Journal of Theoretical and Applied Information Technology
15th May 2017. Vol.95. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

COMPUTER BASED GENOMIC SEQUENCES ANALYSIS USING LEAST MEAN FORTH ADAPTIVE ALGORITHMS

[1]SRINIVASAREDDY PUTLURI, [2]MD ZIA UR RAHMAN, [3] SHAIK YASMIN FATHIMA

[1&2] DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING, K. L. UNIVERSITY, GREEN FIELDS, VADDESWARAM, GUNTUR- 522502, ANDHRA PRADESH, INDIA.
[3] DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING, RVR&JC COLLEGE OF ENGINEERING, GUNTUR-522017, INDIA.

E-mail: sriniputluri@gmail.com, mdzr55@gmail.com, skyf488@gmail.com

ABSTRACT

True prediction of protein coding regions in a deoxyribonucleic acid (DNA) is a major task in the field of Bioinformatics. Study of regions which code for proteins is a key aspect of disease identification and designing drugs. The sections of DNA that include protein coding information are known as exons. Mainly exon regions in the genes show three base periodicity (TBP), which serves as a base for all exon locating methods. For locating the exon regions many techniques have been applied successfully, but development is still needed in this area. Using signal processing methods, TBP can be easily determined. Adaptive signal processing techniques found to be apt due to their diverse ability to alter filter co-efficients depending on the genomic sequence. In this paper, we propose efficient an adaptive exon predictor (AEP) based on these deliberations for DNA sequence analysis and computing. In order to increase the exon locating capability, we develop various AEPs using normalized least mean forth algorithm (NLMF) and its variants. These proposed AEPs notably reduces computational complexity and provides better performance in terms of performance measures like sensitivity, specificity, and precision. It was shown that variable normalized least mean forth (VXENLMF) based AEP is found to be superior than NLMS in exon identification applications based on performance measures with Specificity 0.7468, Sensitivity 0.7562, and Precision 0.7523 at a threshold of 0.8 for a genomic sequence with accession AF009962. Also, this algorithm performs better with respect to convergence by normalization of step size. Finally the exon locating capability of various AEPs is tested using several real DNA sequences obtained from National Center for Biotechnology Information (NCBI) database and compared with existing LMS method. It was shown that proposed AEPs are more efficient for locating exon regions in a DNA sequence.

Keywords: *Adaptive Exon Predictor, Computational Complexity, Deoxyribonucleic Acid, Disease Identification, Exons, Three Base Periodicity*

1. INTRODUCTION

The extensive area of research in the field of bioinformatics is locating the exon regions in a genomic sequence by DNA sequence analysis and computing. Vital genes form a subset in organisms which are needed for development, survival and fertility [1]-[2]. Hence, identification of exons has pragmatic importance to spot human diseases [3] and drug targets discovery in new pathogens [4]-[5] through analysis of DNA sequences. The protein coding regions and non-protein coding regions are present in a genomic sequence. The Sub section of genomics that foucsses on locating the protein coding regions in a genomic sequence is known as gene prediction. The study of prime protein region structure helps the secondary and tertiary structure of protein coding regions for detection of all anomalies, cure diseases and design drugs, as soon as the entire structure of protein regions is analyzed. These studies support in knowing the assessment of phylogenic trees [6] - [7]. Based on the elemental structure of molecules, the living organisms are divided into two types termed as prokaryotes and eukaryotes. The sections which code for proteins are continuous and long in prokaryotes; examples of prokaryotes are bacteria and archaea. The genes are a combination of coding sections divided by long non-protein coding sections in eukaryotes. These sections which code for proteins are also called as exons, whereas the non-protein coding sections are termed as introns. All living organisms other than bacteria and archaea come under this category. The coding sections reside in human eukaryotes are only 3% of the sequence and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the remaining 97% are non-coding regions. Hence the identification of protein coding sections is a significant task [8]-[9]. Almost in all DNA sequences, a three base periodicity (TBP) is exhibited by the protein coding regions. This is obvious by a sharp peak at a frequency f=1/3 in the power spectral density (PSD) plot [10]. Several techniques for predicting exon regions are presented in literature based on various signal processing methods [11] - [13]. But, the length of the sequence in real-time gene sequence is extremely long and also the location of exons varies from sequence to sequence. Existing signal processing techniques are not so accurate in prediction of protein coding regions. Adaptive signal processing techniques are found to be favorable techniques to to process very long sequences in several iterations and can change weight coefficients in accordance to the statistical behavior of the input sequence [13]. In this paper, efficient Adaptive Exon Predictors (AEPs) are developed using adaptive algorithms for DNA sequence analysis and locating protein coding sections. Least mean square (LMS) algorithm is the fundamental adaptive technique. This algorithm is popular because of its simplicity in implementation. But this algorithm suffers problems like gradient noise amplification, weight drift and poor convergence. So, we put forward to normalized least mean forth (NLMF) adaptive algorithms to improve the performance of AEP. NLMF algorithm overcomes the drawbacks of LMS and improves exon locating ability and faster convergence when the error is high [14]. This also leads to reduced excess EMSE in the process of exon prediction. To cope up the computational complexity of an AEP in real time applications, proposed normalized least mean forth adaptive algorithms are combined with variable excess mean square. Sign based algorithms apply signum function and minimizes multiplication operations. The proposed LMF algorithm overcomes the hitches of LMS and NLMS methods by improving exon locating ability and speed of convergence [15]. This also leads to reduced excess mean square error (EMSE) in the process of exon prediction. In real time applications, the computational complexity of an adaptive algorithm plays a key role. Particularly when the sequence length is very large, if the computational complexity of the signal processing technique is large the samples overlap on each other at the input of the exon predictor. These leads to inaccuracy in the prediction and causes inter symbol interference (ISI). Also, the large computational complexity tends to bigger circuit size and large operations. Hence, to cope up with the computational complexity of an AEP in real time applications we combine the adaptive algorithms with sign based algorithms. Sign based algorithms apply signum function and lessen the number of multiplication operations [16] - [17]. The three signum based simplified algorithms are sign regressor algorithm (SRA), sign algorithm (SA) and sign sign algorithm (SSA) [18] - [19]. Therefore, in order to minimize the computational complexity and for faster convergence in DNA sequence analysis and computing, we propose normalized least mean forth and its variants. Normalized LMF algorithm enjoys the advantages of better stability and faster convergence performance resulting due to normalization. The resulting algorithms are normalized least mean forth (NLMF) algorithm, excess mean square error normalized least mean forth (XENLMF) algorithm and variable excess mean square error normalized least mean forth (VXENLMF) algorithm. In these algorithms, the step size is normalized with respect to signal and noise power. When the tap length is larger, which is common in real time applications the large tap length causes an additional computational burden on the AEP. Based on the proposed normalized least mean forth algorithms, we develop various AEPs and the performance is tested using real genomic sequences taken from National Center for Biotechnology Information (NCBI) data base [20]. We consider sensitivity (sn), specificity (sp), precision (pr), convergence characteristics, and computational complexity (O) as performance characteristics to evaluate the performance of the various AEPs. These performance measures of proposed AEPs are compared with existing LMS method in terms of exon locating capability. It was shown that proposed AEPs are more better than existing method for exon prediction. The theory of the adaptive algorithms, discussion on the performance of various AEPs and results of AEPs are presented in the following sections.

2. ADAPTIVE ALGORITHMS FOR EXON PREDICTION

In the AEP proposed, the input genomic sequence is converted into binary representation. This is a significant task in genomic signal processing, as signal processing techniques can be applied only on digital or discrete signals. At this point, we use the binary mapping to convert the input DNA sequence into binary data [14]. This mapping method is used to represent an input DNA sequence as four binary indicator sequences. Using this binary mapping, the nucleotide occurrence at a location is indicated by 1 and absence by 0. Now the resulting sequence is appropriate to give as an input to an adaptive

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

algorithm. Four binary indicator sequences are used as input to the adaptive filter [15]. Now, we consider an adaptive exon predictor (AEP) to be applied on converted binary sequences. Let S(n) be the DNA sequence, M(n) is the binary mapped sequence, R(n) is the TBP obeyed genomic sequence, Y(n) is the output from the adaptive algorithm and F(n) is the feedback signal to update weight coefficients of the algorithm. Consider an LMS adaptive algorithm of length 'N'. In this algorithm, the next weight coefficient can be predicted based on the current weight coefficient, step size parameter 'P', input sequence sample value S(n) at the instance and the feedback signal F(n) generated in the feedback loop. The mathematical expression and analysis of LMS algorithm is presented in [16]. A typical block diagram of proposed AEP is shown in Figure 1.

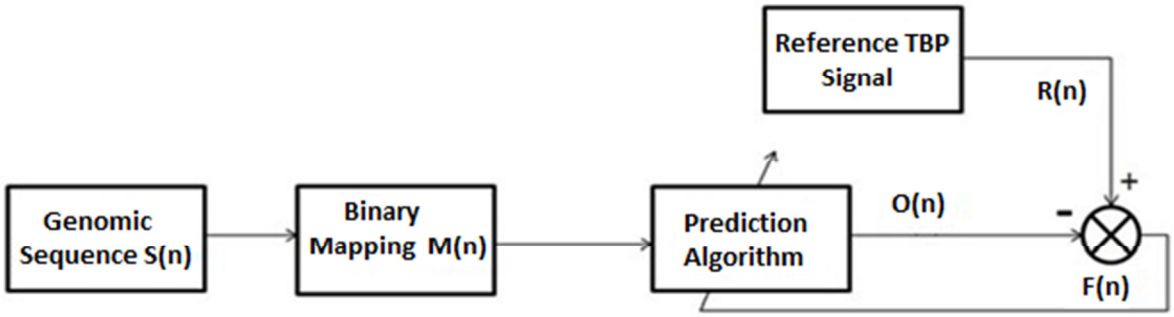

Figure 1. Block Diagram Of An Adaptive Exon Predictor.

Because of its simplicity and robustness, the conventional LMS algorithm may be used in exon prediction applications. For Stability and convergence, the LMS filter needs a prior knowledge of the input power level to select the step size parameter for stability and convergence[17]. Since the input power level is usually one of the statistical unknowns, it is normally estimated from the data before beginning the adaptation process. But the LMS algorithm suffers with two drawbacks in practical situations. It is clear that the input data vector is directly proportional to the weight update mechanism, by observing the weight update recursion of LMS algorithm. Another one is the fixed step size. In practice, an algorithm has to be designed such that, it has to tackle both strong and weak signals. Hence, the tap coefficients should be adjusted accordingly depending upon the filter input and output fluctuations. Therefore, LMS algorithm suffers from a gradient noise amplification problem, when the input

vector at each iteration, adjusted filter weight vector coefficient is normalized with respect to squared euclidian norm of the input vector at each iteration.

The weight update relation of the LMS adaptive algorithm is given by

To reduce the computational complexity compared with an adaptive LMS algorithm, sign regressor algorithm (SRA), sign algorithm (SA) and sign sign algorithm (SSA) algorithms are considered. The

data vector is very large. To avoid this problem normalization has to be applied. The adjusted weight
To further reduce computational complexity of LMS algorithm and for faster convergence, the sign algorithm is to be normalized with adjusted filter weight vector coefficients with respect to input.

Less computational complexity of the adaptive algorithm is highly desirable in exon prediction applications for developing nano devices. This reduction is generally obtainable by clipping either the input data or feedback signal or both. The algorithms based on clipping of error or data are presented in [18]-[19]. Among the adaptive algorithms, the signed algorithms have a convergence rate and a steady-state error that is slightly inferior to those of the LMS algorithm for the same parameter setting. The signum function is written as follows.

$$C\{\boldsymbol{F}(n)\} = \begin{Bmatrix} 1: \boldsymbol{F}(n) > 0 \\ 0: \boldsymbol{F}(n) = 0 \\ -1: \boldsymbol{F}(n) < 0 \end{Bmatrix} \quad (1)$$

$$h(n+1) = h(n) + P\, F(n) S(n) \quad (2)$$

Due to normalization, the step size P varies iteratively and it is proportional to the inverse of the total expected energy of the instantaneous values of the coefficients of the input data vector.

advantage of here is that the step size can be chosen independent of the input signal power and the number of tap weights. On the other hand, some additional computations are required to compute F(n).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The weight update equations of SRA, SA and SSA algorithms are given by

$$h(n+1) = h(n) + P\,F(n)C[S(n)] \quad (3)$$

$$u(n+1) = u(n) + P\,C[F(n)]S(n) \quad (4)$$

$$u(n+1) = u(n) + P\,C[F(n)]C[S(n)] \quad (5)$$

Further, to reduce the computational complexity of the algorithms we apply least mean forth algorithm to develop AEPs. The LMF algorithm possesses faster convergence when the error is high and it reduces when the error decreases. LMF algorithm offers better performance if the weights are initialized nearer to optimum value. One of the advantage of LMF over LMS and NLMS is its faster convergence. But the stability of LMF algorithm depends on input signal power, noise power, and wight initizlization vector.

Thus, the weight update equation of the least mean forth (LMF) algorithm becomes

$$h(n+1) = h(n) + P\,F(n)^3 S(n) \quad (6)$$

A normalized LMF is more advantageous than LMF in addition to stability obtained from the normalization. The idea behind normalization is to normalize the step size with respect to signal power. The basic equations of normalized LMF are taken from the work presented in [13]. Compared with other normalized algorithms, the NLMF algorithm requires a small number of computations.

Thus, the weight update equation of the normalized least mean forth (NLMF) algorithm becomes

$$h(n+1) = h(n) + \frac{P\,F(n)^3 S(n)}{\varepsilon + \max(S(n))^4} \quad (7)$$

where ϵ is a small constant used to avoid the system from becoming unstable when the signal approaches zero. Also, α is usually unity. In order to overcome the dependency on mixing parameter in [20], a variable XE-NLMF algorithm is proposed. So, α value is varied according to the step size.

The weight update equation of the variable normalized least mean forth (XE-NLMF) algorithm becomes

$$h(n+1) = h(n) + \frac{P\,F(n)^3 S(n)}{\varepsilon + (1-\alpha(n))(S(n))^2 + S(n)F(n)^2} \quad (8)$$

Generally, in addition to the even powered signal in the denominator, it is customary to place a very small value ϵ to prevent the system from becoming unstable when the signal approaches zero. As a whole, these terms are making the filter to be variable step size. But this above equation has disadvantages in terms of its stability based on the signal power. Here, both the signal and error power are used in normalization along with a mixed power parameter ϵ which manages the convergence rate in maintaining stability. The MSE performance clearly indicates the improvement of it over NLMF. It is called as XE-NLMF.

The weight update relation of XE-NLMF algorithm is given as -

$$h(n+1) = h(n) + \frac{P\,F(n)^3 S(n)}{\varepsilon + (1-\alpha)(S(n))^2 + \epsilon\,F(n)^2} \quad (9)$$

As long as the error is small in the above equation, the mixed term will be small and the steady state error is small. If the error is more then the mixed term parameter will tend towards unity thereby the stability is achieved. In order to observe the efficiency of LMF over LMS, the normalized least mean square algorithm is used.

The weight update relation of XE-NLMF algorithm is written as -

$$h(n+1) = h(n) + \frac{P\,F(n)S(n)}{\varepsilon + \max(S(n))^2} \quad (10)$$

3. COMPUTATIONAL COMPLEXITY AND CONVERGENCE ISSUES

In general, to estimate and compare algorithm complexity, number of multiplications required to complete the operation is taken as a measure. However, most of the DSP's have a built in hardware support for multiplication and accumulation (MAC) operations. Usually they perform this operation in a single instruction cycle as well as addition or subtraction. In this paper, we concentrate on presenting a comparison between different adaptive algorithms in terms of the computational complexities as summarized in Table 1. Further, as these sign based algorithms are largely free from multiplication operation, these algorithms provide an elegant means for adaptive exon prediction applications. For example, LMS algorithm P+1 MAC operations are required to compute the weight update equation. In case of variable normalized least mean forth algorithm(VXENLMF) only one multiplication is required to compute 'S.F(n)'. Whereas other NLMS, NLMF, XENLMF and VXENLMF based algorithms does not require multiplications if we choose 'S' value a power of 2. In these cases

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

multiplication becomes shift operation which is less complex in practical realizations. In SSA we apply signum to both data and vector, and then we add 'S' to weight vector with addition with sign check (ASC) operation. Among all the algorithms the NLMS adaptive algorithm is more complex, as they require 2P+1 MACs and 1 division operations to implement the weight updating equation (10) on a DSP processor. Among the proposed AEPs, VXENLMF algorithms provide less computational complexity with 1 MAC and 1 division operations for DNA computing and sequence analysis. However, by using a maximum normalization approach, we can minimize multiplications in the denominator from 'P' to '1'.

Compared with other normalized algorithms, the VXENLMF algorithm requires a small number of computations. To compute the variable step with minimum computational complexity, the error value produced in the first iteration is squared and stored. The error value in the second iteration is squared and added to the previously stored value. Then, the result is stored in order to be used in the next iteration, and so on.

Table 1: Computational Complexities Of Various Algorithms Used For The Development Of Aeps.

S.No.	Algorithm	MACs	Add	Divisions	Shifts
1	LMS	P+1	Nil	Nil	Nil
2	NLMS	2P+1	Nil	1	Nil
3	NLMF	P+2	P	1	Nil
4	XENLMF	P	P+1	1	Nil
5	VXENLMF	1	Nil	1	Nil

The VXENLMF algorithm provides significant improvements in minimizing signal distortion. It is clear that variable XENLMF algorithm is able to handle the noises that occurs during transmission in the power spectral density of exons. The parameter α actually controls the convergence. When there is a large error, then α will tend towards unity and the convergence will be very fast. Similarly if the error is small then α will be small and convergence will be slow making the step size small. This actually occurs when the adaptive filter is reaching the steady state. The convergence curves results from plotting the MSE over several samples. It was observed that MSE is reducing over samples and iterations and specifically over the change in α value. It shows that as α value is increasing, error term is weighted more and as a result the noise is effectively suppressed.

In order to cope up with both the complexity and convergence issues without any restrictive tradeoff, the corresponding normalized least mean forth adaptive algorithms are normalized least mean forth (NLMF) algorithm, excess mean square error normalized least mean forth (XENLMF) algorithm and variable normalized least mean forth (VXENLMF) algorithm. These algorithms provide less computational complexity, good filtering capability and faster convergence. The convergence characteristics of the error normalized and maximum error normalized adaptive algorithms are shown in Figure 2. From these characteristics, it is clear that XENLMF is just inferior to its variable normalized least mean forth version. Hence, among the algorithms considered for the implementation of the AEPs VXENLMF algorithm is found to be better with reference to computational complexity and convergence characteristics.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

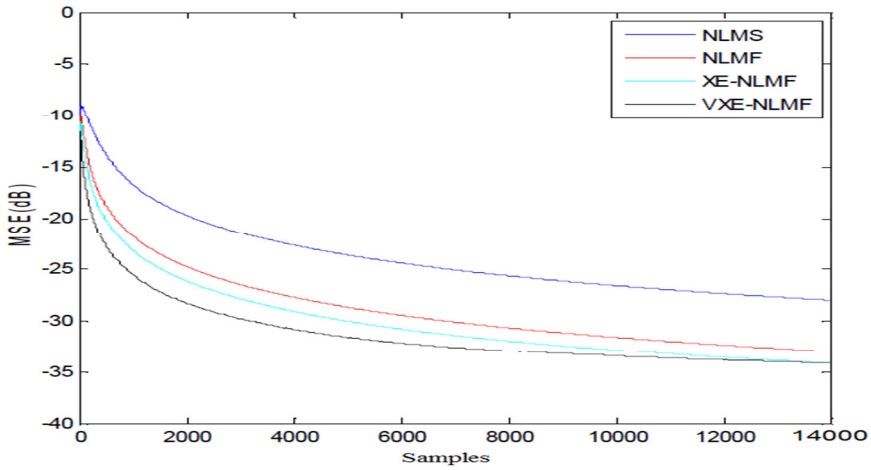

Figure 2: Convergence Characteristics Of Error Normalized LMS With Its Signed Based Variants.

4. RESULTS AND DISCUSSIONS

In this section, the performances of various AEPs are compared with existing LMS method using real DNA sequences. The structure of proposed AEP is shown in Figure 1. The normalized least mean forth (NLMF) algorithm and its variants are used to implement various AEPs. For purpose of comparison, we also developed LMS and normalized based AEPs. For evaluation purpose, we obtained ten DNA sequences from NCBI database [20]. For consistency of results, to evaluate the performance of various algorithms we considered ten DNA sequences as our data set. The description of the dataset considered is shown in Table 2. The performance measure is carried using parameters like sensitivity (Sn), specificity (Sp) and precision (Pr). The theory and expressions for these parameters are given in [11]. The exon prediction results for sequence 5 are shown in Figure 3. The performance measures Sn, Sp and Pr are measured at threshold values from 0.4 to 0.9 with an interval of 0.05. The exon prediction seems to be better at threshold 0.8. Hence at threshold 0.8 the values are tabulated in Table 3.

Table 2: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
6	X59065.1	H.sapiens human acidic fibroblast growth factor (FGF) gene
7	AJ223321.1	Homo sapiens transcriptional repressor (RP58) gene
8	X92412.1	H.sapiens titin (TTN) gene
9	U01317.1	Human beta globin sequence on chromosome 11
10	X51502.1	H.sapiens gene for prolactin-inducible protein (GPIPI)

The steps in adaptive exon prediction are as follows:

1. DNA sequences are chosen from genome data base [18]. Binary mapping technique is used to convert the DNA sequence to binary data.
2. The obtained binary data is given as input to AEP arrangement shown in Figure 1.
3. A DNA sequence that obeys three base periodicity is given as reference to the AEP.
4. As shown in Figure 1, a generated feedback signal is used to update filter coefficients.
5. When a minimum feedback signal is obtained, the adaptive algorithm accurately predicts the location of the protein coding region sequence
6. The exon location is plotted using power spectral density. The performance measures like Sn, Sp and Pr are measured.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

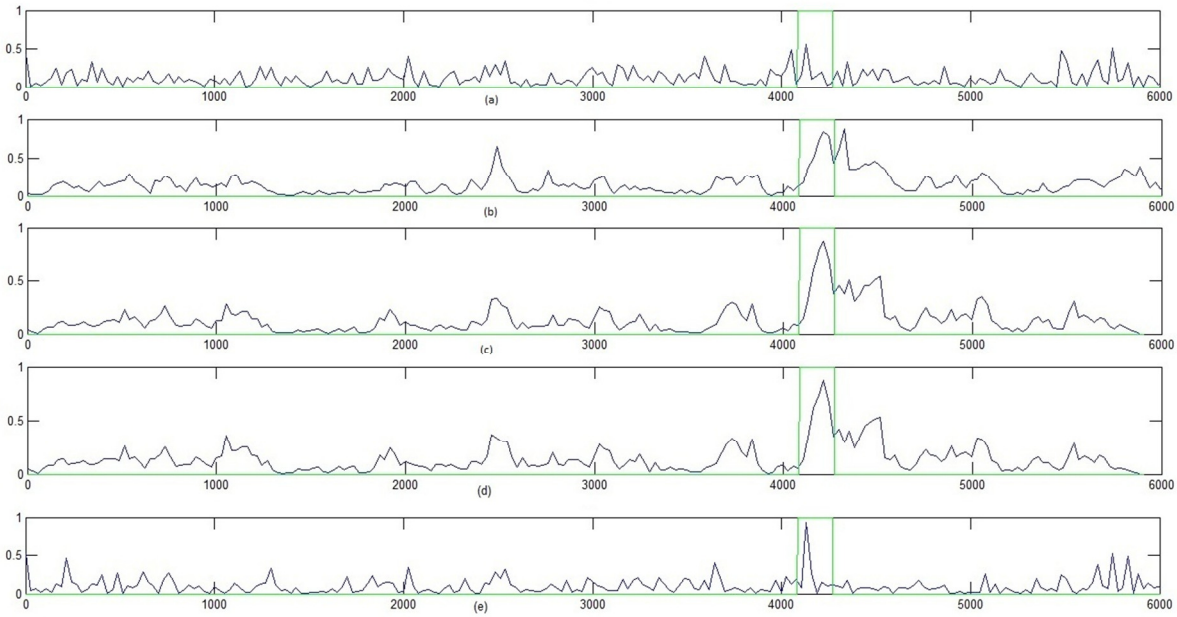

Figure 3: Locations of exons predicted using various adaptive algorithms for genomic sequence with accession AF009962 (a). LMS based AEP, (b). NLMS based AEP, (c). NLMF based AEP, (d). XENLMF based AEP, (e). VXENLMF based AEP.

Figure 3 shows power spectrum for the predicted exon locations of sequence 5 presented in Table 2 by applying various adaptive algorithms. From these plots it is clear that the LMS based AEP has not predicted the coding regions accurately. This algorithm causes some ambiguities in location prediction by identifying some non protein coding regions from the input genomic sequence. In Figure 3 (a) some unwanted peaks are identified at locations 1200th, 2300th and 3700th sample values using LMS based AEP. At the same time the actual exon location 4084-4268 is not predicted accurately. Similar kind of results using LMS based AEP and other signal processing methods for DNA computing and sequence analysis have been presented in the literature [11]–[14]. But, using proposed normalized least mean forth based AEP versions of NLMF, XENLMF, and VXENLMF algorithms for DNA computation and analysis, exactly predicted the exon locations at 4084-4268 with good intensity of PSD. These PSDs are shown in Figure 3 (b), (c) and (d). Because of the normalization involved in these algorithms the tracking capability of these algorithms, sensitivity, specificity and accuracy are much better than LMS and NLMS algorithms. Among these three proposed algorithms, VXENLMF is found to be better with reference to its convergence characteristics and computational complexity. This algorithm needs only two multiplications, the number of multiplications involved in VXENLMF algorithm are independent of tap length of AEP. The convergence characteristics of XENLMF are just inferior to VXENLMF, but due to a large number of reduced multiplications for DNA computation and analysis this inferior behavior in convergence can be tolerable. Therefore, based on computational complexity, convergence characteristics, exon prediction plots, Sn, Sp and Pr calculations, it is found that VXENLMF based AEP is found to be the better candidate in realistic applications for the development of SOCs, LOCs and nano devices in future research.

Table 3: Performance Measures Of Various Aeps With Respect To Sn, Sp And Pr Calculations For Geneomic Sequence With Accession AF009962

Seq. No.	Parameter	LMS	NLMS	NLMF	XENLMF	VXENLMF
1	Sn	0.6286	0.7085	0.7284	0.7364	0.7594
	Sp	0.6435	0.7267	0.7094	0.7247	0.7436
	Pr	0.5922	0.6954	0.7159	0.7327	0.7582
2	Sn	0.6384	0.7137	0.7223	0.7362	0.7572
	Sp	0.6628	0.7458	0.7084	0.7214	0.7483
	Pr	0.5894	0.7027	0.7141	0.7312	0.7556

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3	Sn	0.6457	0.7227	0.7273	0.7327	0.7576
	Sp	0.6587	0.7321	0.7053	0.7245	0.7472
	Pr	0.5934	0.6962	0.7193	0.7385	0.7562
4	Sn	0.6273	0.7086	0.7263	0.7322	0.7576
	Sp	0.6405	0.7278	0.7042	0.7262	0.7482
	Pr	0.5858	0.7096	0.7146	0.7336	0.7588
5	Sn	0.6481	0.7240	0.7246	0.7326	0.7562
	Sp	0.6518	0.7378	0.7045	0.7234	0.7486
	Pr	0.5904	0.6927	0.7134	0.7356	0.7523
6	Sn	0.6162	0.7162	0.7252	0.7334	0.7534
	Sp	0.6324	0.7284	0.7115	0.7215	0.7433
	Pr	0.5786	0.6857	0.7145	0.7383	0.7572
7	Sn	0.6193	0.7192	0.7223	0.7338	0.7545
	Sp	0.6529	0.7396	0.7034	0.7216	0.7446
	Pr	0.5896	0.6904	0.7112	0.7312	0.7593
8	Sn	0.6241	0.7282	0.7223	0.7382	0.7578
	Sp	0.6289	0.7274	0.7145	0.7298	0.7487
	Pr	0.5856	0.6857	0.7134	0.7353	0.7545
9	Sn	0.6268	0.7285	0.7265	0.7383	0.7587
	Sp	0.6452	0.7393	0.7054	0.7275	0.7484
	Pr	0.5814	0.6896	0.7132	0.7334	0.7523
10	Sn	0.6202	0.7286	0.7212	0.7337	0.7492
	Sp	0.6465	0.6976	0.7035	0.7294	0.7466
	Pr	0.5786	0.6825	0.7173	0.7346	0.7568

5. MERITS AND LIMITATIONS

Accuracy in prediction of exon locations in DNA sequences is crucial for disease diagnosis and therapy. The merits of proposed AEPs include more accuracy in exon prediction and less computational complexity when compared with existing techniques. Less computational complexity of proposed AEPs found to better techniques in realistic applications for the development of SOCs, LOCs and nano devices in future research.

When the exon length is short, increasing the accuracy in prediction of exon locations in DNA sequences will become a challenging task for DNA computing and analysis. The limitation of proposed AEPs is improvement of accuracy in

prediction of short exons is desirable which need to be considered in future research.

6. CONCLUSION

In this paper, the problem of identifying exons in a DNA sequence is illustrated. The concept of finding exact location of exons has several applications in current health care technology such as disease diagnosis. At this point, we considered adaptive exon identification technique using novel AEPs. To fulfill this we considered normalized least mean forth adaptive algorithms. In order to reduce computational complexity of the proposed implementations, we introduced the concept of normalization of step size with respect to signal power instead by using normalized least mean forth algorithms of data normalization. To further minimize the computational complexity, the proposed NLMF algorithm is combined with its variable normalized variants. As a result three new hybrid algorithms come into the scenario of exon prediction. The hybrid variants includes NLMF, XENLMF and VXENLMF are considered for present implementation. Different AEPs are developed and tested using NLMS algorithm and these three algorithms on real DNA sequences obtained from NCBI database. It is apparent that NXENLMF based AEP is better in exon prediction applications, based on the convergence characteristics shown in Figure 2, computational complexities shown in Table 1, and based on performance measures with Sensitivity 0.7562, Specificity 0.7486 and precision 0.7523 obtained at a threshold value of 0.8 for genomic sequence with Accession AF009962. This is also clear from the performance measures tabulated in Table 3 and PSD of exon locations shown in Figure 3 where exactly predicted the exon locations at 4084-4268 using proposed AEPs. Therefore, proposed AEP realizations are appropriate for practical genomic applications for the development of SOCs, LOCs and nano devices for future research.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REFERENCES

[1] L.W. Ning, H. Lin, H. Ding, J. Huang, N. Rao & F.B. Guo, "Predicting bacterial essential genes using on sequence composition information," *Genetics and Molecular Research*, 13(2014), 2014, pp. 4564 - 4572.

[2] Min Li, Qi Li, Gamage Upeksha Ganegoda, JianXin Wang, FangXiang Wu, Yi Pan, "Prioritization of orphan disease-causing genes using topological feature and go similarity between proteins in interaction networks," *SCIENCE CHINA Life Sciences*, 57(2014), 2014, pp. 1064–1071

[3] Dickerson JE, Zhu A, Robertson DL, Hentges KE, "Defining the role of essential genes in human disease," *PloS One*, 6(11), 2011, e27368.

[4] Inbamalar T M, Sivakumar R, "Study of DNA Sequence Analysis Using DSP Techniques," *Journal of Automation and Control Engineering*, 1(2013), 2013, pp. 336–342.

[5] Cole S, "Comparative myco bacterial genomics as a tool for drug target and antigen discovery," *The European Respiratory Journal*, 20(36 suppl), 2002, pp. 78s–86s.

[6] S. Maji, D. Garg, "Progress in gene prediction: principles and challenges," *Current Bioinformatics*, 8(2013), 2013, pp. 226–243.

[7] Hamidreza Saberkari, Mousa Shamsi, Hamed Heravi, Mohammad Hossein Sedaaghi, "A Novel Fast Algorithm for Exon Prediction in Eukaryotes Genes using Linear Predictive Coding Model and Goertzel Algorithm based on the Z-Curve," *International Journal of Computer Appications*, 67(2013), 2013, pp. 25–38

[8] S. Maji and D. Garg, "Progress in gene prediction: principles and challenges," *Current Bioinformatics*, 8(2), 2013, pp. 226– 243.

[9] Wazim Mohammed Ismail, Yuzhen Ye, Haixu Tang, "Gene finding in metatranscriptomic sequences," *BMC Bioinformatics*, 15(2014), 2014, pp. 01–08

[10] Mahin Ghorbani, Hamed Karimi, "Bioinformatics Approaches for Gene Finding," *International Journal of Scientific Research in Science and Technology*, 1(2015), 2015, pp. 12–15.

[11] Gangchen Liu, Yihui Luan, "Identification of Protein Coding Regions in the Eukaryotic DNA Sequences based on Marple algorithm and Wavelet Packets Transform," *Abstract and Appied Anaysis*, 2014, 2014, pp. 1-14.

[12] Yusuke Azuma, Shuichi Onami, "Automatic Cell Identification in the Unique System of Invariant Embryogenesis in Caenorhabditis elegans," *Biomedical Engineering Letters*, 4(2014), 2014, pp. 328–337

[13] BurraVenkataSrikanth, Md Zia Ur Rahman, "Efficient ECG Signal Conditioning Techniques using Variable Step Size Least Mean Forth Algorithms," *International Journal of Engineering and Technology'*, 8(2), 2016, pp. 660-668.

[14] Srinivasareddy Putluri, Md Zia Ur Rahman, "New Adaptive Exon Predictors For Identifying Protein Coding Regions In DNA Sequence," *ARPN Journal of Engineering and Applied Sciences,* 11(2016), 2016, pp. 13540 - 13549

[15] Guangchen Liu and Yihui Luan(2014), "Identification of Protein Coding Regions in the Eukaryotic DNA Sequences based on Marple algorithm and Wavelet Packets Transform," *Abstract and Applied Analysis*, Vol. 2014, 2014, pp. 01-14.

[16] Simon O. Haykin, Adaptive Filter Theory, 5th edition, *Pearson Education Ltd.*, 2014

[17] Md. Zia Ur Rahman, Rafi Ahamed Shaik, D. V. Rama Koti Reddy, "Efficient and Simplified Adaptive Noise Cancellers for ECG Sensor Based Remote Health Monitoring", IEEE Sensors Journal, 12(3): 2012, pp. 566-573.

[18] Srinivasareddy Putluri, Md Zia Ur Rahman, "Simplified Adaptive Exon Predictors for extracting protein coding regions in genomic sequences," *Journal of Theoretical and Applied Information Tecnhology* 93(1), 2016, pp. 143 - 151.

[19] Paula S. R. Diniz, Adaptive Filtering, Algorithms and Practical Implementation, Third edition, *Springer Publishers,* 2014.

[20] National Center for Biotechnology Information, www.ncbi.nlm.nih.gov/.