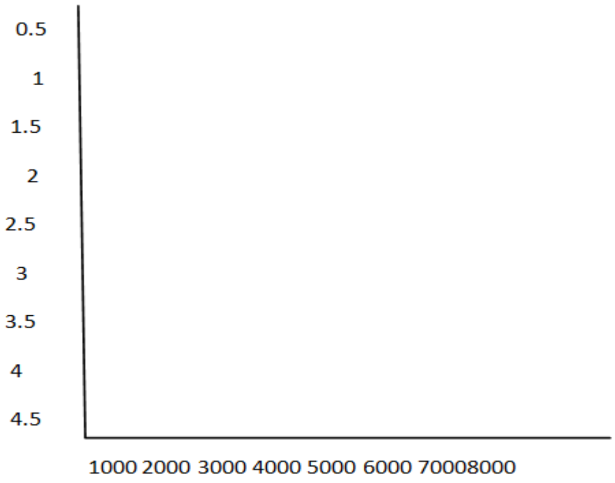
Algorithm

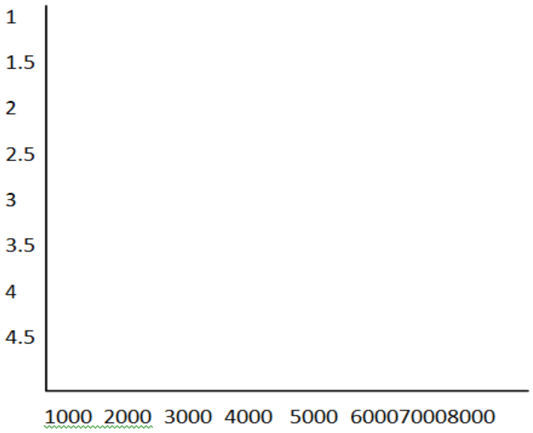
The algorithm of the spectrum content measure for coding region prediction

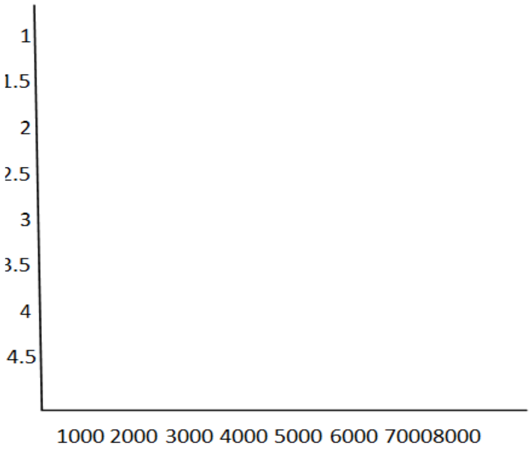
1. for each j, where j=0, 1…..N-L, a sliding window of length L starting at position j of the whole sequenc is analyzed according to formula (1) by computing the spectrum at frequencyL/3.
2. For each window, compute the local SNR, PL (j), according to formula (2).
3. By sliding the window along the sequence, the graphs of PL (j) versus j isgenerated.
4. Initial prediction of probable genes is performed based on the thresholdPL>4
5. to determine the end points of the prediction reasons, the sequence is scanned to an L nucleotide length around the predicted region to locate the stare and stopcodons.
6. The total spectrum of the predicted gene is computed gene is computed by formula (1) to verify the prominence of the period-3 spectrum at frequency N/3 of a codingregion.

Experimental Result









Shows output in MATLAB, above figure exon prediction results in DNA sequence fro a proteins region; the colour line RED, Blue, GREEN of a graph clearly shows the EXON regions of the DNA sequence for the region. In DNA sequence, the resolution of EXON region is improved depending upon the method. The exon regions are clearly identified by exploitiing The period-3 property of the DNA sequenced methods are used to predict the EXON regions of DNA sequence. We are improved faster prediction algorithm of gene prediction i.e. find the Exon Region of DNA sequence in color form. DSP technique using FFT (fast Fourier transformation) which are used to improved the quality of Exon region of DNA sequence.DSP method we are used to exploit this inherent of DNA sequence. The Exon regions we are identified experimentally, we are observed to match those available in NCBI online.

Conclusion and Discussion

On the basis of result DSP with FFT Technique which was found to be the best gene prediction technique among the A, T, C, and G can be further used in the identification of protein region. FFT values can be represented by numerical sequence of the values. By computing the DFT (discrete Fourier transformation) of such sequence, the every distribution periodicities can be observed in terms of the frequency components in the FS (Fourier spectrum) [7]. The Fourier spectra of protein sequences having similar biological functions have common frequency components. This component, called characteristic frequency can be identified by taking the product of the amplitude spectra of the protein sequences belong to a particular functional group. Above Discussion and result proposed methods is fast and accurate prediction of genes from nucleotides sequences.

References

1. *Zemin Ning, Anthony J. Cox and James C. Mullikin, "SSAHA: A Fast Search Method for Large DNA Database", Genome Research, Vol 11, pp.1725-1729, 2001.*
2. *D.Anastassiou,"Genomic Signal Processing," IEEE signal Processing Magazine, PP.8-20, July2001.*
3. *Rayogi T., Takashi N., Takuro L., Akihiro A., Satoshi M.,"The opencl programmingbook".*
4. *Qiu Chen 1*, Koji Kotani 2, Feifei Leel, and Tadahiro Ohmil," A fast search method for dna sequence database using histogram information", International journal of Bioinformatics research, vol. 3, pp. 161, 2011.*
5. *Joao Setubal and Joao Medianis," Introduction to computational molecularbiology".*
6. *http://www.ncbi.ncbi.nlm.gov/*
7. *J.W.Fickett,"the gene prediction problem: an overview for developers", computers chem…, vol.20, no.1 pp.103-118.1996.*