Adaptive Exon Prediction using Maximum Modified Normalized Algorithms

Md. Zia Ur Rahman, Farmanulla Shaik, Srinivasareddy Putluri

Abstract: Exact identification of exon fragments in a deoxyribonucleic acid (DNA) sequence is a critical task in the field of genomics. This is a crucial part in finding health disorders and design drugs. Exons are the infoessentialin coding of proteins in DNA. HenceforwardfindingsuchDNA sectionsremains important part of genomics. In DNA arrangement, nucleotides form the key elementary units. Three base periodicity (TBP) is a basic property displayed by only exon fragments, and is not shown in other DNA sections that could beforecastedeasily with techniques of signal processing. Frommanymethods, adaptive methodswerefavorablebecause of theircompetencein altering weight coefficients depending on gene sequence. Hence, an adaptive exon predictor (AEP) is proposedwithMaximum Modified Normalized Least Mean Square (MMNLMS) algorithm. TheAEP derived using MMNLMS is combined with its sign versions to decrease complexity in computations. Also, this was clear thatModified Normalized Sign Regressor LMS (MMNSRLMS) based AEPwas more effective in exon identification applications withmetricsalikeSpecificity, Sensitivity, and Precision. Thus, computational complexity is greatly minimized, and AEPs proposedweresuitablefor use in nano devices. Lastly exon findingcapabilitywithdiverse AEPs stands verified with DNA datasets from National Center for Biotechnology Information (NCBI) gene databank.

Index Terms: adaptive exon predictor, disorders, deoxyribonucleic acid, exon fragments, three base periodicity
.

I. INTRODUCTION

Finding the exon sections in DNA is an intense part of research in bio-informatics.Actual tracing of such sectionsstands critical to find health disorders along with drug design.Fragments those remain responsible also those not involved in protein coding are part of DNA [1]. In the area of bio-informatics, gene identification focuses on tracing the protein coding fragments. Learning of mainarrangement of protein coding segmentshelps to know abouttheir tertiary also ancillary structure. The sooner study related to whole protein structure is completed; entire health disorders can be found, design drugs and cure them. Likewise, the investigations help to know about assessment of phylogenic trees [2]-[3]. Whole living beings were classified depending on the elementary structure of molecules.

These are prokaryotes and eukaryotes. Coding segments in prokaryotic cells are continuous and long; instances include archaea and bacteria. Arrangement of coding segments in genes is alienated by lengthy non-protein coding sections of eukaryotes. Coding regions responsible for protein synthesis are exons, while rest of segments is introns. Whole living beings excluding archaea and bacteria remain fall under this classification. The coding sections in eukaryotes of human beings comprise only around 3 % of gene sequence whereas introns comprise the rest of 97%. Therefore, locating the protein coding segments in a gene sequence is a significant job [4]-[5]. Three base periodicity (TBP) is pragmatic in relatively all gene sequences. A sharp peak is clearly shown in power spectral density (PSD) at frequency f1=1/3 [6]. Abundantmethodsfrom literature to trace pretein coding fragmentsbased on manytechniques related to signal processing are discussed in [7] – [11]. Nevertheless, in real time, DNA sequence length is too long alsoexon fragments place alters insidediverse sequences. Techniques that follow adaptive strategyare used in more number of iterationsthrumodifying coefficients of weight depending on itsnumericalbehavior [12]. Different AEPs are derived using such techniques. LMS is widely used as it is simple and more easiness for implementation. This facestechnical hitchesfor instance poor convergence, weight drift, and noise amplification of gradient [13]. Subsequently, normalization concept is used in current work. Data normalized version of LMS is normalized LMS (NLMS) technique. This unravelshindrancesrelated toLMSpresents good ability of exon trackingaccompanied by convergence rate and also minimizesexcess mean square error (EMSE) in exon position identification. Complexity involved in computations for an adaptive strategy is a critical aspect explicitly for sequences of larger lengthson account of overlap of samples provided to AEP. Problems alike Inter symbol interference (ISI) furthermoreinexactnessin tracing exon fragments. Besides, complexity is high with respect to computations results in larger size of circuitry to implement AEP on nano device or VLSI circuit.So, presented adaptive methodswith sign function were usedso as to minimize operations related to number of multiplications. The three signum based simplified algorithms involves signed regressor (SRA), signed error (SEA) also signed signed algorithms (SSA) are combined with MNLMS algorithm.

Revised Manuscript Received on 30 May 2019.

* Correspondence Author

Md Zia Ur Rahman*, Professor Department of Electronics and Communication Engineering, Koneru Lakshmaiah Educational Foundation Guntur, India.

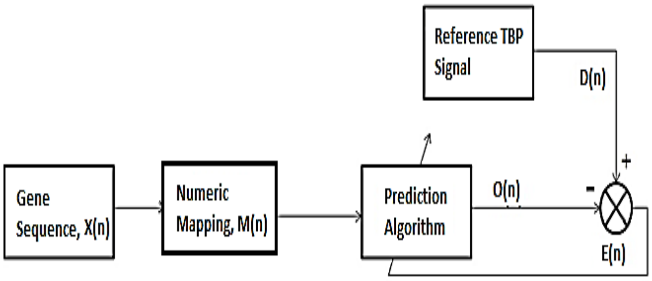
Farmanullah Shaik, Assistant Professor Department of Electronics and Communications Engineering, Eswar College of Engineering, Kesanupalli, Narasaraopeta, Guntur, A.P., India.

Srinivasareddy Putluri Department of Electronics and Communication Engineering, K L University, Guntur, A.P., India.

With normalization, higher length of the tap is minimized to one, using an approach termed as maximum normalization neverthelessof length of the tap. Depending on maximum normalized forms of MNLMS algorithm, several AEPs were developed also they are analyzed by actual DNA of National Center for Biotechnology Information (NCBI) gene bank [14]. Hybrid versions obtained includesmaximum modified normalized signed regressor LMS (MMNSRLMS), maximum odifiednormalized signed error LMS (MMNSLMS) also maximum modified normalized signed signed LMS (MMNSSLMS) techniques. Differentmetrics such as complexity in computations, rate of convergence plots, precision (p_r), specificity (s_p), and sensitivity (s_n) were consideredto verify many AEPs. Results of AEPs, theory of adaptive techniques, and discussion with respect toseveral AEPs performance are deliberated in succeeding sections.

II. ADAPTIVE ALGORITHMS FOR EXON IDENTIFICATION

Translation of alphabetic DNAintonumerical form is a first step of proposed AEP remainscriticalas adaptive methods areapt only for numerical or discrete signals. At the moment, voss representation is presented for this purpose. With this representation, nucleotide presence isrepresented as 1 and absence as 0is explained in [12]. Converted sequence now becomes apt to give as AEP input. Consider AEP with X(n) as input DNA, M(n) as numeric mapped signal, D(n) as reference TBP signal, O(n) as output attained from adaptive technique also E(n) as signal of feedback for weight updating of adaptive technique and length of filter as ‘L’. Expression along with study of LMS was explained in [13].



Block representation of an AEP is depicted in Figure 1.
Figure 1: Block illustration of AEP.

The expression for mass updation of LMS adaptive technique is stated as

$$u(n+1) = u(n) + S\,X(n)E(n) \qquad (1)$$

Adaptive techniques exhibit less complexity in computations in exon identification applications makes them suitable for developing nano devices. Such reduced value is probable by applying clipping to input information otherwise signal of feedback or both. Techniques for clipping of error or data demonstrated in [13]. The signum representation is given below: -

$$\text{sign}\{X(n)\} = \begin{cases} 1: X(n) > 0 \\ 0: X(n) = 0 \\ -1: X(n) < 0 \end{cases} (2)$$

SRA, SA and SSA adaptive techniques are used for minimizing complexity in computations than LMS. LMS has added computational complexity than proposed techniques. SRA remains derived using LMS recursion thru change of tap input vector. X(n) is replaced by means of the vector sign[X(n)].

Mass update expression for SRLMS algorithm remains represented as

$$u(n+1) = u(n) + S\,\text{sign}[X(n)]E(n) \qquad (3)$$

Mass renovate relation for SLMS algorithm is

$$u(n+1) = u(n) + S\,\text{sign}[E(n)]X(n) \qquad (4)$$

Similarly, mass revise expression for SSLMS algorithm derived via applying sign function to X(n), E(n) as

$$u(n+1) = u(n) + S\,\text{sign}[X(n)]\text{sign}[E(n)] \qquad (5)$$

To overwhelm gradient noise problem of LMS, normalized form of LMS creates a own problem, namely small input tap vector. Numerical problems may rise due to which then we have to partition by a little amount of the squared norm. In order to overcome this problem, we change the above recursion by inducing a small positive constant ε. This parameter ε eludes less value from divisor with larger size of step.
The step size parameter can be expressed to be,

$$S(n) = \frac{S}{\varepsilon + ||X(n)||2} \qquad (6)$$

where S(n) is normalized size of step having 0 <S< 2. Alternating S of LMS vector for weight renovate expression with S(n) tends to DNLMS stated as

$$u(n+1) = u(n) + \frac{S}{\varepsilon + ||X(n)||2} X(n).E(n) \qquad (7)$$

In DNLMS, error reduces and multiply computations increases due to squared value of X(n) in the divisor thereby rate of convergence is faster. To minimize complexity in computations, MNLMS is used.
MNLMS is mathematically represented as,

$$u(n+1) = u(n) + \frac{q\,S}{\varepsilon + ||X(n)||2} X(n)E(n) \quad (8)$$

where q = diag {Q} and Q = {1 if x > xmax}. The term q will be either zero or one, based on the value of x. In case the value of x is higher compared to the threshold value, then the q will be set to one otherwise it is set to zero, thus reducing the entire numerator to zero and number of calculations reduces. Here, signed forms of MNLMS are considered for this purpose. Also, all proposed AEPs offers precise tracing of protein coding fragments and better convergence. Hence, to reduce the computational complexity of MNLMS algorithm, maximum forms of MNLMS along with sign based algorithms were derived. The hybrid versions are named as MMNSRLMS, MMNSLMS and MMNSSLMS algorithms.
The mass renovate equations of MMNSRLMS, MMNSLMS added to MMNSSLMS algorithms are numerically expressed as,

$$u(n+1) = u(n) + \frac{q\,S}{\varepsilon + \max(X(n))^2}\text{sign}[X(n)]\,E(n) \qquad (9)$$

$$u(n+1) = u(n) + \frac{q\,S}{\varepsilon + max(X(n))^2} X(n)\,\text{sign}[E(n)] \qquad (10)$$

$$u(n+1) = u(n) + \frac{q\,S}{\varepsilon + \max(X(n))^2}\text{sign}[X(n)]\,\text{sign}[E(n)] \quad (11)$$

Hence, finally the algorithms to develop four AEPs are chosen also compared to LMS. Computational complexities of projected AEPs along with LMS were shown in Table III.

Convergence plots for modified normalized algorithms are shown in Figure 3. From Figure 3, all proposed modified normalized adaptive algorithms have a faster convergence rate than LMS and other AEPs. Hence, among thealgorithms considered for the implementation of AEPs, the MMNSRLMS based AEP is considered to be better, with respect to convergence characteristics and complexity in computations compared to other normalized algorithms.

III. RESULTS AND DISCUSSION

Here, several AEPs were compared also analyzed for performance. Figure 1 presents block illustration of AEP. Modified normalized LMS algorithm and its sign forms are used to derive different AEPs. AEP using LMS also derived for comparison. Five sequences of DNA with description shown in Table I were used from NCBI databank for analysis [14]. The theory and expressions of performance measures like sensitivity (Sn), specificity (Sp) also precision (Pr) parameters are given in [11]. PSD plots along with metrics like Sn, Sp and Pr on values of threshold as of 0.4 to 0.9 thru interval of 0.05 with sequence 5 were depicted in Figure 2. Identification of exon fragment is better at 0.8 threshold value. Therefore values of measures at 0.8 are presented in Table II.

Steps for adaptive exon prediction were as follows:

1. Choose DNA sequences from data base [14]. Using Voss numeric representation, transform DNA sequence to binary data.
2. Give transformed numeric input to AEP shown in Figure 1.
3. A reference signal that confirms TBP property is given to the AEP.
4. A signal for feedback as depicted in Figure 1 is produced is used for filter co-efficient updating.
5. When feedback signal becomes minimum, the adaptive algorithm predicts location of coding region sequence accurately.
6. Location of exons is plotted using PSD also metrics Sn, Sp and Pr were derived.

Figure 2 depicts the traced protein coding fragments using different AEPs. LMS based AEP not identified exon fragments precisely with some ambiguities by tracing few intron regions, which was evident from Figure 2.

In Figure 2 (a) few undesirable peaks were recognized at positions 1200th, 2300th and 3500th values of samples without tracing actual exon location 3934-4581. In the case of modified normalized variants, the MMNLMS, MMNSRLMS and MMNSSLMS algorithms exactly predicted protein coding fragment at 3934-4581 thru high intensity on PSD plot. These PSDs are shown in Figure 2 (b), (c) and (d). Because of normalization exon finding ability is better compared to LMS algorithm.

Therefore, based on computational complexity, convergence characteristics, exon prediction plots, Sn, Sp and Pr calculations, AEP using MMNSRLMS is a better choice in real time applications. Lower computational complexity leads to simplified architecture for lab on chip (LOC) and system on chip (SOC) applications.

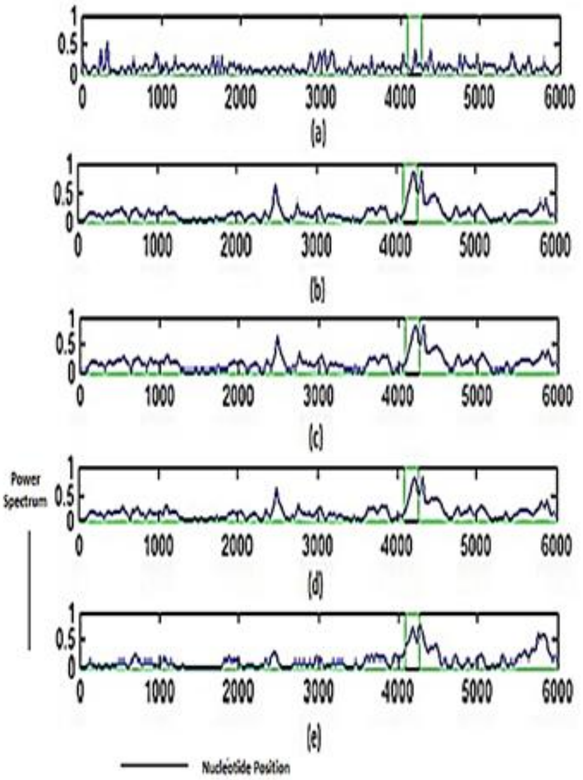


Figure 2: PSD with the location of exon (3934-4581) for a DNA sequence of accession AF009962 predicted using various AEPs, (a). AEP using LMS, (b). MMNLMS based AEP, (c). MMNSRLMS based AEP, (d). MMNSLMS based AEP, (e). MMNSSLMS based AEP

Table I. Dataset of DNA sequences from NCBI database.

Seq. No.	Accession No.	Sequence Definition
1	E15270.1	Human gene for osteoclastogenesis inhibitory factor (OCIF) gene
2	X77471.1	Homo sapiens human tyrosine aminotransferase(tat) gene
3	AB035346.2	Homo sapiens T-cell leukemia/lymphoma 6(TCL6) gene
4	AJ225085.1	Homo sapiens Fanconi anemia group A(FAA) gene
5	AF009962	Homo sapiens CC-chemokine receptor (CCR-5) gene

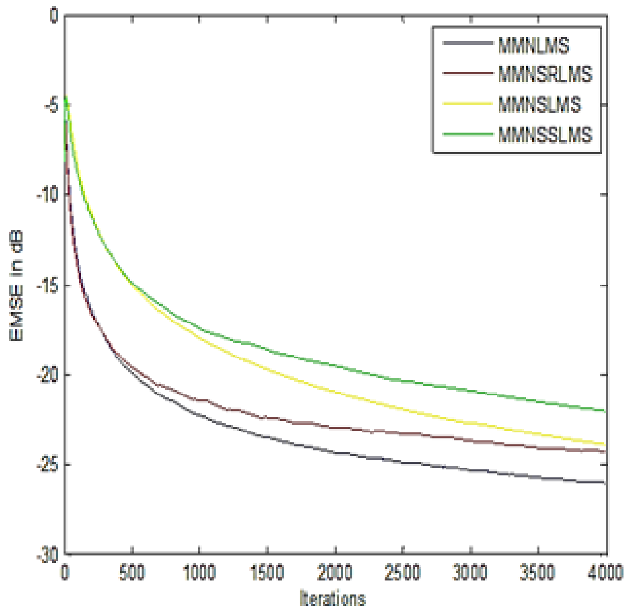


Figure 3: PSD with the location of exon (3934-4581) for a DNA sequence

Table II. Performance measures of various modified normalized based AEPs with respect to Sn, Sp and Pr calculations.

Seq. No.	Parameter	LMS	MMN LMS	MMN SRLMS	MMN SLMS	MMN SSLMS
1	Sn	0.6481	0.7292	0.7087	0.6846	0.6702
	Sp	0.6518	0.7382	0.7123	0.6965	0.6812
	Pr	0.5904	0.7264	0.7052	0.6856	0.6723
2	Sn	0.6286	0.7391	0.7056	0.6832	0.6718
	Sp	0.6435	0.7385	0.7143	0.6976	0.6811
	Pr	0.5922	0.7263	0.7142	0.6957	0.6806
3	Sn	0.6384	0.7292	0.7087	0.6846	0.6702
	Sp	0.6628	0.7382	0.7123	0.6965	0.6812
	Pr	0.5894	0.7264	0.7052	0.6856	0.6723
4	Sn	0.6457	0.7238	0.7157	0.6974	0.6814
	Sp	0.6587	0.7391	0.7056	0.6832	0.6718
	Pr	0.5934	0.7385	0.7143	0.6976	0.6811
5	Sn	0.6273	0.7645	0.7336	0.7035	0.6857
	Sp	0.6405	0.7524	0.7235	0.6989	0.6797
	Pr	0.5858	0.7537	0.7241	0.7075	0.6886

Table III. Computational complexities of proposed AEPs

S.No.	Algorithm	Multiplications
1	LMS	T+1
2	MMN LMS	T+3
3	MMN SRLMS	3
4	MMN SLMS	T+2
5	MMN SSLMS	3

IV. CONCLUSION

In current work, we have addressed key problem of tracing exon fragments of DNA which has several applications in modern health care technology. Here, we considered adaptive exon prediction technique. For this, data normalization adaptive algorithms are considered. In order to lower complexity in computations, we introduced the concept of modified data normalization. To further minimize computational complexity, sign based variants of MMNLMS are used. Resulting hybrid variants are MMNSRLMS, MMNSLMS also MMNSSLMS algorithms. Thus, four AEPs were derived also verified with actual DNA sequences obtained using NCBI databank. Based on computational complexity shown in Table III and convergence characteristics shown in Figure 3, it is clear that AEP derived using MMNSRLMS remains better in applications related to exon identification. This is also clear from performance metrics in Table II also plots of PSD for predicted exon shown in Figure 2. Therefore, AEP using MMNSRLMS is suitable for genomic applications in real time for the development of LOCs, SOCs and nano devices.

REFERENCES

1. Jian Cheng, Wenwu Wu, Yinwen Zhang, Xiangchen Li, Xiaoqian Jiang, Gehong Wei & Shiheng Tao. (2013). "A new computational strategy for predicting essential genes," BMC Genomics, 14, 1-13.
2. L.W. Ning, H. Lin, H. Ding, J. Huang, N. Rao and F.B. Guo, "Predicting bacterial essential genes using on sequence composition information. (2014)." Genetics and Molecular Research, Vol. 13, 4564 - 4572.
3. Sajid A. Marhon & Stefan C. Kremer, "Gene prediction based on DNA spectral analysis: a literature review. (2011)." J.Comput. Biology, 18, 639–676.
4. S. Maji & D. Garg, "Progress in gene prediction: principles and challenges. (2013)." Curr. Bioinformatics, 8, 226– 243.
5. N. Goel, S. Singh, & T. C. Aseri, "A review of soft computing techniques for gene prediction. (2013)." ISRN Genomics, 2013, 1-8.
6. S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, & R. Ramaswamy. (1997). "Prediction of probable genes by Fourier analysis of genomic sequences," Comput. Applications in the Biosci., 13 (1997), 263–270.
7. Aazim Mohammed Ismail, Yuzhen Ye, and Haixu Tang, "Gene finding in metatranscriptomic sequences. (2014)." BMC Bioinformatics, Vol.15, 01–08.
8. Trevor W. Fox & Alex Carreira. (2004). "A digital signal processing method for gene prediction with improved noise suppression," EURASIP J. Appl. Signal Process., 1 (2004), 108-114.
9. N. Rao, X. Lei, J. Guo, H. Huang, & Z. Ren. (2009). "An efficient sliding window strategy for accurate location of eukaryotic protein coding regions," Comput. Biology and Medicine, 39, 392–395.
10. P Ramachandran, Wu-Sheng Lu, & Andreas Antoniou. (2012). "Filter-Based Methodology for the Location of Hot Spots in Proteins and Exons in DNA," IEEE Trans. Biomed. Eng., 59, 1598-1609.
11. Guangchen Liu & Yihui Luan. (2014). "Identification of Protein Coding Regions in the Eukaryotic DNA Sequences based on Marple algorithm and Wavelet Packets Transform", Abstract and Appl. Anal., 2014, 1-14.
12. R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. (1992)." Phys. Rev. Lett., vol. 68, no. 25, pp. 3805–3808.
13. Simon O. Haykin. (2002). "Adaptive Filter Theory," Pearson Educ Ltd., 4, 320-380.
14. National Center for Biotechnology Information, www.ncbi.nlm.nih.gov/

AUTHORS PROFILE

MD ZIA UR RAHMAN (M’09) (SM’16) received M.Tech. and Ph.D. degrees from Andhra University, Visakhapatnam, India. Currently, he is a Professor with the Department of Electronics and Communication Engineering, Koneru Lakshmaiah Educational Foundation Guntur, India. His current research interests include adaptive signal processing, biomedical signal processing, array signal processing, MEMS, Nano photonics. He published more than 100 research papers in various journals and proceedings. He is serving in various editorial boards in the capacity of Editor in Chief, Associate Editor, reviewer for publishers like IEEE, Elsevier, Springer, IGI, American Scientific Publishers, Hindawai etc.

Farmanullah Shaik is currently working as Assistant Professor in the Department of Electronics and Communications Engineering, Eswar College of Engineering, Kesanupalli, Narasaraopeta, Guntur, A.P., India. His areas of interests are genomic signal processing, biomedical signal processing and signal processing applications

Srinivasareddy Putluri is a Ph.D Scholar in the Department of Electronics and Communication Engineering, K L University, Guntur, A.P., India. His interesting areas are Genomic Signal Processing and Adaptive Signal Processing.