doi:10.4236/ajmb.2011.12010 Published Online June 2011 (http://www.SciRP.org/journal/ajmb/).

A reduced computational load protein coding predictor using equivalent amino acid sequence of DNA string with period-3 based time and frequency domain analysis

J. K. Meher[1], G. N. Dash[2], P. K. Meher[3], M. K. Raval[4]

[1]Department of Computer Science and Engineering, Vikash College of Engineering for Women, Bargarh, Orissa, India;
[2]School of Physics, Sambalpur University, Orissa, India;
[3]Department of Embedded System, Institute for Infocomm Research, Singapore;
[4]PG Department of Chemistry, G.M. College, Sambalpur, Orissa, India.
E-mail: jk_meher@yahoo.co.in, gndash@ieee.org, pkmeher@ieee.org, mraval@yahoo.com

Received 12 May 2011; revised 14 June 2011; accepted 29 June 2011.

ABSTRACT

Development of efficient gene prediction algorithms is one of the fundamental efforts in gene prediction study in the area of genomics. In genomic signal processing the basic step of the identification of protein coding regions in DNA sequences is based on the period-3 property exhibited by nucleotides in exons. Several approaches based on signal processing tools and numerical representations have been applied to solve this problem, trying to achieve more accurate predictions. This paper presents a new indicator sequence based on amino acid sequence, called as *aminoacid indicator sequence*, derived from DNA string that uses the existing signal processing based time-domain and frequency domain methods to predict these regions within the billions long DNA sequence of eukaryotic cells which reduces the computational load by one-third. It is known that each triplet of bases, called as codon, instructs the cell machinery to synthesize an amino acid. The codon sequence therefore uniquely identifies an amino acid sequence which defines a protein. Thus the protein coding region is attributed by the codons in amino acid sequence. This property is used for detection of period-3 regions using amino acid sequence. Physico-chemical properties of amino acids are used for numerical representation. Various accuracy measures such as exonic peaks, discriminating factor, sensitivity, specificity, miss rate, wrong rate and approximate correlation are used to demonstrate the efficacy of the proposed predictor. The proposed method is validated on various organisms using the standard dataset HMR195, Burset and Guigo and KEGG. The simulation result shows that the proposed method is an effective approach for protein coding prediction.

Keywords: Genomics; Bioinformatics; Codon; Coding Region; Amino Acid Sequence; Fourier Transform; Antinotch Filter; Periodicity-3; Indicator Sequence

1. INTRODUCTION

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an exponential growth of genomic sequences. An important step in genomic annotation is to identify protein coding regions of genomic sequences, which is a challenging problem especially in the study of eukaryote genomes. In eukaryote genome, protein coding regions (exons) are usually not continuous [1]. Due to the lack of obvious sequence features between exons and introns, distinguishing protein coding regions effectively from noncoding regions is a challenging problem in bioinformatics. Gene Prediction refers to detecting locations of the protein-coding regions of genes in a long DNA sequence. For most prokaryotic DNA sequences, the problem is to determine which segments, in the given sequence, are really coding sequences coding for proteins. For eukaryotic DNA sequences, the problem is to determine how many exons and introns (non-coding regions) are there in the given sequence and what are the exact boundaries between the exons and introns [2].

For the last few decades, the major task of DNA and protein analysis, has been on string matching, either with a goal of obtaining a precise solution, e.g., with dynamic programming, or more commonly a fast solution, e.g., with heuristic techniques such as BLAST and several versions of FASTA [3]. But any of the string matching methodologies could not lead to satisfactory results. A variety of computational algorithms have been developed to predict exons. Most of the exon-finding algorithms are based on statistics methods, which usually use training data sets from known exon and intron sequences to compute prediction functions. As examples, GenScan algorithm [1,2] measured distinct statistics features of exons and introns within genomes and employed them in prediction via hidden Markov model (HMM).

Signal processing techniques offer a great promise in analyzing genomic data because of its digital nature. Signal processing analysis of bio-molecular sequences plays important role for their representation as strings of characters [4,5]. If numerical values are assigned to these characters, the resulting numerical sequences are readily applicable to digital signal processing. During recent years, signal processing approaches have been attracting significant attentions in genomic DNA research and have become increasingly important to elucidate genome structures because they may identify hidden periodicities and features which cannot be revealed easily by conventional statistics methods [6,7]. After converting symbol DNA sequences to numerical sequences, signal processing tools, typically, discrete Fourier transform (DFT) or digital filter can be applied to the numerical vectors to study the frequency domain of the sequences [8]. For most of DNA sequences, one of the principal features is the periodic 3-nucleotide pattern which has been known phenomenon for eukaryotic exons. DNA periodicity in exons is determined by codon usage frequencies. There has been a great deal of work done in applying signal processing methods to DNA recently. The discrete Fourier transform and antinotch filter are applied based on the period-3 property.

The DFT of a given input DNA sequence exhibits a peak at the frequency 2π/3 due to periodicity in the sequence [9]. The DNA sequence consisting of indicator sequence $\{x(n)\}$ of the four bases can be represented in corresponding binary sequences $x_A(n)$, $x_T(n)$, $x_C(n)$ and $x_G(n)$. The DFT of length N for input binary sequence $x_A(n)$ is defined by

$$X_A(k) = \sum_{n=0}^{N-1} x_A(n) \cdot e^{-j2\pi kn/N} \qquad 0 \le k \le N-1 \tag{1}$$

Similarly, $X_T[k]$, $X_C[k]$ and $X_G[k]$ can be found out and the total power at frequency k then be expressed as

$$S(k) = |X_A(k)|^2 + |X_T(k)|^2 + |X_C(k)|^2 + |X_G(k)|^2 \tag{2}$$

The frequency spectrum of $S[k]$, is found to exhibit a peak at $k = N/3$ which indicates the presence of a coding region in the gene.

In digital filtering, for each indicator sequence $x_A(n)$, $x_T(n)$, $x_C(n)$ and $x_G(n)$, a corresponding filter output $Y_A(n)$, $Y_T(n)$, $Y_C(n)$ and $Y_G(n)$, respectively are computed. The sum of the square of magnitude of these filter outputs is expressed as

$$Y(n) = |Y_A(n)|^2 + |Y_T(n)|^2 + |Y_C(n)|^2 + |Y_G(n)|^2 \tag{3}$$

A plot of $Y(n)$ has been used to extract the period-3 region of the DNA effectively [9]. This principle has been applied in antinotch filter and multistage filter. The notch filter is a bandpass filter with passband centered at ω = 2π/3 and minimum stop-band attenuation of about 13 dB. The antinotch filter is a power complementary of notch filter.

In Ref. [6], Tiwari, *et al.* utilized Fourier analysis to detect the probable coding regions in DNA sequences, by computing the amplitude profile of this spectral component which is evidenced as a sharp peak at frequency f = 1/3 in the power spectrum. The strength of the peak depends markedly on the gene. Anastassiou proposed a mapping technique to optimize gene prediction using Fourier analysis and introduced color spectrogram for exon prediction [7]. Although this mapping technique produced comparatively good results than DFT but it was DNA sequence dependent and thus requires computation of the mapping scheme before processing for gene prediction. To improve the filtering through DFT computation, P. P. Vaidyanathan, in [9], proposed digital resonator (antinotch filter) to extract the period-3 components. Short time Fourier transform (STFT) with entropy based methods is incorporated to increase its efficacy to identify the homogeneous regions. [10]. Identification of protein coding regions was developed using modified Gabor-Wavelet transform [11] for the having advantage of being independent of the window length. Entropy minimization criterion in DNA sequences is discussed by Galleani and Garello [12]. Tuqan and Rushdi [13] had explained 3-periodicity related to the codon bias using two stage digital filter and multirate DSP model. Criteria to select the numerical values to represent genomic sequences are discussed by Akhtar *et al*. [14,15].

Genomic information is digital in a real sense; it is represented in the form of sequences of which each element can be one out of a finite number of entities. Such sequences, like DNA and proteins, have been represented by character strings, in which each character is a letter of an alphabet. The first step in gene prediction principle in genomic signal processing involves conversion of string space into signal space of binary numbers called as the indicator sequence. Voss binary representation [16] is the fundamental approach of numerical representation. Various DNA numerical signal representations have been adopted using z-curve [17,18], complex numbers [19], quaternion [20], Gailos field assignment [21], EIIP [22, 23], paired numeric [14] to make indicator sequence in DSP methods to improve the accuracy of exons prediction. Another four-indicator sequence called as relative frequency indicator sequence based on various coding statistics like single-nucleotide, dinucleotide and trinucleotide biases are incorporated into the algorithm to improve the selectivity and sensitivity of filter methods [24]. Real-number representation maps $A = 1.5$, $T = –1.5$, $C = 0.5$, and $G = –0.5$ similar to the complementary property of the complex method are used in [14].

Despite many progresses being made in the identification of protein coding regions by computational methods the performances and efficiencies of the prediction methods still need to be improved. It is indispensable to develop new prediction methods to improve the prediction accuracy. The existing numerical encoding methods can be classified into four-indicator sequences, three-indicator sequences and single-indicator sequences based on computational overhead. The single-indicator sequ- ence reduces the computational overhead by 75% in compared to four-indicator sequence.

A new method to predict protein coding regions is developed in this paper based on the amino acid indicator sequence obtained from DNA string that exon sequences have a 3-base periodicity, while intron sequences do not have this unique feature. The method computes the 3-base periodicity and the background noise of the stepwise amino acid segments of the target amino acid sequences using distributions in the codon positions of the amino acid sequences. The proposed single indicator sequence based on amino acids reduces further the computational load by one-third.

The rest of the paper is organized as follows. Section-2 presents amino acid indicator sequence approach for identification of protein coding regions using Fourier transform and digital filter. Section-3 focuses on the results of the proposed methods with accuracy measures and validated with standard datasets such as HMR195, Burset and Guigo and KEGG. Section-4 presents the conclusions of this paper.

2. PROPOSED AMINO ACID INDICATOR SEQUENCE

It is known that each triplet of bases, called as codon, instructs the cell machinery to synthesize an amino acid. The codon sequence therefore uniquely identifies an amino acid sequence which defines a protein. Thus the protein coding region is attributed by the codons in amino acid sequence [2]. This property is used for detection of period-3 regions using amino acid sequence. The period-3 property is related to difference in the statistical distributions of codon sequence between protein-coding and non-coding sections. This periodicity reflects correlations between residue positions along coding sequences.

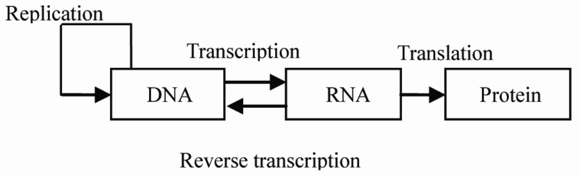


Figure 1. Central Dogma of molecular biology.

The genetic information contained in DNA sequences, RNA sequences, and proteins is extracted in Genomic signal processing. A DNA sequence is made from an alphabet of four elements, namely *A, T, C,* and *G* molecules called nuclotides or bases. This quarternary code of DNA contains the genetic information of living organisms. Similarly protein is also a discrete-alphabet sequences that imparts genetic information and large number of functions in living organism. A protein can be represented as a sequence of amino acids. There are twenty distinct amino acids, and so a protein can be regarded as a sequence defined on an alphabet of size twenty. The twenty letters used to denote the amino acids are the letters from the English alphabet such as ACDEFGHIKLMNPQRSTVWY. It is common that some letters representing amino acids are identical to some letters representing bases. For example the *A* in the DNA is a base called adenine, and the *A* in the protein is an amino acid called alanine. It is known that each gene is responsible for the creation of a specific protein when expressed and this is called as central dogma of molecular biology [2] as shown in **Figure 1**.

The information of expression of particular protein from a gene is contained in a code which is common to all life. The gene gets duplicated into the mRNA molecule which is then spliced so that it contains only the exons of the gene. Each triplet of three adjacent bases of mRNA is called a codon. There are 64 possible codons. Thus the mRNA is nothing but a sequence of codons. Each codon instructs the cell machinery to synthesize a protein using the genetic code. When all the codons in the mRNA are exhausted we get a long chain of amino acids. This is the protein corresponding to the original gene.

In practice numerical values are assigned to the four letters in the DNA sequence to perform a number of signal processing operations such as Fourier transformation, digital filtering, time-frequency plots such as wavelet transformations. Similarly, once we assign numerical values to the twenty amino acids in protein sequences we can do useful signal processing.

The new proposed predictor is based on the analysis of

Table 1. The genetic code.

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

amino acid sequence. In this work the DNA sequence is converted to amino acid sequence *i.e.*, the *A*, *T*, *C*, *G* language is converted to amino acid language [14]. Three characters consisting of nucleotides are represented as codon consisting of twenty alphabets of aminoacids. The mapping from amino acids to codons is many-to-one (**Table 1**). For a given DNA sequence $x_B(n)$, where *B* is nucleotide bases, the corresponding amino acid sequence is obtained as $x_R(n)$, where *R* represents 20 amino acids. For example

$$x_B(n) = \begin{Bmatrix} \text{ATGGGTCCAGCTCCAGTTTTCCC} \\ \text{AAATTCGCGGAAGCCGGCGACACT} \end{Bmatrix}$$

$$x_R(n) = \{\text{MGPAPVFPNSRKPAT}\}$$

The most relevant for the application of signal processing tools is the assignation of properties of amino acid alphabets to form *amino acid indicator sequence*. There are several approaches to convert genomic information in numeric sequences using different representations. Physico-chemical properties of amino acids such as volume, charge, area, EIIP, dipole moment, alpha etc obtained from Hyperchempro 8.0 software of HyperCubeInc, USA are used in this paper for analysis of the proteins (**Table 2**). The resulting numerical sequence by substituting these values is called *amino acid indicator sequence*.

Each amino acid is associated with a unique number of alpha propensities. The indicator sequence is obtained by spreading the numerical value on the amino acid sequence.

$$x_{AA} = \{1.501\ 1.058\ 0.519\ 1.409\ 0.519\ 1.694\ 1.966\ 0.519\ 0.434\ 0.774\ 0.240\ 0.181\ 0.519\ 1.409\ 0.828\}$$

Table 2. Physico-chemical properties of amino acids.

Amino acid	Alpha	EIIP	Dipole moment
A	1.409	0.0373	5.937
R	0.240	0.0959	37.5
N	0.434	0.0036	18.89
D	0.192	0.1263	29.49
C	1.069	0.0829	10.74
Q	0.333	0.0761	39.89
E	0.175	0.0058	42.52
G	1.058	0.0050	0.0
H	0.558	0.0242	20.44
L	1.702	0.0000	3.782
I	1.990	0.0000	3.371
K	0.181	0.0371	50.02
M	1.501	0.0823	8.589
F	1.966	0.0946	5.98
P	0.519	0.0198	7.916
S	0.774	0.0829	9.836
T	0.828	0.0941	9.304
W	1.314	0.0548	10.73
Y	0.979	0.0516	10.41
V	1.694	0.0057	2.692

One of the advantages of using amino acid indicator sequences lies in reducing computational load by one-third as compared to processing DNA indicator sequence.

This technique has been used to identify the coding region which can predict whether a given sequence frame, limited to a specific length *N*, belongs to a coding region or not. This is done by sliding frame in which the amino acids of length *N* of the frame are rated. After that the frame is shifted through a fixed number of samples of residues downstream. The output of every rated window belongs to residues at the specific position. The existence of three-base periodicity exhibited by the sequence as a sharp peak at frequency $f = 1/3$ in the power spectrum in the protein coding regions helps in the prediction of exons.

The discrete Fourier transform (DFT) has been used to predict coding regions in equivalent amino acid sequences of DNA string. As a consequence of the non-uniform distribution of codons in coding regions, a three-periodicity is present in most of genome coding regions, which show a notable peak at the frequency component *N*/3 when calculating their DFT. The DFT of length *N* for input amino acid indicator sequence $x_{AA}(n)$ is defined by

$$X_{AA}(k) = \sum_{n=0}^{N-1} x_{AA}(n) \cdot \mathrm{e}^{-j2\pi kn/N}, \quad 0 \le k \le N-1 \tag{4}$$

for *AA* = amino acids. The absolute value of power of DFT coefficients is given by

$$S(k) = \sum_{k=0}^{N-1} |X_{AA}(k)|^2 \tag{5}$$

The plot of $S(k)$ against *k*, results in peak at $k = N/3$ due to the period-3 property, that indicates the presence of coding regions.

Taking into account the validity of this result the antinotch filter has been applied to amino acid sequences to predict coding regions, using a sliding frame along the sequence. In digital filtering method for indicator sequence $x_{AA}(n)$, corresponding filter output $Y_{AA}(n)$ is computed where *AA* represents 20 amino acids. The sum of the square of magnitude of these filter outputs is expressed as

$$Y(n) = \sum_{n=0}^{N-1} |Y_{AA}(n)|^2 \tag{6}$$

A plot of $Y(n)$ has been used to extract the period-3 region of the of the sequence effectively. Prediction of protein coding regions can be summarized as the following sequence of steps.

1. Convert DNA string to equivalent amino acid sequence with three character code.
2. Substitute physico-chemical properties of amino acid to construct indicator sequence.
3. Apply this sequence to DFT or digital filter to detect period-3 regions.
4. Observe peaks for determining protein coding regions.
5. Evaluate assessment parameters to check accuracy.

3. RESULT AND DISCUSSION

In this paper we propose the technique of using *amino acid indicator sequence* for prediction of protein coding region in gene sequence. We have used digital filtering techniques, such as antinotch filter to detect the protein coding segments using the existing indicator sequences as well as the proposed single indicator sequences based on physico-chemical properties for several organisms. Mainly, three data sets Burset and Guigo [25], HMR195 [26] and KEGG [27] are used for validation of proposed method. The proposed methods performed well in a good number of cases.

The accuracy measures for evaluating the different methods used in this paper are exon-intron discrimination factor *D* [23], sensitivity (S_N), specificity (S_P), miss rate (M_R), wrong rate (W_R) [3,15] and approximate correlation [28]. The discriminating factor is defined as

$$D = \frac{\text{Lowest of exon peaks}}{\text{Highest peak in noncoding regions}} \tag{7}$$

The miss rate and wrong rate are defined as

$$M_R = \frac{ME}{AE} \tag{8}$$

$$W_R = \frac{WE}{PE} \tag{9}$$

where *ME* = missing exons, *AE* = actural exons, *WE* = wrong exons, *PE* = predicted exons.

Table 3. Summary of performance evaluation of amino acid indicator sequence.

Dataset	Assessment Parameters					
	D	S_N	S_P	W_R	M_R	AC
Burset and Guigo	3.8	1	0.85	0	0.33	0.93
HMR195	3.5	1	0.82	0	0.25	0.91
KEGG	2.2	1	0.75	0	0.28	0.89

We define T_P (true positives) as the number of coding regions predicted as coding; T_N (true negatives) as the number of noncoding regions predicted as noncoding, F_P (false positives) as the number of noncoding regions predicted as coding, and F_N (false negatives) as the number of coding regions predicted as noncoding. Based on these parameters, sensitivity and specificity are defined as

$$S_N = \frac{T_P}{T_P + F_N} \tag{10}$$

$$S_P = \frac{T_P}{T_P + F_P} \tag{11}$$

These are widely used measures of accuracy for gene prediction programs. Another measure that captures both specificity and sensitivity is *AC* (approximate correlation). *AC* is defined by

$$AC = \left(\left\{ \frac{1}{4} \left(\frac{T_P}{T_P + F_N} + \frac{T_P}{T_P + F_P} + \frac{T_N}{T_N + F_P} + \frac{T_N}{T_N + F_N} \right) \right\} - 0.5 \right) \times 2 \tag{12}$$

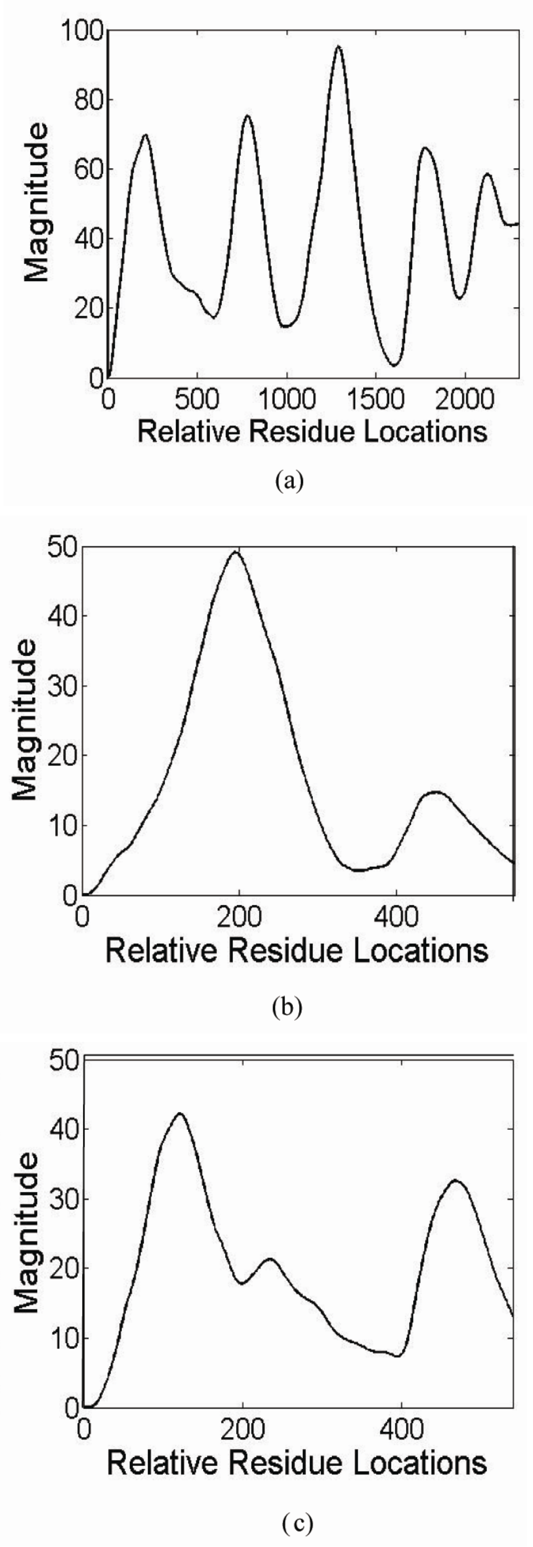
If *D* is more than one (*D* > 1), all exons are identified. High sensitivity and specificity are desirable for higher accuracy. Low miss rate and wrong rate are desirable for better result. The list of genes of organisms is processed with the proposed single-indicator sequences using filtering method and corresponding gene prediction measures have been evaluated. **Table 3** summarizes the observations of eight genes from Burset and Guigo dataset, HMR195 and KEGG dataset. In all the examples cited, the proposed encoding methods show better discrimination compared to the method using multiple indicator sequences. The simulation result shows high discriminating factor, sensitivity and specificity with low miss rate and wrong rate for the proposed methods.

Table 3 summarizes the average performance of proposed method on each dataset. The simulation results using filtering approach on list of selected genes from three datasets are shown in **Table 4**. It is found that the single-indicator sequences based on amino acid sequence show high peak at protein coding locations.

Table 4. Simulation results on selected genes from Burset and Guigo dataset, HMR195 and KEGG dataset.

Gene Name, Acc. No.	Numerical Representations	Accuracy Measures					
	Voss	*D*	S_N	S_P	M_R	W_R	AC
HSODF2,	Real numbers	2.75	1	0.66	0	0.5	0.84
X74614,	Raltive frequency	2.1	1	0.66	0	0.5	0.84
Homo Sapiens	EIIP	3	1	0.66	0	0.5	0.84
ODF2 gene	Amino acid	2	1	0.66	0	0.5	0.84
	Voss	3.5	1	0.75	0	0.33	0.89
	Real numbers	11	1	1	0	0	1
PP32R1,	Raltive frequency	12	1	1	0	0	1
AF00A216,	EIIP	14	1	1	0	0	1
Homo Sapiens	Amino acid	20.6	1	1	0	0	1
	Voss	22	1	1	0	0	1
Humbetgloa,	Real numbers	1.2	1	0.75	0	0.25	0.9
26462,	Raltive frequency	1	1	0.66	0	0.5	0.83
human	EIIP	1.04	1	0.66	0	0.5	0.83
betaglobin	Amino acid	1.5	1	0.75	0	0.25	0.91
	Voss	1.8	1	0.75	0	0.25	0.91
CLDN3,	Real numbers	1.45	1	0.66	0	0.33	0.89
AF007189,	Raltive frequency	1	1	0.66	0	0.33	0.89
Homo sapiens	EIIP	1.04	1	0.5	0	0.5	0.78
Claudin 3	Amino acid	4	1	0.5	0	0.5	0.78
	Voss	1.1	1	0.66	0	0.33	0.86
D p19,	Real numbers	2.2	1	0.66	0	0.5	0.86
AFO61327,	Raltive frequency	1.33	1	0.66	0	0.5	0.86
Homo sapiens	EIIP	3	1	0.66	0	0.5	0.86
cyclin-dependent	Amino acid	1.33	1	0.66	0	0.5	0.86
kinase 4 inhibitor	Voss	2.5	1	0.66	0	0.5	0.86
GalR2,	Real numbers	2	0.66	0.66	0.5	0.5	0.66
AF042784,	Raltive frequency	1.33	1	0.66	0	0.5	0.86
Musculus galin	EIIP	3.2	1	0.66	0	0.5	0.86
receptor	Amino acid	5	1	1	0	0	1
type 2 gene	Voss	5.2	1	1	0	0	1
	Real numbers	2	1	0.66	0	0.5	0.86
NC_002650 Tre-	Raltive frequency	1.3	1	0.66	0	0.5	0.86
ponema Denticola	EIIP	1.8	1	0.66	0	0.5	0.86
U9b Plasmid pTS1	Amino acid	2	1	1	0	0	1
	Voss	2.2	1	1	0	0	1
	Real numbers	1.1	1	0.6	0	0.5	0.86
NC_004767 Heli-	Raltive frequency	1.3	1	0.6	0	0.5	0.86
cobacter pylory	EIIP	1.3	1	0.75	0	0.33	0.89
plamid pHP51	Amino acid	1.4	1	0.75	0	0.33	0.89
		1.8	1	0.75	0	0.33	0.89

The gene sequences "F56 F11.4a" from "Chromosome III" of the organism "C.elegans" (Accession Number AF099922), HUMELAFIN (D13156) of Homo sapiens and ODF2 of Homo sapiens are used for detecting protein coding regions. All the exons of three genes mentioned above are correctly identified as shown in **Figure 2.** In particular **Figure 2(a)** shows the exon prediction results for gene F56 F11.4a showing five peaks corresponding to the exons locations. The simulation result using MATLAB 7.0 shows that of the proposed technique identifies even short sequence. This is observed in first peak of gene F56 F11.4a, whereas it is not pronounced in traditional methods. Similarly **Figure 2(b)** shows two peaks for two exons in gene Humelafin and **Figure 2(c)** shows two peaks for two exons in gene ODF2. The length of amino acid sequence is one-third of that



(a)

(b)

(c)

Figure 2. Gene prediction using *Amino acid indicator sequence* of genes (a) F56F11.4a of C.Elegans chromosome III showing five exons (b) HUMELAFIN of Homo sapiens showing two exons (c) ODF2 of Homo sapiens showing two exons.

of DNA sequence. Hence the exon locations need to be mapped due to reduction of size of the string.

The proposed indicator sequence consisting of alpha propensity, dipole moment and EIIP of amino acids are used for numerical representation and produce sharp peaks at exon locations as well as suppresses the false exons. False exons are the peaks observed in intron locations which do not take part in protein coding. Thus the proposed method is more sensitive to detect true exons which take part in protein coding. Again the execution of reduced sequence due to representation of codons *i.e.*, amino acid sequence reduces the computation time to one-third as compared to the execution of whole sequence of original DNA sequence. Thus the proposed method in not only fast but also efficient.

4. CONCLUSIONS

The new proposed predictor for protein coding regions based on the *amino acid indicator sequence* has good efficacy. The efficacy of the proposed predictor was evaluated by means of accuracy measures such as exonic peaks, discriminating factor, sensitivity, specificity, approximate correlation, wrong rate and miss rate which shows better performance in coding regions detection when compared to the existing methods. The execution of reduced sequence due to representation of codons *i.e.*, amino acid sequence reduces the computation time to one-third as compared to the execution of whole sequence of original DNA sequence. Again the filtering technique with amino acid indicator sequence enables to detect smaller exon regions by showing high peak and minimizes the power in introns giving more suppression to the intron regions. Thus the proposed method is not only fast but also more sensitive.

REFERENCES

[1] Burge, C.B. and Karlin, S. (1998) Finding the genes in genomic DNA. *Current Opinion in Structural Biology*, **8**, 346-354. doi:10.1016/S0959-440X(98)80069-9

[2] Gusfield, D. (1997) Algorithms on strings, trees, and sequences: Computer science and computational biology. Cambridge University Press, Cambridge. doi:10.1017/CBO9780511574931

[3] Wang, Z., Chen, Y.Z. and Li, Y.X. (2004) A brief review of computational gene prediction methods. *Genomics Proteomics Bioinformatics*, **2**, 216-221.

[4] Fickett, J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research*, **10**, 5303-5318. doi:10.1093/nar/10.17.5303

[5] Silverman, B.D. and Linsker, R. (1986) A measure of DNA periodicity. *Journal of Theoretical Biology*, **118**, 295-300. doi:10.1016/S0022-5193(86)80060-1

[6] Tiwari, S., Ramachandran, S. and Bhattachalya, A. (1997) Prediction of probable gene by Fourier analysis of genomic sequences. *CABIOS*, **13**, 263-270.

[7] Anastassiou, D. (2000) Frequency-domain analysis of biomolecular sequences. *Bioinformatics*, **16**, 1073-1081. doi:10.1093/bioinformatics/16.12.1073

[8] Anastassiou, D. (2001) Genomic Signal Processing. IEEE Signal Processing Magazine, 8-20. doi:10.1109/79.939833

[9] Vaidyanathan, P.P. and Yoon, B.J. (2002) Digital filters for gene prediction applications. Proceedings of the 36th Asilomar Conference on Signals, Systems and Computers, 3-6 November 2002, 306-310.

[10] Fuentes, A., Ginori, J. and Abalo, R. (2008) A new predictor of coding regions in genomic sequences using a combination of different approaches. *International Journal of Biological*, *Biomedical and Medical sciences*.

[11] Jesus, P., Chalco, M. and Carrer, H. (2008) Identification of protein coding regions using the modified gabor-wavelet tranform. *IEEE*/*ACM Transaction on Computational Biology and Bioinformatics*, **5**, 198-207.

[12] Galleani, L. and Garello, R. (2010) The minimum entropy mapping spectrum of a dna sequence. *IEEE Transaction on Information Theory*, **56**, 771-783. doi:10.1109/TIT.2009.2037041

[13] Tuqan, J. and Rushdi, A. (2008) A DSP approach for finding the codon bias in dna sequences. *IEEE Journal of Selected Topics in Signal Processing*, **2**, 343-356. doi:10.1109/JSTSP.2008.923851

[14] Akhtar, M., Epps, J. and Ambikairajah, E. (2007) On DNA numerical representations for period-3 based exon prediction. *Proceedings of IEEE International Workshop on Genomic Signal Processing and Statistics*, Tuusula, 1-4. doi:10.1109/GENSIPS.2007.4365821

[15] Akhtar, M., Epps, J. and Ambikairajah, K. (2008) Signal processing in sequence analysis:Advances in eukaryotic gene prediction. *IEEE Journal of Selected Topics in Signal Processing*, **2**, 310-321. doi:10.1109/JSTSP.2008.923854

[16] Voss, R. (1992) Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Physical Review Letters*, **68**, 3805-3808. doi:10.1103/PhysRevLett.68.3805

[17] Zhang, R. and Zhang, C.T. (1994) Z curves, an intuitive tool for visualizing and analyzing the DNA sequences. *Journal of Biomolecular Structure & Dynamics*, **11**, 767-782.

[18] Rushdi, A. and Tuqan, J. (2006) Gene identification using the Z-curve representation. *Proceedings of IEEE International Conference on Acoustics*, *Speech and Signal Processing*, Toulouse, 14-19 May 2006, 1024-1027.

[19] Cristea, P.D. (2002) Genetic signal representation and analysis. *Proc. SPIE Conference*, *International Biomedical Optics Symposium* (*BIOS*'*02*), **4623**, 77-84.

[20] Brodzik, A.K. and Peters (2005) Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences. *Proceedings of IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, **5**, 373-376.

[21] Rosen, G.L. (2006) Signal processing for biologically-inspired gradient source localization and DNA sequence analysis. Ph.D. Thesis, Georgia Institute of Technology, Atlanta.

[22] Nair, T.M., Tambe, S.S. and Kulkarni, B.D. (1994) Application of artificial neural networks for prokaryotic

transcription terminator prediction. *FEBS Letters*, **346**, 273-277. doi:10.1016/0014-5793(94)00489-7

[23] Nair, A.S. and Sreenathan, S.P. (2006) A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). *Bioinformation*, **1**, 197-202.

[24] Nair, A.S. and Sreenathan, S.P. (2006) An improved digital filtering technique using frequency indicators for locating exons. *Journal of the Computer Society of India*, **36**.

[25] Burset, M. and Guigo, Â.R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353-367. doi:10.1006/geno.1996.0298

[26] Rogic, S., Mackworth, A. and Ouellette, F. (2001) Evaluation of genefinding programs on mammalian sequences. *Genome Resarch*, **11**, 817-832. doi:10.1101/gr.147901

[27] Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acid Research*, **28**, 27-30. doi:10.1093/nar/28.1.27

[28] Biju, I. and Gajendra P.S.R. (2004) EGPred: Prediction of eukaryotic genes using ab initio methods after combining with sequence similarity approaches. *Genome Research*, **14**, 1756-1766. doi:10.1101/gr.2524704