2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies

Genetic algorithm based indicator sequence method for exon prediction

V.G.BIJU
Department of ECE, College of Engineering Munnar
Kerala, India
Email:-bvgpillai@gmail.com

P. Mydhili
Department of ECE, School of Engineering CUSAT
Kerala, India
Email:-mythili@cusat.ac.in

ABSTRACT

Considerable research effort has been devoted in predicting the exon regions of genes. The binary indicator (BI), Electron ion interaction pseudo potential (EIIP), Filter method are some of the methods. All these methods make use of the period three behavior of the exon region. Even though the method suggested in this paper is similar to above mentioned methods , it introduces a set of sequences for mapping the nucleotides selected by applying genetic algorithm and found to be more promising .

Keywords and phrases:
Exon, genetic algorithm, Binary indicator (BI) Electron ion interaction pseudo potential (EIIP)

1. INTRODUCTION

The exons are responsible for protein coding of genes in a DNA strand. It has been observed that the protein-coding regions of DNA sequences exhibit period-three behavior, which can be exploited to predict the location of coding regions (exons) within genes, because of these periodicity the discrete Fourier transforms (DFT)-based methods have been used for the identification of coding regions [1].

Typically a DNA molecule contains millions to hundreds of millions of elements [2]. The problem of finding exons in DNA sequence is well suited to computers because DNA sequences can be represented by data that is easily processed by a computer. DNA strands can be represented by sequences of letters from a four-character alphabet. Convention dictates the use of the letters A, T, G, and C in each element to represent each of the four distinct nucleotides [2].

The exons, coding regions within genes, are denoted by start and stop codons. Codons are a subsequence of three letters within 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 [2]. 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 a protein [2].

In this paper we have also utilized the period three property of the protein coding region. In our method similar to the EIIP values we have introduced a set of four random values for mapping A T G & C, whose magnitude lies between 0 & .15. These random values are best in predicting the exon region and are selected by applying genetic algorithm.

2. GENETIC ALGORITHM INDICATOR SEQUENCE METHOD (GAIS)

A genetic algorithm indicator sequence method (GAIS) to obtain indicator sequence for predicting the exon regions has been proposed in this paper. This method is similar to the Binary indicator (BI) and the Electron ion interaction pseudo potential (EIIP) method except for the coefficients which are used to obtain the indicator sequence. In BI method four sequences are generated from the original sequence by mapping one to a nucleotide and zeros to the other then DFT is calculated followed by power spectrum. The power spectrum is obtained by adding the product of the four DFT values with its conjugate. In the case of EIIP method instead of four sequences only one sequence will be there and is obtained by mapping the EIIP values of the nucleotide. A=0.1260, T=0.1335, G=0.0806, C=0.1340. [3]

978-0-7695-3915-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ACT.2009.217

The coefficients for mapping A T G & C are having values which lie between 0 & 0.15. For generating these coefficients we randomly generated twenty set of binary values and each is of 20 bit, these 20 bit is again divided into 4 groups of 5 bits and these 5 bits are converted to decimal values after dividing it into suitable values so that the decimal valued should lie below 0.15. Now we are having 20 set of random decimal values and each set consist of 4 values. We call it as W1, W2, W3 & W4 and these are used to map A T G & C respectively as shown in table 2.

A study was done using genetic algorithm to find suitable coefficients using this set which are more fit in predicting exon regions for large number of DNA sequences. Hence in our method the coefficients are assigned arbitrary weights as shown in table.2 So for a sequence x(n)=[A A T G C A T C] can be represented as [W1 W1 W2 W3 W4 W1 W2 W4].

Table. 2 GAIS coefficients pattern

A	T	G	C
W1	W2	W3	W4

For the selection of best coefficients using genetic algorithm we first generated a reference sequence, the reference sequence in our experiment is linearly increasing and decreasing values at exon defined area as shown in fig.1 of a known gene whose peak is maximum at the middle of the exon. A known gene means a gene whose exon position is known. In our experiment by using Genscan for Humelafin (D13156) gene the exons are defined in the nucleotide positions 247 to 325 (exon-1) and 1185 to 1459(exon-2).so the reference sequence is created for finding the fitness value. These fitness values are used for applying genetic algorithm [4].

All the generated random sequences after mapping the DNA sequence is used to find the Sliding window DFT and then the power spectrum S[k]. The values of S[k] sequence are values at K=N/3 for each sliding of the window. N is taken as 351. This is to be compared with a reference sequence mentioned above to find the error matrix for the 20 different set of random coefficients. The fitness matrix is obtained from this error matrix.

In the equation shown below R is the power sequence of reference signal, S is the power sequence of a signal whose fitness is to be calculated. E1 is sequence given by

$$E1= \sqrt{(|\mathbf{R-S}|)}$$

The error signal (e) is given by the equation

$$e= \sum E1(i)/L$$

Where L is the total length of the sequence
The fitness (f) is given by the equation

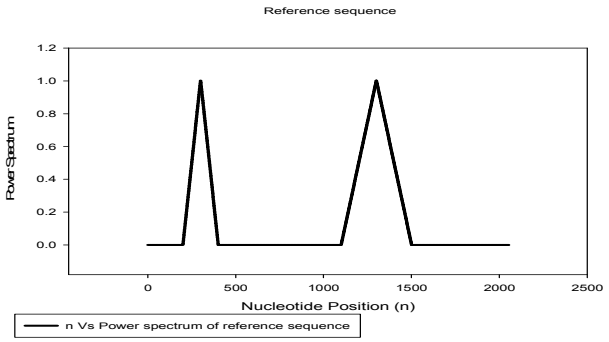
$$f=(1/e) \text{ for } e\neq 0 \text{ \& } 10 \text{ for } e=0$$

The genetic algorithm uses fitness matrix to find order of the binary set of random sequences for crossing by using rowlet wheel selection method. Cross over step followed by mutation we get a new 20 set of random coefficients which are better than the first set. The process mentioned above is repeated N times until optimum set of random values is obtained. These random values which appeared to be best were used to map other DNA sequences whose exon regions are to be predicted. [4]

The authors have tested a large number of DNA sequences and the results were found better than the other methods [5][6]. An optimal window size of 351 is adopted by the authors in this experiment. The authors have compared these results with Binary indicator (BI) and Electron ion interaction pseudo potential (EIIP) method and results were found to be more promising. The table 3 shows 5 such values obtained. The authors have generated and tested a large number of such values for different DNA sequences.

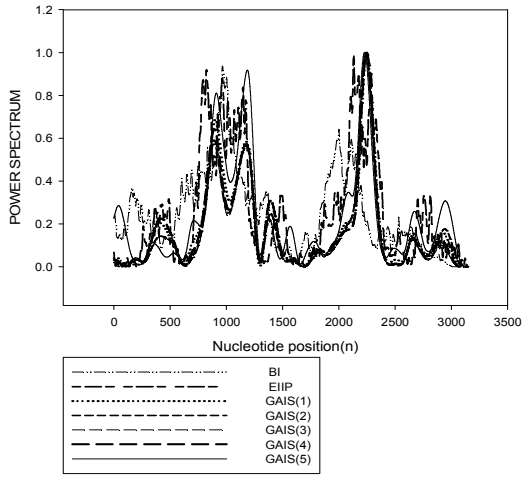
Table. 3 random coefficients selected using GAIS

	A	T	G	C
GAIS(1)	0.1238	0.1429	0.0143	0.1476
GAIS(2)	0.0524	0.0476	0.0905	0.0524
GAIS(3)	0.1429	0.1476	0.0810	0.1381
GAIS(4)	0.1381	0.1381	0.0810	0.1381
GAIS(5)	0.1190	0.1000	0.1381	0.1000



Sl No	Gene Name, Acc No.	Exon Regions	Discrimination measure (D)					
			EIIP	GAIS(1)	GAIS(2)	GAIS (3)	GAIS (4)	GAIS (5)
1	Humelafin, D13156	247-325 (E1)	1.5	1.44	1.46	1.376	1.153	2.451
		1185-1459 (E2)						
2.	AB009589	8540-9479 (E1)	1.064	1.15	1.16	1.368	1.285	0.896
		10624-10949 (E2)						
3.	AF019074	3107-3187 (E1)	0.669	0.71	0.52	0.361	0.404	0.753
		3761-4574 (E2)						
		5832-6007 (E3)						
4.	HUMBETGLOA	866-957 (E1)	2.122	1.9	2.32	2.297	1.827	2.561
		1080-1310 (E2)						
		2161-2289 (E3)						
5.	AF015224	1056-1110 (E1)	0.225	0.41	0.40	0.409	0.409	0.331
		1713-2020 (E2)						

POWER SPECTRUM OF HUMBETGLOA GENE USING BI,EIIP & GAIS METHOD



Result and Discussions

The genetic algorithm indicator sequence (GAIS) have been tested for a number of genes and some of the results are given in table4. Here the discrimination (D) is the ratio of minimum value of exon peak to the maximum value of intron peak.

D= (Lowest of the exon peak/Highest peak in non-coding regions (intron))

So a value of D greater or equal to one is said to be good in discriminating exons. The table shows that all the GAIS coefficients are good compared to the EIIP method in discriminating exons. AF019074, AF015224 gene shows D value less than one for EIIP and GAIS method, but GAIS(1) & GAIS(5) shows better result than EIIP.

CONCLUSIONS

Recently proposed genetic algorithm indicator sequence method for exon region prediction provides superior properties for the separation of exon and intron regions as compared with their frequency-domain counterparts such as binary sequence indicator and EIIP method. This method helps to find more and more number of coefficients which give better results and discrimination than the existing coefficients

REFERENCES

[1] P . P. Vaidyanathan and B.-J. Yoon, "Digital filters for gene pre- diction applications," in Proc. Asilomar Conference on Signals, Systems, and Computers, pp. 306–310, Pacific Grove, Calif, USA, November 2002.

[2] D. Anastassiou, "Genomic signal processing," IEEE Signal Processing Magazine, vol. 18, no. 4, pp. 8–20, 2001.

[3] Achuthsanker S. Nair., Sivarama Pillai Sreenathan "A Coding measure scheme employing electron-ion interaction pseudopotential (EIIP) "Bioinformation by bioinformatics publishing group.

[4] David Beasley, David R. Bull, Ralph R. Martin –"An over view of genetic algorithms: part-2 research topics", University computting, 1993, 15(4) 170—181.

[5] http://genome.imim.es/datasets/genomics96

[6] http://www.cs.ubc.ca/~ rogic/evaluation