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 contain 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, identification of 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

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 contain 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, identification of 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

2 METHODOLOGY

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

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

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

\[ x(n) = \ldots A \ldots A \ldots T \ldots C \ldots G \ldots T \ldots C \ldots \]

\[
\begin{array}{c|c|c|c|c|c|c|c}
    DNA Sequence \to X_A[n] \to X_C[n] & X_T[n] & X_G[n] \\
    \hline
    \ldots A \ldots A \ldots T \ldots C \ldots G \ldots T \ldots C \ldots & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
    \hline
    X_A[n] & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
    X_C[n] & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
    X_T[n] & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
    X_G[n] & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\end{array}
\]

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.

\[
\begin{array}{c|c|c|c|c}
    \text{Numerical Mapping} \to A \to C \to T \to G \\
    \hline
    \text{Complex Number} & 1+j & -1+j & 1-j & -1-j \\
    \hline
    \text{Real Number} & 1.5 & 0.5 & 0 & 1.5 \\
    \hline
    \text{EIIP} & 0.1260 & 0.1340 & 0.1335 & 0.0806 \\
    \hline
    \text{Quaternion Technique} & i+j+k & i-j-k & -i+j-k & -i-j+k \\
    \hline
    \text{Atomic Number} & 70 & 58 & 66 & 78 \\
    \hline
    \text{QPSK-based} & j & 1 & -1 & -j \\
\end{array}
\]

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.

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.

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 $X_S[k]$ be the DFT of mapped sequence given by

$$X_S(k) = \sum_{n=0}^{N-1} X_S(n)e^{-2\pi mk/N}$$  \hspace{1cm} (1)

Where $n=0, 1, 2, \ldots, N-1$, $X_S(n) =$ Mapped sequence.

The power spectral density of the sequence is

$$P_s(k) = \sum |X_S(k)|^2$$ \hspace{1cm} (2)

Plot of $P_s(k)$ 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 prediction. Block diagram of the filtering process is shown in the Fig. 5.

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, $wc_1=0.6666$ and $wc_2=0.6667$ (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.

**TABLE 3**

<table>
<thead>
<tr>
<th>Mapping Technique</th>
<th>Numerical Value</th>
<th>Average deviation from period-3 position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary Indicator</td>
<td>x1=1; x2=1;</td>
<td>170.2</td>
</tr>
<tr>
<td></td>
<td>x3=1; x4=1;</td>
<td></td>
</tr>
<tr>
<td>Real Number</td>
<td>a=2; t=0; c=1; g=3;</td>
<td>189.8</td>
</tr>
<tr>
<td>EIIP Code</td>
<td>a=0.1260; t=0.1335;</td>
<td>179.8</td>
</tr>
<tr>
<td></td>
<td>c=0.1340; g=0.0806;</td>
<td></td>
</tr>
<tr>
<td>Molecular Mass</td>
<td>a=134; t=125; c=110; g=150</td>
<td>316.6</td>
</tr>
<tr>
<td>Atomic Number</td>
<td>a=70; t=66; c=58; g=78</td>
<td>151.8</td>
</tr>
<tr>
<td>Paired Nucleotide</td>
<td>a=42; t=42; c=62; g=62</td>
<td>224.6</td>
</tr>
<tr>
<td>Atomic Number</td>
<td>a=1+j; t=1-j; c=1-j; g=1+j;</td>
<td>264.4</td>
</tr>
<tr>
<td>Complex Number</td>
<td>a=j; t=1-j; c=1-j; g=1+j;</td>
<td>139.6</td>
</tr>
<tr>
<td>Pyrimidine Purine Complex</td>
<td>a=j; t=1; c=1; g=1;</td>
<td>139.6</td>
</tr>
</tbody>
</table>

Fig. 6 PSD plots without filter

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 mapping compared to Pyrimidine Purine complex mapping but also noise level is reduces in multirate filter (Table-4).

**TABLE 4 PERFORMANCE COMPARISONS OF DIFFERENT TYPES OF MAPPING TECHNIQUE**

<table>
<thead>
<tr>
<th>Mapping Type</th>
<th>SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>IIR anti-notch with multirate filter using Pyrimidine Purine complex mapping</td>
<td>1.46</td>
</tr>
</tbody>
</table>
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 case in 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