DOI https://doi.org/10.30898/1684-1719.2021.1.11
UDC 615.84:621.39:004.3

DIGITAL SIGNAL PROCESSING-BASED APPROACH TO IDENTIFY SPLICING MUTATIONS FOR DETECTING GENETIC DISEASES

P. Kumar Varadwaj [1], N. Purohit [1], T. Lahiri [1], V. Antisiperov [2]

**[1] Indian Institute of Information Technology Allahabad,
Devghat, Jhalwa, Prayagraj-211015, U. P. India
[2] Kotelnikov Institute of Radioengineering and Electronics,
Mokhovaya 11-7, Moscow, 125009, Russia**

The paper was received on January 12, 2021

Abstract. More than ninety percent of genes in Homo Sapience are reported to exist as discontinuous segments of coding regions called as ‘Exons’ and are separated by intervening non-coding regions, called “Introns”. During the splicing mechanism, the non-coding regions got removed and coding regions are joined together for producing the precursor messenger RNA. The site of these Exon-Intron splicing is called Splice Site. The anomalies caused due to genetic mutation in spice site during the processing of precursor m-RNA into mature m-RNA causes several genetic diseases like Cancer, Dementia, Epilepsy, Hematological Disorders, Parathyroid Deficiency etc. It is estimated that as many as 50% of disease-causing mutations affect splicing. The present invention describes the design of digital signal processing-based approach to detect these Splicing Site. A successful identification of the splice site will help in finding the mutations hence can be used as an inference tool for predicting genetic disease.

Keywords: biomedical signal processing, coding regions, splicing, statistic inference.

Introduction

The article is devoted to the development and design of a new approach to predicting complex structures in biomedical data, including the structures of coding regions in genomic sequences. More specifically, this approach could also be used for detecting the multiple exonic locations in a DNA sequence. The entire approach can be divided into 1) Hybrid Numerical mapping which include three numerical mapping techniques namely novel information theory based variable mapping, Integrated EIIP and Angle based Paired mapping that convert base pair DNA sequence to numerical sequence & 2) Nonparametric based thresholding technique for processing the Multi scaled MGWT transformed data to differentiate exonic and intronic regions in a DNA sequence.

1. Novel Information Theory Based Variable Numerical Mapping for Single Nucleotide and Dinucleotide

The nucleotide occurs in the genomic sequences with different probabilities. The occurrence of nucleotides is biased for exons and introns. So, we have encoded the probability of each nucleotide to find number of bits for each nucleotide in genomic sequences. The value of encoding is varying with sequences. Here we are explaining procedure to encode for single nucleotide for a particular DNA sequence. Same procedure is followed for dinucleotide which will be another new numerical mapping that will help to find out the frequency characteristic of dinucleotide.

A DNA sequence consists of four nucleotides $X[j] \in \{A,T,G,C\}$, $0 \le j < N$, where N represent the length of the given DNA sequence X. The DNA sequence is mapped into four signals depending on numbers of bits required to encode the nucleotide $(UA, U\,T, UG$ and $UC)$. These four signals represent the encoding for each of the nucleotide for particular signals having length N.

$$UA = -\log_2 P(A)$$
$$UT = -\log_2 P(T)$$
$$UG = -\log_2 P(G)$$
$$UC = -\log_2 P(C)$$

where, $P(x)$ represents the probability of occurrence of nucleotide x in the sequence X and $X[j]$ represents mapped sequence.

In a given DNA sequence, we map the sequence with this variable mapping method for the presence and absence of nucleotide. For example, given the DNA sequence

$$ATGCGATGACTA,$$

the mapped sequences for each base A, T, G and C are respectively:

$$X = UA\ UT\ UG\ UC\ UG\ UA\ UT\ UG\ UA\ UC\ UT\ UA,$$

The main advantage of this mapping is that it gives the encoding of each nucleotide and dinucleotide occurrence in the string. In this mapping, biological structural information and statistical information for each DNA sequence will be combined together which helps to improve the accuracy. The mapping into numeric sequences varies with different genomic sequences.

The paper [1] introduced paired numeric mapping which maintain DNA structural property of exon and intron content. It claims that exons are rich in C and G whereas introns are rich in A and T. Based on this complementary mapping indicating $A = T = 1$ and $C = G = -1$ are introduced [1]. As we are interested in Exonic region, instead of giving equal magnitude to all nucleotide (A, C, G and T) with opposite sign, we modified the mapping by giving the angle of $1 - j$ to C and G indicating $C = G = \text{angle}(1 - j)$ and $A = T = \text{angle}(1 + j)$. This helped to increase 1% accuracy of previous paired mapping which replace 1 and -1 for C, G and A, T.

2. Modified Integrated EIIP Values

This mapping is developed based on the concepts of three nucleotide based potential mapping. In EIIP mapping, Electron Potential of single nucleotide is used to carry the biological information. If we combine EIIP values of every three nucleotide in sliding manner, uniqueness of each codon (three nucleotide) for each DNA sequence can be carried from DNA sequence to numerical values.

1. EIIP numerical mapping (A=0.1260, C=0.1340, G=0.0806, T=0.1335) is applied for the given DNA sequence. For example, given the DNA sequence $ATGCGATGACTA$, Integrated EIIP will be

$$\begin{gathered}[0.1260\ 0.1335\ 0.0806\ 0.1340\ 0.0806\ 0.1260\\ 0.1335\ 0.0806\ 0.1260\ 0.1340\ 0.1335\ 0.1260]\end{gathered}$$

2. Add EIIP values of 3 nucleotide and move one by one to add every 3 nucleotide of DNA sequence, and then Integrated EIIP will be

$$\begin{gathered}[0.3401\ 0.3481\ 0.2952\ 0.3406\ 0.3401\ 0.3481\\ 0.2952\ 0.3406\ 0.3935\ 0.3935\ 0.2595\ 0.1260]\end{gathered}$$

Integrated EIIP values of all 64 codons are given in the following table for reference.

Table 1. Integrated EIIP values of all codons

	A (0.1260)	IEIIP	C (0.1340)	IEIIP	G (0.0806)	IEIIP	T (0.1335)	IEIIP	
A 0.1260	AAA	0.3780	ACA	0.386	AGA	0.3326	ATA	0.3885	A
	AAC	0.3860	ACC	0.3940	AGC	0.3406	ATC	0.3935	C
	AAG	0.3326	ACG	0.3406	AGG	0.2872	ATG	0.3401	G
	AAT	0.3885	ACT	0.3935	AGT	0.3401	ATT	0.3930	T
C 0.1340	CAA	0.3860	CCA	0.3940	CGA	0.3406	CTA	0.3935	A
	CAC	0.3940	CCC	0.4020	CGC	0.3486	CTC	0.4015	C
	CAG	0.3406	CCG	0.3486	CGG	0.2952	CTG	0.3481	G
	CAT	0.3935	CCT	0.4015	CTA	0.3935	CTT	0.4010	T
G 0.0806	GAA	0.3326	GCA	0.3406	GGA	0.2872	GTA	0.3401	A
	GAC	0.3406	GCC	0.3486	GGC	0.3935	GTC	0.3481	C
	GAG	0.2872	GCG	0.2952	GGG	0.2418	GTG	0.2947	G
	GAT	0.3401	GCT	0.3481	GGT	0.2947	GTT	0.3476	T
T 0.1335	TAA	0.3855	TCA	0.3935	TGA	0.3401	TTA	0.3930	A
	TAC	0.3935	TCC	0.4015	TGC	0.3481	TTC	0.4010	C
	TAG	0.3401	TCG	0.3481	TGG	0.2947	TTG	0.3476	G
	TAT	0.3930	TCT	0.4010	TGT	0.3476	TTT	0.4005	T

3. Hybrid Mapping Technique

All three numerical mapping (Information based mapping, Integrated EIIP, Angle based paired mapping) individually gives better result. In our protocol, we combine all three numerical mapping and after that MGWT Transform is applied, which carries the important characteristic of all the numerical mapping [2].

4. Proposed Thresholding Technique

Third phase of gene prediction process deals with thresholding technique. Selecting proper threshold will be an extremely critical issue in the analysis of DNA sequence as it deals with the boundary of Exon and Intron so that we can differentiate exon and intron with more accuracy.

Currently we are not having standard procedure to fix the threshold for all DNA sequence, for example, in MGWT based method [19], mention the range 60% to 85% will give the better result. But this range will differ for each DNA sequence.

In this, we propose a new thresholding that is based on potential and frequency variation of Exonic and Intronic region.

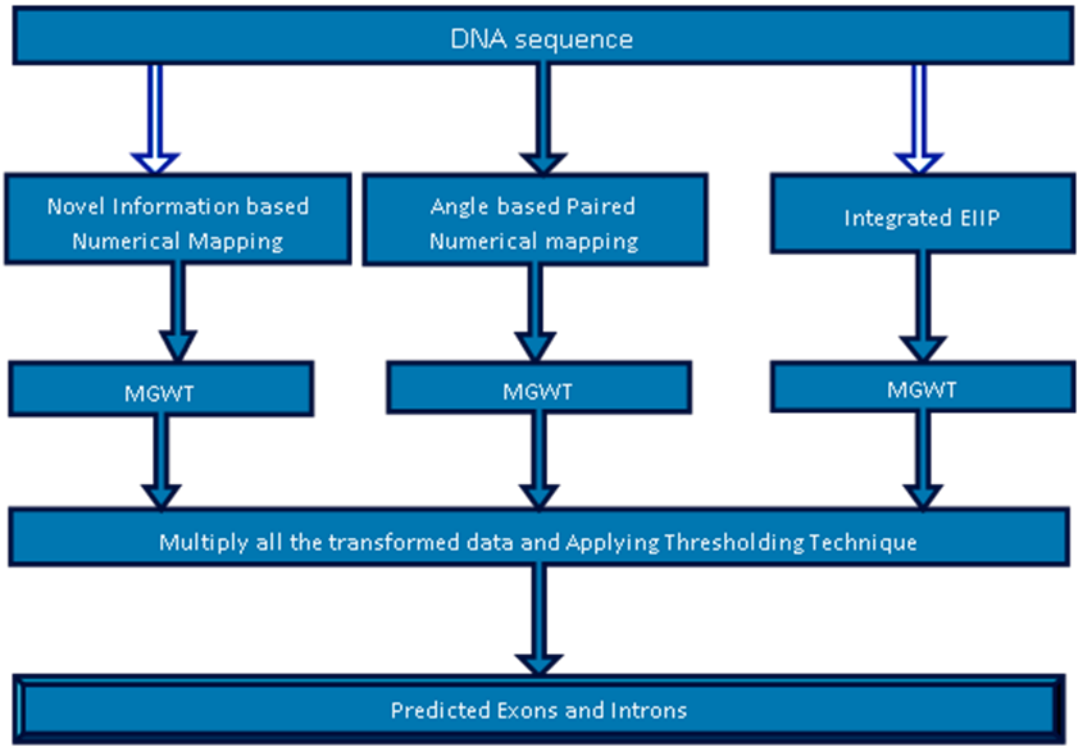
DNA sequence
Novel Information based Numerical Mapping
Angle based Paired Numerical mapping
Integrated EIIP
MGWT
MGWT
MGWT
Multiply all the transformed data and Applying Thresholding Technique
Predicted Exons and Introns

Fig.1. Flowchart of protocol.

This section describes step by step procedure of our proposed protocol:

1. DNA sequence is given as an input in the form of "ACGACCT".
2. Novel Information theory based variable numerical mapping, Angle based paired Mapping, Integrated EIIP is applied to given DNA sequence. (Procedure is already explained).
3. Modified Gabor wavelet transform is applied to all numerically mapped sequence using the formula,

$$\varphi_{MGWT}(x,a,b) = \exp\{-(x-b)^2/2a^2\} * \exp\{-iw_0(x-b)\}$$

4. Multiply all the transformed values.
5. Apply periodogram based thresholding techniques.
6. This will give the set of exons range.
7. Apply moving average filter which is FIR filter that will reduce the smaller fluctuation and helps to improve longer cycles tendency.
8. After applying smoothing moving average filter, we get set of exon range that may contain some intron also which can be identified by checking the CG content and frequency distribution of each dinucleotide. The identified Exons are categorized into optimal and Sub-optimal based on its CG content, frequency distribution of dinucleotide.

5. Result

Our protocol is analyzed with the world standard datasets. HMR195 Dataset gives 92% nucleotide accuracy, Genescan Dataset gives 93% nucleotide accuracy, Burset and Guigó dataset of 570 human genes (BG-570) gives 92 %. Dataset is available at URL ftp://ftp.cse.ucsc.edu/pub/dna/genes. Overall Accuracy for all datasets HMR195, Genescan, Burset and Guigo is given below, in the Table 2

Table 2. Overall Accuracy for 3 datasets namely

Dataset	HMR195	Genescan	Burset and Guigo
Accuracy	92.36%	93%	92.4%

6. Conclusion

The accuracy predicted by our developed protocol is 93.26% on HMR195 dataset which is significantly higher than any other existing protocols. Main advantage of our protocol is its speed: i.e. 1. it works very fast as no training time is required, 2.novel annotation can be possible from new species. 3. Useful for large scale genome annotation in post-genomic era. , 4. can be developed as stand-alone tool which can be interfaced to DNA sequencing machine. Overall, the proposed protocol will be useful in future work for prediction of exon using DSP methods. We can also extend proposed protocol by for different genomic dataset in future by including biological statistical information.

Most of the Gene prediction tools combing approaches are based on a set of rules that are created based on co-relation between different gene prediction tools and require large amount of training data and time. Our new hybrid combining approach in which we used position based nucleotide analysis using multi scaled MGF function with different scales values to overcome this limitation and this approach showed new way to analysis and combines multiple tools. As an initial step, we achieved 10 - 12 % higher exon level accuracy and prediction of missed exons and identification of wrong exons are improved in our hybrid approach by 4 % and 1 % respectively. In future this algorithm will be included with more tools which are designed for different vertebrates and invertebrates so that we make this tool to perform better for any species rather than existing gene finding approaches and we can utilize the advantages of multi scaled analysis using Gabor wavelet function to combine different tools that will provide better nucleotide accuracy and exon level accuracy for all dataset.

References

1. Akhtar, J. Epps, E. Ambikairajah. Signal processing in sequence analysis. *IEEE Journal of Selected Topics in Signal Processing.* 2008, Vol.2(3). P.310-321. https://doi.org/10.1109/JSTSP.2008.923854
2. Saxena A., Pitchaipillai G., Vardawaj P.K. Annotation of Human Genomic Sequence by Combining Existing Gene Prediction Tools Using Hybrid Approach.

In: Parashar M., Kaushik D., Rana O.F., Samtaney R., Yang Y., Zomaya A., editors. Contemporary Computing. IC3 2012. Communications in Computer and Information Science, 2012. Vol 306. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-32129-0_48

For citation:

Varadwaj P.K., Purohit N., Lahiri T., Antisiperov V. Digital signal processing-based approach to identify splicing mutations for detecting genetic diseases. *Zhurnal Radioelektroniki* [Journal of Radio Electronics]. 2021. No.1. https://doi.org/10.30898/1684-1719.2021.1.11