Rabia *et al. (IJECI)* 2023

International Journal for
Electronic Crime Investigation

DOI: https://doi.org/10.54692/ijeci.2023.0703160

(IJECI)
ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

Research Article

Vol. 7 issue 3 Jul-Sep 2023

Genomic Signal Processing Methods in DNA Mapping Schemes for Prediction of Exon in a Gene Using Digital Filters

Rabia Aslam Khan[1], Muhammad Bilal But[2] and Sabreena Nawaz[3]

[1] University of Management and Technology, Lahore

[2] University of South Asia, Lahore

[3] University of Engineering and Technology, Lahore

Corresponding author: f2019288013@umt.edu.pk

Received: June 29, 2023; **Accepted:** August 28,2023; **Published:** September 20, 2023

Abstract:

Genomic signal processing (GSP) is an engineering domain involved with the analysis of genomic data using digital signal processing (DSP) approaches after transformation of the sequence of genome to numerical sequence. One challenge of GSP is how to minimize the error of detection of the protein coding region in a specified deoxyribonucleic acid (DNA) sequence with a minimum processing time. Since the type of numerical representation of a DNA sequence extremely affects the prediction accuracy and precision. The impact of different DNA statistical representations on the identification of coding sequences (exons) was researched. In this study using the IIR inverse Chepyshev filter for twenty benchmark human genes. In order to accomplish this, the sensitivity, specificity, and correlation coefficient of the four most modern DNA numerical representation schemes GCC, FNO, atomic number, and 2-bit binary were measured and contrasted with those of EIIP, the most used technique for locating protein-coding regions

Keywords: Genomic signal processing, DNA, exons, numerical sequence, atomic number

1. Introduction

Short exon detection is a formidable issue for bioinformatics and becomes more complicated as becomes more complicated side of short intron. To categorize these exonic regions accurately, it's essential to create computer methods that are both more efficient and dependable. This is necessary because many of the existing methods do not handle the small exons separated by brief introns effectively. The methods for identifying exons are based on the quest for material, signal or resemblance. For classification of exon disunited by short intron has been divided into two methods; Model independent and model dependent [1]. The DNA coding model frequently relies on probability, enabling the measurement of the likelihood of a DNA sequence because it encodes the sequence.

Although the values (scores) of a specific data code statist are calculable in a variety of different ways in reality, we will measure scores based on this probability for model-based coding statistics. In fact, provided the query sequence, under the coding model and an alternate model or DNA non coding we can determine the likelihood of the sequence. The model-based coding statistics may catch more of the particular DNA-coding characteristics, more as the model is more complex i.e. more parameters dependent. Model based coding statistics can also be more effective in distinguishing against non-coding DNA coding. However, model based coding statistics involve a representative DNA coding sample from the species included in the estimation of model parameters (probabilities). The more intricate the model, the more susceptible it is to sample distortion and dimension. Model independent coding statistics, however, capture only the "universal" characteristics of DNA coding, as no sample is needed and where coding regions of the species being considered are not identified, they may be used[2]. In [3], they have used Markov Chain to identify the sequences in DNA. Markov chain models of DNA and its use for Bayesian gene recognition algorithms for protein coding sequences. Gene Scout is the other method for detecting DNA sequences that used Markov Chain. In recent work, the local spectrum of the first intrinsic mode feature was determined to detect short exons. A technique focused on filters was also documented in order to detect short exons [4]. However, this method is based on the model by evaluating the fictitious EIIP values the fictitious EIIP values the fictitious EIIP values optimised and the weights for the four filtered binary sequences. Depending on the study of the windows form and scale, the efficiency of DSP bases that DFT can be used to analyse the spectral properties of DNA sequence depends [5].

2 Literature Review

A more concise timeframe can detect short exons, but not long exon scan lead to further false alarm. On the other side, wide windows can lead to fewer fake detections, however short exons are lacking. Multiscale analysis was conducted by MGWT-based approach [6]. Marhon & Kremer recently suggested the Broad Range Wavelet Window (WRWW) approach to the forecasting of protein coding areas in a recent work [7]. In order to deal with the problem of window size. A technique to fix the issue of window size selection was also introduced to adapt the window length [8]. The WRWW approach has been shown to operate effectively over a number of exon lengths through simulation experiments. The effectiveness of the methods used for detecting exons has not yet been assessed when there is a brief intron separating two adjacent short exons. Furthermore, no computer model to identify alternate splicing that could occur due to intron retention has been investigated for implementation of the annotation of certain regions in eukaryotic DNA (IR)[9]. In IR, part of the gene is not encrypted and can join premature stop codons in the center of a mature transcription. In an IR, numerous factors such as weak splice sites, short introns within genes, elevated levels of exonic splicing silencing, and lower density can contribute to the occurrence of IR [10].

Additionally, the IR is linked to short introns (274 nucleotides) and, if retention takes place, all neighboring exons, which are about 135 nt long, are linked to the exon retained, creating an exon retention intron (EIE) exon that is 544 nt long. In order to find IR-likely sites, short exons separated by short introns can be identified using computer-based methods.

3 Dna Mapping Scheme

"Deoxyribonucleic acid (DNA)" sequences are important for the understanding of living organisms, and in these macromolecules, much of the knowledge concerning heritable evolution and species growth is stored. Prokaryotes and eukaryotes are possible for organisms. DNA is free inside the cell in prokaryotes while DNA is retained within the nucleus in eukaryotes and is disassociated by a nuclear memebrane from the rest of the cell. Four major chemicals, thymine (T), cytosine(C) guanine (G) and adenine (A) form the DNA chain . The determination of protein coating regions (exons) in eucaryotic gene structures is one of the present problems in studying the DNA sequences. Both probabilistic and deterministic approaches are employed to categorize protein coding regions or exons in eukaryotic cells. Probabilistic methods have high precision, but rely on model and require adequate prediction training data. In the other hand, predictability of detergent methods is comparatively lower but model-independent and best suited for study of uncharacterized genomic sequences, where prior details of the studied species does not exist.

The base-coding region contains a pronounced period-3 segment attributed to the codon structure utilized in the translation of the base sequence into amino acids. Most deterrent techniques use the "Discrete Fourier Transform" to classify the period-3 portion by spectral analysis of the DNA sequences. A variety of algorithms were designed to classify protein-coding regions based on the period-3 property. DFT-based approaches efficiency depends on the duration of the window[11]. In order to classify protein-coding areas, a system based on "Modified Gabor-wavelet transform" (MGWT) was implemented. Depending on window length, the efficiency of the MGWT is higher than the DFT based approaches[6].

There are four significant shortcomings in the present method for representing and aligning new input genomes with the reference genome. To begin, even though several algorithmic implementations are widely used, there is no established standard method for aligning DNA bases from a newly sequenced input genome with positions in the reference genome [12].

Secondly, various mapping procedures encounter a challenge when there are (almost) equally valid mappings to multiple separate positions within the reference genome, a situation often referred to as the "multi-mapping problem." This arises because of the inherent repetition of larger subsequences in the reference genome.

Thirdly, the GRC reference genome encompasses only a limited portion of common segregating genome variations, with the remainder scattered across various formats and data sources like the Single Nucleotide Polymorphism Database (dbSNP) and the

1000 Genomes Project. Consequently, there is presently no singular, all-encompassing resource for common human genome variations, and there is a lack of consistent naming or identification conventions [13].

Lastly, whenever a new reference genome assembly is issued, updates are made to the reference genome's coordinates, necessitating the remapping of all associated data. This remapping process is often the most computationally intensive stage in a genome analysis pipeline. It can be a time-consuming task, taking weeks to complete and consuming substantial computational resources, particularly when dealing with a large set of genomes.

4 Representation Of DNA

The following five representation methods were used to numerically represent the sequences of the selected genes DNA:

4.1 *Genetic Code Context (GCC)*

The following triple codons are found in the various reading frameworks for a particular DNA sequence Y= ACGATTCAGGT: The initial reading phrase is ACG ATT CAG, followed by CGA TTC AGG and finally by GAT TCA GGT.The corresponding encoded amino acids for the first frame are [T, I, Q], [R, F, R], and [D, S, G] for the second and third frames, respectively. Each amino acid is described by a unique complex number, as shown in Table 1.

	Amino Acid	Number Representation
1	Ala (A)	0.61+88.3i
2	Cys (C)	l.07+112.4i
3	Asp (D)	0.46+II 0.Si
4	Glu (E)	0.47+140.Si
5	Phe (F)	2 02+189i
6	Gly (G)	0.07+60i
7	His (H)	0.61+152.6i
8	lie (I)	2.22+168.Si
9	Lys (K)	Ll 5+175.6i
10	Leu (L)	l.53+168.Si
11	Met (M)	Ll 8+162.2i
12	TyT (Y)	l.88+193i
13	Trp (W)	2.65+227i
14	Val (V)	l.32+141.4i
15	Pro (P)	l.95+122.2i
16	Asu (N)	0.06+125.li
17	Arg (R)	0.60+181.2i
18	Ser (S)	0.05+88.7
19	Thr (T)	0.05+118.2i

4.2 Frequency of Nucleotide Occurrence

According to Table 2 given below, A real value is assigned to each nucleotide in the DNA sequence Y= ACGATTCAGGT from two different datasets. As a result, the corresponding DNA numerical sequence from the HMR195 dataset is [0.22750, 0.28312, 0.27600, 0.22750, 0.21336, 0.21336, 0.28312, 0.22750, 0.27600, 0.27600, 0.21336, 0.28312, 0.22750, 0.27600, 0.27600, 0.21336].

Data Set	**Frequency of Occurrence**			
	A	**C**	**G**	**T**
Burset	0.243	0.27215	0.27909	0.20576
HMR 195	0.2275	0.28312	0.276	0.21336
OCTN2	0.243	0.27215	0.27909	0.20576
MTA1-L1	0.2275	0.28312	0.276	0.21336
hCLCA1	0.243	0.27215	0.27909	0.20576
LCC-1 precursor	0.2275	0.28312	0.276	0.21336

4.3 Atomic Number

The molecular signature pattern constants over a certain DNA sequence Y= ACGATTCAGGT are: A=70, G=78, C=58, T=66. As a consequence, the numerical sequence of DNA is [70, 58, 78, 70, 66, 66, 58, 70, 78, 66, 58, 70, 78, 66, 58, 70, 78, 66, 58, 70, 78, 66, 58, 70, 78, 66, 58, 70, 78, 66, 58, 70, 78, 66].

4.4 *Electron Ion Interaction Potential (EIIP)*

The EIIP indicator sequence values for the specific DNA sequence Y= ACGATTCAGGT are A= 0.1260, G= 0.0806, C= 0.1340, and T= 0.1335.As a result, [0.1260, 0.1340, 0.0806, 0.1260, 0.1335, 0.1335, 0.1340, 0.1260, 0.0806, 0.0806, 0] is the numerical sequence for DNA .1335]

4.5 *2-bit Binary*

The values of the DNA the 2-bit digital sign sequencesY= ACGATTCAGGT are A=00, G=10, T=01, C=11 for the DNA sequence Y= ACGATTCAGGT.

5 Results

The detection technique was applied using the IIR inverse Chepyshev electronic filter on 20 human testing genes with single and multiple exons downloaded from the HMR195 dataset.in order to achieve our goal. The accession numbers, gene descriptions, sequence lengths, and true exon locations of the genes are all displayed in Table 3.

Gene Accession No.	**Sequence lengths**	**Gene Description**	**True Exon Location**
One Exon Gene			
AF009731	702	C)' ochrome b (C)' b) gene of Lepussaxatilis	1-702
AF007189	1601	CLDN3 (Homo sapiens ciaudin 3) gene	477-1139
AF071552	1618	Homo sapiens N-acetyitransferase-1 (NATI) gene	44 1-1313
AF055080	2078	Winge.d-heiix transcription factor forkhead 5 gene in Homo sapiens	964-1938
AF009962	7422	CCR-5 (CC-chernokine receptor) gene in Homo sapiens	3934-4581

Two Exon Gene			
AF061327	1812	D pl 9 gene of Homo sapiens cyclin-dependent kinase 4 inhibitor	13-153 1245-1604
AF058762	3036	Homo sapiens galanin receptor subty']le 2 (GAL'lJRl) **gene**	115-482 1867-2662
AF042782	3390	GALR2 (Homo sapiens galanin receptor ty']le 2) gene	305-672 2063-2858
AF058761	3607	S12 ribosomal protein gene in Homo sapiens	1815-1863 2854-3221
AF092047	4477	SIX3 (Homo sapiens homeobox protein) gene	1275-2080 3740-3932
Three Exon Gene			
AF076214	4002	Homo sapiens prophet of PitI (PROPI) gene	310-4 18 1901-2133 3191-3529
AF042001	4034	The zinc finger protein slug (SLUG) gene in Homo sapiens	447-525 1271-1816 2724-2905
AF015224	4206	Homo sapiens mammaglobin gene	1056-1110 1713-1900 3789-3827
AF036329	4498	Gonadotropin-reie.asing hormone in Homo sapiens	2105-2258 2369-2526 3372-3422
AF028233	4575	Homo sapiens distal-less homeobox protein (DLX3)	68- 392 1483-1673 3211-3558
Four Exon Gene			
AF059734	2401	Gene for Homo sapiens homeodomain transcription factor (HESXI).	335-491 1296-149 1756-1857 1953-2051
AF013711	5388	Gene for Homo sapiens 22 kDa actin-binding protein (51122).	3643-3822 3935 4112 44 10- 4512 4843- 4987
AF045999	5895	The rod cGMPphosphodiesterase delta subunit (PDEd) gene of Homo sapiens	159-297 1257-1382 2103-2208 5296-537
AF037062	6330	Homo sapiens retinol dehydrogenasegene	2372-2681 2876-3134 5065-5228 5501-5724
AF055475	9531	Homo sapiens GAGE-7B gene	2226 -2309 2776-2896 5718-5843 8279-8301

5.1 Single Exon Gene

Both the frequency of nucleotide occurrence in exons (FNO) and the 2-bit binary representation schemes showed a distinct and prominent peak at the precise location of true exons (964-1938), without any misleading peaks at the individual exon level, when compared to EIIP, GCC, and atomic number schemes. Additionally, the FNO and 2-bit binary representation schemes demonstrated the highest levels of sensitivity, specificity, and correlation coefficient for various single exon genes, achieving 100 percent, 75.228 percent, and 0.4994, respectively. Notably, the 2-bit binary representation scheme clocked in at 7.38ms, which was the quickest processing time when compared to the other representation methods.

5.2 Two Exonic Region Gene

Different genes with two exonic regions were used to test the predictive accuracy of different representation techniques. Surprisingly, nucleotide location identification and sensitivity for the FNO and 2-bit binary techniques were identical, as shown in Fig. 6. These two techniques successfully located the two authentic exons (at locations 115–482 and 1867–2662) within the (GALNR2) gene.

The FNO and 2-bit binary methods fared better than other schemes, with specificity scores of 56.012 percent and 65.02 percent, respectively, despite having almost half the specificity of single exonic region prediction. Interestingly, among all the representation techniques, the 2-bit binary representation approach had the highest correlation coefficient (0.6838) and the fastest processing speed.

5.3 Three Exonic and Four Exonic Region Gene

When applying five various recognition algorithms to genes with three and four exonic regions, the 2-bit binary representation method consistently beat other representation methods in terms of accuracy. The number of incorrect exons was reduced by this method's accurate detection of actual nucleotide locations in the proper order. As a result, in this particular situation, it achieved the highest levels of sensitivity, correlation, specificity, and CPU run time.

6 Conclusion

The findings demonstrated that the 2-bit binary representation method, when compared to other representation schemes, significantly improved true nucleotide position identification accuracy regardless of the number of exonic regions in the sequences tested, with high levels of sensitivity, correlation coefficient, specificity, and minimal processing time. These results are consistent with other studies that applied the 2-bit binary technique in a different setting. When applied to human DNA sequences for promoter prediction using neural networks, it was found that the 2-bit binary scheme outperformed the 4-bit binary and integer representation methods.

Intriguingly, the 2-bit binary and FNO representation schemes both displayed comparable high levels of sensitivity, correlation coefficient, and specificity when compared to other schemes, especially at the one and two exonic region detection levels, despite using different numerical representation techniques. Notably, the FNO system is based on statistically derived measurements while the 2-bit binary scheme relies on the arbitrary assignment of nucleotide numbers.

The FNO was outperformed by the 2-bit binary representation approach for the detection of three and four exonic areas. These results corroborate a previous study that found that the protein coding region prediction accuracy could be improved by using the DFT base technique by increasing the frequency of nucleotide occurrence and matching numeric recognition schemes.

7 References

[1] A. D. Baxevanis and B. F. F. Ouellette, Bioinformatics_-a-practical-guide-to-the-analysis-of-genes- and-proteins, 2nd ed. Wiley Interscience 2004

[2] R. Guigó, "DNA Composition, Codon Usage and Exon Prediction," Academic Press 1997.

[3] Mx. Borodovsky and J. Mcininch, "GENMARK: PARALLEL GENE RECOGNITION FOR BOTH DNA STRANDS," Computers Chem, vol. 17, pp. 123-133, 1993

[4] N. Y. Song and H. Yan, "Short Exon Detection in DNA Sequences Based on Multifeature Spectral Analysis," EURASIP Journal on Advances in Signal Processing, vol. 2011, no. 1, 2010.

[5] P. Ramachandran, W. S. Lu, and A. Antoniou, "Filter-based methodology for the location of hot spots in proteins and exons in DNA," IEEE Trans Biomed Eng, vol. 59, no. 6, pp. 1598-609, Jun 2012.

[6] J. P. Mena-Chalco, H. Carrer, Y. Zana, and R. M. Cesar, Jr., "Identification of protein coding regions using the modified Gabor-wavelet transform," IEEE/ACM Trans Comput Biol Bioinform, vol. 5, no. 2, pp. 198-207, Apr-Jun 2008.

[7] S. A. Marhon and S. C. Kremer, "Prediction of Protein Coding Regions Using a Wide-Range Wavelet Window Method," IEEE/ACM Trans Comput Biol Bioinform, vol. 13, no. 4, pp. 742-53, Jul-Aug 2016.

[8] D. K. Shakya, R. Saxena, and S. N. Sharma, "An adaptive window length strategy for eukaryotic CDS prediction," IEEE/ACM Trans Comput Biol Bioinform, vol. 10, no. 5, pp. 1241-52, Sep-Oct 2013.

[9] A. J. Matlin, F. Clark, and C. W. Smith, "Understanding alternative splicing: towards a cellular code," Nat Rev Mol Cell Biol, vol. 6, no. 5, pp. 386-98, May 2005.

[10] N. J. Sakabe and S. J. de Souza, "Sequence features responsible for intron retention in human," BMC Genomics, vol. 8, p. 59, Feb 26 2007.

[11] S. D. Sharma, K. Shakya, and S. N. Sharma, "Evaluation of DNA mapping schemes for exon detection," International Conference on Computer, Communication and Electrical Technology – ICCCET, 2011.

[12] W. J. B. Matei Zaharia, Kristal Curtis, Armando Fox, David Patterson, Scott Shenker, Ion Stoica, Richard M. Karp, Taylor Sittler‡∗, "Faster and More Accurate Sequence Alignment with SNAP," 2011.

[13] C. Genomes Project et al., "A map of human genome variation from population-scale sequencing," Nature, vol. 467, no. 7319, pp. 1061-73, Oct 28 2010.