Identification of Exon segments in DNA sequences using Modified Normalized Adaptive Algorithms

Md. Zia Ur Rahman, Farmanulla Shaik, SrinivasareddyPutluri

Abstract: ***A key task of genomics area is precisely tracing protein coding sections in a gene sequence. For identification of ailments and designing the drugs, analysis of these coding segments plays a crucial role. Information required for coding of proteins is present in gene fragments termed as Exons. Henceforth tracing the protein coding fragments of DNA is a key part in genomics. The elementary units in structure of DNA are Nucleotides. Three base periodicity (TBP) remains a typical property displayedthru only protein coding sections and not present withinintron segments of DNA. TBP of exon segments can be easily predicted using Signal processing techniques. Amongst several techniques, adaptive techniques are promising due to their capability to alter coefficients of weight based on deoxyribonucleic acid (DNA) sequence. From these deliberations, we propose an adaptive exon predictor (AEP) using Modified Normalized Least Mean Square (MNLMS) algorithm. To minimize computational complexity of the proposed techniques, we combined MNLMS based AEP with its sign-based variants. It was shown that AEP based on Sign Regressor MNLMS standsmucheffective in applications relation to exon identification using measures like Sensitivity, Specificity and Precision. This greatly reduces computational complexity, so that projected AEPs are attractive in nano devices. Finally, the exon locating ability of different AEPs is verified by gene sequences considered from the renowned genomic data base NCBI databank.***

Index Terms**: adaptive exon predictor, ailments, *computational complexity, deoxyribonucleic acid, disease identification, nucleotide, three base periodicities***

I. INTRODUCTION

Extreme area of research in the area of genomics is tracing the protein coding fragments of DNA. Precise identification is vital aimed at analysis of ailments also designing drugs. Sequence of DNA forms the combination of coding and non-protein coding segments [1]. Gene finding is a key subarea of genomics aimed at finding the exon segments. Study related to principal arrangement of exons aids its ancillary also tertiary structure. We can find all anomalies, drugs design and treat ailments, the moment study of complete protein region structure is done. 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 sections 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 shownpart of power spectral density (PSD)at frequency f1=1/3 [6]. Numerous techniques for locating the exon segments depending on several signal processing methods are presented in the literature [7] – [11]. However, length of DNA sequences in practice is very long and position of coding sections within different sequences changes. To process such sequences adaptive techniques are favorable which are capable for lengthy sequences in several repetitions by changing weight coefficients with respect to statistical behavior of input sequence [12]. From these, AEP is developed with adaptive techniques. Due to its ease to implement, LMS is more used technique. It undergoes hitches alike weight drift, gradient noise amplification, and poor convergence [13]. So, to improve the performance of AEP we propose to use normalization. Data normalized variant of LMS is known as normalized LMS (NLMS) algorithm. NLMS resolves the setbacks of LMS also offers better tracking ability along with speed of convergence. Excess mean square error (EMSE) also reduces part of exon identification[12]. Computational complexityfor an adaptive technique is crucial specifically for lengthy sequences due to overlap of samples to be given to AEP. This results in inter symbol interference (ISI) also inaccuracy in locating exon segments.Moreover, for AEP implementation on VLSI circuit or nano device, more complexity in computations tends to big circuit size also more operations. Henceforth, we combine proposed adaptive techniques using sign based algorithms to reduce multiply operations[13]. The three signum based simplified algorithms involves signed regressor (SRA), signed error (SEA) also signed signed algorithms (SSA) are combined thru MNLMS algorithm. Resulting algorithms are modified

Revised Manuscript Received on June 05, 2019

Md Zia Ur Rahman, Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur-522502, Andhra Pradesh, India.

Farmanullah Shaik, Department of Electronics and Communication Engineering, Kesanupalli, Narasaraopeta, 522601, Guntur, Andhra Pradesh, India.

Srinivasareddy Putluri, Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur-522502, Andhra Pradesh, India.

normalized signed regressor LMS (MNSRLMS), modified normalized signed error LMS (MNSLMS) also modified normalized signed signedLMS (MNSSLMS) algorithms. Based on these MNLMS algorithm, several AEPs were developed also they are analyzed by actualDNAofNational Center for Biotechnology Information (NCBI) gene bank [14]. Various measures such ascomputational complexity, convergence characteristics, specificity (S_p), sensitivity (S_n), and precision (p_r) are deliberated for validating several AEPs. Results of AEPs, theory of adaptive techniques, and discussion related to various AEPs performance are discussed in subsequent sections.

II. ADAPTIVE ALGORITHMS FOR EXON IDENTIFICATION

The first step in proposed AEP is to convertalphabetic gene sequence todigitalform which is crucial because methods based on signal processing were suitable only for signals of discrete or digitalnature. Now,voss numeric notation for this conversion processto represent DNA as four numeric sequences. Using this, existence of base is designatedas 1 and nonexistenceas 0 isillustrated in [12]. At presentit wasappropriateto give as AEP input. Deliberate the AEP withX(n) asinput DNA, M(n) asnumericmapped signal, D(n) asreference TBPsignal, v(n) is the weight vector, O(n) as output attained from adaptive techniquealsoE(n) as signal of feedbackfor weight updating of adaptive technique and length of filter as 'L'. Expression along with study of LMSwasexplained in [13]. Block illustration of an AEP isdepicted in Figure 1.

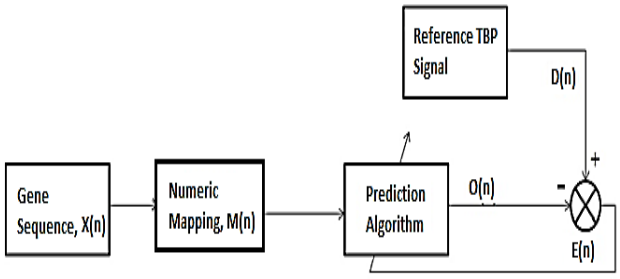

Figure 1: Block illustration of AEP.

The expression for massupdation of LMS adaptive technique is stated as

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

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 that clips the data or error isdemonstrated part of [13]. The signum representation is given below: -

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

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

Mass update expressionforSRLMS algorithm remains represented as

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

Mass renovate relation for SLMS algorithm is

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

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

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

To overwhelm gradient noise problem of LMS, normalized form of LMS creates an 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 εeludes less value from divisorwith 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 DNLMSstated as

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

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,

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

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 offersprecisetracing of protein coding fragments and better convergence. Hence, to reduce the complexity involved in performing computations of MNLMS algorithm, we combine MNLMS with sign-based algorithms. The hybrid versions are named as MNSRLMS, MNSLMS and MNSSLMS algorithms.

The mass renovates equations of MNSRLMS, MNSLMS added to MNSSLMS algorithms are numerically expressed as,

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

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

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

Hence, finally the algorithms to develop four AEPs are chosenalso comparedto LMS. Computational complexities of projected AEPs along with LMSwere 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 the algorithms considered for the implementation of AEPs, the MNSRLMS 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. Different AEPs are developed using modified normalized LMS procedurealsotheir signed forms. 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 alike specificity (Sp), sensitivity (Sn),also precision (Pr) parameters are given in [11]. PSD plots along with metrics like Sn, Sp and Pron values of thresholdas of 0.4 to 0.9 thru interval of 0.05with sequence 5 were depicted in Figure 2. Identification of exon fragment is better at 0.8 threshold value. Therefore, values of measures at 0.8standspresented in Table II.

Process for AEP wasillustrated as below:

1. From NCBI databank, gene datasets are extracted and considered for analysis [14]. Voss numeric representation is used totransforminput gene sequence to digital notation.
2. Now, give the digital form of input to the AEP.
3. A reference signal that conforms TBP property is given to the AEP.
4. A signal for feedback as depicted in Figure 1 is producedis usedfor filter co-efficient updating.
5. Adaptive technique locates the exon segments precisely once signal of feedback turn out to be minimum.
6. PSD plots were derived to depict the exon segments also metrics Sn, Sp and Pr were derived.

Figure 2depictsthe traced protein coding fragmentsusing different AEPs. LMS based AEP not identifiedexon fragments precisely with some ambiguities by tracing few intron regions, which was evident from Figure 2. Someundesirable peaks werepredictable at positions 1200th, 2300thalso 3500th values of samples without tracing actual exon location 3934-4581 from Figure 2 (a).From modified normalized variants, MNLMS, MNSRLMS also MNSSLMS based AEPscorrectlyforecastedprotein coding fragmenton 3934-4581 thruhigh intensity on PSD plot that are depicted in Figure 2 (b), (c) and (d). Exon finding ability is better compared to LMS algorithm due to use of normalization.

As a result, from convergence performance, complexity in performing computations, PSD plots for locating exons, alsoSn, Sp and Pr calculations, AEP using MNSRLMS remainsa better choice in real time applications. Lower computational complexity leads to less complex architecture aimed at system on chip (SOC)also lab on chip (LOC) applications.

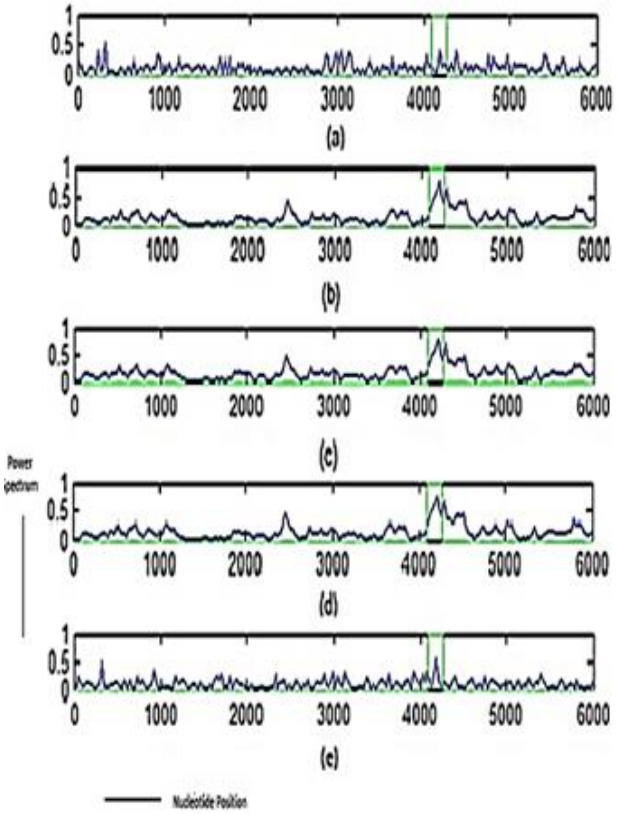

Figure 2: PSD with the location of exon (3934-4581) for a gene sequence of accession AF009962 locatedwith use ofvarious AEPs, (a). AEP using LMS, (b). MNLMS based AEP, (c). MNSRLMS based AEP, (d). MNSLMS based AEP, (e). MNSSLMS based AEP

Table I. Gene datasets from NCBI gene databank

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

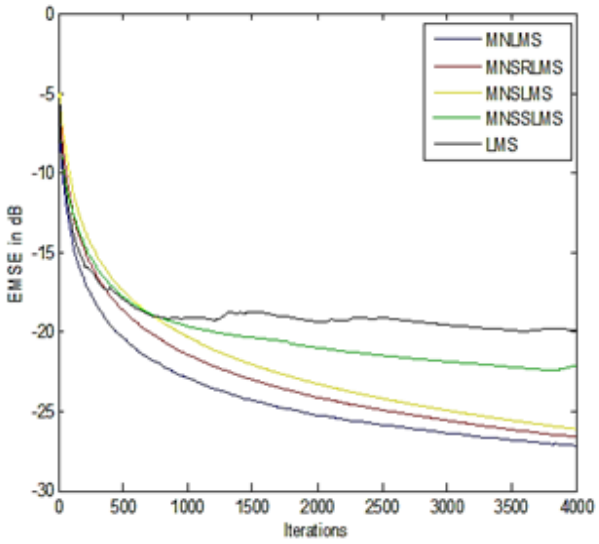

Figure 3: Convergence Characteristics for various algorithms in exon prediction.

Table II. Performance measures of various modified normalized based AEPs pertaining toSp, Sn, and Prcomputations.

Seq. No.	Parameter	LMS	MDN LMS	MDN SRLMS	MDN SLMS	MDN SSLMS
1	Sn	0.6481	0.7634	0.7324	0.7015	0.6826
	Sp	0.6518	0.7428	0.7172	0.6995	0.6781
	Pr	0.5904	0.7521	0.7236	0.7082	0.6865
2	Sn	0.6286	0.7428	0.7172	0.6995	0.6781
	Sp	0.6435	0.7521	0.7236	0.7082	0.6865
	Pr	0.5922	0.7437	0.7123	0.6967	0.6788
3	Sn	0.6384	0.7524	0.7235	0.6989	0.6797
	Sp	0.6628	0.7532	0.7241	0.7075	0.6886
	Pr	0.5894	0.7636	0.7324	0.7015	0.6826
4	Sn	0.6457	0.7428	0.7172	0.6993	0.6781
	Sp	0.6587	0.7521	0.7236	0.7082	0.6865
	Pr	0.5934	0.7434	0.7123	0.6967	0.6788
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	MNLMS	2T+2
3	MN SRLMS	T+3
4	MNSLMS	2T+1
5	MN SSLMS	T+3

IV. CONCLUSION

In current work,we have addressed key problem of tracing exon fragments of DNA whichdevisesnumeroushealth care applications part ofcurrent technology. At this point, AEPs based on data normalization are considered for locating exon sections in DNA. To further lessen the complexity in performing calculations, concept of modified data normalization is used. Towards further minimize computational complexity,sign basedvariants of MNLMS are used. Resulting hybrid variants are MNSRLMS, MNSLMS also MNSSLMS algorithms. Thus, four AEPs werederivedalsoverifiedwithactualgene datasetsacquiredusing NCBI databank. From complexity involved in performing computations in Table III also characteristics related to convergence depicted in Figure 3, AEP usingMNSRLMS remains better in applications related to exon identification. Metrics related to performance in Table IIalsoplots of PSD for predicted exon shown in Figure 2. Therefore, AEP using MNSRLMS is suitable for genomic applications in real timeto develop nano devices, SOCs also LOCs.

REFERENCES

1. S. Nemati, M. E. Basiri, N. Ghasem-Aghaee& M. H. Aghdam, "A novel ACO-GA hybrid algorithm for feature selection in protein function prediction. (2009)." Expert Syst. with Applicat., 36, 12086–12094.
2. 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.
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. Niranjan Chakravarthy, A. Spanias, L. D. Iasemidis, & K. Tsakalis, "Auto regressive modeling and feature analysis of DNA sequences. (2004)." EURASIP J. Appl. Signal Process., 1 (2004), 13–28.
8. 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.
9. 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.
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,KoneruLakshmaiah 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

SrinivasareddyPutluri 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.