Comparison of Different Mapping Technique for Better Identification of Exon Regions

Saikat Singha Roy, Soma Barman

Abstract–Genomic signal processing is a new area of research comprised of gene analysis using signal processing technique. Proper identification of coding and non-coding regions of DNA sequence has become a challenge to the researchers of various fields. Application of DSP needs a mapping rule to convert the alphabetic (A, C, T, G) sequence into its corresponding numerical value. Several fixed mapping technique have been employed so far for the application of DSP. In recent years variable mapping technique has also being introduced for better identification of period-3 peaks in the power spectrum of a DNA sequence. The present paper compares the result between fixed mapping technique and the variable mapping technique in trams of signal to noise ratio (SNR) and exon position which will better identify the coding region of a DNA sequence.

Index Terms: Digital signal processing, Deoxyribo Nucleic Acid, Exon, Codon, Period-3, Signal to noise ratio, Filter.

1 Introduction

Performing genetics research on a computerized platform using signal processing becomes passionate after the discovery of double-helix structure [1] of Deoxyribonucleic acid (DNA) molecule.Genes are the only segments that content genetic information. The genome, made of DNA, is composed of bio molecular components, called nucleotides [2]. Genetic information is stored in DNA in a particular order consisting of Adenine (A), Guanine (G), Thymine (T) and Cytosine (C). The base A always pairs with T and C with G. The two strands of the DNA molecule are therefore complementary to each other and the double helix structure of DNA is made. The DNA sequence is divided into genes and intergenic spaces and gene is sub divided into exons and introns as shown in Fig.1. Among them only the exons are involved in protein coding. Therefore, identificationof the locations of protein-coding regions (exons) and the non-coding regions (introns) in DNA sequences through computational means has become important.

Since codon structure is involved in the translation of the base sequence into amino acids, a strong period-3 component is found in base sequence of the protein coding region [3-4]. A number of authors have developed algorithms for detection of protein coding regions in genomic sequence by finding regions exhibiting period-3 characteristic [5-7].

Vaidyanathan and Yoon [8] applied an anti-notch Infinite Impulse Response (IIR) digital filter to the indicator sequences to detect the period-3 components. To find the period-3 regions in genomic sequences an enormous application of the Discrete

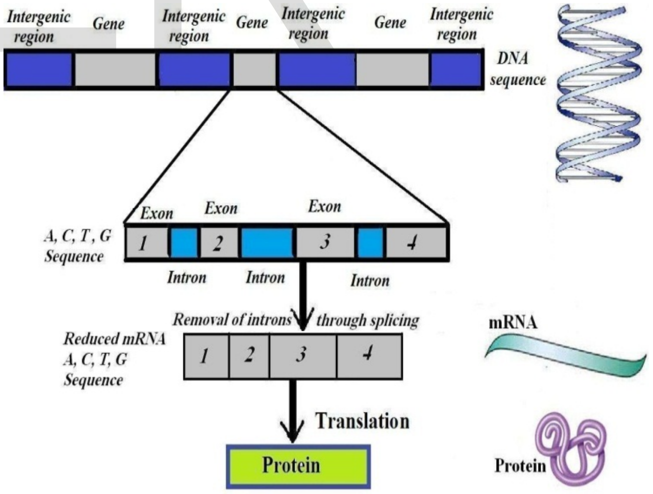

Fig.1 DNA sequence with gene

Matlab environment has been used for filter realization. This paper is organized as introduction, methodology, result discussions and conclusion.

2 Methodology

According to FASTA representation, DNA sequence consists of four nucleotides namely A, C, T and G and the

- Soma Barman (Corresponding author)
  Institute of Radio Physics & Electronics, University of Calcutta,
  Kolkata 700009, India
  email: barmanmandal@gmail.com
- Saikat Singha Roy
  Institute of Radio Physics & Electronics, University of Calcutta,
  Kolkata 700009, India

sequence is in the form of alphabetic representation. Therefore a suitable mapping rule is used to convert the alphabetic sequence to numerical form prior to DSP application. Generally the periodicity of the sequence is examined by applying DFT on the sequence and then by squaring which is called PSD of sequence. But the PSD of DNA sequence is very noisy in nature and difficult to identify period-3 peaks properly. Filtering technique is used to clear the noise. The block representation of the procedure is duplicated in Fig.2.

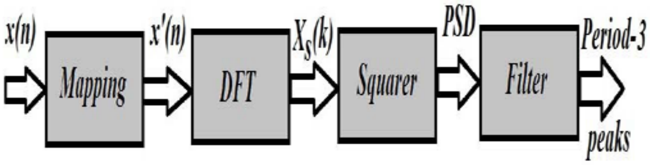

Fig. 2 Block diagram representation of overall method

2.1 Numerical Mapping

2.1.1 Existing Fixed Numerical Mapping Technique

Proper choice of mapping technique makes easier to find the protein coding region of a DNA sequence. Hence different types of conversion methods were proposed by different researcher [9]. The simplest and oldest method for conversion is the binary or Voss [10] representation. Where the nucleotides A, C, T and G map into four binary indicator sequences $x_A(n)$, $x_C(n)$, $x_T(n)$ and $x_G(n)$. It allocates numerical '1' to represent the presence and numerical '0' to represent the absence of the respective nucleotides at particular locations 'n'. For example (Table 1) given a DNA sequence is

x(n)= ….. A G A A T C G T C …..

Table 1

Example of Voss mapping technique

DNA Sequence	..A	G	A	A	T	C	G	T	C…
$X_A[n]$	1	0	1	1	0	0	0	0	0
$X_C[n]$	0	0	0	0	0	1	0	0	1
$X_T[n]$	0	0	0	0	1	0	0	1	0
$X_G[n]$	0	1	0	0	0	0	1	0	0

Where $X_A[n] + X_C[n] + X_T[n] + X_G[n] = 1$ and n represents the base index.

Another numerical representation in complex form [11], the complementary nature of A-T and C- G pairs are reflected. Chakravarthy et al [12] and Cristea [13] have proposed a real number mapping of the DNA sequence. Real number mapping using electron-ion interaction potential (EIIP) [14] of nucleotide is used to map DNA character strings into numerical sequences. Akhtar et al. [15] introduced quaternion mapping technique and used to compute the 3-periodicity in DNA sequences. In atomic number mapping technique [16] a single atomic number indicator sequence is formed where the atomic number in each nucleotide in a DNA sequence is allotted. Z-curve mapping technique [17] represents visualized analysis of a DNA sequence in 3-D. All these mapping rule substitute fixed numerical value to the nucleotide. The authors introduced here a new variable mapping rules and the performance of proposed mapping technique is compared with the existing QPSK-based mapping [18] rule. The values of A, C, T and G for the existing mapping technique mentioned above are listed in Table 2.

Table 2

Values of A, C, T and G for different numerical mapping technique

Numerical Mapping	A	C	T	G
Complex Number	1+j	-1+j	1-j`	-1-j
Real Number	2	1	0	3
Real Number	1.5	0.5	-1.5	-0.5
EIIP	0.1260	0.1340	0.1335	0.0806
Quaternion Technique	i+j+k	i-j-k	-i+j-k	-i-j+k
Atomic Number	70	58	66	78
QPSK-based	j	1	-1	-j

2.1.2 Variable Mapping Technique Based on Twiddle Factor

The authors [19] proposed a new variable mapping technique for better identification of protein coding region. In this technique each and every nucleotide (A, C, T and G) is represented by complex numerical value which varies with position of nucleotide in codon along the DNA sequence. For the present algorithm 16-point DFT is used where 16 different Twiddle factors define 16 different locations on the circle. The circle is divided into 4 quadrants for 4 nucleotides. Where A, C, G and T are represented by the values of Twiddle factor of 1st, 2nd, 3rd and 4th quadrant respectively. Since the proposed technique is variable mapping rule, a circle of different radius is considered for every codon present in a DNA sequence. Unlike fixed mapping technique the values of each nucleotide will vary through the entire sequence depending on the location of codon and the position of nucleotide within the codon. The 1st nucleotide of a codon will be considered in the place of '1' position. Similarly 2nd and 3rd nucleotide of the codon will be placed in '2' and '3' position in Fig. 4.

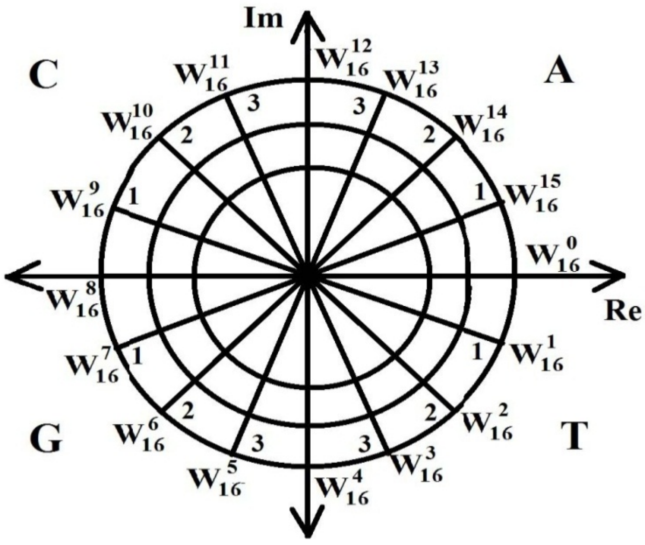

Fig. 3 Variable mapping technique

For example, to represent a DNA sequence AAC GAC TTA …, Fig.4 shows the 1st, 2nd and 3rd position of each nucleotide with their corresponding Twiddle factor.

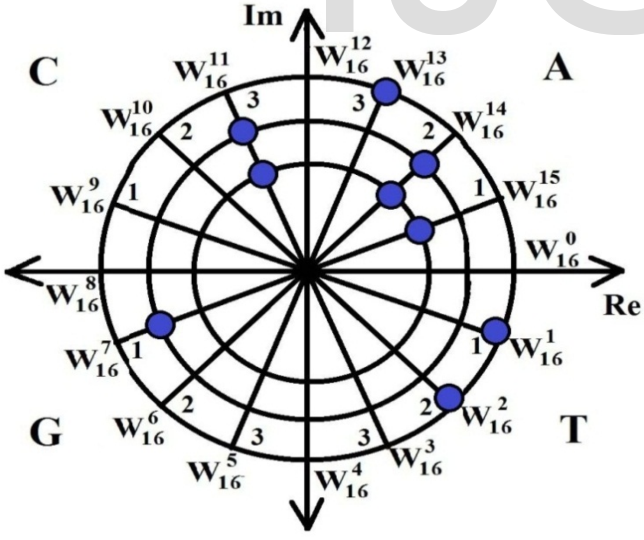

Fig.4 Nucleotide positions of the 1st, 2nd and 3rd codon

2.2 DFT ANALYSIS OF DNA SEQUENCE

After having the mapped sequence a multistage filtering technique has to be applied to find the protein coding region, after obtaining spectrum of the sequence using DFT. Let XS[k] be the DFT of mapped sequence given by

$$X_s(k) = \sum x_s(n)e^{-2\pi nk/N} \tag{1}$$

Where n=0, 1, 2, ………, N-1 , Xs(n) = Mapped sequence.

The power spectral density of the sequence is

$$P_s(k) = \sum |X_s(k)|^2 \tag{2}$$

Plot of Ps(k) may be used as preliminary indicator for detecting probable coding region in DNA sequence. A peak at the frequency k=N/3, where N is the length of the sequence, is observed in a protein coding region of DNA sequence based on period-3 property. From the PSD plot (Fig. 6) of genomic sequence is much noisy; it is difficult to locate exact position of protein coding region. Therefore filtering approach is necessary to find exact location of coding region in presence of noise.

2.3 FILTERING FOR GENE PREDICTION

The authors have attempted multirate filter to suppress the noise in the PSD plots of DNA sequences and compare the performances of filtering in terms SNR and exon position.According to Singha Roy et al[18] multirate filter is the better choice for exon predication. Block diagram of the filtering process is shown in the Fig. 5.

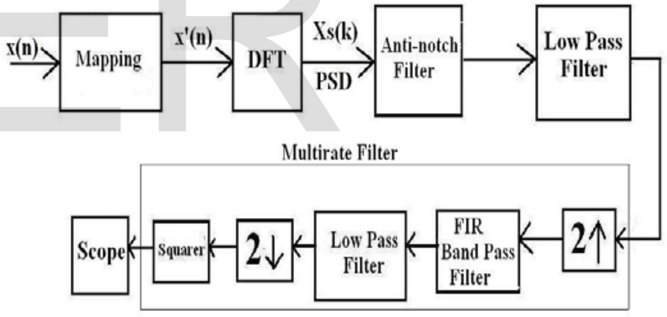

Fig. 5 Block diagram of Multirate Filter.

Here for designing filter we used Up-sampler followed by Down-sampler by N with cascade of LPF and BPF are used in between and N=2 for both the down and up sampling. In order to smooth the filter response FIR low pass filter with Blackman window function having specifications: Direct Form Structure, order=400, FS=8000 Hz and FC=0.003 has been used and FIR band pass filter (BPF) with Blackman function having specifications: Direct form FIR, order=35, wc1=0.6666 and wc2=0.6667 (Normalized) is used to identify the period-3 property.

3 RESULT AND DISCUSSION

The gene C-Elegan Cosmid F56F11.4a [20], Accession No. AF099922.1, sequence length 7990 bp comprising of 1st coding segment: 929-1135 bp, 2nd coding segment: 2528-2857 bp, 3rd coding segment: 4114-4377 bp, 4th coding segment: 5465-5644 bp and 5th coding segment: 7255-7605 bp relative to 7021 is used as sample for comparative performance analysis throughout the article.Fig. 6 shows the PSD plot of the given gene sequence which is very random in nature and the period-3 peaks cannot be detected.

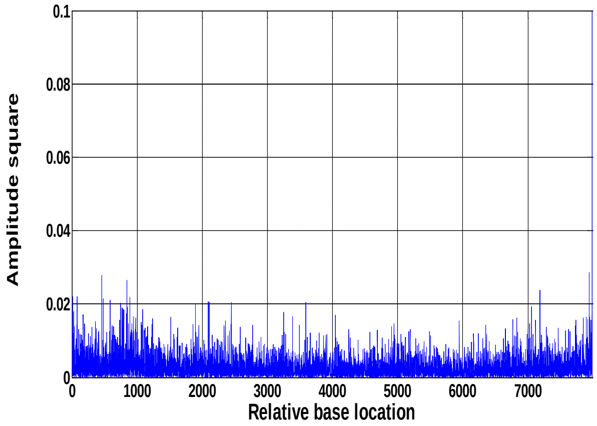

Fig.6 PSD plots without filter

TABLE 3
AVERAGE DEVIATION FROM PERIOD-3 POSITION FOR VARIOUS MAPPING METHOD

Mapping Technique	Numerical Value	Average deviation from period-3 position
Binary Indicator	x_a=1; x_c=1; x_t=1; x_g=1;	170.2
Real Number	a=2; t=0; c=1; g=3;	189.8

EIIP Code	a=0.1260;t=0.1335;
	c=0.1340;g=0.0806
		179.8

Molecular Mass	a=134;t=125;c=110 ;g=150	316.6

Atomic Number	a=70; t=66;
	c=58; g=78
		151.8

Paired Nucleotide Atomic Number	a=42; t=42;
	c=62; g=62
		224.6

Complex Number	a=1+j; t=1-j;
	c=-1-j; g=-1+j;
		264.4

Pyrimidine Purine Complex	a=j; c=1;
	t=-1; g=-j
		139.6

Variable Mapping	Depends on Codon Number	66.6

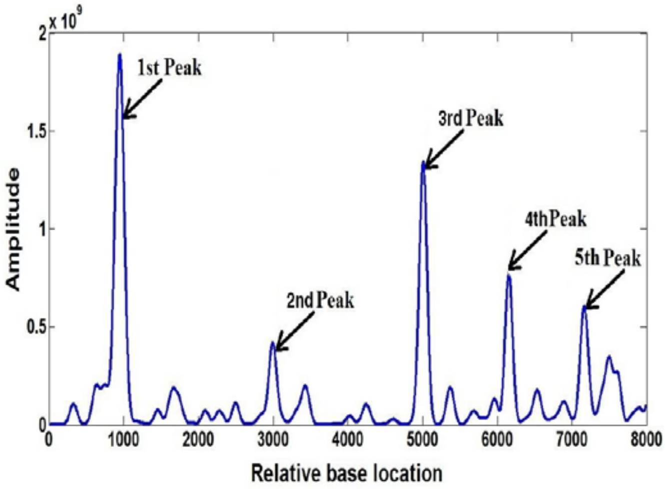

Fig. 7 Output of IIR anti-notch with multirate filter using Pyrimidine Purine complex mapping

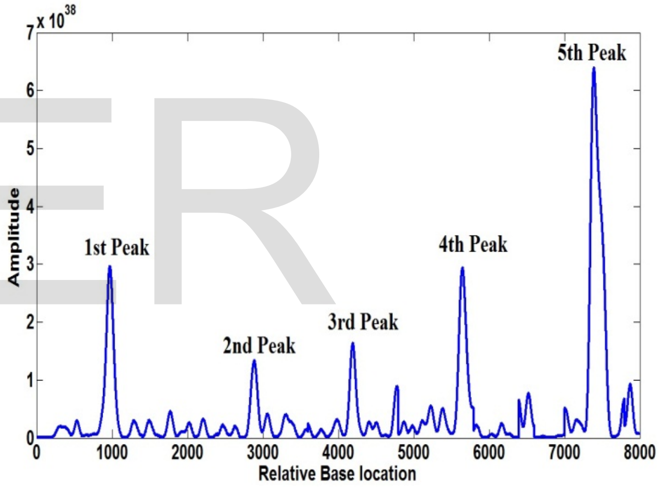

Fig. 8 Output of IIR anti-notch with multirate filter using Variable mapping

The simulated PSD plot of two different types of mapping i.e. IIR anti-notch with multirate filter using Pyrimidine Purine complex mapping and IIR anti-notch with multirate filter using variable mapping are shown in figure 7 and 8 respectively. From Plots, it is clear that spectrum resolution not only clear in case of variable mappingcompared to Pyrimidine Purine complex mappingbut also noise level is reduces in multirate filter (Table-4).

TABLE 4 PERFORMANCE COMPARISONS OF DIFFERENT TYPES OF MAPPING TECHNIQUE

Mapping Type	SNR
IIR anti-notch with multirate filterusing Pyrimidine Purine complex mapping	1.46

IIR anti-notch with multirate filter using variable mapping	1.63

4 Conclusion

The spectrum plots showed variable mapping provides less deviation of coding regions locations than Pyrimidine Purine complex mapping. From table-3 it is clear that the average deviation of period-3 peaks is better in case of variable mapping. Performance of variable mapping is better compared with Pyrimidine Purine complex mapping by measuring SNR (Table 4). Location accuracy and noise suppression which are critical issues in gene prediction both considered in this article and successfully achieved.

Acknowledgment

The author Saikat Singha Roy would like to thank Camellia Institute of Technology & Management, India for giving him the opportunity to pursue his research work.

References

[1] Crick, Francis, and James Watson. "Molecular structure of nucleic acids" Nature 171.4356 (1953): 737-738.

[2] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, Essential Cell Biology. New York: Garland Publishing, 1998.

[3] Trifonov, E.N.: '3-, 10.5-, 200- and 400-base periodicities in genomesequences', Physica A: Stat. Theor. Phys., 1998, 249, pp. 511–516

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

[5] D Anastassiou,. (2000). Frequency –domain analysis of biomolecular sequences Bioinformatics 16,1073-1081.

[6] D Anastassiou,.(2001). DSP in genomics: Processing and frequency domain analysis of character strings. IEEE-7803-7041-2001.

[7] J. Tuqan, and A. Rushdi, "A DSP perspective to the period – 3 detectionproblem.", Proceedings of the IEEE International Workshop onGenomic Signal Processing and Statistics, GENSIPS ,pp.53-54, 2006

[8] P. P. Vaidyanathan, and B. J. Yoon, "The role of signal-processing concepts in genomics and proteomics", Journal of the Franklin Institute, special issue on Genomics, 2004.

[9] Akhtar M, Epps J, and Ambikairajah E, (2008) Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction. IEEE J. Sel. Topics Signal Process, vol. 2, no. 3, pp. 310-321

[10] Voss R F (1992) Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Physical Review Letters; 68(25):3805-3808

[11] CristeaP D (2002) Conversion of nucleotides sequences into genomic signals. J Cell. Mol. Med., vol. 6, pp. 279-303

[12] Chakravarthy N, Spanias A, Iasemidis LD and Tsakalis K (2004) Autoregressive modeling and feature analysis of DNA sequences. EURASIP Journal on Applied Signal Processing, vol. 2004, no. 1, pp. 13–28

[13] Cristea P D (2002) Genetic signal representation and analysis. Proc. SPIEConference, International Biomedical Optics Symposium (BIOS'02), vol. 4623, pp. 77–84

[14] Ning J, Moore C N and Nelson J C (2003) Preliminary wavelet analysis of genomic sequences. Proc. IEEE Bioinformatics Conf. (CSB), pp. 509–510

[15] Akhtar M, Epps J and Ambikairajah E (2007) On DNA numerical representations for period-3 based exon prediction. Proc. of IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS), pp. 1-4

[16] Holden T, Subramaniam R, Sullivan R, Cheng E, Sneider C, Tremberger G, Flamholz J A, Leiberman D H and Cheung T D (1992) A TCG nucleotide fluctuation of Deinococcusradiodurans radiation genes. Proc. of Society of Photo-Optical Nature, San Diego, CA, USA, 168

[17] Zhang R and Zhang C T, Z curves, An Intuitive Tool, for Visualizing and Analyzing the DNA sequences. J BioMol.Struct.Dyn. , vol. 11, pp. 767-782

[18] Singha Roy S, Barman S (2014) Identification of protein coding region of DNA sequence using multirate filter. Computational Advancement in Communication Circuits and Systems, Lecture Notes in Electrical Engineering 335, DOI 10.1007/978-81-322-2274-3_16

[19] Singha Roy S, Barman S(2016), "Polyphase Filtering with Variable Mapping Rule in Protein Coding Region Prediction", Microsystem Technologies, vol-22, Issue-4, DOI: 10.1007/s00542-016-2884-5

[20] National Centre for Biotechnology Information (NCBI): http://www.ncbi.nlm.nih.gov.