Renu Sharma, Ajay Kaushik / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 1, Jan-Feb 2012, pp. 295-298

EXONS IDENTIFICATION USING FILTER

Renu Sharma[*] Ajay Kaushik[**]

[*]Follower, Electronics and communication Engineering
Maharishi Markandeshwar Engineering College, MMU, Mullana, India
** Lecturer, Electronics and communication Engineering, MMU, Mullana, India

ABSTRACT

Gene prediction is critical problem in the area of bioinformatics. Gene is divided into exons and introns. Gene prediction means to locating the location of exons. Many researchers said that Period-3 property of codon structure helps in predicting exons. In this paper the same property is used along with digital FIR filter which not only predict the location of exons but greatly reduces the background noise. To overcome background noise means to suppress the non-coding regions i.e., introns. We are locating only exons because exons are responsible for protein synthesis.

Keywords: **Codon Structure, Exons, Gene Prediction, Introns, Period-3, Protein.**

1. INTRODUCTION

DNA is made up of genes and intergenic spaces and genes are made up of exons and introns. DNA constitutes of four nucleotides A, T, C, G. These four nucleotides and period-3 behavior of exons are used to predict the genes. DNA sequence is character sequences, so to relate the bioinformatics with signal processing, it is compulsory to convert the sequences. Binary indicator sequences are used for this purpose.

A large amount of literature carried out on this subject. Taher et al. [1] purposed www server for homology-based gene prediction. The user enters a pair of evolutionary related genomic sequences, for example from human and mouse. Alignment of the input sequence is calculated using CHAOS and DIALIGN and then searches for conserved splicing signals and start/stop codons around regions of local sequence similarity. Stanke and Waack [2] proposed a program which is based on a Hidden Markov Model and integrates a number of known methods and submodels. Chakravarthy et al. [3] presented a parametric signal processing approach for DNA sequence analysis based on autoregressive modelling. Autoregressive modelling model residual errors and autoregressive model parameters are used as features. Rao and Shephard, in [4], proposed an AR technique as an alternative tool for this purpose, due to its improved coding region resolution for small data records. Theoretical analysis and experimental results show that the detection resolution for the AR technique is higher than that of Fourier methods for small DNA sequences. Fox and Carrerira [5] introduced a new technique (a single digital filter operation followed by a quadratic window operation) that suppresses nearly all of the non-coding regions. Vaidyanathan [6] said that the digital filtering techniques, transform domain methods, and Markov models have played important roles in gene identification, biological sequence analysis, and alignment. He described the problem of gene finding using digital filtering and the use of transform domain method in the study of protein binding spots. Allen and Salzberg [7] designed a new gene finding system JIGSAW to automate the process of predicting gene structure from multiple sources of evidence, with results that often match the performance of human curators. Its sensitivity and specificity are 92% and 72% respectively. Epps and Akhtar [8] introduced two new techniques TDP and AMDF to this application. They also present an indicative comparison of time domain and existing frequency domain techniques, from which the AMDF appears to be the most promising technique. Mahmood Akhtar [9] introduced new methods frequency' domain techniques and 'time' domain techniques, for gene and exon prediction. He presented a detailed comparison of time-domain and frequency-domain techniques for the detection of both short and long coding regions that are both closely and widely spaced. Rather than performing classification, the features of the various techniques are compared using the receiver operating characteristic (ROC) curve. Nair and Sreenadhan [10] abandoned the four sequences all together and adopted a single 'EIIP indicator sequence' which is formed by substituting the electron-ion interaction pseudopotentials (EIIP) of the nucleotides A, G, C and T in the DNA sequence, reducing the computational overhead by 75%. Gross and Brent [11] proposed N-SCAN used to model the phylogenetic relationships between the aligned genome sequences, context dependent substitution rates, and insertions and deletions. Nair and Mahalakshmi [12] said that a novel symbol-to-

signal mapping for DNA sequences, based on the concept of categorical periodograms. It is observed that the spectral signatures in CCP are functionally equivalent to the established *N*/3 peak in the spectrum of indicator sequences of genomes. Akhtar et al. [13] improved the prediction accuracy of frequency-domain methods by proposing a new algorithm known as the paired and weighted spectral rotation (PWSR) measure, which exploits both period-3 behaviour and another useful statistical property of genomic sequences. Vinson et al.[14] said that Conditional Random Fields (CRFs), directly model the conditional probability P(y|x) of a vector of hidden states conditioned on a set of observations, provide a unified framework for combining probabilistic and non-probabilistic information and have been shown to outperform HMMs on sequence labeling tasks in natural language processing. Bernal et al. [15] described CRAIG, a new program for ab initio gene prediction based on a conditional random field model with semi-Markov structure that is trained with an online large-margin algorithm related to multiclass SVMs. Gunawan et al. [16] presented a signal boosting technique for gene and exon identification of a DNA sequence. Signal boosting technique is used to enhance the coding region and improve the likelihood of correctly identifying the coding region. Hamdani et al. [17] developed a project ueing development tools such as Perl and PHP. The project will identify stretches of sequence for genomic DNA that is biologically functional including protein coding regions. They used Hidden Markov Model which is a mathematical functional will to predict the DNA sequence. Akhtar et al. [18] used DNA symbolic-to-numeric representations and compared with existing techniques in terms of relative accuracy for the gene and exon prediction problem. Novel signal processing-based gene and exon prediction methods are then evaluated together with existing approaches at a nucleotide level using the Burset/Guigo1996, HMR195, and GENSCAN standard genomic datasets. Tomar et al. [19] presented a Harmonic Suppression filter and parametric Minimum Variance Spectrum estimation technique for gene prediction. Hota and Srivastava [20] observed that complex indicator sequence provides strong spectral component compared to EIIP indicator sequence. They observed that windowed DFT taking complex indicator sequence provides better exon prediction compared to windowed DFT taking EIIP indicator sequence and digital filters methods. Computational overhead is reduced by 75% in complex indicator sequence compared to binary indicator sequence. Akhtar et al. [21] investigated the effect of window lengths on selected signal processing-based gene and exon prediction and these methods were then optimized to improve their prediction accuracy by employing the best DNA representation, a suitable window length, and boosting the output signals to enhance protein coding and suppress the non-coding regions. Cai et al. [22] introduced an integrating gene finder, which combines the results of several existing gene finders together, to improve the accuracy of gene finding. Four integration schemes, based on majority voting, are developed for the analysis of two datasets – the basic dataset and the testing dataset. Ahmad et al. [23] focused on the novel solution for nucleotide range estimation. It incorporates denoising DNA signal with discrete wavelet transforms and indicator sequence. Upsampling and downsampling of signal in conjunction with suitable nucleotide choice greatly removes 1/f noise. Ahmad et al. [24] proposed an enhanced and robust technique for exonic prediction by introducing a novel UTP (University Technology PETRONAS) indicator sequence with denoising target DNA signal using discrete wavelet of third order. Ahmad et al. [25] proposed that discrete wavelet transform for noise reduction in DNA sequence and a novel indicator sequence for better signal mapping.

2. WORK METHODOLOGY

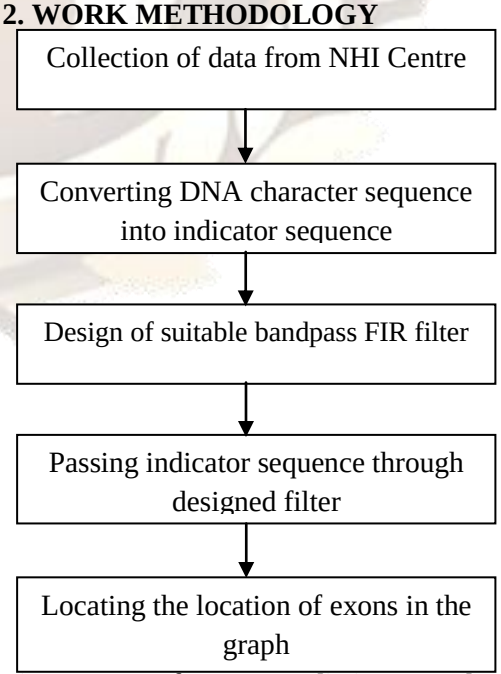

Figure 1 Steps for gene prediction procedure

Steps to carried gene prediction are as shown in figure 1. The data has been collected from NHI centre, in this paper chromosomes elegan F56F11.4 is used. Signal processing only deals with the indicator sequences rather than character sequence, so next step is to convert the DNA character sequences into indicator sequences. Many indicator sequences are used for this purpose for eg. Binary indicator sequence, paired indicator sequence, EIIP indicator sequence and so on. In this paper we have used binary indicator sequences. Binary indicator sequences for four nucleotides are as follows

Let x(n)= { ATCGAAATCCGAATT..}

Then

X_A= 100011100001100..,,,.

X_T =010000010000011.....

X_c =001000001100000.....

X_G =000100000010000.....

After converting the sequences the next step is to design suitable FIR filter using FDA tool in MATLAB. We have designed bandpass FIR filter with normalized frequency. The next step is to pass the converting indicator sequence through designed filter. For this special function filtfilt is used which perform zero phase shifting in both forward and reverse direction.

3. RESULT

Technique which is discussed in this paper was simulated using windowed technique. In [26], Tiwari et al. suggested that the threshold power for C.elegans' chromosome III is 0.4 on a Normalized scale.

Kaiser window

The response of Kaiser window is shown in figure 2. The design specification is selected as wc1= 0.661, wc2=0.667 and order 170.

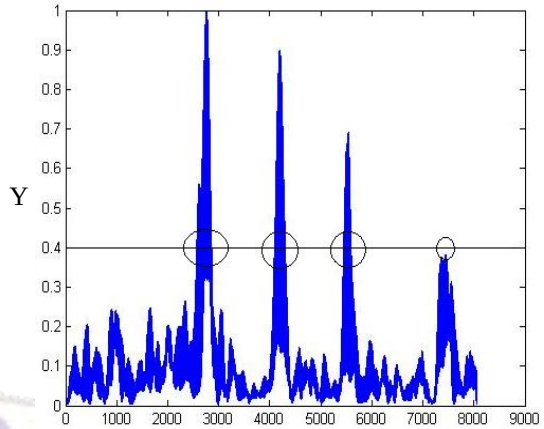

Figure 2 Response of Kaiser Window

Hamming Window

The response of Hamming window is shown in figure 2. The design specification is selected as wc1= 0.665, wc2=0.667 and order 160.

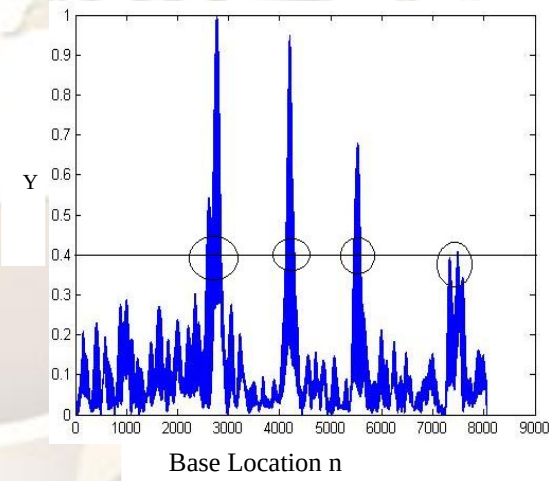

Figure 3. Response of Hamming Window

4. CONCLUSION

a) Identification of the period-3 regions helps in predicting gene location. The peaks in the figures give us the location of exons corresponding to base location i.e. protein-coding regions.

b) Fourier transform and time and frequency domain which is used for this purpose can be replaced by indicator sequence approach.

c) As shown by results that hamming window FIR technique gives better result than Kaiser Window FIR technique. Although noise is very high in case of hamming window technique.

d) There is future scope of converting character string into grouped number indicator sequence and also in Real number indicator sequence and then passing through the filter.

REFERENCES

[1] Leila Taher Oliver Rinner , Saurabh Garg1, Alexander Sczyrba, Michael Brudno, Serafim Batzoglou and Burkhard Morgenstern, " homology-based gene prediction", Vol. 19 no. 12, pp 1575–1577, 2003.

[2] Mario Stanke and Stephan Waack," Gene prediction with a hidden Markov model and a new intron submodel", Bioinformatics, Vol. 19 Suppl. 2, pp 215–225 2003.

[3] N. Chakravarthy, A. Spanias, L. D. lasemidis, and K. Tsakalis, "Autoregressive modelling and feature analysis of DNA sequences,"EURASIP JASP, vol. 1, pp. 13-28, 2004.

[4] N. Rao and S.J. Shepherd, "Detection of 3-periodicity for small genomic sequences based on AR technique", International Conference on communications, Circuits and Systems, ICCCAS, vol. 2, pp. 1032- 1036, June 2004.

[5] T.W.Fox and A.Carreira, "A Digital Signal Processing Method for Gene Prediction with Improved Noise Suppression", EURASIP journal on Applied Signal processing, vol.1, no. 1, pp 108-114, 2004.

[6] P.P. Vaidyanathan, "Genomics and Proteomics: A Signal Processor's Tour" IEEE Circuits and Systems Magazine, vol. 4, no. 4, pp. 6-29, 2004.

[7] Jonathan E. Allen and Steven L. Salzberg," integration of multiple sources of evidence for gene prediction", Vol. 21 no. 18, pp 3596–3603, 2005.

[8] E. Ambikairajah, J. Epps, and M. Akhtar, "Gene and exon prediction using time-domain algorithms," IEEE 8th Int. Symp. On Sig. Proc. and its Appl., pp. 199-202, 2005.

[9] Mahmood Akhtar, "Comparison of Gene and Exon Prediction Techniques for Detection of Short Coding Regions", International Journal of Information Technology Vol. 11, No.8, 2005.

[10] Achuthsankar S. Nair and Sivarama Pillai Sreenadhan, "A coding measure scheme employing electron-ion interaction pseudo potential (EIIP)," Bioinformation vol. 6, pp 197-202, 2006.

[11] Samuel S. Gross and Michael R. Brent, "Using Multiple Alignments to Improve Gene Prediction", Journal of Computational Biology Vol. 13, pp 379-393, 2006.

[12] Achuthsankar S. Nair and T. Mahalakshmi, "Are Categorical Periodograms and Indicator Sequences of Genomes Spectrally Equivalent?", Silico Biology, vol. 6, pp 215-222, 2006.

[13] M. Akhtar, J. Epps, and E. Ambikairajah, "Time and frequency domain methods for gene and exon prediction in eukaryotes", IEEE ICASSP, pp. 573−576, 2007.

[14] Jade P. Vinson, David DeCaprio, Matthew D. Pearson, Stacey Luoma, James E. Galagan, "Comparative Gene Prediction using Conditional Random Field", Advances in Neural Information Processing Systems, vol no. 19, pp 1466-1472, 2007.

[15] Axel Bernal, Koby crammer, Artemis Hatzigeorgiou, Fernando Pereira, "Global Discriminative Learning for Higher Accuracy Computational Gene Prediction", PLOS computational biology, vol.3, no. 3, pp 0488-0497, March 2007.

[16] T. S. Gunawan, E. Ambikairajah, J. Epps, "A signal boosting technique for gene prediction," IEEE ICICS, 2007.

[17] Hazrina Yusof Hamdani,; Siti Rohkmah Mohd Shukri,"Gene Prediction System,"ITSIM, vol.1, pp. 1-7, 2008.

[18] Mahmood Akhtar, Julien Epps, and Eliathamby Ambikairajah, "Signal Processing in Sequence Analysis:Advances in Eukaryotic Gene prediction", IEEE Journal of Selected Topics in Signal Processing, vol. 2, no. 3, pp 310-321, 2008.

[19] Vikrant Tomar, Dipesh Gandhi, C. Vijaykumar, "Digital Signal Processing for Gene Prediction", pp 1-5, 2008.

[20] M.K. Hota and V.K.Srivastava, "DSP technique for gene and exon prediction taking complex indicator sequence", IEEE TENCON, vol. 1, no. 6, pp. 1-6, 2008.

[21] M. Akhtar, J. Epps, and E. Ambikairajah, "Optimizing period-3 methods for eukaryotic gene prediction", IEEE ICASSP, vol. 18, no. 4, pp. 621-624, 2008.

[22] Yudong Cai, Zhisong He, Lele Hu, Bing Li, Yi Zhou, Han Xiao, Zhiwen Wang, Kairui Feng, Lin Lu, Kaiyan Feng, Haipeng Li, "Gene finding by integrating gene finders", Journal of Biomedical Science and Engineering, vol. 3, no. 11, pp 1061-1068, 2010.

[23]Muneer Ahmad, Azween Abdullah and Khalid Burraga, "Optimal Nucleotides Range Estimation in Diffused Intron-exon Noise", World Applied Sciences Journal, vol. 11, no. 2, pp 178-183, 2010.

[24] Muneer Ahmad, Azween Abdullah and Khalid Buragga, "DNA Splicing in Eukaryotes by an Enhanced and Robust Technique", Australian Journal of Basic and Applied Sciences, vol. 5, issue 3, pp 158-170, 2011.

[25] Muneer Ahmad, Azween Abdullah and Khalid Buragga, "A Novel Optimized Approach for Gene Identification in DNA Sequences", Journal of Applied Sciences, vol. 11, no. 5, pp 806-814, 2011.

[26] Tiwari S., Ramachandran S. and Bhattachalya A., et al. "Prediction of probable gene by Fourier analysis of genomic sequences", CABIOS, vol.13, no.3, 1997, pp. 263- 270.