Optimal Nucleotides Range Estimation in Diffused Intron-exon Noise

Muneer Ahmad, Azween Abdullah and Khalid Burraga

King Faisal University, College of Computer Science, Saudi Arabia
2Department of Computer Science, University Technology PATRONAS, Malaysia

Abstract: Nucleotides range estimation in diffused intron-exon noise is a critical and challenging problem whose solution may bring fruitful results for drug design and genetic disorders research. Many solutions have been proposed for approximate bound estimation but an optimal solution is still required. This paper focuses 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. We performed a comprehensive comparative analysis of proposed approach with existing ones over real datasets from NCBI and found significant improvement in prediction accuracy in diffused intron-exon noise and 75% reduction in computations than Binary method. This profound achievement in results may help in achieving the optimal solution for the problem.

Key words: Nucleotide range %Intron-exon noise %Downsampling %Upsampling %1/f noise

INTRODUCTION

DNA sequence contains genes and gene comprises genic and intergenic regions. RNA translation from DNA is an important and critical task because exact identification of protein helps in knowing information regarding protein structure and cell functions. Exons are the regions in gene that are translated to protein and exons boundaries are diffused in intron-exon noise. Optimal identification of exons from 1/f noise needs careful attention and adoption of suitable methodology. Recently statistical and DSP techniques have been proposed for maximization of prediction accuracy in identification of genic regions.

Protein is composed of small scale units called amino acids. There are 20 types of amino acids and the sequence of these units determines the type and function of individual protein molecule.

There are 64 possible codon (tri-nucleotide structure of bases) values that transcribe the DNA chains to protein chains at regions known as exon in several clusters of non genic regions introns. Exons are the regions responsible for carrying nucleotide bases for protein translation. A codon "ATG" identifies the start of the sequence that contains the protein coding regions and codons "TAA", "TGA" and "TAG" are stop codons of this sequence where T is normally replaced with U (called Uracil). It is worth mentioning that mere start codon may not help in protein sequence identification, perhaps some other factors are also required in certain species. Learning the exact location of coding regions leads to the provision of optimal solution of the underlying problem.

According to the concepts of Fourier transforms, a signal can be expressed in the form of summation over sine and cosine which only narrates the frequency components of signal (frequency domain analysis) without any depiction of time domain analysis. All frequency components of a digital signal can be obtained but when these components are present and at which time frame (period of time), this information is lacking in Fourier analysis. The restriction is due to inability to cut the signal into pieces and perform the analysis piecewise over the chopped signal. This problem can be stated as Heisenberg uncertainty principal which stated that it is impossible to get the time information of frequency components and also the occurrence of these components in the specified time duration. A more improved solution can be achieved using wavelet transforms.

Tina P George et al., [1] have performed that discrete wavelet transform can be used for better minimization of noise and maximization of prediction accuracy. Roy et al., [2] proposed a generic algorithm for frequency distribution of various spectral values relevant to individual nucleotide bases. Guo Shuo et al., [3]

Kakumani et al., [6] proposed a method using statistically optimal null filter to maximize the SNR (signal to noise ratio). The presence of exonic region in DNA strands has been detected by employment of least square optimization criteria. Akhtar et al., [7] have demonstrated an optimization for DFT based methods relying on effect of window lengths for signal processing based exonic identification. Hota et., al [8] have proposed a complex indicator sequence methodology for exon prediction in DNA. Akhtar et al., [9] have described a digital signal processing methodology for exonic and intronic region prediction with comparisons to the existing techniques. Grandhi et al., [10] has proposed 2-simplex mapping method for identification of exon regions in DNA. Mena-Chalco et al., [11] have used Modified Gabor-Wavelet Transform for exonic analysis. Gupta et al., [12] have proposed a time series approach for exon and intron prediction. Changchuan et al., [13] have predicted the exonic regions based on period three property of exons. DFT has been used for extraction of Fourier coefficients from four indicator sequences made from the DNA stretch. Vaidyanathan et al., [14] suggested an approach based on antinoch IIR filter and compared results against traditional approaches applying windowed Discrete Fourier Transforms. Datta et al., [15] formulated a fast DFT based methodology for genetic region search in DNA. Hang Chen et al., [16] have predicted protein secondary structure using continuous wavelet transforms and Chou-Fasman method. Suprakash et al., [17] has employed a DFT based algorithm for detection of exonic regions in DNA strand. Suprakash et al., [18] presented an empirical observation of DFT approach for protein regions. Mahmood et al., [19] have compared the existing techniques of exon prediction and formulated the comparison between existing and proposed technique. Vaidyanathan et al., [20, 21] described digital filters for gene prediction in DNA. The designed filter has been used for the prediction of period 3 components and elimination of background noise 1/f spectrum shown by all DNA sequences. Al Dadi et al., [23] used wavelet transforms for forecasting volatility in experimental results. M. Hashemi et al., [24] provided Identification of Escherichia coli O157:H7 Isolated from Cattle Carcasses in Mashhad Abattoir by Multiplex PCR.

Proposed Approach: We have developed a novel criterion for nucleotides range estimation based on a new indicator sequence and introducing the wavelet transforms for denoising the DNA signal. Initially we set the suitable thresholds for the sequence. This sequence is mapped to a DNA signal. The signal is decomposed and synthesized with discrete wavelet transforms. Later we calculate the magnitude, power spectral density that leads to genic regions bounds estimation.

Fig. 1 presents the architecture of proposed approach. It is important to threshold the range of raw dataset for clear comprehension of system and significant results. For the purpose, we have used dataset Sus Scrofa domesticus mitochondrion (Accession: NC_012095 with 7500 base pairs containing four genic regions) for comparative analysis of results between existing and proposed approach.

The novel indicator sequence in conjunction with denoising DNA signal minimizes the 1/f noise to achieve significant results for genic regions identification. We have used dataset Sus Scrofa domesticus mitochondrion (Accession: NC_012095 with 7500 base pairs) for comparative analysis of results between existing and proposed approach.

Fig. 1 Architecture of proposed approach
The proposed indicator sequence is defined as by setting the values for base pairs as follows,

- Adenine (A) = X(A) = 0.260
- Thymine (T) = X(T) = 0.375
- Guanine (G) = X(G) = 0.125
- Cytosine (C) = X(C) = 0.370

The defined novel indicator sequence is a result of minute examination for the formation of clusters containing the tri-nucleotide codon composition in exonic regions.

The discrete wavelet transforms have been used for denoising the signal in terms of decomposition and synthesis.

Mathematically the up-sampling (convolution) can be written as

\[ a(n) = \sum_i cA_i(k)h_i(n-2k) \]
\[ b(n) = \sum_i cD_i(k)h_i(n-2k) \]

(Equation 1)

Repeating the same process as above in inverse and we reach at level 1 of the original transform as

\[ A_i(n) = \sum_k cA_i(k)h_i(n-2k) \]
\[ D_i(n) = \sum_k cD_i(k)h_i(n-2k) \]

The denoising of DNA signal helps in better estimation of discrimination measure and suppresses 1/f noise.

We tested different combinations of window functions for selection of an appropriate window for this analysis and found Kaiser Window of size 351 bp for better minimization of spectral leakage of DNA signal frequency components.

\[ w(n) = I_m\left(\beta\left(1-\left((n-\alpha)/\alpha\right)^2\right)^{\frac{1}{2}}\right) / I_m(\beta) \quad 0 \leq n \leq M - 1 \]
\[ 0 \quad \text{otherwise} \]

(Equation 3)

We found Kaiser Window being the most suitable for frame generation.

Absolute value of Frame = \(|\text{Frame}| = A_x(f) = |X(f)|\)

Where \(X_1(f)\) (Equation 1) calculates the absolute value \(A_x(f) = \sqrt{X^2 + iX^2}\)

Power of Frame = \(|\text{Frame}|^2 = P_x(f) = |X(f)|^2\)

The frequencies are normalized by \(P_x(f) = |X(f)|^2 \frac{1}{fL}\), where \(f_s\) is the sampling frequency and \(L\) is the length of original signal.

**RESULTS AND DISCUSSIONS**

We calculated the power spectral density for proposed and existing approaches over the dataset Sus Scrofa mitochondrion.

Binary indicator sequence method showed the pitfalls concerning diffusion of exons in 1/f noise. From calculations, it is clear that an intron is more visible in regions 400-800 bp. Such an intron can't be seen in other PSD plots. Exon E_2 is not prominent in its entire range; rather it shows its peak from 3800-4200 bp and then a sudden discontinuity occurs with a rejoin between 4200-4900 bp. This phenomenon prevents from correct and transparent calculation of discrimination measure and range of coding regions bounds.

There is more comprehensive plot for EII method than Binary method. The intron peaks are reduced in the range 0 to 1000 bp. The last exon E_4 is having same range as Binary method (6800-7600) bp. Exonic peaks are not larger as compared to first method but the coding regions are more prominent for a comparable range to NCBI.

Complex method contains sharp intron peak in the same regions as Binary method but the coding regions boundaries are more visible that helps for a suitable analysis of discrimination measure and calculations of standard range of nucleotide pairs.

Proposed method in demonstrates the reduction of noise (as well as spectral leakage) and improves the coding regions peaks and boundaries measures for nucleotides. There are no discontinuities found in exonic range and more comparable results are obtained (as set by the standard NCBI).

Table 1 describes the exons range calculated in PSD’s of different approaches.
Table 1: Genic region boundaries for different approaches with NCBI range

<table>
<thead>
<tr>
<th>Method</th>
<th>E₁</th>
<th>E₂</th>
<th>E₃</th>
<th>E₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary Method</td>
<td>2600-3800</td>
<td>3800-4200</td>
<td>4200-4900</td>
<td>5500-6800</td>
</tr>
<tr>
<td>ELIP Method</td>
<td>2600-3500</td>
<td>3600-5000</td>
<td>5100-6800</td>
<td>6800-7600</td>
</tr>
<tr>
<td>Complex Method</td>
<td>2800-3800</td>
<td>3800-4980</td>
<td>6000-7000</td>
<td>7200-7500</td>
</tr>
<tr>
<td>Filter 1</td>
<td>2600-3800</td>
<td>3800-4800</td>
<td>5500-6750</td>
<td>6800-7600</td>
</tr>
<tr>
<td>Filter 2</td>
<td>2600-3950</td>
<td>3600-4900</td>
<td>5000-6750</td>
<td>6800-7550</td>
</tr>
<tr>
<td>UTP Method</td>
<td>2650-3650</td>
<td>3800-4990</td>
<td>5050-6800</td>
<td>6800-7600</td>
</tr>
<tr>
<td>NCBI Range</td>
<td>2746-3702</td>
<td>3911-4954</td>
<td>5335-6879</td>
<td>7027-7722</td>
</tr>
</tbody>
</table>

Fig. 2 demonstrates the power spectral density comparisons between Complex method and proposed approach. It is visible that Complex method has more prominent peaks for introns shown in red curve. Two intergenic regions in the range 0-1000 bp and 1000-2000 bp contain higher values compared with green curve. All genic region peak heights in green curves are higher than red curves. This phenomenon is the obvious demonstration of distinguishing factors between the two methods.

**CONCLUSION**

We have proposed a novel approach for nucleotides range estimation in diffused intron-exon noise. We incorporated the notion of discrete wavelet transforms for denoising our DNA signal along with approximate mapping of signal with new indicator sequence. We have calculated the bound estimation for nucleotide in power spectral density estimation graphs. The sharp curves for genic regions identify the nucleotide range of exons in 1/f noise. A comparative analysis of results tested over DNA sequence Sus Scrofa domesticus mitochadrion (Accession: NC_012095) revealed the significant gain in prediction measures for proposed architecture as against the existing solutions. This outperformance of proposed system is bestowed by hybridization of wavelet and DNA mapping. In future, the approach will be extended by introducing the...
concepts of discrimination measure to reveal the degree of difference in PSD estimation in terms of genic and intergenic peaks.

REFERENCES


