International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020
ISSN: 1475-7192

Genomic Analysis using Higher Order Adaptive Exon Predictors

[1]Srinivasareddy Putluri, [2]Nagesh Mantravadi, [3]Md. Zia Ur Rahman

***ABSTRACT**--In genomics, true identifying exon regions in deoxyribonucleic acid (DNA) sections are an important activity for the identification and development of disease medications. All exon identification techniques are based on three basic periodicity (TBP) properties of exons. The techniques of adaptive sign processing have been successful compared to various other methods. This paper uses the least mean fourth (LMF) algorithm also its signed variants that includes SRLMF, SLMF also SSLMF algorithms to develop multiple adaptive exon predictors (AEPs) with less computational complexity. Eventually, a performance evaluation is performed for different AEPs using various standard gene data sequences derived from National Biotechnology Information Centre (NBI) genomic sequence database, such as Sensitivity (Sn), Precision (Pr) and Specificity (Sp) measurements.*

***Keywords**--adaptive exon predictor, computational complexity, deoxyribonucleic acid, disease medications, exon, three base periodicity*

I. INTRODUCTION

Genomics remains as an immense field in which areas that code for proteins are identified using smart AEP based system presented here. In determining diseases and for drug design, exon areas have a role to play. DNA sequence contains intergenic and genic sections [1]. In order to sustain both tertiary and secondary exon segments the structure of the principal protein segments is examined. After this has been determined for the entire exon segments, all anomalies and health problems are likely to be detected [2] [3]. The prokaryotes and eukaryotes remain segregated from all living organisms. Among the eukaryotic segments, exons remain the protein coding segments, whereas introns remain considered as non-protein coding sections. Just 3% of the eukaryotic human gene sequence has coding areas and rest remain non-coding areas. Hence, it is an important task to detect coded components in a gene sequence [4] [5]. Some literary methods therefore rely on specific techniques of signal processing [6]-[10].

Adaptive methods using AEP based system in a number of iterations may process more lengthy sequences. In our research we are currently introducing a new AEP with use of Least Mean Fourth techniques. LMF is considered to minimize AEP performance over LMS with its signed variants. LMS drawbacks are resolved by LMF algorithm, thereby increases speed and ability of exon tracking. Excess mean square error (EMSE) also minimizes during exon identification [11]. Sign-based algorithms also reduce the sign function by quantity of multiplication calculations [12]. However, several errors do not meet monitoring criteria because of data-

[1] *Department of ECE, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur – 522002*

[2] *Department of ECE, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur – 522002*

[3] *Department of ECE, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur – 522002, mdzr55@gmail.com*

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

independent steady step-size techniques [13]. Lower EMSE and larger step size are necessary for the best convergence rate. Disadvantages of LMS are overcome by use of LMF based techniques.

In cases where the step size of various adaptive algorithms in [14]-[18] is prohibited, this error in the iteration technique is seen. These techniques are better than the variants of the LMS method. In order to lessen computational complexity, LMF techniques remain combined with sign based algorithms. Hybrid variants with presented AEPs comprise error sign regressor LMF (SRLMF), sign LMF (SLMF), as well as sign sign LMF (SSLMF) methods. Proposed AEPs are accurately evaluated using the actual genomic database data set [19]. Performance metrics remains precision (Pr), convergence characteristics, specificity (Sp), computational complexity, also sensitivity (Sn) are employed to determine the effectiveness of various AEPs. Multiple methods are discussed in the literature for exon detection [20]–[22]. The theory and studies of AEPs are also discussed in the following pages, which address the efficacy of different AEPs.

II. EXON IDENTIFICATION USING ADAPTIVE TECHNIQUES

First move remains to interpret the NCBI database DNA sequence dependent on nucleotide densities of dimer, then translated as numerical notation and processed using the suggested AEP. It remains a chief job of genomic processing because only digital or discrete signals can be used for signal processing. DNA sequence here is translated to binary information describing four binary streams with binary mapping. Digital transformation is a crucial job for the analysis of gene sequences, because only such signals can be used with techniques for signal processing.

Consequently, 1 shows the nucleotide's presence and 0 is its absence. The consequent sequence is now suitable as an input for the adaptive algorithm. AEP remains deliberate also developed using methods for adaptive signal processing.

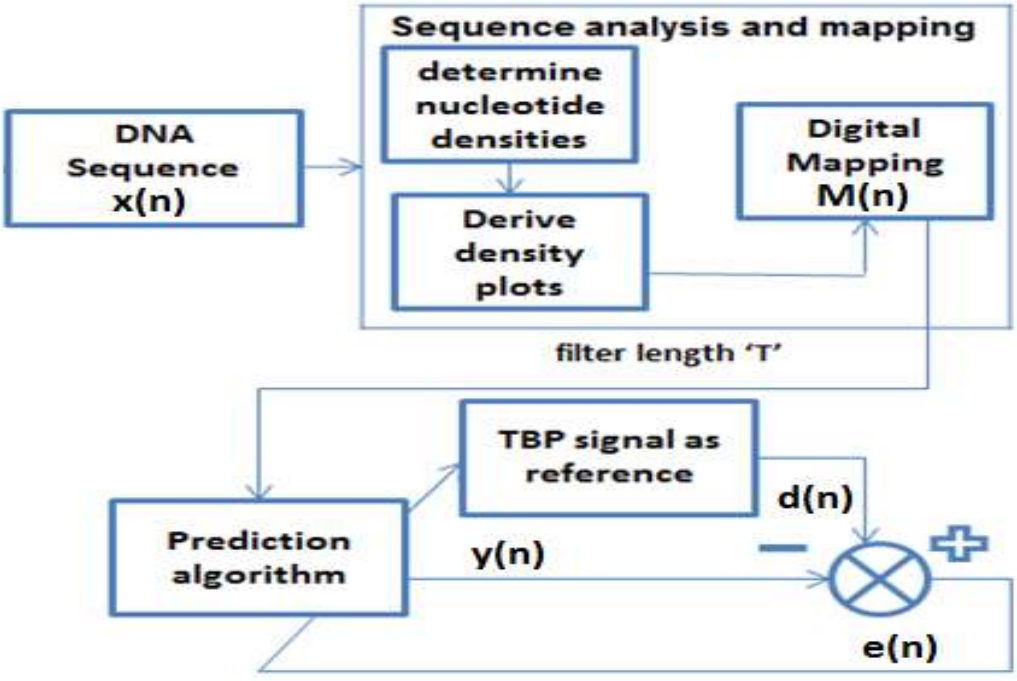

Figure 1: Proposed AEP block diagram

Let, M(n) as a mapped digital sequence, X(n) as a DNA sequence, D(n) as the gene sequence that is TBP, y(n) show results obtained through the use of adaptive technique, e(n) also reflect feedback signal to alter coefficients of weight produced within the feedback loop. Length is still considered as "T" in LMS technique.

The current weight coefficient of w(n), can be calculated at the moment as the next weight coefficient is the mapped input binary sequence indicated as M(n). LMS technique remains interpreted as well as evaluated mathematically on the basis of the actual weight coefficient step size 'μ' in [12]. Figure 1 displays a typical AEP block diagramWeight expression of LMS remains written as

$$w(n+1) = w(n) + \mu\, x(n)\, e(n) \qquad (1)$$

Applications related to locate exons in order to promote nano-based bioinformatics applications have limited computational difficulties. Through applying clipping, a feedback signal or both is added to the input information of gene, this lowering is viable. The techniques for this reason are discussed in [18]. These approaches cover three signed versions.

Signum notation is expressed as

$$C\{x(n)\} = \begin{Bmatrix} 1: x(n) > 0 \\ 0: x(n) = 0 \\ -1: x(n) < 0 \end{Bmatrix} \qquad (2)$$

The complexities of LMS computations are decreased by these variants. Similar to these versions, LMS is more difficult to compute. The technique for Data Clipped LMS (DCLMS) can be represented by changing the input tap vectors as the LMS recursion. In that case, the average vector values C[x(n)] will be replaced with x(n) and sign function C will be used for x[n] on the component-by-component basis.

Mass update relation for DCLMS method remains as

$$w(n+1) = w(n) + \mu\, C\{x(n)\}\, e(n) \qquad (3)$$

The relationship of weight for ECLMS remains achieved by modifying e(n) by signed notation as

$$w(n+1) = w(n) + \mu\, x(n)\, C\{e(n)\} \qquad (4)$$

Also, the DECLMS weight relation results in x(n), e(n) replacing by the application of signed forms as

$$w(n+1) = w(n) + \mu\, C\{x(n)\}\, C\{e(n)\} \qquad (5)$$

Due to its robustness as well as simplicity, the standard adaptive LMS technique is suitable for exon forecast. In order to choose parameter of step size for the convergence as well as stability, understanding of preceding input power level rate is required for LMS filter. As one of the statistical unknown levels is generally the input power level, it will normally be assessed thru information prior start of adaptation process. The vector of the input information is proportionate to weight update process. Other one being its step size is fixed. Both these remain two setbacks of LMS

An algorithm must be designed so that weak as well as strong signals can be handled in real time. Therefore, the tap coefficients must be adapted accordingly on the basis of filter changes in input as well as output. For the LMS method, therefore, the gradient noise amplification setback is affected for a large input data vector.

Adaptive filters are more efficient than the mean square estimate in specific situations used for LMS algorithm relying on higher order statistics. To investigate this, we have created various AEPs to locate exon locations in gene sequences. The Least Mean Fourth (LMF) algorithm remains one such instance, because of the minimization of fourth moment of output estimate error. This algorithm remains closely related to LMS algorithm. In the fourth order, LMF algorithm seeks to increase the strength of the error signal. The LMF algorithm overwhelms LMS constraints and improves convergence speeds and ability to track exons. Here, we have used LMF and its adaptive algorithm based on SRA to enhance AEP efficiency. The LMF algorithm overcomes the LMS disadvantages and increases the ability of exon identification and quicker convergence when error is high. This also reduces the surplus EMSE in the exon identification process. These LMF adaptive algorithms are used for developing AEPs in order to cope with computing difficulty of an AEP in practical applications. The sign function is often used to reduce complexity. Three simplified SRLMF, SLMF and SSLMF signed variants are derived from the application of sign function to LMF.

The weight relations for SRLMF, SLMF, and SSLMF techniques becomes

$$w(n+1) = w(n) + \mu e^3(n)\, C[x(n)] \quad (9)$$

$$w(n+1) = w(n) + \mu\, C[e^3(n)]\, [x(n)] \quad (10)$$

$$w(n+1) = w(n) + \mu\, C[e^3(n)]\, C[x(n)] \quad (11)$$

Finally, these algorithms generated four AEPs, whose results were opposed to AEPs with LMS. Performance review of metrics sensitivity, precision also specificity shows that SRLMF is just below the sign-based non-regressor variants. Thus, among the different techniques selected for use, the SRLMF is superior to other signed versions.

III. COMPUTATIONAL COMPLEXITIES AND CONVERGENCE ISSUES

To calculate and compare the algorithmic complexity, the amount of multiplications necessary is calculated. Emphasis remains not on accurate calculation analysis, but on the evaluation of different adaptive methods based on LMF. Furthermore, the sign-based methods are without multiplication calculations that are needed for the applications of exon identification. For example, when calculating the mass update equation, LMS requires T+1 multiplication calculations with an addition. During computation of 'S. e(n)', for an SRLMF-based adaptive version, only 2T+1 multiplications are needed. 2T+1 multiplication calculations remain important in the case of two other signed LMF algorithms. With SRLMF, it needs fewer multiplications with less computing difficulty compared to other LMF based techniques. Table 1 displays LMS and LMF-based variants computational complexity.

Table 1: Computations required for LMS and various LMF based AEPs

S. No.	Algorithm	Multiplications	Additions
1	LMS	T+1	T+1
2	LMF	T+3	T+1
3	SRLMF	3	T+1
4	SLMF	T	T+1
5	SSLMF	T+1	T+1

The LMF-dependent AEPs proposed give less complexity to determine the position of the desired gene for the genomic input sequence. Figure 2 displays convergent characteristics related to suggested LMF and its signed variants. Clearly, all suggested Adjustable LMF algorithms converge faster than LMS dependent AEP. Therefore, the SRLMF adaptive algorithm is considered better, based on computing difficulty as well as convergence efficiency in contrast to LMS and its other signed algorithms, among the algorithms considered for AEP implementation. It was obvious that SRLMF converges quicker compared to SLMF and SSLMF based AEPs from convergence features.

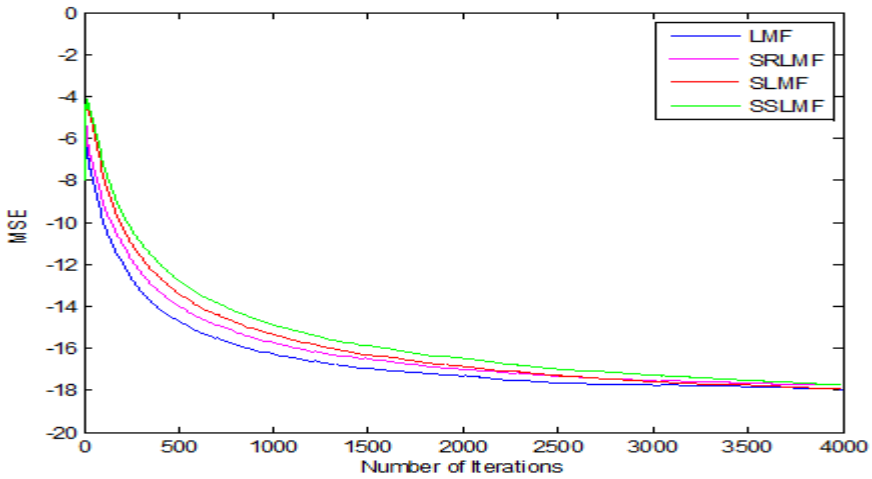

Figure 2: Convergence Curves of LMF and its signed variants

IV. RESULTS AND DISCUSSION

The discussion has to do with the contrast of different AEPs. The AEP contour is shown in the Figure 1. The LMF-based algorithms with sign variants remain derived from many AEPs. We have produced a comparative AEP analysis focused on LMS. Ten NCBI database genomic datasets are used for assessment purposes. The measurements, such as precision (Pr), sensitivity (Sn), even specificity (Sp) from [13] pose their expressions and theory, are taken into account in the determination of the output performance. The results from the different algorithms in Table 3 decide the performance.

In order to evaluate the exon sections using DSP methods, there are few measures based on changes in the threshold level in the output spectrum used for comparison. The amount of nucleotides found in exon positions as introns is still defined as True Negative (TN) and correctly identified exonal areas for example, remain as True Positive (TP). Furthermore, maximum quantities of exon areas found as intron areas are stated to be False Negative

(FN), as opposed to quantities of introns that are currently expected to be False Positive (FP) as exon areas. In ten NCBI gene datasets, the efficiency of different algorithms will continue to be evaluated. The accession for these sequences remains X59065.1, E15270.1, U01317.1, X77471.1, AF009962, X92412.1, AB035346.2, AJ223321.1, AJ225085.1, and X51502.1 respectively as shown in Table 2.

Expressions for performance metrics are

Pr= (TP+TN)/ (TP+FP+TN+FN)

Sp = TP/ (TP + FP)

Sn= TP/ (TP+FN)

Table 2: Computations required for LMS and various LMF based AEPs

S. No.	Accession No.	Sequence Definition
1	E15270.1	Human gene inhibitory factor of Osteoclastogenesis (OCIF)
2	X77471.1	Human gene of tyrosine aminotransferase(TAT)
3	AB035346.2	Human gene T-cell leukemia/lymphoma 6(TCL6)
4	AJ225085.1	Human gene Fanconi anemia group A(FAA)
5	AF009962	Human gene CC-chemokine receptor (CCR-5)
6	X59065.1	Human gene human acidic fibroblast growth factor(FGF)
7	AJ223321.1	Human gene ns transcriptional repressor(RP58)
8	X92412.1	Human gene titin (TTN)
9	U01317.1	Human beta globin sequence on chromosome 11
10	X51502.1	Human gene for prolactin-inducible protein (GPIPI)

Specificity (Sp) remains as number of exons found in part of the exons, while the quantity of exons which remains correctly predicted is computed as sensitivity (Sn). Figure 3 includes exon identification results in the gene sequence 5 using LMF-dependent techniques. Threshold values are selected at an interval of 0.05 from 0.4 to 0.9. The efficiency of metrics Pr, Sn, and Sp is evaluated by using these values. The exon prediction is accurate at a threshold of 0.8.

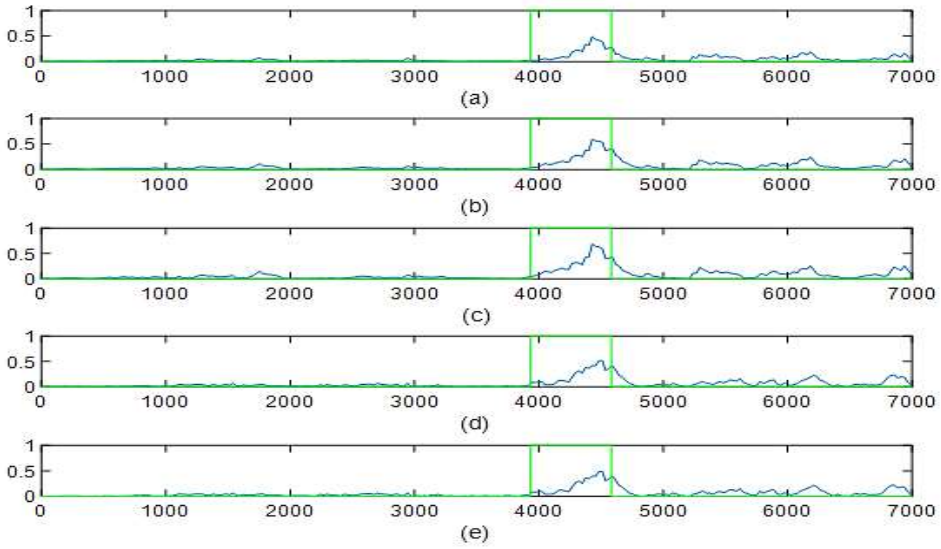

Figure 3: PSD plots for genomic sequence with accession AF009962 via various AEPs with Exon location (3934-4581), (a). AEP with LMS, (b). AEP with LMF, (c). AEP with SRLMF, (d). AEP with SLMF, (e). AEP with SSLMF.

(*Relative Base Location is taken on x-axis and Power Spectrum on y-axis*)

As a consequence, the performance indicators at 0.8 are shown in Table 3. Intergenic sequence components typically reveal sections with significant percentages of A+T nucleotides from a DNA sample, whilst low A+T and greater G+C nucleotides show potential genes. Mostly, high CG dinucleotide content is often found ahead for a gene. Functions of statistics for a gene sequence remains beneficial for determining whether the input gene sequence has protein-coding segments.

Table 3: Metrics for Performance using Numerous LMF and its Signed based AEPs with respect to Sn, Sp also

Algorithm	Metric	Gene Sequence Serial Number									
		1	2	3	4	5	6	7	8	9	10
LMS	Sn	0.6286	0.6384	0.6457	0.6273	0.6481	0.6162	0.6193	0.6241	0.6268	0.6202
	Sp	0.6435	0.6628	0.6587	0.6405	0.6518	0.6324	0.6529	0.6289	0.6452	0.5965
	Pr	0.5922	0.5894	0.5934	0.5858	0.5904	0.5786	0.5896	0.5856	0.5814	0.5761
LMF	Sn	0.6694	0.6708	0.6686	0.6647	0.6705	0.6596	0.6546	0.6572	0.6682	0.6718
	Sp	0.6676	0.6791	0.6692	0.6635	0.6717	0.6632	0.6732	0.6594	0.6592	0.6592
	Pr	0.6412	0.6586	0.6518	0.6597	0.6512	0.6404	0.6514	0.6576	0.6535	0.6484
SRLMF	Sn	0.6513	0.6523	0.6594	0.6573	0.6649	0.6445	0.6432	0.6482	0.6548	0.6683
	Sp	0.6561	0.6743	0.6648	0.6568	0.6585	0.6475	0.6685	0.6516	0.6563	0.6486
	Pr	0.6394	0.6427	0.6494	0.6486	0.6451	0.6343	0.6431	0.6482	0.6465	0.6373
SLMF	Sn	0.6426	0.6492	0.6536	0.6431	0.6519	0.6327	0.6356	0.6325	0.6441	0.6523
	Sp	0.6487	0.6716	0.6614	0.6468	0.6485	0.6375	0.6602	0.6483	0.6518	0.6385
	Pr	0.6249	0.6372	0.6385	0.6386	0.6352	0.6252	0.6317	0.6337	0.6362	0.6293
SSLMF	Sn	0.6311	0.6422	0.6536	0.6375	0.6649	0.6241	0.6232	0.6412	0.6345	0.6457
	Sp	0.6361	0.6682	0.6597	0.6368	0.6385	0.6274	0.6567	0.6364	0.6496	0.6286
	Pr	0.6195	0.6227	0.6288	0.6286	0.6251	0.6143	0.6231	0.6283	0.6265	0.6182

Pr Calculations

Following steps of AEP are listed below:

a) For the purposes of determining the existence of gene positions based on the nucleotides density base pairs for G+C and also A+T dimers, we have analyzed gene data sets for NCBI's input using density plots in Figure 5. Following the assessment, this sequence is then converted into digital notation after analysis using the digital mapping technique, while input of AEP remains to be binary information from Figure 1.

b) After the assessment, the following sequence shall be considered as an input to the presented AEP. The obedient biological sequence of TBP is provided as a reference signal for the proposed LMF-based AEPs.

c) For updating filter coefficients, derived e(n) feedback signal from Figure 1 has been used.

d) When this signal is reduced the DNA sequence genes with a plot of PSD are correctly identified.

e) The PSD shows plots of the needed exon areas. Sp, Pr and Sn are further compared and acquired.

The MATLAB software uses sequence 5 metrics with AF009962 accession, for all LMF variants. Table 3 lists the performance metrics of LMF dependent AEPs that are less iterated than SRLMF because of their low complexity and exon placement. The plot precisely located the exon at 3934-4581 and has a high intensity and sharp PSD peak. In such cases, the SRLMF algorithm becomes efficient due to low complexity in performing the

computations and in terms of exon locating ability. The signum function present in all signed versions of LMF reduces the computational complexity and thus all signed versions predict the exon locations more accurately.

Of all these algorithms, SRLMF based AEP is effective in terms of accurate exon prediction when compared to LMS, LMF and its other signed variants with Specificity Sp, 0.6585 (65.85%), Sensitivity Sn 0.6649 (66.49 %), also Precision, Pr 0.6451 (64.51%) respectively. At 0.8 threshold value, the exon prediction appears to be better for SRLMF based AEP. These PSD curves are provided in Figures 4 (b), (c) and (d) respectively, for LMF and their signed variants. Finally, all proposed LMF based AEPs are more effective to discover exon areas in genomic sequences compared with the prevailing LMS technique.

V. CONCLUSION

This paper addresses the issue of defining the exon location within the gene sequence. Novel approach remains proposed for adaptive exon detection using a bioinformatics adaptive AEP based system. Multiple DNA sequences to solve this issue are interpreted using LMF-based adaptive algorithms. In Table 3 and plots of PSD are shown in Figure 4, measurements for exon sites are evident. The AEPs presented were correctly located the exon location in PSD plot at 3934-4581. SRLMF offers better computing performance, performance metrics obtained with a gene sequence 5 having a 0.8 threshold accession AF009962 remain just under the AEP values based on LMF. However, it is also easier to find exons specifically because of its reduced compute complexity. Therefore, in Nano-bioinformatics and cloud based exon prediction applications based on SOC and LOC, SRLMF-based AEP can be used.

REFERENCES

1. L.W. Ning, H. Lin Ding, J. Huang, N. Rao, and F.B. Guo, “Predicting bacterial essential genes using only sequence composition information,” Genet. Mol. Res., vol. 13, pp. 4564–4572, June 2014.
2. Li. Min, Li. Qi, G. Gamage Upeksha, W. Jian Xin, Wu. Fang Xiang, and Yi. Pan, “Prioritization of orphan disease-causing genes using topological feature and go similarity between proteins in interaction networks,” Sci. China Life Sci., vol. 57, pp. 1064–1071, November 2014.
3. T. M. Inbamalar, and R. Sivakumar, “Study of DNA sequence analysis using DSP techniques,” J. Autom. Control Eng., vol. 1, pp. 336–342, December 2013.
4. S. Maji, and D. Garg, “Progress in gene prediction: principles and challenges,” Curr. Bioinform., vol. 8, pp. 226–243, April 2013.
5. P. Srinivasareddy, and Md. Zia Ur Rahman, “New adaptive exon predictors for identifying protein coding regions in DNA sequence,” ARPN J. Theor. Appl. Sci., vol. 11, pp. 13540–13549, December 2016.
6. H. Saberkari, M. Shamsi, H. Hamed, and M. H. Sedaaghi, “A novel fast algorithm for exon prediction in eukaryotes genes using linear predictive coding model and goertzel algorithm based on the Z-curve,” Int. J. Comput. Appl., vol. 67, pp. 25–38, April 2013.
7. M. Wazim Ismail, Ye. Yuzhen, and T. Haixu, “Gene finding in metatranscriptomic sequences,” BMC Bioinform., vol. 15, pp. 01–08, September 2014.

8. M. Ghorbani, and K. Hamed, “Progress in gene prediction: principles and challenges,” Bioinformatics approaches for gene finding, vol. 4, pp. 12–15, September 2015.
9. S. Devendra Kumar, S. Rajiv, and S. Narayan Sharma, “An adaptive window length strategy for eukaryotic CDS prediction,” Trans. Comput. Biol. Bioinform., vol. 10, pp. 1241–1252, September 2013.
10. Y. Azuma, and S. Onami, “Automatic cell identification in the unique system of invariant embryogenesis in caenorhabditis elegans,” Biomed. Eng. Lett., vol. 4, pp. 328–337, December 2014.
11. Liu. Guangchen, and Luan. Yihui, “Identification of protein coding regions in the eukaryotic DNA sequences based on marple algorithm and wavelet packets transform,”Abstr. Appl. Anal., vol. 2014, pp. 1–14, July 2014.
12. O. Simon Haykin, “Adaptive filter theory,” 5th ed., Pearson Education Ltd., 2014, pp. 320-380.
13. H. Saberkari, M. Shamsi, H. Hamed, and M. H. Sedaaghi, “A Fast Algorithm for Exonic Regions Prediction in DNA Sequences,” J. Med. Signals Sens., vol. 3, pp. 139–149, July 2013.
14. M. Nagesh, S.V.A.V. Prasad, and M.Z. Rahman, “Efficient cardiac signal enhancement techniques based on variable step size and data normalized hybrid signed adaptive algorithms,” Int. Rev. Comp. Soft., vol. 11, pp. 1–13, October 2016.
15. M. O. Sayin, N. D. Vanli and S. S. Kozat, "A Novel Family of Adaptive Filtering Algorithms Based on The Logarithmic Cost,” IEEE Trans. Signal Process., vol. 62, no. 17, pp. 4411–4424, September 2014.
16. V. C. Gogineni and S. Mula, "A Family of Constrained Adaptive filtering Algorithms Based on Logarithmic Cost", IEEE Trans. Signal Process., pp. 1–14, July 2017.
17. S. Mula, V. C. Gogineni and A. S. Dhar, “Algorithm and Architecture Design of Adaptive Filters with Error Non-linearities,” in IEEE Trans. VLSI Syst., vol. 25, no. 9, pp. 2588-2601, September 2017.
18. S. R. Paula Diniz, “Adaptive filtering, algorithms and practical implementation,” 4th ed., Springer Publishers, 2013.
19. National Center for Biotechnology Information. Accessed: January 25, 2019. [Online]. Available: www.ncbi.nlm.nih.gov/
20. P. Srinivasareddy, Md. Zia Ur Rahman, A. Chandra Sekhar, and P. Nagireddy, "New Exon Prediction Techniques Using Adaptive Signal Processing Algorithms for Genomic Analysis," IEEE Access, Vol.7, pp. 80800-80812, 2019.
21. S. R. Putluri, and Md. Zia Ur Rahman, “Identification of Protein Coding Region in DNA Sequence Using Novel Adaptive Exon Predictor”, J. Sci. Ind. Res., Vol. 77, pp. 1 - 5, 2018.
22. P. Srinivasareddy, Md. Zia Ur Rahman, and S. Y. Fathima, “Cloud Based Adaptive Exon Prediction for DNA Analysis”, IET Healthc. Technol., Vol. 5, No. 1, pp. 1 - 6, 2018.

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020